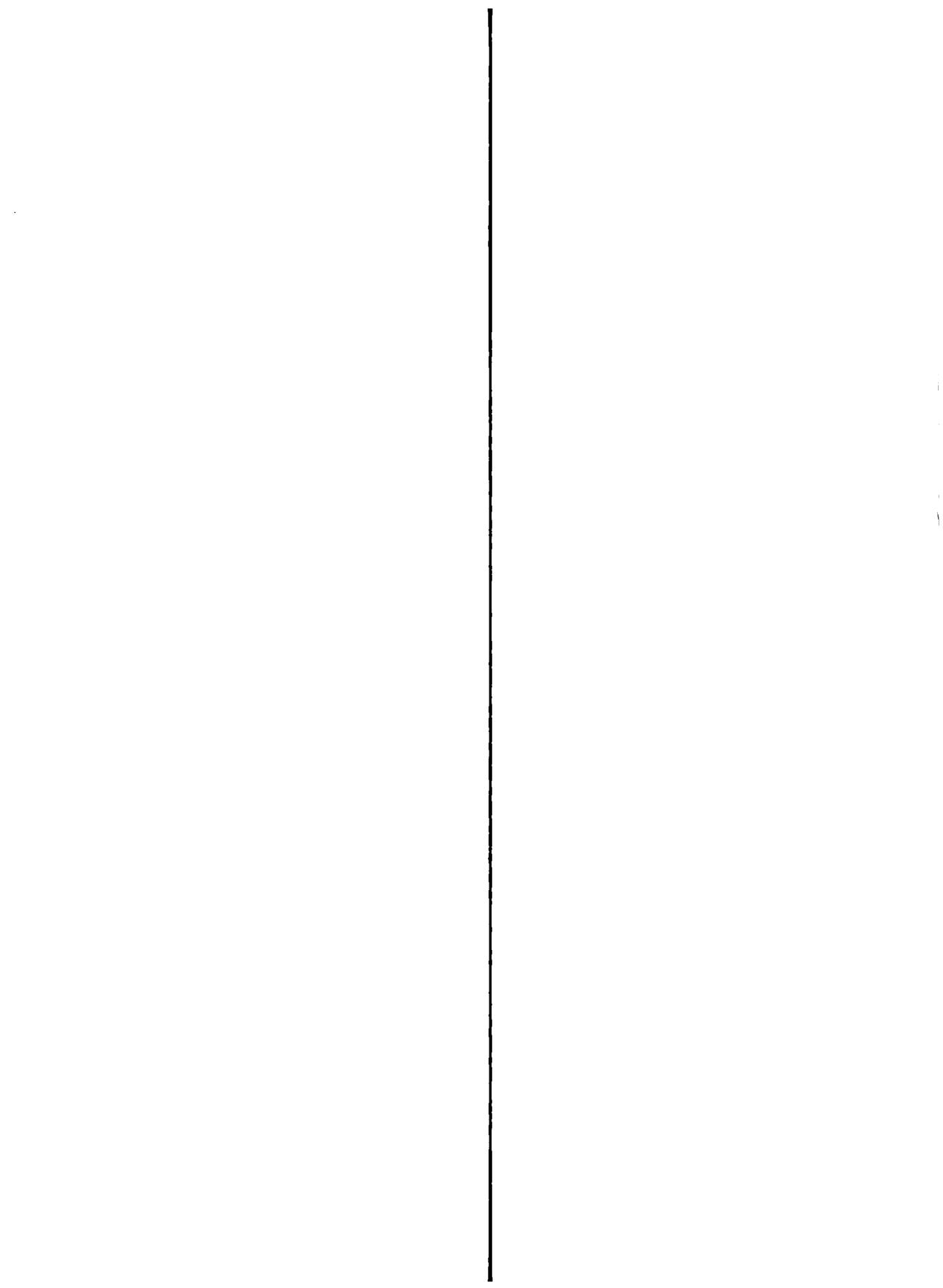


# **Mathematical Methods in Chemical Engineering**

**S. PUSHPAVANAM**





# **Mathematical Methods in Chemical Engineering**

**S. PUSHPAVANAM**

*Associate Professor*

*Indian Institute of Technology Madras  
Chennai*

**Prentice-Hall of India Private Limited**  
**New Delhi-110001**

**2005**

**This One**



**46AX-HTJ-D8CE**

**Rs. 275.00**

**MATHEMATICAL METHODS IN CHEMICAL ENGINEERING**  
by S. Pushpavanam

© 1998 by Prentice-Hall of India Private Limited, New Delhi. All rights reserved. No part of this book may be reproduced in any form, by mimeograph or any other means, without permission in writing from the publisher.

**ISBN-81-203-1262-7**

The export rights of this book are vested solely with the publisher.

**Third Printing**                    ...                    ...                    **February, 2005**

Published by Asoke K. Ghosh, Prentice-Hall of India Private Limited, M-97, Connaught Circus, New Delhi-110001 and Printed by Rajkamal Electric Press, B-35/9, G.T. Karnal Road Industrial Area, Delhi-110033.

# Contents

---

<i>Preface</i>	<i>vii</i>
<b>1. Models in Chemical Engineering</b>	<b>1–19</b>
1.1 Modelling and Simulation	1
1.2 Linear Equations	2
1.3 Nonlinear Equations	11
<i>Problems</i>	17
<i>References</i>	19
<b>2. Vector and Vector Spaces</b>	<b>20–44</b>
2.1 Vectors	21
2.2 Vector Spaces	22
2.3 Metrics, Norms and Inner Products	23
2.4 Linear Dependence and Dimension	30
2.5 Gram-Schmidt Orthonormalisation	39
<i>Problems</i>	42
<i>References</i>	44
<b>3. Matrices, Operators and Transformations</b>	<b>45–65</b>
3.1 Matrices	45
3.2 Eigenvalues and Eigenvectors	49
3.3 Fredholm Alternative (Solvability Conditions)	58
3.4 Rayleigh's Quotient	60
<i>Problems</i>	63
<i>References</i>	64
<b>4. Applications to Chemical Engineering Systems</b>	<b>66–93</b>
4.1 Linear Algebraic Equations	66
4.2 First Order System of Homogeneous Ordinary Differential Equations (Initial-Value Problems)	68
4.3 Nonhomogeneous First Order Ordinary Differential Equations (Initial-Value Problems)	70
4.4 Geometric Basis of the Method	78
4.5 Implications in Process Control	80
4.6 Non Self-adjoint Systems (A Special Approach)	81
<i>Problems</i>	89
<i>References</i>	93

<b>5. Partial Differential Equations</b>	<b>94–112</b>
5.1 Fundamental Concepts: A Review	94
5.2 Classification of Second Order Partial Differential Equations	97
5.3 Linearity and Superposition	102
<i>Problems</i>	110
<i>References</i>	112
<b>6. Sturm-Louiville Theory</b>	<b>113–146</b>
6.1 Infinite Dimensional Spaces	113
6.2 Eigenvalue Problems	118
6.3 Classical Eigenvalue Problems	128
6.4 Fourier Series	138
6.5 Rayleigh's Quotient	141
<i>Problems</i>	143
<i>References</i>	145
<b>7. Separation of Variables and Fourier Transforms</b>	<b>147–180</b>
7.1 Rectangular Cartesian Coordinates	147
7.2 Cylindrical Coordinates	155
7.3 Spherical Coordinate Systems	159
7.4 Fourier Series and Finite Fourier Transforms	162
7.5 Fourier Transforms Unbounded Domains	168
7.6 Laplace Transform	174
<i>Problems</i>	176
<i>References</i>	179
<b>8. Green's Functions</b>	<b>181–213</b>
8.1 Ordinary Differential Equations	182
8.2 Green's Function for Partial Differential Equations	192
8.3 Unbounded Domains	210
<i>Problems</i>	211
<i>References</i>	213
<b>9. Uniqueness Conditions for Linear and Nonlinear Systems</b>	<b>214–238</b>
9.1 Maximum Principles	215
9.2 Energy Methods	223
9.3 Fredholm Alternative	225
9.4 Monotone-Iteration Methods	227
<i>Problems</i>	237
<i>References</i>	238
<b>10. Steady State Characteristics of Nonlinear Dynamical Systems</b>	<b>239–260</b>
<i>Introduction</i>	239
10.1 Dynamic Systems	240
10.2 Steady States (Numerical Evaluation)	245
10.3 Continuation Methods	248
<i>Problems</i>	259
<i>References</i>	260

<b>11. Linear Stability and Limit Cycles</b>	<b>261–299</b>
11.1 Linear Stability of Dynamical Systems	261
11.2 Bifurcation Theory	273
11.3 Maps	289
<i>Appendix: Numerical Computation of Unstable Limit Cycle</i>	294
<i>Problems</i>	296
<i>References</i>	298
<b>12. Secondary Bifurcations and Chaos</b>	<b>300–323</b>
12.1 Landau-Hopf Scenario	301
12.2 Period-Doubling Cascades	302
12.3 Ruelle-Takens Scenario	311
12.4 Characterisation of Trajectory	316
12.5 Concluding Remarks on Nonlinear Dynamical Systems	319
<i>Appendix: Floquet Theory</i>	321
<i>References</i>	322
<b>Index</b>	<b>325–327</b>



# Preface

---

---

This book is an outgrowth of my several years of teaching the subject at Indian Institute of Technology Kanpur. It is aimed as a text for the first year master's level students. The student at this level has already come across the various concepts of linear algebra, ordinary differential equations and partial differential equations. He is also introduced to solving equations numerically. He has seen how these equations arise in different contexts of chemical engineering in individual courses on unit operations, kinetics, and so on.

The primary objective of this book is to show the connection that underlies the different mathematical disciplines. The methods of analyzing the different kinds of equations at the undergraduate level are usually dealt with in separate mathematical courses. These methods are modified and tailor-made to suit certain requirements and to render them more efficient. The student at the end of all this believes that the different mathematical techniques are distinct from each other. My endeavour in this book is to establish that for every concept in linear algebra there is an analogous concept in linear partial differential equations. The aim is to present a unified approach, that will enable the student to have a better philosophy towards mathematical modelling and analysis.

In this book, an overall mathematical structure is established. It enables one to have a general approach, that can be used for solving different classes of problems. The mathematical methods presented here are primarily analytical.

In this modern era, digital computers are being used to analyze system behaviour extensively. This has led to a preponderance of different software packages. There is a tendency for the student to use these packages as "black-boxes" blindly. He has no understanding of the basis of the different numerical techniques. A thorough understanding of the various analytical concepts presented in this book is vital even for the student interested purely in the numerical methods of solving models. In this book we bring out how the different computational techniques traditionally used are based on the concepts presented here.

The different concepts are presented in such a way as to enable the student to have an intuitive understanding of the subject. This is done keeping in mind the engineering student as the audience. The aim is not to wear down the student with mathematical detail but present at least two different perspectives—the mathematical and the physical, the algebraic and the geometric. This does not mean that the book has been written in a half-hearted manner. The student interested in a more rigorous treatment of the various concepts will find it in those works cited in the Bibliography.

The book starts with a study of finite-dimensional systems. In Chapter 1, examples of equations from different areas of chemical engineering are presented. The basics of vectors and vector spaces are introduced in Chapter 2. All definitions and concepts are such that they can be extended easily to partial differential equations, i.e. infinite dimensional systems, discussed later. Chapter 3 introduces the matrix as a linear operation. The properties of eigenvalues and eigenvectors are discussed in

detail. This forms the basis of the method of solution for linear algebraic equations and ordinary differential equations (initial-value problems) presented in Chapter 4.

Chapter 5 opens with an introduction to linear partial differential equations and the principle of linearity and superposition. Chapter 6 on Sturm-Louiville Theory discusses the properties of eigenvalues, and eigenfunctions of differential operators. It is structured along the same lines as Chapter 3. In Chapter 7 we see how the technique of separation of variables is a method of solution similar to that presented in Chapter 4. We also see how this naturally generates the concepts of Fourier and Laplace transforms. The Green's function is introduced in Chapter 8 to bring out the similarity with the matrix inverse.

The conditions of uniqueness of solutions to linear and nonlinear equations are presented in Chapter 9. Chapter 10 explains the homotopy continuation method and how it can be used to study the effect of a system parameter on a solution. Chapter 11 discusses the fundamentals of bifurcation theory and linear-stability analysis. We see the origin of oscillatory solutions here for autonomous nonlinear systems. Chapter 12 introduces the basics of chaos, its characteristics and implications.

The section on nonlinear equations has been written under the recommendation (in fact, insistence) of Dr. S.K. Gupta. He went through the entire first draft of this manuscript. His constructive critical comments have been instrumental in bringing the book to this shape. I am happy to acknowledge his support and the interest he took in this effort.

I also thank Dr. Keshava Rao of the Indian Institute of Science, Bangalore and Dr. R.K. Jain of the Mathematics Department of Indian Institute of Technology Kanpur for going through the manuscript and for their inputs. Dr. Jeff Harmon during his visit to India went through the entire second draft of the manuscript. I thank him for his valuable suggestions, which I believe have made the presentation more lucid. Dr. Shobha Madan went through parts of the manuscript involving the concepts of Fredholm alternative, completeness, and so on.

The students of IIT Kanpur have played an important role in the development of this manuscript. My own student Pavan Shukla has helped with the figures, editing as well as solving the various numerical examples. It is my pleasure to acknowledge his support. I am happy to acknowledge the discussions I had with Shirshedu De and Ritwik Bhatia. Joydeb Mukherjee went through the entire final manuscript enthusiastically and checked it for clarity of presentation.

Professors Ranga Narayanan and Lewis Johns of the University of Florida deserve special mention here for kindling my interest in the area of applied mathematics.

This effort would never have been possible but for the support provided by my wife Geetha. She managed my sons Karthik and Vishnu single handedly on all those numerous late nights I was at work and encouraged me throughout.

Thanks are also due to Mr. S.L. Yadav for typing the manuscript patiently. Mr. Panesar was very prompt with all the drawings. Mr. N.K. Metia helped get the numerous voluminous printouts whenever I needed them. The financial assistance for the preparation of the manuscript came from the Quality Improvement Programme (QIP) of Indian Institute of Technology Kanpur.

**S. Pushpavanam**

# Models in Chemical Engineering

---

---

The behaviour of a system in the different engineering disciplines and the sciences is studied either experimentally or theoretically. Experimental study involves conducting experiments on the actual system and characterising its behaviour by making quantitative measurements of the dependent variables. The steady state of a multistage distillation column, for example, is determined by measuring the time independent compositions of the liquid and vapour phases in different stages. The temperature profile of a packed bed reactor or a heat exchanger in this approach is measured directly by using thermocouples at fixed locations. This approach yields an accurate and a realistic idea of the system behaviour (should our measurements be free of any errors). It, however, is time consuming, expensive and labour intensive.

## 1.1 MODELLING AND SIMULATION

The theoretical approach for analysing a system consists of two steps: (i) modelling, and (ii) simulation. Modelling is the abstract representation of a physical system by equations. These depict the different physical interactions in a system and are based on the fundamental laws of physics, typically the conservation of mass, energy and momentum. The equations accurately represent or model the system behaviour. This approach is useful when the different processes and the physical interactions occurring are well characterised. It enables us to quantify the different processes occurring in a system and its response to different inputs. In a chemical reactor, for example, the reaction kinetics should be well understood to generate a good model. Similarly, an accurate description of the vapour liquid equilibrium (VLE) characteristic of a multicomponent system is a must to be able to simulate a distillation column. A detailed exposition of the fundamental principles of modelling can be found in Aris (1978).

Simulation entails solving the modelling equations either numerically or analytically. The solution represents the system behaviour and its performance. This book deals with analytical methods for solving equations which arise in modelling systems. The model (i.e. the describing equations) is assumed to be available. Throughout this book we will be interested only in the simulation of the system behaviour, and not in modelling aspects. The analytical methods developed here can be applied to solve a wide class of linear problems—linear algebraic equations, linear ordinary differential equations and partial differential equations.

The theoretical approach is inexpensive and effective in studying a system comprehensively. The advent of high-speed computers has facilitated the numerical solution of models of complex systems. This has given a big boost to this approach. The disadvantage here is that mispredictions can occur due to incorrect modelling or due to wrong assumptions made in modelling. These usually arise because of a poor understanding of the processes occurring in the system. The role

of simulations in understanding system behaviour in various disciplines is discussed extensively in Aburdeen (1988).

Simulation can be classified broadly into two categories:

**(i) Steady state simulation.** This involves studying the time invariant characteristics of a system. Chemical plants and various units are operated in a continuous mode, usually under steady conditions. A steady state operation allows us to control the system efficiently and elegantly at a desired set-point. The behaviour and performance of such systems is analysed by steady state simulation.

**(ii) Dynamic simulation.** The evolution of the state of a system with time is studied here. This finds applications in the study of batch operations and other transient phenomena associated with start-up and shut-down of plants. The dynamic mode of operation of a plant is sometimes preferred as it can maximise selectivity in a reactor and productivity in a plant (see Lee et al., 1980). A second example arises in biochemical fermentation processes which are typically characterised by substrate inhibition. The optimum performance of the reactor here is obtained when the process is carried out in a fed-batch mode. This is a dynamic mode where the substrate is added to the reactor over a period of time (see San and Stephanopoulos, 1984).

The equations generated in modelling can be either linear or nonlinear. Linear equations can be solved elegantly by analytical methods. This forms the subject of Chapters 1–8. Here we discuss solution techniques in a generalised framework using the concept of operators.

Nonlinear equations are solved normally using numerical techniques. In the third section, analytical methods and semi-analytical methods for nonlinear equations are discussed. Here the emphasis is not on methods of solution (unlike in the first two sections), but on obtaining insight into different kinds of system behaviour that can arise due to nonlinearities. Bifurcation theory is used for this purpose. This is a theory which is sufficiently general and can be applied to algebraic equations at one end of the spectrum and to partial differential equations at the other end. It is based on the analysis of the linearised equations, and helps us study how a parameter can influence system behaviour.

The approach adopted in this book, it is believed, will enable the student to develop a deeper understanding and appreciation of the fundamental concepts behind the mathematics associated with a problem. The various concepts are developed intuitively. This will enable the aspiring engineer, with some mathematical inclination, to readily extend the concepts and methods discussed in this book to the analysis of the behaviour of complex systems which he might encounter.

The remainder of this chapter is devoted to discussing different situations where equations of varying degrees of complexity arise in modelling of chemical engineering systems. Examples of general engineering systems can be found in Kersten (1969).

## 1.2 LINEAR EQUATIONS

A system of equations is linear when the dependent variables and their derivatives with respect to the independent variables occurring in it, are raised to the power of unity. A linear equation should also not contain the product of the dependent variable and its derivatives.

A **linear equation** such that every term in the equation has the dependent variable or its derivative is said to be a **homogeneous equation**. A **nonhomogeneous equation** is one that has

a term independent of the dependent variable. This classification as homogeneous or nonhomogeneous is relevant only for linear systems. Consider the equations

$$\left(\frac{du}{dt}\right)^2 + u = 0 \quad (1.1a)$$

$$u\left(\frac{du}{dx}\right) + \nabla^2 u = 0 \quad (1.1b)$$

$$\frac{du}{dt} + e^u = 0 \quad (1.1c)$$

$$\nabla^2 u + \frac{\partial u}{\partial x} = 0 \quad (1.1d)$$

$$\nabla^2 u + \frac{\partial u}{\partial x} + \sin x = 0 \quad (1.1e)$$

The first three examples, (1.1a)–(1.1c), are nonlinear. The fourth (1.1d) is an example of a linear homogeneous equation, and the fifth (1.1e) an example of a linear nonhomogeneous equation (as ‘ $\sin x$ ’ is independent of the dependent variable ‘ $u$ ’).

### 1.2.1 Linear Algebraic Equations

Such systems normally arise in the modelling of the steady states of lumped parameter systems, i.e. systems characterised by variables which are spatially independent. Here we have a system of ‘ $n$ ’ coupled linear equations in ‘ $n$ ’ unknowns. These equations can be written compactly as

$$Au = b \quad (1.2)$$

where  $A$  is an  $n \times n$  matrix,  $b$ ,  $u$  are vectors with  $n$  coordinates each.  $Au$  denotes the matrix  $A$  multiplying the vector  $u$ . This operation will be formally defined later. Here we are interested in determining ‘ $u$ ’, for a given  $A$ ,  $b$ , such that (1.2) is satisfied. This equation is a nonhomogeneous equation if  $b \neq 0$  (now the right-hand side has  $b$  which is a vector independent of  $u$ ).

We will see a few situations where such equations arise. A typical chemical plant consists of several interlinked units. The effluent stream from a unit serves as the inlet to the next unit downstream. These units act as junctions/nodes where different streams combine or separate. The flowsheet in Fig. 1.1 has nine liquid streams connecting five units. An experimental approach would involve measuring the nine flow-rates to describe the state of the plant. Neglecting density variations across each stream, the mass balance equations across each node at steady state can be written as

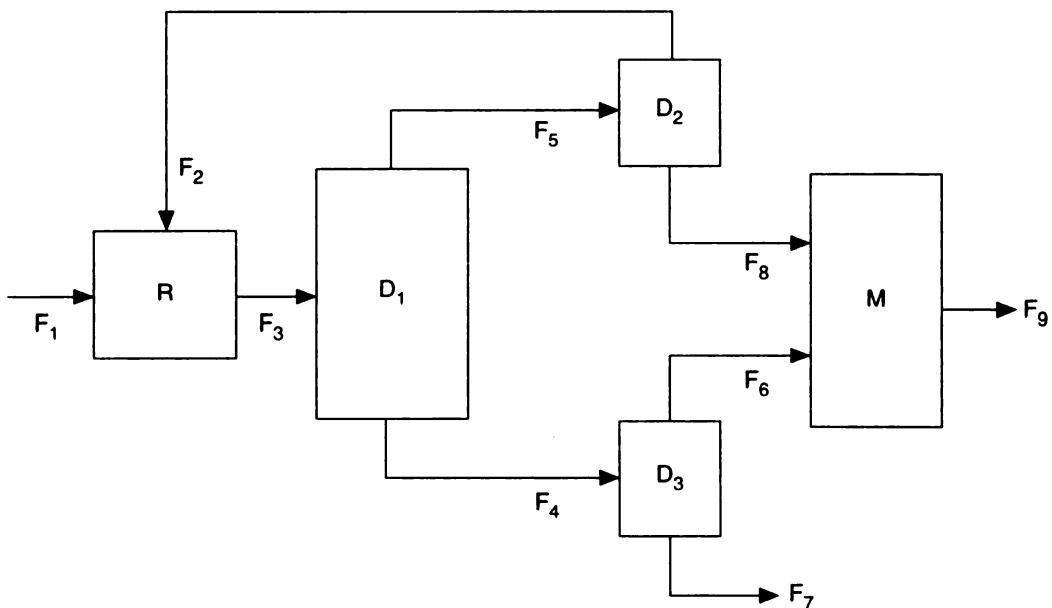
$$F_3 - F_2 = F_1, \quad F_3 - F_4 = F_5, \quad F_4 - F_7 = F_6, \quad F_2 + F_8 = F_5, \quad F_8 = F_9 - F_6 \quad (1.3a)$$

Here  $F_i$  represents the volumetric flow rate of the  $i$ th stream (see Reklaitis (1983) and Himmelblau (1962)). In (1.3a) at least four variables have to be specified or determined experimentally. The remaining five can then be estimated from the equations (model) which are generated by applying the principle of conservation of mass to each unit. This is in contrast to the purely experimental approach where we would have to measure all nine flow rates. Using the modelling approach can therefore result in a lower investment in the instrumentation of a process plant. Equations (1.3a) have been written such that  $F_1, F_5, F_6, F_9\dots$  are experimentally measured, i.e. they occur on the right-hand side. In vectorial form, this system reads as

$$\begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} F_2 \\ F_3 \\ F_4 \\ F_7 \\ F_8 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_5 \\ F_6 \\ F_5 \\ F_9 - F_6 \end{bmatrix} \quad (1.3b)$$

or  $Au = b$ .

The accuracy of our predictions using (1.3) can be tested by independently measuring the flow rates  $F_2, F_3, F_4, F_7, F_8$ . A discrepancy between the measured value and the predicted value may occur. Two plausible explanations for this could be: (i) Significant density variations in the different streams exist which we have neglected. (ii) Leaks are present in the plant network which we have considered to be nonexistent.



**Fig. 1.1** Typical process flowsheet ( $R$ : reactor;  $M$ : mixer;  $D_1, D_2, D_3$ : separator units; and  $F_i$ : flow rates).

The state of the plant in Fig. 1.1 cannot be described by measuring **any** set of four variables. The specification of  $F_8, F_5, F_2, F_1$ , for instance, does not allow us to determine the other five flow rates uniquely. The first three variables are “dependent” as they satisfy the fourth equation in (1.3). The problem hence reduces to determining five variables from the remaining four equations. This concept of “dependence” has wide applications in sensor placement in a chemical plant (see Madron and Veverka, 1992). This concept is discussed in detail in Chapter 2.

The system described above is similar to electrical networks, where the electric current flowing through each branch is the quantity of interest. Equations of the form (1.3) stem from an application of Kirchhoff's first law at a junction (see Del Toro, 1986). This analogy leads us to our next

example where we come across linear algebraic equations. It is based on Ohm's law. The voltage drop  $V$  across a resistor  $R$  follows the linear relation

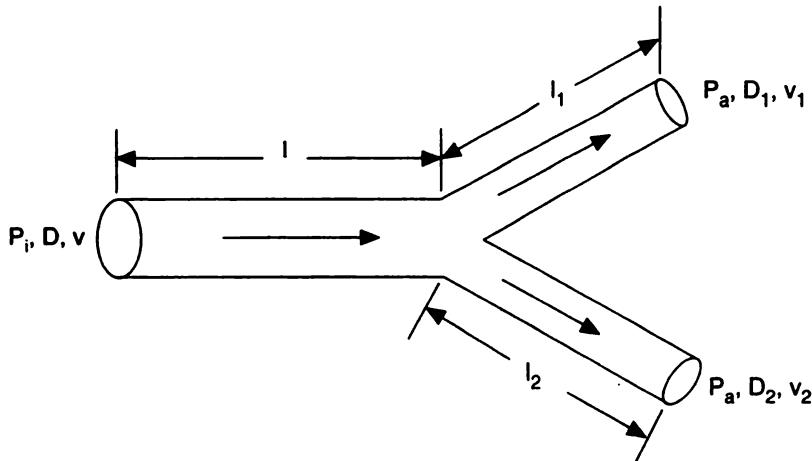
$$V = iR \text{ (Ohm's law)}$$

The hydrodynamic analogue of this equation relates the pressure drop to the mean velocity  $v$  for laminar flow in a pipe by

$$\Delta p = v \left( \frac{32\mu l}{D^2} \right)$$

This is the classical Hagen-Poiseulle equation (see Gupta and Gupta, 1984). The resistance offered to the flow ( $32\mu l/D^2$ ) is proportional to the pipe length and varies inversely as its area. This is analogous to the situation prevailing in electrical networks, where the electrical resistance has the same features (Halliday and Resnick, 1991).

Consider a liquid entering into a pipe of length  $l$  and diameter  $D$  at a fixed pressure  $P_i$ . The flow distributes itself into two pipes each of length  $l_1$  ( $l_2$ ) and diameter  $D_1$  ( $D_2$ ), see Fig. 1.2.



**Fig. 1.2** Flow distribution in a pipe network.

Neglecting pressure losses at the junction, and assuming the flow is laminar in each pipe, the macroscopic momentum balance and the mass balance at the junction yield

$$\left. \begin{aligned} P_i - P_a &= \left( \frac{32\mu l}{D^2} \right) v + \left( \frac{32\mu l_1}{D_1^2} \right) v_1 \\ P_i - P_a &= \left( \frac{32\mu l}{D^2} \right) v + \left( \frac{32\mu l_2}{D_2^2} \right) v_2 \\ D^2 v &= D_1^2 v_1 + v_2 D_2^2 \end{aligned} \right\} \quad (1.4a)$$

Here  $P_a$  is the pressure at which the fluid leaves the system at the two outlets. The set of three

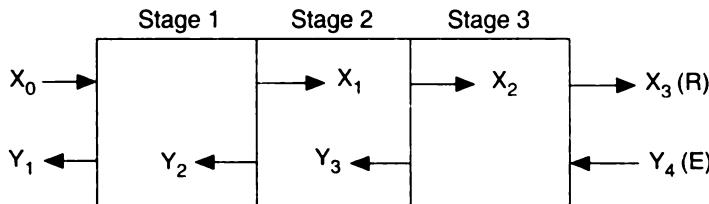
equations in (1.4a) can be solved for  $v$ ,  $v_1$ ,  $v_2$  for a fixed  $(P_i - P_a)$ . They can be recast vectorially as

$$\begin{bmatrix} \frac{32\mu l}{D^2} & \frac{32\mu l_1}{D_1^2} & 0 \\ \frac{32\mu l}{D^2} & 0 & \frac{32\mu l_2}{D_2^2} \\ -D^2 & D_1^2 & D_2^2 \end{bmatrix} \begin{bmatrix} v \\ v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} P_i - P_a \\ P_i - P_a \\ 0 \end{bmatrix} \quad (1.4b)$$

or  $Au = b$ .

Should the flow be turbulent, the friction factor  $f$  is not equal to  $64/Re_d$  anymore and is obtained from a nonlinear correlation or the Moody diagram (see Gupta and Gupta, 1984). This renders the equations (1.4a) nonlinear and prevents us from recasting the set (1.4a) in the vectorial form (1.4b) as done above.

Another example where linear algebraic equations occur is in an idealised extraction unit. A counter-current three-stage extraction unit is shown in Fig. 1.3. The component A is present in phase E (extract) along with a nondiffusing substance as a binary mixture. It is extracted into



**Fig. 1.3** A three-stage counter-current extraction unit ( $X_i$ ,  $Y_i$  are mole ratios leaving the  $i$ th stage).

phase R (raffinate) by a nondiffusing solvent. The streams leaving the  $i$ th stage are denoted by the subscript  $i$ . The mole ratio of A in phase E(R) is denoted by  $Y(X)$ . A material balance of component A across each extraction stage (see Treybal, 1981) yields

$$\left. \begin{array}{l} E_s Y_4 + R_s X_2 = R_s X_3 + E_s Y_3 \\ E_s Y_3 + R_s X_1 = E_s Y_2 + R_s X_2 \\ E_s Y_2 + R_s X_0 = E_s Y_1 + R_s X_1 \end{array} \right\} \quad (1.5a)$$

$Y_i(X_i)$  = moles of A/moles of non-A in phase E(R) coming out of stage  $i$  (this is a mole ratio and not a mole-fraction).

$E_s(Y_s)$  refers to flow rates in terms of moles of non-A in phase E(R) of each stage. This is a constant and does not vary between the different stages as only species A diffuses from one phase to another. The assumption of a linear equilibrium relationship (in terms of mole ratios) for the compositions leaving the  $i$ th stage yields three more equations

$$Y_i = KX_i \quad \text{for } i = 1, 2, 3 \quad (1.5b)$$

These, six equations in eight compositions ( $X_0, X_1, \dots, X_3, Y_1, Y_2, \dots, Y_4$ ) can be solved by specifying two variables. It is most convenient to specify  $Y_4$  and  $X_0$  the inlet compositions of the two phases.

The above set of equations can be written vectorially as

$$\begin{bmatrix} R_s & E_s & 0 & -E_s & 0 & 0 \\ K & -1 & 0 & 0 & 0 & 0 \\ -R_s & 0 & R_s & E_s & 0 & -E_s \\ 0 & 0 & K & -1 & 0 & 0 \\ 0 & 0 & -R_s & 0 & R_s & E_s \\ 0 & 0 & 0 & 0 & K & -1 \end{bmatrix} \begin{Bmatrix} X_1 \\ Y_1 \\ X_2 \\ Y_2 \\ X_3 \\ Y_3 \end{Bmatrix} = \begin{Bmatrix} R_s X_0 \\ 0 \\ 0 \\ 0 \\ E_s Y_4 \\ 0 \end{Bmatrix} \quad (1.5c)$$

or  $\dot{A}u = b$ .

The vector  $\{X_1, Y_1, X_2, Y_2, X_3, Y_3\}$  can be obtained for a given  $E_s, R_s, K$ . The assumption of a linear equilibrium relationship between  $Y_i$  and  $X_i$  (1.5b) could be invalid. This can give rise to a discrepancy in the predicted and actual system behaviour.

Systems of equations of the form (1.2) also arise while solving linear elliptic partial differential equations (see Chapter 5) by finite differences. In such situations the dimension of the system, i.e. number of equations 'n', can be very large.

### 1.2.2 Ordinary Differential Equations (Initial-Value Problems)

This class of equations describe the dependence of different quantities on one independent variable. Such problems usually occur in modelling the transient behaviour of spatially homogeneous systems. The independent variable here is time. Alternatively, in a steady state problem, this could be a spatial coordinate. Equations of this kind can be cast in the vectorial form

$$\dot{u} = Au + b \quad (1.6a)$$

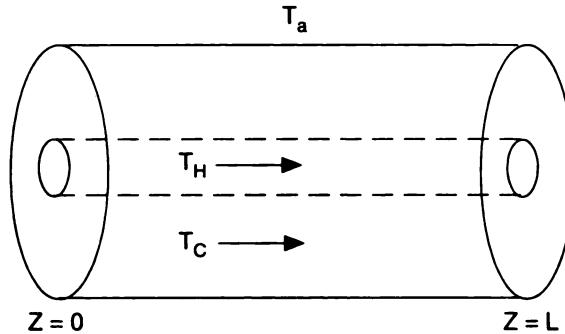
where the dot over  $u$  represents the derivative with respect to the independent variable. This equation represents a system of coupled first order linear ordinary differential equations as long as  $b$  is independent of  $u$ . In (1.6a), "b" is a nonhomogeneity. In a transient problem (when the independent variable is time), the evolution of each of the coordinates or elements of  $u$  with time is to be determined such that (1.6a) is satisfied subject to some conditions on  $u$ . In an initial-value problem, these conditions specify all the coordinates of  $u$  at the same value of the independent variable, usually time  $t$ . When the condition is specified at the point where the independent variable is zero, these take the form,

$$u(0) = u_0 \quad (1.6b)$$

Equation (1.6b) is called the initial condition of  $u$  (since usually the independent variable is time, and  $t = 0$  denotes the initial instant of time. The problem (1.6a)–(1.6b) is called an *initial value problem*.

A boundary-value problem, on the other hand, has some coordinates of  $u$  specified at one instant of time and the remaining ones at another instant of time. The treatment of boundary value problems is taken up in Chapters 5–8 and 9–12. In this section we restrict ourselves to initial value problems.

A co-current shell and tube heat exchanger at steady state is shown in Fig. 1.4. This is an example of a system where the model yields an initial value problem. The hot (cold) stream is



**Fig. 1.4** A co-current shell and tube heat exchanger.

flowing through the tube (shell) side (see Kern (1950) and Holman (1972)). Assuming that the heat transfer between the fluids (across the tube) is governed by an overall heat transfer coefficient  $U$ , and occurs across the perimeter  $P$ , the steady axial variation of temperature of the two streams is given by

$$\left. \begin{aligned} \rho_H C_{pH} q_H \frac{dT_H}{dz} &= -UP(T_H - T_C) \\ \rho_C C_{pC} q_C \frac{dT_C}{dz} &= UP(T_H - T_C) \end{aligned} \right\} \quad (1.7a)$$

Here  $\rho$ ,  $C_p$ ,  $q$  denote the density, specific heat and flow-rate of the streams. Subscripts H, C denote hot and cold streams, respectively. These equations are subject to initial conditions which specify  $T_H$ ,  $T_C$  at  $z = 0$ , the entrance point (since the heat exchanger is co-current).

$$T_H(z = 0) = T_{H0}, \quad T_C(z = 0) = T_{C0} \quad (1.7b)$$

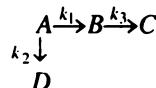
The system (1.7a) can be recast in vectorial form as

$$\begin{bmatrix} \dot{T}_H \\ \dot{T}_C \end{bmatrix} = \begin{bmatrix} -UP \\ \rho_H C_{pH} q_H \end{bmatrix} \begin{bmatrix} UP \\ \rho_C C_{pC} q_C \end{bmatrix} \begin{bmatrix} T_H \\ T_C \end{bmatrix} \quad (1.8)$$

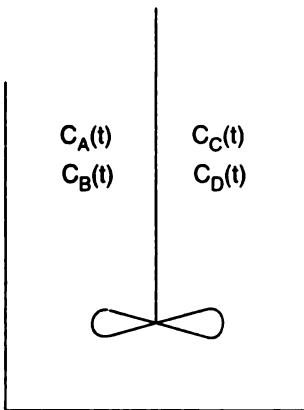
or  $\dot{u} = Au$ .

The independent variable in this problem is “ $z$ ”, the axial coordinate. This is an example of a homogeneous equation as it has no term independent of  $u$ , i.e.  $b = 0$  in (1.6a). This is a steady state model as  $T_H$ ,  $T_C$  are independent of time “ $t$ ”. The two conditions (1.7b) are both specified at  $z = 0$ . The system (1.8) is hence mathematically an initial-value problem, though the independent variable is not time there.

Another example where a similar system of equations is encountered is in the modelling of a batch reactor (Fig. 1.5) sustaining a network of single-phase first order reactions:



The reactor is assumed to be isothermal. The rate constants  $k_i$  can then be treated as constants.



**Fig. 1.5** A well-stirred batch reactor sustaining a network of reactions.

The time variation of the concentration  $C_i$  of the  $i$ th species in the reactor is given by the mass balance equations (see Fogler, 1992) as

$$\left. \begin{aligned} \frac{d}{dt} C_A &= -k_1 C_A - k_2 C_A, & \frac{d}{dt} C_B &= k_1 C_A - k_3 C_B \\ \frac{d}{dt} C_C &= k_3 C_B, & \frac{d}{dt} C_D &= k_2 C_A \end{aligned} \right\} \quad (1.9a)$$

These are subject to initial conditions which specify the initial state of the reactor

$$C_A(t = 0) = C_{A0}, \quad C_B(t = 0) = C_{B0}, \quad C_C(t = 0) = C_{C0}, \quad C_D(t = 0) = C_{D0} \quad (1.9b)$$

Equation (1.9a) can be recast in vectorial form as

$$\begin{bmatrix} \dot{C}_A \\ \dot{C}_B \\ \dot{C}_C \\ \dot{C}_D \end{bmatrix} = \begin{bmatrix} -k_1 - k_2 & 0 & 0 & 0 \\ k_1 & -k_3 & 0 & 0 \\ 0 & k_3 & 0 & 0 \\ k_2 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} C_A \\ C_B \\ C_C \\ C_D \end{bmatrix} \quad (1.10)$$

$$\dot{u} = Au.$$

This is an example of a dynamic system as the different variables evolve with time. The steady state problem of the heat exchanger is not a dynamic simulation. However, since it is a co-current system with all conditions specified at  $z = 0$ , it is an initial value problem. The solution method for both these systems are hence identical. The two systems (1.8), (1.10) are homogeneous since they are independent of the constant term 'b' in (1.6a), i.e. all terms contain the dependent variables and its derivatives linearly.

We next see instances of a system where nonhomogeneous equations arise. The equations of the heat exchanger, i.e. (1.7a), assume there is no loss from the insulated shell surface. Improper insulation results in heat loss to the ambient at  $T_a$ . This modifies the equation for  $T_C$  in (1.7a) as

$$q_C \rho_C C_{PC} \frac{dT_C}{dz} = UP(T_H - T_C) - U_1 P_1 (T_C - T_a)$$

where  $U_1, P_1$  is the overall heat transfer coefficient and perimeter of the shell surface respectively. In vectorial form the system (1.10) now becomes modified as

$$\begin{bmatrix} \dot{T}_H \\ \dot{T}_C \end{bmatrix} = \begin{bmatrix} -UP \\ \rho_H C_{PH} q_H \\ UP \\ \rho_C C_{PC} q_C \\ -UP - U_1 P_1 \\ \rho_C C_{PC} q_C \end{bmatrix} \begin{bmatrix} T_H \\ T_C \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{U_1 P_1 T_a}{r_C C_{PC} q_C} \end{bmatrix} \quad (1.11a)$$

or

$$\dot{u} = Au + b \quad (1.11b)$$

This modified system is clearly nonhomogeneous, as here the term “ $b$ ” of (1.6a) is nonzero.

The concentration of a reactant undergoing a first order reaction  $A \rightarrow B$  in an ideal isothermal continuous stirred tank reactor (CSTR), see Fig 1.6, is governed by the mass balance equation

$$V \frac{dC_A}{dt} = q(C_{Af} - C_A) - V k C_A \quad (1.12)$$

where

$V$  = reactor volume

$q$  = volumetric flow rate of reactant stream

$C_{Af}$  = concentration of feed stream

$k$  = reaction rate constant

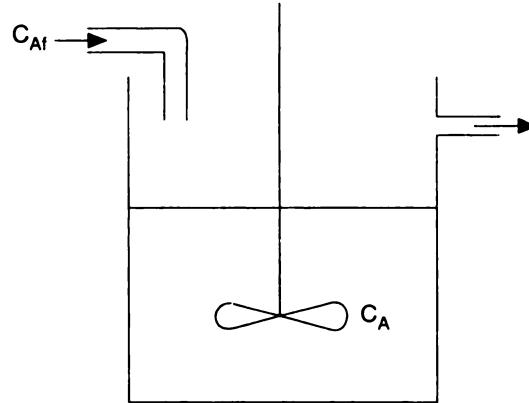


Fig. 1.6 A well-stirred continuous tank reactor.

The term  $qC_{Af}$  is independent of  $C_A$ , the dependent variable. This renders the system (1.12) nonhomogeneous.

Systems of the form (1.6) also arise while solving linear parabolic partial differential equations (see Chapter 5) using finite differences. The variables are discretised in the spatial direction. The system behaviour now is represented by a finite number of variables governed by equations which can be recast in the form (1.6a).

### 1.2.3 Partial Differential Equations

These equations occur widely in modelling systems in different areas of chemical engineering.

There is more than one independent variable in such problems. These are also more generically called distributed parameter systems as now the variables are dependent on spatial coordinates.

Consider a spherical pellet of radius  $R$  at a uniform temperature  $T_i$  dipped in a well-stirred bath maintained at a constant temperature  $T_a$ . The pellet temperature varies with time and position until it reaches the equilibrium value  $T_a$  uniformly inside. The variation with position occurs due to the finite value of thermal conductivity. The time and spatial dependence of pellet temperature  $T$  is governed by the energy balance equation

$$\frac{\partial T}{\partial t} = \alpha \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial T}{\partial r} \right) \quad (1.13)$$

Here,  $\alpha$  represents the thermal diffusivity of the pellet (see Bird et al., 1960). This is a typical example of a homogeneous linear parabolic partial differential equation.

Consider next a finite sized cylindrical pellet, for example ( $0 < z < 1$ ,  $0 < r < R$ ) which is being heated electrically. Assuming the internal rate of heat generation per unit volume  $q$  to be a constant, the steady temperature  $T(r, z)$  profile is obtained from the energy balance equation as

$$k \left( \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T}{\partial r} \right) + \frac{\partial^2 T}{\partial z^2} \right) + q = 0 \quad (1.14)$$

This is a nonhomogeneous elliptic partial differential equation. The presence of the constant source term  $q$ , which is independent of  $T$ , renders this equation nonhomogeneous. Similar equations also occur in modelling fluid flow problems.

Viscosity effects are normally neglected in fluids flowing far away from surfaces of bodies, conduits, etc. A valid assumption under these conditions is that the flow is irrotational. For an incompressible, irrotational fluid flow, the velocity field can be expressed as the gradient of a potential  $\phi$ . The potential for such flows satisfies the linear homogeneous elliptic Laplace's equation (see Gupta and Gupta, 1984)

$$\nabla^2 \phi = 0 \quad (1.15)$$

The drag force on a sphere immersed in a liquid flowing across it is a classical problem in fluid mechanics. The creeping flow (low Reynolds number or large viscosity limit) approximation for flow around a sphere is obtained by describing the flow in terms of  $\psi$ , the stream function. The stream function for this flow is obtained (see Bird et al., 1960) as the solution to

$$\left( \frac{\partial^2}{\partial r^2} + \frac{\sin \theta}{r^2} \frac{\partial}{\partial \theta} \left( \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \right) \right)^2 \psi = 0 \quad (1.16)$$

The partial differential equations (1.13–1.16) are to be solved subject to boundary conditions and/or initial conditions. We discuss these in detail in Chapter 5.

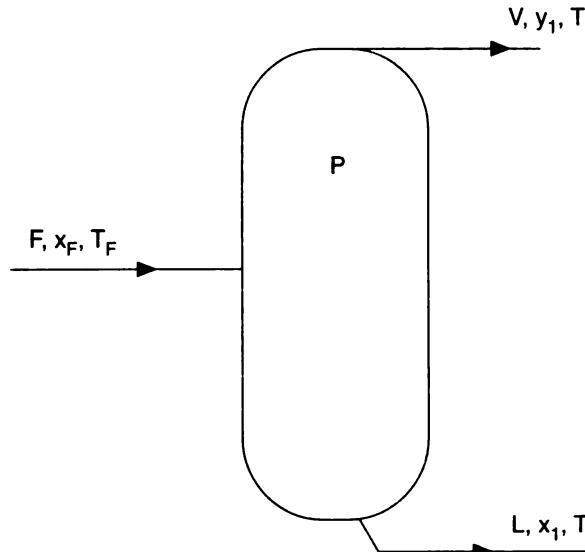
### 1.3 NONLINEAR EQUATIONS

The linear models we have presented so far are valid under certain restrictions or assumptions, i.e. laminar flow conditions, isothermal first order reactions, etc. These restrictions are very severe. Most systems in chemical engineering violate these assumptions and are consequently nonlinear. The presence of a single nonlinear interaction renders the entire system nonlinear. We shall now give some examples of realistic systems which are modelled by nonlinear equations.

### 1.3.1 Algebraic Equations

These equations arise in the steady state modelling of well-mixed systems. The steady state assumption eliminates the time dependence and the well-mixed assumption eliminates the spatial dependence. Alternatively, they also occur while numerically solving steady state models of spatially varying systems. Here the spatial dependence is discretised, and this results in a system of coupled nonlinear equations.

As a first example, consider the steady state of an adiabatic flash unit (Fig 1.7). A binary mixture is fed to the flash at a molar flow rate of  $F$  mols/hr feed and composition (mole-fraction) of  $x_F$ .



**Fig. 1.7** A typical single stage flash unit.

of A)  $x_F$ . The effluent liquid and vapour flow rates are  $L$ ,  $V$  mols/hr, respectively at a composition of  $x_1$ ,  $y_1$  (mole fraction of A). The conservation of mass and energy equations yield

$$F = L + V \quad (1.17a)$$

$$Fx_F = Lx_1 + Vy_1 \quad (1.17b)$$

$$FH_F(T_F) = LH_L(T) + VH_V(T) \quad (1.17c)$$

The total number of unknowns are  $L$ ,  $V$ ,  $x_1$ ,  $y_1$ ,  $T$ . To obtain these five unknowns, we need two more equations. These arise from the equilibrium relationships

$$y_1 = k_1 x_1 \quad (1.18a)$$

$$y_2 = k_2 x_2 \quad (1.18b)$$

along with

$$y_1 + y_2 = 1 \quad (1.18c)$$

$$x_1 + x_2 = 1 \quad (1.18d)$$

The system of equations (1.17) and (1.18) is nonlinear as it contains products of unknowns, i.e.,

$Lx_1, Vy_1, LH_L(T), VH_V(T) \dots$ . In addition to these, the dependence of the equilibrium constant  $k_i$  on  $x, T$ , and the enthalpy  $H_L, H_V$  on  $T$  can generate further nonlinearities.

Consider the pipe network problem discussed earlier, described by (1.4), see Fig. 1.2. The equations for the mean velocities were linear as we had assumed the flow to be laminar. When the flow is turbulent, the pressure drop in a pipe is given by the relation

$$\Delta p = f \frac{2l \rho v^2}{D} \quad (1.19a)$$

where the friction factor  $f$  is obtained from empirical correlations or Moody diagrams. Using the correlation

$$f = 0.3164 Re_D^{-0.25} \quad \text{for } Re_D < 10^5 \quad (1.19b)$$

where  $Re_D$  is the Reynolds number based on the pipe diameter, the equations (1.4) get modified as

$$P_i - P_a = 0.3164 \left( \frac{\mu}{\rho} \right)^{25} \frac{v^{1.75}}{D^{1.25}} \cdot \frac{l}{2g} + 0.3164 \left( \frac{\mu}{\rho} \right)^{25} \frac{v_1^{1.75}}{D_2^{1.25}} \frac{l_1}{2g} \quad (1.20a)$$

$$P_i - P_a = 0.3164 \left( \frac{\mu}{\rho} \right)^{25} \frac{v^{1.75}}{D^{1.25}} \frac{l}{2g} + 0.3164 \left( \frac{\mu}{\rho} \right)^{25} \frac{v_2^{1.75}}{D_1^{1.25}} \frac{l_2}{2g} \quad (1.20b)$$

Now a set of nonlinear equations (1.20a) and (1.20b) has to be solved along with the third equation in (1.4a) for the resulting flow distribution.

The presence of nonlinearities in this system prevents us from recasting the equations in a vectorial form as

$$Au = b$$

where  $b$  is a known vector and  $A$  an  $n \times n$  matrix.

### 1.3.2 Ordinary Differential Equations

Systems as mentioned above arise normally in dynamic simulations where the independent variable is time. We will now see two examples of systems modelled by coupled nonlinear first order ordinary differential equations.

The evolution of the reactant concentration  $C_A$  and temperature  $T$  in a CSTR (Fig. 1.6), sustaining an exothermic irreversible first order reaction is governed by the mass and energy balance equations. This (see Fogler 1992) takes the form

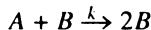
$$V \frac{dC_A}{dt} = q(C_{Af} - C_A) - V k_0 e^{-E/RT} C_A \quad (1.21a)$$

$$V \rho C_P \frac{dT}{dt} = q \rho C_P (T_f - T) + (-\Delta H) V k_0 e^{-E/RT} C_A - U A (T - T_C) \quad (1.21b)$$

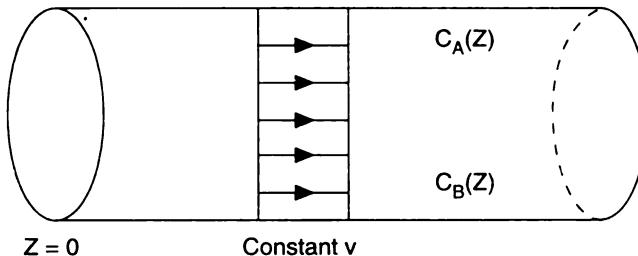
The subscript “ $f$ ” is used to denote feed conditions and  $T_C$  represents the coolant temperature. The heat transfer from the reactor to the coolant (not shown in Fig. 1.6) occurs over a surface area  $A$  and with an overall heat transfer coefficient  $U$ . The nonlinearity arises from the Arrhenius temperature dependency of the reaction rate  $e^{-E/RT}$ . This can be clearly seen by remembering that this exponential term has a Taylor series expansion in  $T$ . It is hence equivalent to an infinite order polynomial. A

similar situation (i.e. nonlinear interactions) prevails in an isothermal reactor sustaining a reaction whose order is different from one. In a network of reactions even if one reaction has an order other than one, and all others are first order, the resulting system of equations is nonlinear.

Consider next a plug-flow reactor (Fig. 1.8) sustaining the reaction



Here the velocity of the liquid is assumed to be invariant across the cross-sectional area.



**Fig. 1.8** A plug flow reactor.

For an isothermal operation, the variation of the concentration of  $A$ ,  $B$  with the axial position  $z$  is given by the mass balance equations

$$v \frac{dC_A}{dz} = -kC_A C_B \quad (1.22a)$$

$$v \frac{dC_B}{dz} = kC_A C_B \quad (1.22b)$$

The system of equations (1.22) is nonlinear due to the presence of the product term  $C_A C_B$ . It is solved by specifying the values of  $C_A$ ,  $C_B$  at  $z = 0$ . As both the conditions are specified at the same  $z$ , this is an initial value problem, like the co-current shell and tube-heat exchanger problem. The independent variable here is again a spatial direction.

Setting the time derivative to zero in the CSTR example, we obtain a system of algebraic equations. The solutions to this represent the steady state of the system. In the PFR example we cannot set the space derivative to zero. Equations (1.22) are representative of the steady state solution and neglecting the space derivative yields no meaningful result.

### 1.3.3 Partial Differential Equations

In most systems in fluid mechanics the velocity field is obtained by applying the principle of conservation of momentum and conservation of mass. The former is represented by the Navier-Stokes equation. For a Newtonian fluid where the shear stress varies linearly with the velocity gradient, the conservation of momentum is given by (see Bird et al., 1960).

$$\rho \left( \frac{\partial v}{\partial t} + v \cdot \nabla v \right) = \mu \nabla^2 v - \nabla p + \rho g \hat{k} \quad (1.23)$$

This is a vectorial equation and consists of a system of three coupled equations. This is nonlinear due to the second term occurring on the left. It contains the product of the dependent variable (the velocity) with its derivatives. This term is called the inertial term or the advective term. In the limit of low Reynolds number, the viscous effects dominate. The inertial terms can be neglected and the

resulting simplified equations become linear. For high Reynolds number, the inertial terms are dominant in comparison to the viscous term ( $\mu \nabla^2 v$ ). The effect of the advective nonlinearity cannot be neglected in this limit.

The heat conduction equation in the sphere (1.13) or any other body is rendered nonlinear if we include the effect of the temperature dependence of thermal conductivity. This is normally given by a power-law dependence (see Pushpavanam and Narayanan, 1988). Equation (1.13) gets modified now as

$$\rho c_p \frac{\partial T}{\partial t} = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 k(T) \frac{\partial T}{\partial r} \right) \quad (1.24)$$

where  $k(T) = aT^m$ .

This equation is nonlinear because it contains the product of the first derivative with another function of  $T$  and not due to the  $r^2$  multiplying the temperature derivative. If the thermal conductivity were to be dependent only on  $r$  and not on  $T$  (if the material is nonhomogeneous, and so its properties are nonuniform), (1.24) would be linear.

Every differential equation is solved subject to initial conditions and/or boundary conditions (see Chapter 5). These are dictated by the physical considerations of the situation prevailing in the system. A system with a **fixed boundary** is said to be linear if and only if both the differential equation and the boundary conditions are linear. If even one of them is non-linear, the system is termed nonlinear. An example of a situation where the differential equation is linear but the boundary condition is not, is a furnace brick transferring heat from its surface by radiation. The boundary condition here is of the form (see Bird et al., 1960)

$$-k \frac{\partial T}{\partial x} = e\sigma(T^4 - T_a^4) \quad (1.25)$$

where ' $\sigma$ ' represents the Stephan-Boltzman constant and  $e$  the emissivity of the surface.

The temperature profile in the brick is governed by the linear heat conduction equation similar to (1.13). Such a problem is nonlinear due to the boundary condition of the form (1.25) and cannot be solved by the methods presented in Chapters 5–8.

**Example 1.1** Consider the steady radial flow of air between two long parallel co-axial cylinders. The temperature of the inner cylinder is  $T_i$  and of the outer cylinder is  $T_o$ . Are the governing equations for the velocity and temperature profiles in the annular region linear?

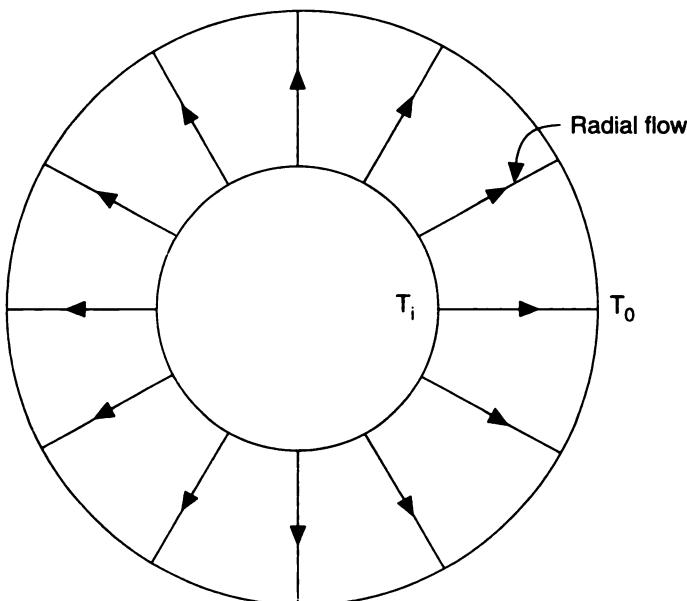
The equations of continuity and the energy balance yield

$$\frac{d}{dr}(r v_r) = 0 \quad (1.26a)$$

$$\rho c_p v_r \frac{dT}{dr} = \frac{k}{r} \frac{d}{dr} \left( r \frac{dT}{dr} \right) \quad (1.26b)$$

Here the dependent variables are  $v_r$ ,  $T$ , and the independent variable is  $r$ . Clearly the first equation is linear and the second nonlinear, since the latter contains the product of  $v_r$  and  $dT/dr$ . So if we were solving the two equations simultaneously we would be solving a system of nonlinear equations.

The first equation can be solved directly for  $v_r$ . The solution from this can be used to obtain  $T$  from the second equation (1.26b). Since now  $v_r$  is known, this method renders the second equation linear. This approach results in solving two linear equations sequentially.



**Fig. 1.9** Radial flow of air in an annular cylinder.

Moving boundary problems occur in applications to systems like crystal growth and noncatalytic gas-solid reactions. The domain of the problem over which these equations are valid varies with time. The growth of the crystal leads to an increase in the domain over which the governing differential equations hold. The progress of the reaction in a non-catalytic gas-solid reaction results in the depletion of the solid reactant. The domain over which the differential equations are valid decreases now. If the variation of the domain is determined by the dependent variable and is not predetermined due to external considerations we have a nonlinear problem. Problems of this kind are beyond the scope of this text.

The steady state behaviour of separation systems like distillation columns, extraction units, absorption columns are governed by algebraic equations. These characteristics can be studied by employing the methods presented in this text. The dynamic behaviour of such systems is, however, modelled by differential algebraic equations. The differential equations arise from the principles of conservation of mass and energy and the algebraic equations from equilibrium conditions. The methods discussed in this text cannot be directly and easily extended to analyse DAEs.

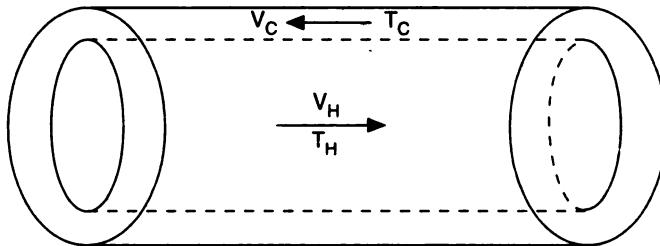
The examples seen in this chapter so far illustrate the different kinds of equations which arise in various chemical engineering contexts. The complexity of the various classes of equations seems to be different. The partial differential equations that we will be dealing with in this book have one dependent variable and more than one independent variable, while the algebraic systems have no independent variable but many dependent variables.

In this text, we develop a theory and a universal analytical method of solution to these apparently "different" physical and mathematical systems. This is rendered possible by using the notion of operators. Chapters 1–4 deal with finite dimensional systems, i.e. systems of linear algebraic equations, and ordinary differential equations (initial value problems). The different concepts are presented in general terms in the operator framework. This allows their extension to Chapters 5–8 which deal with infinite dimensional systems, i.e. partial differential equations. Chapters 9–12

cover nonlinear equations. Here again we demonstrate how the techniques developed on algebraic systems can be extended to differential equations. The theory of linear equations developed in the earlier sections is applied to gain insight into the behaviour of these nonlinear systems.

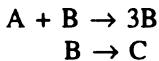
## PROBLEMS

1. A counter-current heat exchanger is shown in Fig. 1.10. It is at steady state and the inlet temperature of both streams are specified. Write down the differential equation governing the temperature of the two streams. Is this an initial value problem?



**Fig. 1.10** A counter-current heat exchanger.

2. The liquid phase autocatalytic reaction



- (i) occurs in a batch reactor, and
- (ii) a continuous stirred tank reactor.

Derive the governing equations for the two systems. Assume that the inlet concentrations of A, B are given by  $C_{A_f}$ ,  $C_{B_f}$  and initial concentrations by  $C_{A_0}$ ,  $C_{B_0}$ . Assume  $q$  to be the volumetric flow rate into the reactor,  $V$  to be the reactor volume. Is this system linear or nonlinear? Is it an initial-value problem? Would the equations change if the reaction were to be in the gas phase. Explain qualitatively.

3. An immersion heater generates  $q$  watts and is immersed in an insulated bucket containing  $V$  litres of well-stirred water. Obtain the equation determining the evolution of temperature in the bucket. Is this equation linear, is it nonlinear? Is this homogeneous? Does the system have a steady state? Discuss. Can you render it homogeneous?

4. The hot water (refer Problem 3) is cooled by removing the insulation on the bucket. The overall heat transfer coefficient is  $U$ . This heat loss occurs across a surface area  $A$ , to the ambient air (at  $T_a$ ). Write the equation governing the temperature profile of the bucket. Is this linear? Is this homogeneous. Can this equation be rendered homogeneous? Does the system have a steady state? Discuss.

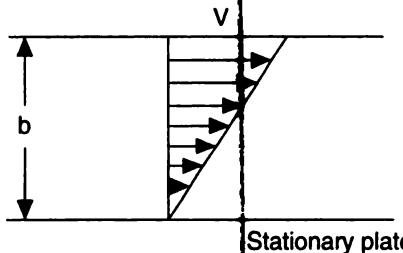
5. In process control we study the effect of disturbances on the system performance. The frequency response is obtained by varying a parameter periodically. The transient response to such an input reveals dynamic features of the system. The equation that arises in such a study is of the form

$$\frac{dx}{dt} = k \sin \omega t - x$$

Classify this equation as linear/nonlinear, homogeneous/nonhomogeneous.

6. Consider a two-dimensional flow field. Here the stream function  $\psi$  is defined to satisfy the equation of continuity. Write down the governing equation for  $\psi$  an irrotational flow field. Is this a linear homogeneous equation?

7. Consider the Couette flow of a fluid confined between two parallel plates. The lower plate is at rest and the upper is moving at a velocity  $V$  to the right (Fig. 1.11). Heat is generated in the



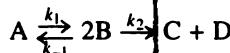
**Fig. 1.11** Couette flow between two parallel plates.

fluid due to viscous dissipation. The temperature 'T' here is determined from

$$-k \frac{d^2 T}{dx^2} = \mu \left( \frac{V}{b} \right)^2$$

Classify this equation. Is it linear? Is it homogeneous? List all assumptions made in deriving the equation. Here  $b$  is the separation between the plates;  $k$  and  $\mu$  represent the thermal conductivity and viscosity of the fluid.

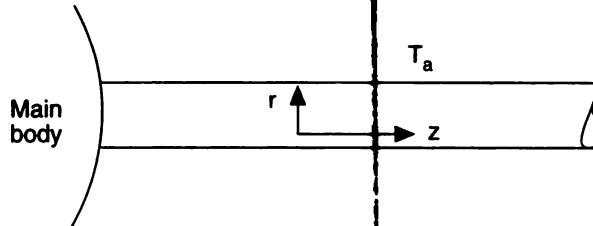
8. Consider a well-stirred continuous reactor sustaining the reactions (elementary)



Feed to the reactor is pure A at a concentration of  $C_{A0}$ . Write down the equations that describe the evolution of the concentrations of A, B, C when the residence time in the reactor is ' $\tau$ '.

9. (i) Consider the heat transfer occurring across a fin which is cylindrical in cross-sectional area. Write down the heat conduction equation assuming the temperature to vary axially and with the radial co-ordinate. Is this a linear homogeneous equation? Write the boundary conditions, when Newton's law of cooling is valid on the surface (Fig. 1.12) exposed to ambient air at  $T_a$ .

(ii) Assuming the temperature to vary only in the axial direction, determine the equation for the temperature profile. Is this a homogeneous equation. Under what conditions is this assumption (of no radial dependence) justified? Explain physically.



**Fig. 1.12** Heat transfer across a fin.

**REFERENCES**

- Aburdeen Maurice, F., Computer Simulation of Dynamical Systems, Wm. C. Brown Publishers, Dubuque (Iowa) (1988).
- Aris, R., Mathematical Modelling Techniques, Pitman, London (1978).
- Bird, R.B., Stewart, W.E. and Lightfoot, E.N., Transport Phenomena, Wiley (International Edition), New York (1960).
- Del Toro, V., Electrical Engineering Fundamentals, Prentice-Hall of India, New Delhi (1986).
- Fogler, H.S., Elements of Chemical Reaction Engineering, Prentice-Hall of India, New Delhi (1992).
- Gupta, V. and Gupta, S.K., Fluid Mechanics and Its Applications, Wiley Eastern, New Delhi (1984).
- Himmelblau, D.M., Basic Principles and Calculations in Chemical Engineering, 6th Ed., Prentice-Hall of India, New Delhi (1989).
- Holman, J.P., Heat Transfer, McGraw-Hill, New York (1972).
- Kern, D.Q., Process Heat Transfer, McGraw-Hill, New York (1950).
- Kersten, R.D., Engineering Differential Systems, McGraw-Hill, New York (1969).
- Lee, C.K., Yeung, S.Y.S. and Bailey J.E., Experimental studies of consecutive-competitive reaction in steady-state and forced periodic CSTRs, *The Canadian Journal of Chemical Engineering*, **58**, 212 (1980).
- Madron, F. and Veveka, V., Optimal selection of measuring points in complex plants by linear models, *AIChE*, **38**, 227 (1992).
- Pushpavanam, S. and Narayanan, R., Uniqueness conditions for steady state solutions of  $m$ th order reaction-non-isothermal pellets with variable transport coefficients, *Chemical Engineering Science*, **43**, 394 (1988).
- Reklaitis, G.V., Introduction to Energy and Material Balances, Wiley, New York (1983).
- Resnick, R. and Halliday, D., Physics, Part 2, Wiley Eastern, New Delhi (1991).
- San, K.Y. and Stephanopoulos, A note on the optimality criterion for maximum biomass production in a fed-batch fermentor, *Biotechnology and Bioengineering*, **XXVI**, 1261 (1984).
- Treybal, R.E., Mass-Transfer Operations (International Student Edition), McGraw-Hill, New York (1981).

## 2

# Vector and Vector Spaces

---

We came across several systems in chemical engineering which were described by linear equations in Chapter 1. Such equations are typical and they arise in modelling many systems in chemical engineering as well as other disciplines. The first two sections of this book are concerned with analytical methods of solutions to linear equations which arise in modelling of systems. We develop these methods in a general setting and apply it to a wide class of linear equations—algebraic equations, ordinary differential equations, and partial differential equations. This demonstrates the universal nature and the common basis of the traditional methods used in solving these systems.

The universal technique we discuss is based on concepts from linear algebra and their generalisation. In Chapters 2–4, we restrict ourselves to problems with finite degrees of freedom. These are governed by linear algebraic equations and linear ordinary differential equations. These arise in the steady state modelling and the transient analysis of lumped systems, respectively. The spatial dependence of the variables in such systems is neglected, but the time dependence is retained. Examples of such systems include distillation columns, stirred tank reactors, and so on. The number of dependent variables in all these systems is finite. Such systems also occur in the solution of partial differential equations using numerical techniques as finite differences etc. The dependent variables describe a physical system and hence are restricted to be necessarily real throughout this book. In this section we investigate equations of the form

$$\frac{du}{dt} = Lu + b \quad (2.1)$$

where  $u$ ,  $b$  are vectors each with  $n$  coordinates and  $L$  is an  $n \times n$  matrix.

Our universal method of solution of these different systems is rooted deeply in concepts of vector spaces and their generalisation. These are developed in the rest of this chapter. In Chapter 3 we discuss the matrix as a linear operator. The properties and definitions of the linear operator are introduced in a general setting. This enables their ready extension to infinite dimensional operators which occur when dealing with partial differential equations in the next section. The method of solution for equations of the form (2.1) is detailed in Chapter 4, with reference to typical problems arising in chemical engineering.

A word on system dimension is in order at this stage. The dimension or order of a physical system is used to denote the number of dependent variables which interact with each other and determine system behaviour. It does not represent the geometrical dimension, i.e. the number of spatial coordinates. The dimension of a vector space, on the other hand, is used to denote the number of coordinates of each element belonging to that space.

## 2.1 VECTORS

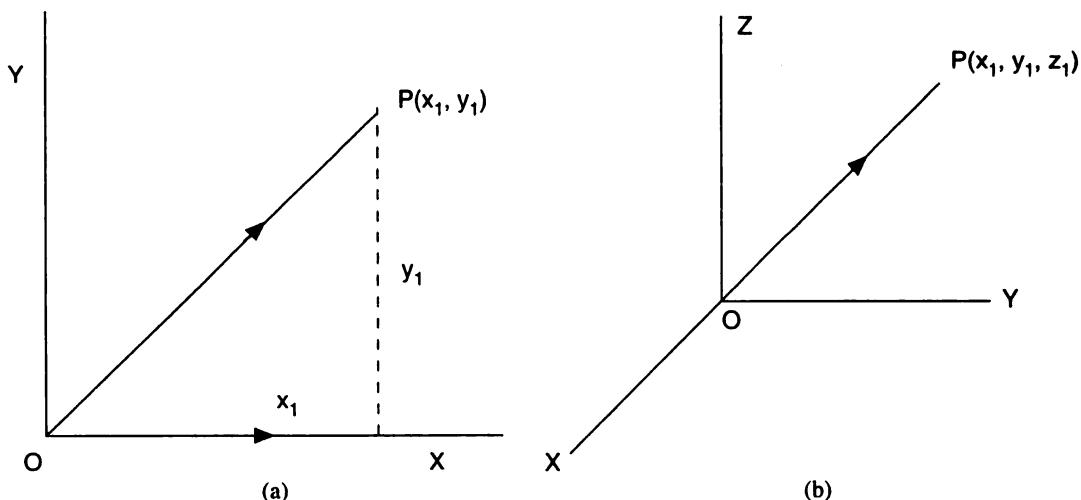
Students of physics like to think of vectors as a directed line segment, i.e. quantities which have magnitude as well as direction. Mathematics students visualise vectors as an ordered  $n$ -tuple of numbers written as a column or a row. These are different but equivalent ways of representing the same quantity. The former is called the geometric representation and the latter the algebraic representation. Each has its own advantages. The former helps in visualising and understanding the different concepts like distance, length, etc. in a vector space. The extension of these concepts and their generalisation to higher dimensional spaces is possible only because of the latter. We discuss both these methods of representing a vector.

Consider a plane with two perpendicular axes  $OX$ ,  $OY$  (Fig. 2.1a). The origin  $O$  is the point of intersection of the two axes. A vector  $OP$  in this plane is represented by a directed line starting from  $O$ . Its magnitude and direction are given by the length and direction of the line segment  $OP$ . This is the geometric representation of a vector in the plane which is a two-dimensional space.

The same vector can also be represented by the coordinates of the point  $P(x_1, y_1)$ . The two coordinates  $(x_1, y_1)$  represent the distance of  $P$  from  $O$  measured parallel to the two axes  $OX$ ,  $OY$ , respectively. A vector in this space is an ordered pair (2-tuple) of numbers written as a column  $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$  or a row  $[x_1, y_1]$ . This is the algebraic representation of the vector  $OP$ .

Consider three mutually perpendicular axes  $OX$ ,  $OY$ ,  $OZ$  in the three-dimensional space we live in (Fig. 2.1b). A point  $P$  in this space is represented by its coordinates  $(x_1, y_1, z_1)$  or an ordered 3-tuple. The coordinates measure the distance of  $P$  from  $O$  measured parallel to the three axes. The ordering is important as the first coordinate always denotes the distance parallel to the  $OX$  axis, the second, the distance parallel to the  $OY$  axis, and so on. This is the algebraic representation of the vector. The geometrical representation here again is the directed line segment  $OP$ .

We have discussed the geometric and algebraic representations of vectors so far, for low dimensional vector spaces. We next describe the algebraic properties in detail. This necessitates the introduction of many formal mathematical concepts. A detailed formal discussion of these concepts can be found in Murdoch (1965) and Kreyszig (1982). These definitions are important since they



**Fig. 2.1** Vector space illustration of geometric and algebraic representation of vectors: (a) Two-dimensional; (b) Three-dimensional space.

allow us to easily extend the methods and techniques developed in the context of finite dimensional vector spaces to infinite dimensional systems, or partial differential equations. The theory we develop allows us to discuss developments in general terms, irrespective of whether the operator  $L$  in (2.1) is a matrix or a differential operator.

## 2.2 VECTOR SPACES

The notion of vector space will now be formally defined. This is based on the concept of closure.

**Definition 2.1: Closure.** A set of numbers is said to be closed under an operation when any two elements of the set subject to the operation yields a third element belonging to the same set. For example, the set of integers is closed under the operation of addition as the sum of two integers yields a third integer. Closure under other operations as subtraction and multiplication is defined in a similar way. A set of numbers is closed under division if for every  $a, b$  belonging to the set with  $b \neq 0$ ,  $a/b$  is a member of the set. The set of all integers I (positive, negative and zero) is not closed under division. The integers 2, 3 give rise to 2/3 or 3/2 which are both nonintegers and do not belong to the set I. They are, however, closed under addition.

**Definiton 2.2: Field.** A field is a set of numbers closed under addition, subtraction, multiplication and division. The set of integers is not closed under division. It is hence not a field. The set of real (complex) numbers forms a field and is denoted by  $\mathbf{R}(\mathbf{C})$ . More details on the properties of the elements of a field can be found in Murdoch (1970), Noble and Daniel (1977).

**Definition 2.3: Scalar multiplication.** A scalar  $k$  multiplying a vector  $u$  yields another vector whose length is increased (or decreased) by a factor of  $|k|$ . If  $k$  is positive (negative), the direction of the vector is unchanged (reversed) in the geometric representation. In the algebraic representation, each coordinate of the new vector is  $k$  times the corresponding previous coordinate.

**Definition 2.4: Vector addition.** Two vectors  $OP$  and  $OQ$  can be added in the geometric representation using the parallelogram law. The first vector is drawn starting at the origin  $O$ . The second vector is drawn parallel to itself, starting from the terminating point of the first vector ( $P$ ) and terminating at  $R$ . The line segment  $OR$  represents the vectorial sum in the geometric representation (Fig. 2.2).

In terms of coordinates if  $(x_1, y_1)$  represent the coordinates of the first vector,  $(x_2, y_2)$  that of the second vector then  $(x_1 + x_2, y_1 + y_2)$  represent the coordinates of the vector sum. The coordinates of the sum of two vectors are equal to the sum of their coordinates. This can be readily verified from Fig. 2.2. Vector addition also obeys the commutative and associative laws

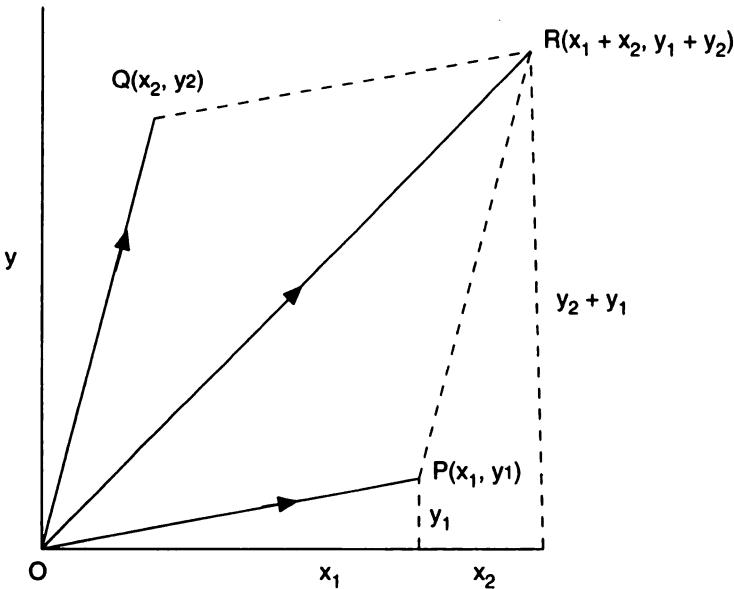
$$u + v = v + u$$

$$u + (v + w) = (u + v) + w$$

The algebraic representation of a vector as an  $n$ -tuple allows the easy verification of these laws. We have now established all the concepts required to define a vector space (see Murdoch (1970) and Stakgold (1979)).

A vector space or a linear space has two kinds of elements:

1. A field  $\mathbf{F}$  which consists of scalars. These could be real or complex, i.e.  $\mathbf{F}$  can be  $\mathbf{R}$  or  $\mathbf{C}$ .
2. A set of vectors  $\mathbf{V}$ , which is closed under the operations of scalar multiplication and vector addition.



**Fig. 2.2** Geometric representation of vector addition in two-dimensional spaces.

This implies that, for  $c_1, c_2 \in \mathbb{F}$  and  $u^1, u^2 \in \mathbf{V}$ , the vectors  $c_1u^1$ ,  $c_2u^2$  and  $c_1u^1 + c_2u^2$  belong to  $\mathbf{V}$ . This is true for all  $c_i \in \mathbb{F}$  and  $u^i \in \mathbf{V}$ . In this section of the book we use the first few letters of the alphabet  $b, c$  to denote scalars in  $\mathbb{F}$  and the last few alphabets  $u, v$  to denote vectors in  $\mathbf{V}$ . The superscript ‘ $i$ ’ in the vector indicates the  $i$ th vector.

$n$ -dimensional real (complex) vectors are said to belong to the vector space denoted by  $\mathbf{R}^n$  ( $\mathbf{C}^n$ ). Every vector in  $\mathbf{R}^n$  ( $\mathbf{C}^n$ ) has  $n$ -coordinates, each of which is a real (complex) number. More specifically, the vectors in the plane belong to  $\mathbf{R}^2$ . For the present we take the dimension of a vector space to be synonymous with the number of coordinates of a vector in a vector space. The formal definition of dimension of a vector space is based on the algebraic concept of linear dependence and will be introduced later.

### 2.3 METRICS, NORMS AND INNER PRODUCTS

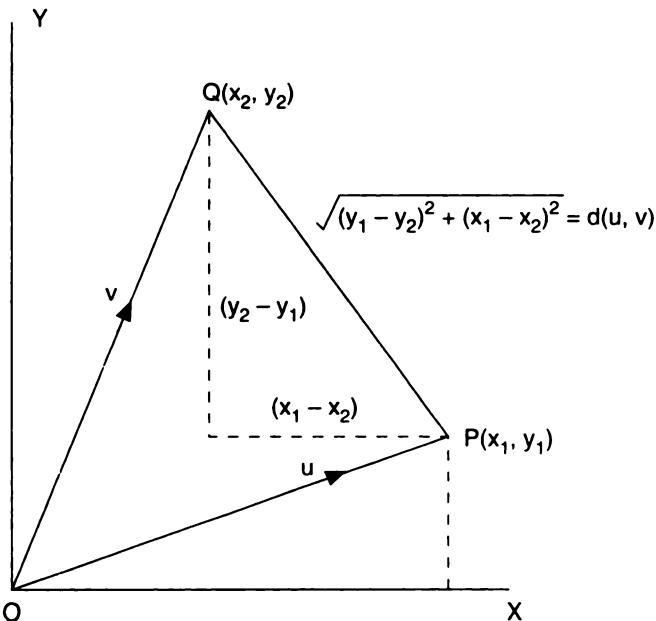
Concepts like distance between vectors, length of a vector, angle between vectors will now be explained in two- and three-dimensional spaces using the geometric representation (Murdoch (1965)). They will be extended to higher dimensional spaces using the algebraic representation (see Stakgold (1979), Naylor and Sell (1971)).

**Distance between points (Vectors):** Let  $u(x_1, y_1)$ ,  $v(x_2, y_2)$  be two vectors in  $\mathbf{R}^2$  as shown in Fig. 2.3. Each vector has its starting point at the origin  $O$ . The distance between the vectors denoted by  $d(u, v)$  is defined *classically* as

$$d(u, v) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2.2a)$$

This definition can be extended directly to a three-dimensional space  $\mathbf{R}^3$  as

$$d(u, v) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \quad (2.2b)$$



**Fig. 2.3** Geometric representation of distance or metric between vectors,  $d(u, v)$ .

$d(u, v)$  is a scalar and is also called the metric. The definition in (2.2a) and (2.2b) is called the Euclidean metric. It represents the geometric distance between the terminal points of the vectors.

**Length.** The geometric (Euclidean) length of a vector  $u$  is the distance of its terminal point from the origin. This is also called the norm of the vector and is denoted by  $\|u\|$ . Since the origin has zero as its coordinates, we have

$$\|u\| = \sqrt{x_1^2 + y_1^2} \quad \text{in } \mathbb{R}^2 \quad (2.3a)$$

$$= \sqrt{x_1^2 + y_1^2 + z_1^2} \quad \text{in } \mathbb{R}^3 \quad (2.3b)$$

This is identical to the length of a line in Euclidean geometry and so (2.3) is called the Euclidean norm.

**Angle.** The angle between two vectors ( $u$  and  $v$ ) is indirectly given by the inner-product between them. This is a measure of the projection of a vector  $u(x_1, y_1)$  in the direction of the vector  $v(x_2, y_2)$  and is denoted by  $\langle u, v \rangle$ . For vectors in a plane, this is defined as

$$\langle u, v \rangle = x_1 x_2 + y_1 y_2 \quad (2.4a)$$

The inner-product is a scalar. Equation (2.4a) is analogous to (in fact, a scalar multiple of) the dot product between vectors in physics. The inner-product directly yields the cosine of the angle  $\theta$  between the two vectors as given by

$$\cos \theta = \langle u, v \rangle / \|u\| \|v\| \quad (2.4b)$$

The length of a vector, the distance and the angle between two vectors have been defined using the algebraic representation (i.e. their coordinates). The geometric representation allows us

to visualise these in the lower dimensional spaces  $\mathbf{R}^2$ ,  $\mathbf{R}^3$ . These concepts can be extended to higher dimensions easily, using the algebraic representation as column vectors.

For notational convenience, from now on we represent the  $i$ th coordinate of the row vector  $u$  by  $u_i$  and of  $v$  by  $v_i$ , i.e.

$$u = (u_1, u_2, \dots, u_n), \quad v = (v_1, v_2, \dots, v_n)$$

This differs from the notation used so far, where  $x$  represents the first coordinate,  $y$  the second coordinate, etc. The change in the notation, to which we will adhere for the rest of the book, is to introduce mathematical elegance and facilitate generalisation. We do not restrict ourselves to a maximum dimension of 26, which is what we would have to if we had used the English alphabets to represent the different coordinates. A further word on notation is necessary before we continue any further. A subscript  $i$  denotes the  $i$ th coordinate and a superscript  $j$  the  $j$ th vector. So  $x^j$  is the  $j$ th vector, whereas  $x_i$  is a scalar representing the  $i$ th coordinate of the vector  $x$ . Capital letters will be used to denote operators, i.e. matrix operators, differential operators or boundary operators.

We digress from this notation when we talk of maps. While dealing with maps and the Newton-Raphson method (see Chapter 10) the subscript ' $n$ ' is used to refer to the  $n$ th iterate of the map, and not the  $n$ th coordinate of a vector. Most of our discussion will centre on one-dimensional maps. Hence there is no vector. This change in notation is necessitated by the need for consistency with literature. We make specific mention of this again in Chapter 10 to eliminate any confusion that may exist in the mind of the reader. Also, by dimension of a system, we mean the number of independent state variables needed to describe the system completely. A finite-dimensional system is modelled by a finite number of ordinary differential equations or algebraic equations. We are primarily concerned with such systems in Chapters 1–4 and 9–12. An infinite-dimensional system, on the other hand, is modelled by an infinite number of equations. These occur while solving boundary value problems. We would like to emphasise that by "infinite dimensional" we do not mean spatially unbounded or extending to infinity in space. Our systems can be spatially bounded and still be infinite dimensional. Here the system state is determined by specifying the values of a variable in an interval (or an infinity of points). These variables interact with each other to determine system behaviour. This will be discussed again in Chapter 6.

The ordered  $n$ -tuple, i.e. a vector can be written as a row vector or a column vector. A distinct identity is maintained for a row vector as opposed to a column vector. This is necessary to carry out algebraic operations on the vectors. Transposing a column vector, i.e. writing the column as a row while preserving the order of the coordinates, yields a row vector. Similarly, a row vector can be transposed to yield a column vector. The transpose operation is denoted by a superscript  $t$  as

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}' = (x_1, x_2, \dots, x_n) \quad (2.5a)$$

and

$$(x_1, x_2, \dots, x_n)' = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad (2.5b)$$

### 2.3.1 Metric Space

The Euclidean distance defined in (2.2) can be extended to higher dimensional spaces. The distance  $d(u, v)$  between two vectors  $(u, v)$  in  $\mathbb{R}^n$  is the positive scalar defined as

$$d(u, v) = \left[ \left( \sum_{i=1}^n (u_i - v_i)^2 \right) \right]^{1/2} \quad (2.6a)$$

In higher dimensional spaces we lose the geometric significance of the Euclidean distance as we cannot visualise the space. Equation (2.6a) is a general and abstract representation of the distance between  $u, v$ . The loss of the physical significance of distance prompts us to seek other permissible definitions in  $\mathbb{R}^n$ . These can be allowed as long as they satisfy certain axioms. The axioms ensure that the definitions do not violate the basic properties which we expect a function such as distance to satisfy. The definitions are a representative measure of the distance and will not correspond to the Euclidean (geometrical) distance between vectors. These should also be valid for lower, i.e. two- and three-dimensional spaces.

A metric  $d(u, v)$  is defined so as to satisfy the following axioms:

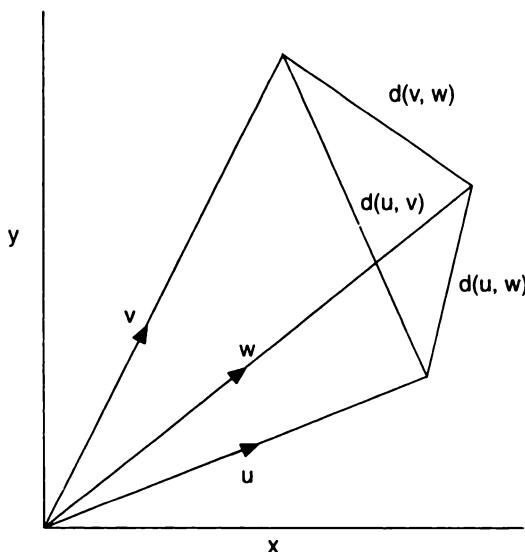
D1  $d(u, v) \geq 0$

D2  $d(u, v) = 0$  implies  $u = v$

D3  $d(u, w) \leq d(u, v) + d(v, w)$

D4  $d(u, v) = d(v, u)$

These axioms are mathematical statements of the properties the distance between two vectors must possess. D1 states that the distance between two points cannot be negative. D2 states, if the distance between two points is zero, then the two points are identical. D3 is the mathematical representation of the triangle inequality, i.e. two sides of a triangle are always greater than the third side (Fig. 2.4). The last axiom says that the distance between  $u$  and  $v$  is the same as that between  $v$  and  $u$  (see Stakgold, 1979).



**Fig. 2.4** Illustration of the triangle inequality in a two-dimensional space.

A metric space consists of a set of elements where a metric is defined. This does not have to be a linear space or a vector space as the axioms (D1–D4) do not involve concepts of scalar multiplication or vector addition. The Euclidean metric defined on  $\mathbb{R}^n$  in (2.6a) is denoted by  $d_2(u, v)$ . Examples of other metrics on  $\mathbb{R}^n$  are

$$d_1(u, v) = \sum_{i=1}^n |u_i - v_i| \quad (2.6b)$$

$$d_\infty(u, v) = \max_{i \in 1, n} |u_i - v_i| \quad (2.6c)$$

**Example 2.1** Determine if  $d(u, v) = \sum_{i=1}^4 (u_i - v_i)$  is a suitable metric on  $\mathbb{R}^4$ .

$d(u, v) = \sum_{i=1}^4 (u_i - v_i)$  must satisfy the four axioms for all vectors in  $\mathbb{R}^4$ . It is easy to demonstrate that the axioms are violated by taking an example. Let  $u = (0, 1, 0, 0)'$  and  $v = (0, 2, 0, 0)'$ .

$$\begin{aligned} d(u, v) &= -1 \\ &< 0, \text{ for } u \neq v \end{aligned}$$

The first axiom is violated, hence this is not an appropriate metric.

**Example 2.2** Compute  $d_1(u, v)$ ,  $d_2(u, v)$ ,  $d_\infty(u, v)$  for

$$u = (1 \ 2 \ 3 \ 4)', v = (3 \ 5 \ 2 \ 1)'$$

$$d_1(u, v) = \sum_{i=1}^4 |u_i - v_i| = |1 - 3| + |2 - 5| + |3 - 2| + |4 - 1| = 9$$

$$d_2(u, v) = \left( \sum_{i=1}^4 (u_i - v_i)^2 \right)^{1/2} = \sqrt{23}$$

$$d_\infty(u, v) = \max_{i \in 1, 4} |u_i - v_i| = 3$$

For the same two vectors  $(u, v)$ , the three metrics yield different values for distance. They are a measure of the separation, the difference between  $u$  and  $v$ . All three metrics, however, become identically zero when  $u$  equals  $v$  as they must in order to satisfy axiom D2.

The concept of a metric is very important as it generates the notion of convergence. In many problems the solution is obtained numerically through an iterative process. The iterates generate a sequence of vectors. The sequence is said to converge to the solution if the metric between two successive iterates is less than a desired tolerance (see Chapter 10). The tolerance determines the accuracy desired. The metric employed to check convergence can be any of the above valid definitions (see Hlavacek and Kubicek, 1983).

### 2.3.2 Normed Linear Space

The norm of a vector is an abstract generalisation of its length. A normed linear space is a linear space in which a norm is defined. The norm is a non-negative number denoted by  $\|u\|$ . The axioms a norm has to satisfy are

- N1:  $\|u\| > 0$  for  $u \neq 0$
- N2:  $\|u\| = 0$  if and only if  $u = 0$
- N3:  $\|\alpha u\| = |\alpha| \|u\|$
- N4:  $\|u + v\| \leq \|u\| + \|v\|$

These axioms are mathematical statements of the properties we expect a norm to possess. The first two axioms state that the length of a non-zero vector is always positive. The only vector that can have a zero norm is the zero vector and only the zero vector has zero norm. The third axiom tells us that the length of a scalar multiple of a vector  $u$  is the scalar multiple times  $\|u\|$ . The absolute value on the scalar is necessary to satisfy the first axiom. The fourth axiom is the triangle inequality which we saw earlier in D4. The third and fourth axioms involve the concepts of scalar multiplication and vector addition. A norm has to be hence necessarily defined on a linear space. This ensures that  $\alpha u$ ,  $u + v$ , occurring in axioms N3, N4 belong to the same vector space as  $u$ ,  $v$  (see Stakgold, 1979).

A metric is generated by a norm on a linear space. This is obtained from the relation

$$d(u, v) = \|u - v\| \quad (2.7)$$

The metric so generated satisfies the axioms D1–D4. Thus, a normed linear space is necessarily a metric space.

A norm like the metric can be defined on a linear space in many ways. The various definitions have the significance of a length. They are not equal to the Euclidean or geometrical length. The Euclidean norm in  $\mathbb{R}^n$  is given by

$$\|u\|_2 = \left( \sum_{i=1}^n u_i^2 \right)^{1/2} \quad (2.8a)$$

This norm generates the Euclidean metric via (2.7). It has the significance of the geometric length of a vector in  $\mathbb{R}^2$ ,  $\mathbb{R}^3$ .

Other valid definitions of the norm in  $\mathbb{R}^n$  are

$$\|u\|_1 = \sum_{i=1}^n |u_i| \quad (2.8b)$$

$$\|u\|_\infty = \max_{i \in 1, n} |u_i| \quad (2.8c)$$

Norms (2.8b) and (2.8c) generate the metrics defined in (2.6b) and (2.6c) via (2.7).

**Example 2.3** Determine if  $\|u\| = \max_{i \in 1, n} x_i$  is a valid definition of a norm in  $\mathbb{R}^n$ .

We consider an element  $u$  in  $\mathbb{R}^n$   $(-2, -3, -3, \dots, -3)$ . This has the last  $n - 1$  coordinates as  $-3$ , and the first coordinate as  $-2$ .

$$\|u\| = -2 < 0$$

This violates axiom N1 and hence is not a valid norm.

### 2.3.3 Inner-Product Space

The inner-product between two vectors  $u, v$  is a measure of the angle between them. It represents

the projection of one vector in the direction of the second vector. This is denoted as  $\langle u, v \rangle$  and is defined on a **real linear space** so as to satisfy the following axioms:

$$I1: \langle u, u \rangle > 0 \text{ for } u \neq 0$$

$$\langle u, u \rangle = 0 \text{ for } u = 0$$

$$I2: \langle u, v \rangle = \langle v, u \rangle$$

$$I3: \langle u, \alpha v \rangle = \langle \alpha u, v \rangle = \alpha \langle u, v \rangle$$

$$I4: \langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$$

The inner-product is a scalar quantity. A suitable candidate satisfying the above axioms in  $\mathbb{R}^n$  is

$$\langle u, v \rangle = \sum_{i=1}^n u_i v_i \quad (2.9)$$

The axioms involve scalar multiplication and vector addition. So, the inner-product like the norm can be defined only on a linear space. An inner-product generates a norm of the form

$$\|u\| = \langle u, u \rangle^{1/2} \quad (2.10a)$$

This norm satisfies all axioms *N1–N4*. A metric can be generated using this norm from

$$d(u, v) = \|u - v\| = \langle u - v, u - v \rangle^{1/2} \quad (2.10b)$$

An inner-product space is hence necessarily a normed linear space and a metric space (Stakgold, 1979). A normed linear space may not be an inner-product space. This indicates a hierarchy in the different concepts of angle, length and distance. The concept of an angle generates the notion of length which in its turn defines distance. If distance is defined in a space, it does not imply length or angle is defined in it. The axioms *I1–I4* and Definition (2.9) are valid only for real vector spaces. The field  $F$  is the set of real numbers and the vectors have only real numbers as their coordinates. The reason for this can be best understood from the example below.

**Example 2.4**  $u = [2i, 2i]^t$  is a nonzero element of the two-dimensional complex vector space. Using (2.9) we obtain

$$\langle u, u \rangle = 2i \cdot 2i + 2i \cdot 2i = -8$$

This violates axiom *I1*. The inner-product as defined in (2.9) is invalid. This is overcome by redefining the inner-product more generally in a complex vector space as

$$\langle u, v \rangle = \sum_{i=1}^n \bar{u}_i v_i \quad (2.11)$$

where the bar over  $u$  denotes complex conjugate. Using this we have

$$\langle u, u \rangle = 8$$

The axioms an inner-product must satisfy in a complex vector space are modified as

$$IC1: \langle u, u \rangle > 0 \text{ for } u \neq 0$$

$$\langle u, u \rangle = 0 \text{ for } u = 0$$

$$IC2: \langle u, v \rangle = \langle \bar{v}, u \rangle$$

$$IC3: \langle u, \alpha v \rangle = \alpha \langle u, v \rangle$$

$$\langle \alpha u, v \rangle = \alpha \langle u, v \rangle$$

$$IC4: \langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$$

The inner-product defined in (2.9) is similar to the dot product arising in vector algebra and physics. It is a measure of the cosine of the angle between two vectors.

Two vectors  $u, v$  are said to be **orthogonal** when  $\langle u, v \rangle = 0$ . This is an important concept as it is often convenient to deal with an orthogonal set of vectors. Two vectors are said to be **orthonormal** if they are orthogonal to each other and each of them possesses a unit norm. In an orthogonal set of vectors, each vector is orthogonal to every other member of that set.

## 2.4 LINEAR DEPENDENCE AND DIMENSION

### 2.4.1 Linear Dependence

Two vectors  $u^1, u^2$  from a vector space are linearly dependent if constants  $c_1, c_2$  one or both nonzero, exist such that

$$c_1 u^1 + c_2 u^2 = 0 \quad (2.12a)$$

For nonzero  $u^1, u^2$ , a nonzero  $c_1$  implies  $c_2$  is also nonzero and we have  $u^1 = -c_2 u^2 / c_1$ . The linear dependence of two nonzero vectors implies one is a scalar multiple of another. The vectors are independent if the relation (2.12a) is satisfied only for  $c_1 = c_2 = 0$ . We now extend this definition to a set of  $r$  vectors.

A linear relation of  $r$  vectors  $\{u^1, u^2, \dots, u^r\}$  is an equation of the form

$$\sum_{i=1}^r c_i u^i = 0 \quad (2.12b)$$

where the  $c_i$ 's are scalars. A linear relation with all  $c_i$ 's identically zero is called the trivial relation. A nontrivial relation has some nonzero  $c_i$ 's. A vector set  $\{u^i\}$  is dependent if it satisfies a nontrivial relation (Stakgold, 1979). A set of vectors which satisfies only the trivial relation and no nontrivial relation is independent.

For a set of  $r$  dependent vectors there is at least one nonzero  $c_i$  such that (2.12b) is satisfied. Without any loss of generality we assume this to be  $c_1$ . This allows  $u^1$  to be written as a linear combination of the other vectors as

$$u^1 = -\sum_{i=2}^r \frac{c_i u^i}{c_1} \quad (2.12c)$$

The linear relation (2.12b) for vectors with  $n$  coordinates each, yields  $n$  equations in  $r$  unknowns  $c_1, c_2, \dots, c_r$ . The vector  $u^i$  is denoted  $(u_1^i, u_2^i, \dots, u_n^i)^T$ . Here  $u_j^i$  represents the  $j$ th coordinate of vector  $u^i$ . The unknowns  $c_i$  are obtained by equating each coordinate of the vector sum in (2.12b) to zero. This yields

$$\sum_{j=1}^n u_j^i c_j = 0 \quad \text{for } i = 1, \dots, n \quad (2.13)$$

The coordinates  $u_j^i$  are known and the unknowns are the  $c_i$ 's. The set is independent if the only solution to the system (2.13) is  $c_i = 0$  for all  $i$ . It is dependent if a nontrivial solution, such that at least one  $c_i \neq 0$  satisfies (2.13).

The linear relation (2.12b), as we have seen, yields  $n$  equations in  $r$  unknowns. If  $r > n$  the  $c_i$ 's can be determined, assigning nonzero values to  $r - n$  of them and evaluating the rest from (2.13). Therefore, a set with  $r > n$  is always dependent. For  $r \leq n$ , we have to solve for the  $c_i$ 's and evaluate the nature of the set.

**Example 2.5** Is the set  $(2 \ 1 \ 3)', (4 \ 1 \ 5)', (2 \ 0 \ 2)'$  a dependent set?

The linear relation  $\sum_{i=1}^3 c_i u^i = 0$  results in

$$2c_1 + 4c_2 + 2c_3 = 0$$

$$c_1 + c_2 = 0$$

$$3c_1 + 5c_2 + 2c_3 = 0$$

A possible solution to this system is  $c_1 = c_2 = c_3 = 0$ . This trivial solution is always a candidate as the equations are homogeneous. We have to determine if this system can admit a nonzero solution. We have  $c_1 = -c_2$  from the second equation.

The first and third equations yield

$$2c_2 + 2c_3 = 0$$

when we eliminate  $c_1$  from them. Assigning  $c_2 = -1$ , a nonzero arbitrary number we obtain  $c_1 = 1, c_3 = 1$  as a set of nonzero solutions. We conclude that the three vectors form a dependent set. There is an infinite number of such nonzero solutions. These can be obtained by assigning different values arbitrarily to  $c_2$ .

**Example 2.6** Show  $(2 \ 1 \ 3)', (4 \ 1 \ 5)',$  and  $(2 \ 0 \ 4)'$  form a linearly independent set.

Equating coordinates of  $\sum_{i=1}^3 c_i u^i = 0$ , we get

$$2c_1 + 4c_2 + 2c_3 = 0$$

$$c_1 + c_2 = 0$$

$$3c_1 + 5c_2 + 4c_3 = 0$$

This is reduced to

- (a)  $c_1 = -c_2$
- (b)  $2c_2 + 2c_3 = 0$
- (c)  $2c_2 + 4c_3 = 0$

This system of equations has only the zero solution and the given set is an independent set.

**Example 2.7** Find whether the following sets are dependent or independent:

- (a)  $(2, 0, 4)', (2, 0, 8)', (2, 1, 3)', (4, 1, 5)'$
- (b)  $(2, 1, 4)', (6, 3, 12)'$
- (c)  $(2, 1, 4)', (3, 3, 12)'$

(a) We obtain three equations in the four unknown  $c_i$ 's:

$$2c_1 + 2c_2 + 2c_3 + 4c_4 = 0$$

$$c_3 + c_4 = 0$$

$$4c_1 + 8c_2 + 3c_3 + 5c_4 = 0$$

We assign  $c_3 = 1$  arbitrarily, and solve for  $c_1, c_2, c_4$  to yield  $c_1 = 3/2, c_2 = -1/2, c_4 = -1$ . This is a dependent set, as we expect, since the number of vectors exceeds the number of coordinates of each vector ( $r > n$ ).

(b) The linear relation yields

$$2c_1 + 6c_2 = 0$$

$$c_1 + 3c_2 = 0$$

$$4c_1 + 12c_2 = 0$$

An admissible solution is  $c_1 = -3, c_2 = 1$ . Therefore, this is a dependent set.

(c) Equating the coordinates of the vectors to zero yields

$$2c_1 + 3c_2 = 0$$

$$c_1 + 3c_2 = 0$$

$$4c_1 + 12c_2 = 0$$

The only solution here is the trivial solution and the given two vectors are independent.

The concept of linear dependence plays a vital role in the solution of linear algebraic equations as we will see in Chapter 3. This also formally defines the dimension of a vector space.

## 2.4.2 Dimension of a Vector Space

So far we have identified the dimension of a real vector space as being equal to the number of coordinates of the vector in it. Thus in  $\mathbb{R}^3$  each vector has three coordinates and we said that the vectors belong to a three-dimensional space. We will now formally define the dimension of a vector space. This is based on the algebraic concept of linear independence (see Stakgold (1983), Noble and Daniel (1977)). This approach also generates the notions of a subspace and a basis. It also enables us to understand the concept of an infinite dimensional space in Chapter 5.

**Definition 2.5: Dimension.** The dimension of a vector space  $V$  is equal to the maximum number of linearly independent vectors in  $V$ .

**Theorem 2.1** The dimension of  $\mathbb{R}^n$  is  $n$ .

*Proof.* A set of  $r$  vectors with  $r > n$  in  $\mathbb{R}^n$  forms a linearly dependent set as seen earlier. The dimension of  $\mathbb{R}^n \leq n$ . If at least one set of  $n$  independent vectors in  $\mathbb{R}^n$  can be obtained, this will be the maximum number and the dimension of  $\mathbb{R}^n$  will be  $n$ . Let  $e^i$  denote the vector with  $i$ th coordinate 1 and all other coordinates zero. So  $e^3$  is  $(0, 0, 1, 0, 0, \dots, 0)^T$ . The linear relation

$$\sum_{i=1}^n c_i e^i = 0 \quad (2.14)$$

implies  $[c_1, c_2, \dots, c_n]^T = 0$ . The only solution to the linear relation is  $c_i = 0$  for all  $i$ .

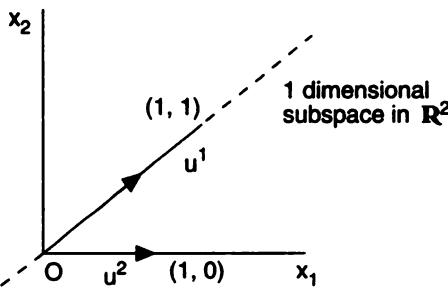
The set  $(e^1, e^2, \dots, e^n)$  is hence linearly independent.  $n$  is the maximum number of linearly

independent vectors in  $\mathbb{R}^n$ . Consequently the dimension of  $\mathbb{R}^n$  is  $n$ . If the dimension of a space is  $n$  it does not imply that the space is  $\mathbb{R}^n$ . The elements can belong to  $\mathbb{R}^p$ , where  $p > n$ . This gives rise to the notion of subspace.

### 2.4.3 Subspace

The set  $\{u^1, \dots, u^m\}$  is called a **generating set** of a vector space  $V$  if every vector  $u$  in  $V$  can be expressed as a linear combination of this set.  $V$  consists of all scalar multiples and linear combinations of the members of the generating set. If each  $u^i$  has  $n$  coordinates and the number of linearly independent vectors in the set  $\{u^i\}$  is less than  $n$ , the space  $V$  is a subspace of  $\mathbb{R}^n$  (see Ramkrishna and Amundson, 1985). We explain this concept of a subspace in detail with illustrations.

Every vector in  $\mathbb{R}^2$  can be written as a linear combination of two independent vectors  $u^1, u^2$ . These generate the space  $\mathbb{R}^2$ .  $u^1$  by itself generates many vectors which are scalar multiples of  $u^1$ . These vectors lie on the directed line segment  $u^1$ . They form a one-dimensional subspace in the two-dimensional space generated by  $u^1, u^2$  (see Fig. 2.5). Every vector in this subspace is



**Fig. 2.5** A one-dimensional subspace in  $\mathbb{R}^2$ , consisting of all scalar multiples along  $u^1$  (indicated by dashed line).

specified by two coordinates. Similarly, all scalar multiples of  $u^2$  generate another one-dimensional subspace.

The three-dimensional space  $\mathbb{R}^3$  is generated by  $u^1, u^2, u^3$ .  $u^1$  generates a one-dimensional subspace, as discussed earlier, as do  $u^2, u^3$ . A two-dimensional subspace consists of all scalar multiples of  $u^2, u^3$  and their linear combinations. Similarly, linear combinations of  $u^1, u^2$  and of  $u^1, u^3$  generate at least two other two-dimensional subspaces of  $\mathbb{R}^3$ .

**Example 2.8** Vectors  $(1, 0)', (1, 1)'$  generate  $\mathbb{R}^2$ . Vectors  $(2, 2)', (1, 1)'$  generate a one-dimensional subspace in  $\mathbb{R}^n$ . The subspace consists of all scalar multiples of  $(1, 1)'$  (see Fig. 2.5).

**Theorem 2.2** Let the  $m$  vectors  $\{u^1, u^2, \dots, u^m\}$  be a linearly independent generating set of vector space  $V$ . A set of  $r$  vectors,  $r > m$  is linearly dependent in  $V$ .

**Proof.** Let  $(v^1, v^2, \dots, v^r)$  be a set of  $r$  vectors in  $V$ . Each of the  $v^j$ 's can be written as a linear combination of the  $u^i$ 's as the  $u^i$ 's are a generating set. So

$$v^j = \sum_{i=1}^m a_{ji} u^i \quad \text{for } j = 1 \dots r \quad (2.15a)$$

The  $c_i$ 's in the linear relation

$$\sum_{i=1}^r c_i v^i = 0 \quad (2.15b)$$

decide if the set  $\{v^1, v^2, \dots, v^r\}$  is either dependent or independent. Expressing each  $v^i$  in (2.15b) using (2.15a), we obtain

$$\sum_{i=1}^r c_i \sum_{k=1}^m a_{ik} u^k = 0 \quad (2.16)$$

or

$$\sum_{k=1}^m \left( \sum_{i=1}^r c_i a_{ik} \right) u^k = 0, \text{ changing the order of the summations.}$$

As the  $u^k$ 's form an independent set, this relation among  $u^k$ 's hold if and only if

$$\sum_{i=1}^r c_i a_{ik} = 0 \quad \text{for } k = 1 \dots m \quad (2.17)$$

This is a system of  $m$  equations in the  $r$  unknown  $c_i$ 's ( $r > m$ ). This admits a nontrivial solution ( $c_1, c_2, \dots, c_r$ ) as seen earlier, which renders  $\{v_i\}$  a dependent set from (2.15b). Since  $\{v_i\}$  was chosen arbitrarily, the theorem holds for any set of  $r$  vectors ( $r > m$ ).

**Example 2.9** Consider the set of four vectors in  $\mathbb{R}^3$ :

$$(1 \ 2 \ 3)^t \ (2 \ 1 \ 3)^t \ (3 \ 1 \ 2)^t \ (5 \ 3 \ 7)^t$$

The dimension of  $\mathbb{R}^3$  is 3. Hence we can have at most three vectors which form an independent set. The given set of vectors must therefore be necessarily dependent. Consider

$$\sum_{i=1}^4 c_i u^i = 0$$

This yields a system of three equations (one for each coordinate) in the four unknowns  $c_i$ ,  $i = 1, 4$ .

$$\begin{aligned} c_1 + 2c_2 + 3c_3 + 5c_4 &= 0 \\ 2c_1 + c_2 + c_3 + 3c_4 &= 0 \\ 3c_1 + 3c_2 + 2c_3 + 7c_4 &= 0 \end{aligned}$$

A nonzero solution can be determined by assigning some value to one of the variables. Let us assign  $c_4 = 1$ . (We could have chosen any other  $c_i$  or assigned any other value.) This then gives a system of three nonhomogeneous linear equations in three unknowns. The system has a feasible nonzero solution when the first three vectors form an independent set. This assures us that a nontrivial linear relation exists between the  $u^i$ 's and the given set is dependent.

If the nonhomogeneous system of three equations in  $c_1, c_2, c_3$  (when we assign  $c_4 = 1$ ) has no feasible nonzero solution, it implies that the set of vectors  $u^1, u^2, u^3$  are dependent. Here we can assign  $c_4$  the value zero and seek a nonzero solution for  $c_1, c_2, c_3$ . This will again yield the result such that the set of vectors is dependent. We conclude that a set of  $p$  vectors in  $\mathbb{R}^n$ , where  $p > n$ , is necessarily dependent.

### 2.4.4 An Application of the Concept of Linear Dependence

The concept of linear dependence plays a vital role in sensor placement in a chemical plant. It enables us to determine the variables which need to be measured, which will allow us to determine the state of the plant completely.

The system of Fig. 1.1 has five equations in nine unknowns and can be written as

$$\left[ \begin{array}{ccccccccc} -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -1 \end{array} \right] \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \\ F_6 \\ F_7 \\ F_8 \\ F_9 \end{bmatrix} = 0 \quad (2.18a)$$

The five rows of the matrix  $A$  are vectors in  $\mathbb{R}^9$ . These five vectors are linearly independent. They span a five-dimensional subspace in the nine-dimensional space  $\mathbb{R}^9$ . One has to specify four variables to determine all the others uniquely. This however cannot be any four variables as we will see now.

The system (2.18a) has nine flow-rates  $\{F_1, F_2, \dots, F_9\}$ . The mass balances across each unit yields a system of five equations, viz.

$$\begin{aligned} F_3 - F_2 - F_1 &= 0 \\ F_3 - F_4 - F_5 &= 0 \\ F_4 - F_7 - F_6 &= 0 \\ F_2 + F_8 - F_6 &= 0 \\ F_8 - F_9 + F_6 &= 0 \end{aligned} \quad (2.18b)$$

When four of the flow-rates are measured and specified, the remaining five can be predicted from (2.18a) or (2.18b). Specifying  $F_1, F_5, F_6, F_9$ , we can recast the equations in vectorial form as

$$\left[ \begin{array}{ccccc} -1 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{array} \right] \begin{bmatrix} F_2 \\ F_3 \\ F_4 \\ F_7 \\ F_8 \end{bmatrix} = \begin{bmatrix} F_1 \\ F_5 \\ F_6 \\ F_5 \\ F_9 - F_6 \end{bmatrix} \quad (2.18c)$$

or  $Au = b$ .

It can be verified that the five rows of  $A$  are linearly independent and we can solve for  $[F_2, F_3, F_4, F_7, F_8]$ . Thus the state of the entire plant can be uniquely determined by experimentally measuring the four variables  $[F_1, F_5, F_6, F_9]$ .

Consider now a situation where  $[F_1, F_2, F_5, F_8]$  are the four flow-rates measured. The system of equations (2.18b) for the unknowns can be recast as

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} F_3 \\ F_4 \\ F_6 \\ F_7 \\ F_9 \end{bmatrix} = \begin{bmatrix} F_1 + F_2 \\ F_5 \\ 0 \\ F_5 - F_2 - F_8 \\ -F_8 \end{bmatrix} \quad (2.18d)$$

The five rows of matrix  $A$  are linearly dependent. This can be seen directly as the fourth row vector is the zero vector (every set containing the zero vector is dependent). This is a system of four independent equations in five unknowns. The state of the plant cannot therefore be obtained uniquely by specifying  $F_1, F_2, F_5, F_8$ .

As a second example, consider the splitter-mixer network shown in Fig. 2.6. The system is characterised by five flow-rates  $\{F_1, F_2, \dots, F_5\}$ . The mass balance across each unit can be written to yield the three equations

$$F_1 - F_2 - F_3 = 0, \quad F_2 - F_4 = 0, \quad F_3 + F_4 - F_5 = 0$$

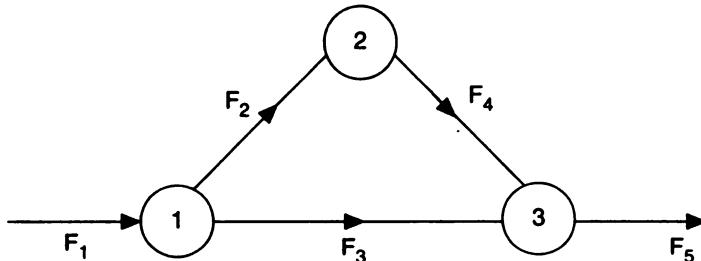


Fig. 2.6 A splitter-mixer network.

The mass balance across units 1, 2 yield

$$F_1 = F_4 + F_3$$

and across units 2, 3 result in

$$F_2 + F_3 = F_5$$

It would appear that we now have a system of five equations for the five  $F_i$ 's. This can be cast in vectorial form as

$$\begin{bmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 1 & -1 \\ 1 & 0 & -1 & -1 & 0 \\ 0 & 1 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \\ F_5 \end{bmatrix} = 0 \quad (2.19)$$

The rows of the matrix are vectors in  $\mathbb{R}^5$ . The five row vectors  $\{r^1, r^2, \dots, r^5\}$  are, however, not independent. Clearly,

$$r^4 - r^1 - r^2 = 0, \quad r^5 - r^2 - r^3 = 0$$

where  $r^i$  represents the  $i$ th row vector.

The five row vectors in (2.19) span a three-dimensional subspace in  $\mathbb{R}^5$ . Generating more equations, without including any additional information, does not enable us to obtain all five  $F_i$ 's. One has to specify two of the  $F_i$ 's (such that the resulting three equations are independent), to estimate the remaining  $F_i$ 's.

#### 2.4.5 Basis

**Definition 2.6** A linearly independent generating set of a vector space  $V$  is called a **basis** of  $V$ .

**Theorem 2.3** Consider  $\mathbb{R}^n$  a vector space of dimension  $n$ . Every basis of  $\mathbb{R}^n$  contains exactly  $n$  vectors and these are linearly independent. Conversely, a set of  $n$  linearly independent vectors forms a basis of  $\mathbb{R}^n$ .

We refer the interested reader to any standard text of linear algebra for a proof of this theorem (Murdoch, 1965).

**Theorem 2.4** Every vector in a vector space can be expressed in one and only one way as a linear combination of its basis.

*Proof.* Let  $\{e^i\}_{i=1,n}$  be a basis set for an  $n$ -dimensional space  $\mathbb{R}^n$ . A vector  $u$  in  $\mathbb{R}^n$  can be expressed as a linear combination of  $\{e^i\}$  as

$$u = \sum_{i=1}^n c_i e^i \quad (2.20a)$$

If this representation is nonunique, there is another set of scalars  $b_i$  such that

$$u = \sum_{i=1}^n b_i e^i \quad (2.20b)$$

Subtracting, we get

$$0 = \sum_{i=1}^n (c_i - b_i) e^i \quad (2.20c)$$

Since  $e^i$  is a basis, it is a linearly independent set, and (2.20c) is satisfied only for  $c_i = b_i$ .

An arbitrary vector  $u$  in  $\mathbb{R}^n$  can be expressed in terms of a basis  $\{u^i\}_{i=1,n}$  uniquely as

$$u = \sum_{i=1}^n c_i u^i \quad (2.21)$$

Equating the  $n$  coordinates on both sides, we obtain  $n$  equations in the  $n$  unknown  $c_i$ 's. The determination of the  $c_i$ 's in the representation (2.21) is hence cumbersome and tedious. The  $c_i$ 's now do not have the significance of distance measured parallel to a basis vector as the basis set may not be orthogonal.

Consider now a situation where the basis is orthogonal or orthonormal. Here every vector in

the set is orthogonal to every other vector of the set. Consider an orthogonal basis  $\{e^i\}_{i=1..n}$ . An arbitrary  $u$  can be expressed as a linear combination of this basis as

$$u = \sum_{i=1}^n b_i e^i \quad (2.22)$$

The  $b_i$ 's can be found taking the inner-product of both sides with  $e^j$ . Using the definition of the real inner-product this yields

$$\langle u, e^j \rangle = \sum_{i=1}^n \langle b_i e^i, e^j \rangle = b_j \|e^j\|^2$$

as  $\langle e^i, e^j \rangle = 0$  for  $i \neq j$  (since they are orthogonal) or

$$b_j = \langle u, e^j \rangle / \|e^j\|^2 \quad (2.23)$$

**Example 2.10** Express the vector  $(8 \ 5 \ 11)'$  in terms of the basis  $(1 \ 2 \ 3)', (2 \ 1 \ 3)', (3 \ 1 \ 2)'$ . We seek

$$u = \sum_{i=1}^3 c_i u^i$$

The coefficients  $c_i$  are obtained by solving

$$c_1 + 2c_2 + 3c_3 = 8$$

$$2c_1 + c_2 + c_3 = 5$$

$$3c_1 + 3c_2 + 2c_3 = 11$$

This system has the solution  $c_1 = c_3 = 1$ ,  $c_2 = 2$ . Every vector in  $\mathbb{R}^3$  can be similarly represented uniquely in terms of this basis. The coefficients  $c_i$  are uniquely obtained by solving a set of linear equations as we have shown. This is an inefficient and a tedious exercise, especially for higher-dimensional systems. We illustrate with an example how this problem is overcome when the basis is orthogonal.

**Example 2.11** The set  $u^1 = (1 \ 2 \ 3)', u^2 = (3 \ -3 \ 1)', u^3 = (-11 \ -8 \ 9)'$  forms an orthogonal basis in  $\mathbb{R}^3$ . Represent the vector  $b = (4 \ 5 \ 6)'$  in terms of this basis.

We write  $b = \sum_{i=1}^3 c_i u^i$ . Taking the inner-products of both sides with  $u^1$ , we have

$$\langle u^1, b \rangle = \langle u^1, c_1 u^1 \rangle + \langle u^1, c_2 u^2 \rangle + \langle u^1, c_3 u^3 \rangle$$

$$\langle u^1, u^2 \rangle = \langle u^1, u^3 \rangle = 0$$

as the basis set is orthogonal. We now obtain

$$c_1 = \frac{\langle u^1, b \rangle}{\langle u^1, u^1 \rangle} = \frac{32}{14}$$

Similarly,

$$c_2 = \frac{\langle u^2, b \rangle}{\langle u^2, u^2 \rangle} = \frac{3}{19}$$

$$c_3 = \frac{\langle u^3, b \rangle}{\langle u^3, u^3 \rangle} = -\frac{15}{133}$$

**Example 2.12** Express the vector  $u(7, 3, 9)'$  in  $\mathbb{R}^3$  as a linear combination of the following basis sets:

- (a)  $u^1(1, 1, 1)', u^2(2, 1, 3)', u^3(3, 0, 4)'$
- (b)  $u^1(1, 1, 1)', u^2(1, 0, -1)', u^3(1, -2, 1)'$

(a) The basis set (a) is not an orthogonal basis as  $\langle u^i, u^j \rangle = 0$  for  $i \neq j$  is not satisfied. We write

$$u = \sum_{i=1}^3 c_i u^i$$

The scalars  $c_i$  are obtained by equating the coordinates on both sides. This yields,

$$\begin{aligned} c_1 + 2c_2 + 3c_3 &= 7 \\ c_1 + c_2 &= 3 \\ 2c_1 + 3c_2 + 4c_3 &= 9 \end{aligned}$$

Solving  $c_1 = 2$ ,  $c_2 = c_3 = 1$ .

- (b) The basis set (b) is orthogonal as  $\langle u^i, u^j \rangle = 0$  for  $i \neq j$ . The inner-product we work on is

$$\langle u, v \rangle = \sum_{i=1}^n u_i v_i$$

We represent

$$u = \sum_{i=1}^3 b_i u^i$$

The  $b_i$ 's are obtained by taking the inner-product with  $u^j$  on both sides. This yields

$$b_j = \frac{\langle u, u^j \rangle}{\langle u^j, u^j \rangle}$$

Substituting for  $u$ ,  $u^j$ , we obtain  $b_1 = 19/3$ ,  $b_2 = -1$ ,  $b_3 = 5/3$ . Each coefficient  $c_i$  in the representation (2.19) can be obtained in a relatively elegant manner, when the basis set is orthogonal. This is particularly effective for higher dimensional systems. Here each coefficient is determined independently, i.e.  $b_i$  is independent of  $b_j$  for  $i \neq j$ . This is an important feature when we work in infinite dimensional spaces (as we will see in the next section).

The coordinates of a vector represented in terms of a basis  $\{u^i\}$  are the coefficients of the linear combination of the basis set in which it is expressed. The vector  $u$  in (2.21) in the basis  $\{u^i\}$  has coordinates  $(c_1, c_2, \dots, c_n)'$  and in the basis  $\{e^i\}$  (2.22) has coordinates  $(b_1, b_2, \dots, b_n)'$ .

It is thus preferable to work in an orthogonal basis to working in a general linearly independent basis. We will now discuss a technique which will convert a linearly independent basis to an orthonormal basis.

## 2.5 GRAM-SCHMIDT ORTHONORMALISATION

The representation of a vector in terms of an orthogonal basis is very elegant as we have just seen.

The Gram-Schmidt orthonormalisation technique enables us to obtain an orthonormal set from a linearly independent set. We explain the method geometrically in  $\mathbb{R}^2$  and generalize it to  $\mathbb{R}^n$  for large  $n$  using the algebraic concepts developed so far.

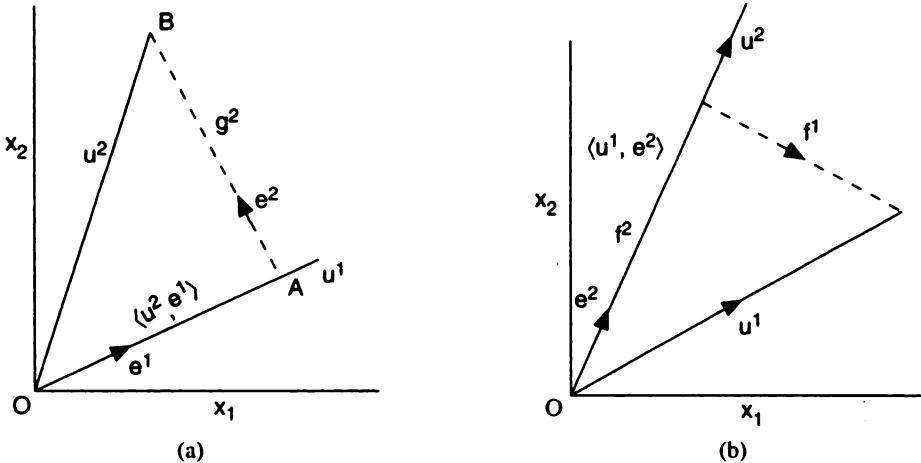
Let  $u^1, u^2$  be two independent vectors in  $\mathbb{R}^2$  as shown in Fig. 2.7(a). Therefore, the vectors are not collinear.  $e^1$  is the unit vector along  $u^1$ .  $OA$  is the projection of  $u^2$  on  $u^1$ . Its magnitude, i.e. length, is given by  $\langle u^2, e^1 \rangle$ . The line segment  $AB$ , i.e. vector  $g^2$ , is orthogonal to  $u^1$  and clearly is the vectorial difference between  $u^2$  and  $OA$  or  $\langle u^2, e^1 \rangle e^1$ . That is,

$$g^2 = u^2 - \langle u^2, e^1 \rangle e^1 \quad (2.24a)$$

The unit vector in the direction of  $g^2$  is

$$e^2 = g^2 / \|g^2\| \quad (2.24b)$$

A second set of orthogonal vectors  $f^1, f^2$  can be constructed from  $u^1, u^2$  by reordering the vectors as  $u^2, u^1$ . This is illustrated in Fig. 2.7(b). In an  $n$ -dimensional space from an independent basis it is possible to generate  $n$  orthonormal basis, by changing the order of the vectors.



**Fig. 2.7** Geometric illustration of Gram-Schmidt orthonormalisation in  $\mathbb{R}^2$ : (a)  $u^1$  is chosen as first vector; (b)  $u^2$  is chosen as the first vector.

We extend the above geometric method to an  $n$ -dimensional vector space  $\mathbf{V}$ , algebraically with the basis  $\{u^1, u^2, \dots, u^n\}$ . The first

$$e^1 = u^1 / \|u^1\| \quad (2.25)$$

$e^1$  clearly has unit norm.  $e^2$  is found such that it is orthogonal to  $e^1$  and has unit norm. We construct  $g^2$  as we did earlier, by subtracting from  $u^2$ , its projection on  $e^1$ . We normalise  $g^2$  to obtain  $e^2$ .

$$g^2 = u^2 - \langle u^2, e^1 \rangle e^1 \quad (2.26a)$$

$$e^2 = g^2 / \|g^2\| \quad (2.26b)$$

$g^2$  is clearly not the zero vector, since it is a nontrivial combination of two independent vectors.  $g^2$  is orthogonal to  $e^1$  as

$$\langle g^2, e^1 \rangle = \langle u^2, e^1 \rangle - \langle u^2, e^1 \rangle \langle e^1, e^1 \rangle = 0$$

A nonzero vector  $g^3$  is obtained by subtracting from  $u^3$ , its projections on  $e^1$  and  $e^2$ . This yields

$$g^3 = u^3 - \langle u^3, e^1 \rangle e^1 - \langle u^3, e^2 \rangle e^2 \quad (2.27a)$$

The unit vector  $e^3$  is defined as

$$e^3 = g^3 / \|g^3\| \quad (2.27b)$$

Clearly,  $\langle g^3, e^1 \rangle = \langle g^3, e^2 \rangle = 0$ . This construction process is continued for all  $n$  vectors (see Noble and Daniel, 1977). The general relationship to obtain  $g^k, e^k$  is

$$g^k = u^k - \sum_{i=1}^{k-1} \langle u^k, e^i \rangle e^i \text{ for } k = 1, \dots, n \quad (2.28a)$$

$$e^k = g^k / \|g^k\| \quad (2.28b)$$

**Example 2.13** Generate an orthonormal set from the linearly independent set  $(2, 0, 1)', (2, 1, 3)', (4, 1, 2)'$  in  $R^3$ .

We use the inner-product definition

$$\langle u, v \rangle = \sum_{i=1}^n u_i v_i$$

$$g^1 = u^1 = (2, 0, 1)'$$

This yields

$$e^1 = \frac{1}{\sqrt{5}} (2, 0, 1)'$$

$$g^2 = u^2 - \langle u^2, e^1 \rangle e^1$$

$$= (2, 1, 3)' - \frac{7}{\sqrt{5}} \frac{1}{\sqrt{5}} (2, 0, 1)'$$

$$= \frac{1}{\sqrt{5}} (-4, 5, 8)'$$

$$e^2 = \frac{1}{\sqrt{105}} (-4, 5, 8)'$$

$$g^3 = u^3 - \langle u^3, e^1 \rangle e^1 - \langle u^3, e^2 \rangle e^2$$

$$\begin{aligned}
 &= (4, 1, 2)' - \frac{10}{\sqrt{5}} \frac{1}{\sqrt{5}} (2, 0, 1)' - \frac{1}{21} \frac{1}{\sqrt{105}} (-4, 5, 8)' \\
 &= \frac{4}{21}(1, 4, -2)' \\
 e^3 &= \frac{1}{\sqrt{21}}(1, 4, -2)'
 \end{aligned}$$

The set

$$\{e^i\}, \quad e^1 = \frac{1}{\sqrt{5}}(2, 0, 1)', \quad e^2 = \frac{1}{\sqrt{105}}(-4, 5, 8)', \quad e^3 = \frac{1}{\sqrt{21}}(1, 4, -2)'$$

is an orthonormal basis for  $R^3$ . Here,

$$\langle e^i, e^j \rangle = \delta_{ij}$$

where  $\delta_{ij}$ , the Kronecker delta function, is defined as

$$\begin{aligned}
 \delta_{ij} &= 1 \quad \text{for } i = j \\
 &= 0 \quad \text{for } i \neq j
 \end{aligned}$$

More details of the method of Gram Schmidt orthogonalisation can be found in Noble and Daniel (1977), Stakgold (1979).

## PROBLEMS

**1.** State which of the following statements are true or false with reasons and/or examples:

- (i) A metric space is necessarily a vector space.
- (ii) An inner-product space is necessarily a vector-space.
- (iii) A normed vector space is necessarily an inner-product space.
- (iv) A normed vector space is necessarily a metric space.
- (v) Every vector space is a metric space.
- (vi) A set of  $n$  linearly independent vectors is a basis for  $R^n$ .
- (vii) If  $V$  is a subspace of  $W$ , then (a) the number of coordinates in a vector in  $V$  is less than  $W$ , and (b) the number of linearly independent vectors in  $V$  is less than those in  $W$ .
- (viii) A basis in a vector space is unique.
- (ix) Every vector space has only one orthonormal basis.

**2.** Consider the set of non-negative integers  $I$ . Is this set a vector space? Is it a metric space?

**3.** Verify that the three definitions (2.8a)–(2.8c) satisfy the axioms of a norm. Calculate all three norms for  $u, v, w$  defined as

$$u = [1 \quad 4 \quad 6]', \quad v = [2 \quad 1 \quad 3]', \quad w = [1 \quad -2 \quad 0]'$$

Find  $d_\infty(u, v)$ ,  $d_1(u, w)$ ,  $d_2(v, w)$ .

**4.** Show that the three norms in (2.8) generate the metrics  $d_1$ ,  $d_2$ ,  $d_\infty$ , respectively.

**5.** Show that for vectors  $u, v \in C^n$ ,  $\langle u, v \rangle = \sum \bar{u}_i v_i$  is a valid definition of an inner-product.

6. Consider  $u, v \in \mathbb{R}^4$ . Is  $d(u, v) = |x_2 - y_2|$  a valid definition of a metric in  $\mathbb{R}^4$ ?

7. Show  $\langle u + v, u - v \rangle = \|u\|^2 - \|v\|^2$

8. Is  $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$  a basis in  $\mathbb{R}^3$ ? Why?

9. Is  $\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 1 \end{pmatrix}$  a basis in  $\mathbb{R}^2$ ? Is this set orthogonal? Obtain two different orthonormal bases

from this. How many orthonormal bases can you have in  $\mathbb{R}^2$ ?

10. Check if the following sets are dependent or independent:

$$(i) \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \\ 4 \end{pmatrix}$$

$$(ii) \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix}$$

$$(iii) \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 2 \\ 6 \end{pmatrix}$$

11. (i) Consider  $(2 \ 1 \ 3)', (4 \ 1 \ 5)', (2 \ 2 \ 4)'$ . Is this a basis in  $\mathbb{R}^3$ ? Is this a generating set in  $\mathbb{R}^3$ ? Is this set orthogonal? From this set obtain an orthonormal set.

(ii) The vectors  $[3 \ 2 \ 1]', [2 \ -3 \ 0]'$  belong to  $\mathbb{R}^3$ . Make an orthogonal basis for  $\mathbb{R}^3$  out of this set.

12. From the basis  $[1 \ 2 \ 3]', [2 \ 1 \ 3]', [4 \ 3 \ 6]'$ , generate an orthonormal basis.

13. Are the following sets dependent:

$$(i) [1, 2, 3]', [4 5 6]', [10 11 12]'$$

$$(ii) [1 2 3]', [4 8 9]'$$

14. Find a basis for the vector space generated by:

$$(1 \ 3 \ 7)', (2 \ -1 \ 0)', (1 \ 10 \ -21)', (4 \ -3 \ 2)'$$

What is the dimension of the space? Does the set  $(1 \ -1 \ 1)', (1 \ 1 \ -3)', (1 \ 2 \ 5)'$  span the same space?

15. Find the dimension of the vector space spanned by:

$$[1 \ -1 \ 1 \ 2]', [2 \ -3 \ -3 \ 2]', [-1 \ 2 \ 3 \ 1]', [1 \ 1 \ 1 \ 7]'$$

16. Obtain an orthonormal set spanning the same subspace as

$$u^1 = [1 \ 1 \ 1 \ -1]', u^2 = [2 \ -1 \ -1 \ 1]', u^3 = [-1 \ 2 \ 2 \ 1]'$$

17. Consider the sets:

$$(i) [1 \ 2 \ 3]', [4 \ 5 \ 6]', [7 \ 8 \ 9]', [2 \ 3 \ 4]'$$

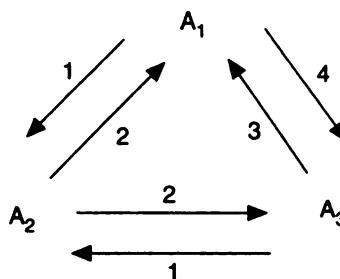
$$(ii) [1 \ 2 \ 3]', [2 \ 1 \ 3]', [3 \ 1 \ 2]', [5 \ 3 \ 7]'$$

Are these dependent? Obtain a basis that spans the space generated by these vectors.

18. Compute the norm of the vector

$$x = (1 + i, 1 - i)'$$

19. Consider the set of reversible elementary molecular reactions shown in Fig. 2.8. Determine



**Fig. 2.8** Reaction scheme of Problem 19.

the algebraic equation governing the equilibrium concentrations of  $A_1$ ,  $A_2$ ,  $A_3$ . Is this uniquely determined? How many concentrations does one need to specify to determine the equilibrium state uniquely? Write the equations in vectorial form. How many independent rows and columns does the matrix have? All rate constants are in  $s^{-1}$ .

#### REFERENCES

- Hlavacek, V. and Kubicek, M., Numerical Solution of Non-linear Boundary-value Problems with Applications, Prentice-Hall, Englewood Cliffs, New Jersey (1983).
- Kreyszig, E., Advanced Engineering Mathematics, Wiley, New York (1982).
- Madron, F. and Veverka, V., Optimal selection of measuring points in complex plants by linear models, *AIChE*, **38**, 227 (1992).
- Murdoch, D.C., Linear Algebra, Wiley, New York (1970).
- \_\_\_\_\_, Linear Algebra for Undergraduates, Wiley, New York (1965).
- Naylor, A.W. and Sell, H., Linear Operator Theory in Engineering and Science, Holt, Rinehart & Winston, New York (1971).
- Noble, B. and Daniel, J.W., Applied Linear Algebra, Prentice-Hall, Englewood Cliffs, New Jersey (1977).
- Ramkrishna, D. and Amundson, N.R., Linear Operator Methods in Chemical Engineering: With applications to transport and chemical reaction systems, Prentice-Hall, Englewood Cliffs, New Jersey (1985).
- Stakgold, I., Green's Function and Boundary-value Problems, Wiley, New York (1979).

# 3

## Matrices, Operators and Transformations

---

The basic concepts of vectors and vector spaces were discussed in Chapter 2. These form the foundation for the theory of linear equations. This chapter deals with the transformations or mappings of vectors to other vectors. A vector multiplied by (operated on by) a matrix is converted into another vector. In this respect the matrix can be viewed as being analogous to a function of a real variable in calculus.

Consider the function of one real variable  $y = f(x)$ . The function  $f$  takes the real independent variable  $x$  and transforms or maps it to the real dependent variable  $y$  according to a predefined rule. This action of a function is similar to that of a matrix. A matrix  $A$  acts on the vector  $u$  and yields a new vector  $v$ . This is represented as  $v = Au$ . The new vector, as we will see, is obtained by matrix multiplication.

In this chapter we define the matrix  $A$  as a linear operator and see how its action is similar to that of a function. The matrix is a linear operator as it satisfies certain axioms which we will define shortly. The domain space and range space of the matrix operator are also identified. We see how the operator generates a natural basis which can be used in construction of solutions to problems of the form (2.1). We define other concepts like Rayleigh's quotient for the matrix and see how they can be extended to infinite dimensional spaces in Chapter 6. We now recall some fundamental concepts from calculus.

The **domain**  $D$  of a function is the set of values  $x$  for which the function is defined. The set of all  $y$  which the function assigns to each  $x$  in  $D$  is called the **range** of the function. This set is denoted by  $R_f$  (see Kreyszig, 1982). The action of a function can be compactly represented as  $f: D \rightarrow R_f$ . This symbolically represents that the function  $f$  acts on elements in  $D$  and generates elements in  $R_f$ .

**Example 3.1** Find  $D$ ,  $R_f$  for  $f(x) = 1/(x^2 + 1)$ ,  $f(x)$  is defined for all real  $x$ . The domain  $D$  is the entire real line  $(-\infty, \infty)$ . The function is always positive and takes on a maximum value of 1 at  $x = 0$ . The range  $R_f$  is the half-closed interval  $(0, 1]$ .

**Example 3.2** Find  $D$ ,  $R_f$  for  $f(x) = \sqrt{x^2 - 1}$ . The domain  $D$  of the function consists of all  $x$  for which the  $x^2 - 1 \geq 0$ , i.e.  $x \geq 1$  or  $x \leq -1$ .  $D$  is  $(-\infty, -1] \cup [1, \infty)$ . The function is not defined for  $x$  in  $(-1, 1)$  as we restrict  $f$  to be a real function of a real variable.  $R_f$ , the range of the function here, is  $[0, \infty)$ .

### 3.1 MATRICES

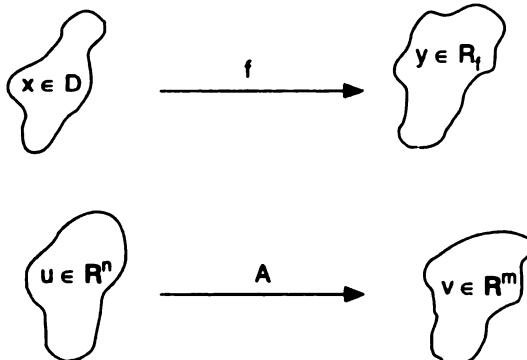
A matrix  $A$  is a rectangular array of numbers of the form

$$\begin{bmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{22} & a_{2n} \\ \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{mn} \end{bmatrix}$$

The  $a_{ij}$ 's are called the elements of the matrix;  $a_{ij}$  represents the element in the  $i$ th row and the  $j$ th column. A real matrix has all real elements. The matrix depicted above has  $m$  rows and  $n$  columns (and is called an  $m \times n$  (read as  $m$  by  $n$ ) matrix. A square matrix has as many rows as columns, i.e.  $m = n$ . An  $m \times n$  matrix  $A$  operates on, i.e. it multiplies, vectors  $u(u_1, u_2, \dots, u_n)'$  in  $\mathbb{R}^n$  and yields vectors  $v(v_1, v_2, \dots, v_m)'$  in  $\mathbb{R}^m$  (see Noble and Daniel, 1977). The relationship between the coordinates  $v_i, u_i$  is given by

$$v_i = \sum_{j=1}^n a_{ij} u_j \quad \text{for } i = 1, \dots, m \quad (3.1)$$

The role of a matrix can be seen to be similar to that of a function. An  $m \times n$  matrix  $A$ , transforms vectors in  $\mathbb{R}^n$  to vectors in  $\mathbb{R}^m$ . This is denoted as  $A: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . Here  $\mathbb{R}^n$  is the domain of  $A$  and  $\mathbb{R}^m$  the range. The analogy between a matrix and a function is shown in Fig. 3.1. The matrix can therefore be viewed as a transformation or a map or an operator (see Kaplan, 1962).



**Fig. 3.1** Analogy between a function and a matrix.

**Example 3.3** Find the vector  $v = Au$ , where  $u = (2, 1, 2)'$ , and

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

Using

$$v_i = \sum_{j=1}^n a_{ij} u_j$$

we have

$$v_1 = 1.2 + 2.1 + 1.2 = 6$$

$$v_2 = 4.2 + 3.1 + 5.2 = 21$$

The vector  $v$  is  $(6, 21)'$  and clearly belongs to  $\mathbb{R}^2$ .

The matrix  $A$  ( $m \times n$ ) is a finite dimensional operator as it operates on elements, i.e. vectors which have a finite number of coordinates  $n$ , and yields vectors with  $m$  coordinates.

**Definition 3.1**  $L$  is said to be a **linear operator** if it satisfies

$$(i) Lu = 0, \text{ for } u = 0$$

$$(ii) L(c_1u^1 + c_2u^2) = c_1Lu^1 + c_2Lu^2, \text{ for arbitrary } c_1, c_2 \text{ belonging to a field } F, \text{ and } u^1, u^2 \text{ are elements in the domain of } L.$$

The functions we saw in the earlier examples do not satisfy these conditions and are nonlinear. The matrix  $A$ , on the other hand, is a linear operator or a linear map. The notion of operator is an important concept as it enables us to generalise concepts from elementary calculus (functions of one variable) to linear algebra (matrices operating on vectors in  $\mathbb{R}^n$ ) and partial differential equations (differential operators on infinite dimensional spaces).

### 3.1.1 Determinant of a Matrix

This concept is valid only for square matrices. Consider an  $n \times n$  matrix. A product of  $n$  elements can be written such that each element in the term is taken from a different row and a different column. This means the occurrence of  $a_{12}$  as an element in a term precludes the remaining  $n - 1$  elements from being chosen from the first row or the second column and so on. Hence the first subscripts of the elements in such a term are distinct (they are not repeated). So is the second subscript. Hence the term can be arranged such that the first or second subscript is in an ascending order. Arranging the second subscript in an ascending order results in the following form:

$$a_{\alpha_11}, a_{\alpha_22}, \dots, a_{\alpha_nn}$$

The first subscript denoted by  $\alpha_i$ 's are distinct and take on values from 1 to  $n$ .  $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$  form a permutation of the first  $n$  integers. An inversion is an interchange between two adjacent numbers in a permutation. Let  $N$  denote the number of inversions necessary to arrange the first subscripts  $\{\alpha_i\}$  in an ascending order. This depends on the permutation, i.e.  $N(\alpha_1, \alpha_2, \dots, \alpha_n)$ . The determinant  $D$  of a matrix is given as the sum of  $n!$  such terms, each term being a product of  $n$  terms.

$$\det A = D = \sum (-1)^{N(\alpha_1, \alpha_2, \dots, \alpha_n)} a_{\alpha_11} a_{\alpha_22} a_{\alpha_nn} \quad (3.2)$$

$N$  determines the sign associated with each term. The sign is positive (negative) if  $N$  is even (odd), see Shilov (1977).

**Example 3.4** Evaluate the determinant of the  $3 \times 3$  matrix

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

The  $3!$  or  $6$  terms are  $a_{11} a_{22} a_{33}$ ,  $a_{11} a_{32} a_{23}$ ,  $a_{12} a_{31} a_{23}$ ,  $a_{12} a_{21} a_{33}$ ,  $a_{13} a_{21} a_{32}$ ,  $a_{13} a_{22} a_{31}$ .

The calculation of  $N$  for the fourth and sixth terms is illustrated now. The fourth term written with the second subscript in ascending order is  $a_{21} a_{12} a_{33}$ . The first subscripts form the permutation 2 1 3. Interchanging the positions of 2 and 1, which are adjacent to each other, we obtain the permutation 1 2 3.

$$2 \ 1 \ 3 \longrightarrow 1 \ 2 \ 3$$

Here  $N = 1$ , and the sign of this term is negative.

The sixth term can be rearranged as  $a_{31}a_{22}a_{13}$ . The number of inversions to arrange the first subscript in ascending order is  $N = 3$ .

$$3 \quad 2 \quad 1 \xrightarrow{N=1} 3 \quad 1 \quad 2 \xrightarrow{N=2} 1 \quad 3 \quad 2 \xrightarrow{N=3} 1 \quad 2 \quad 3$$

This term is again associated with a negative sign. The value of  $N$  calculated for a given permutation is nonunique. It however will always be odd or always be even determining the sign uniquely. Now, evaluating  $N$  for all other terms, we get,

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

## **Definitions 3.2**

1. A matrix is called singular if its determinant is zero.
  2. The trace of a square matrix is the sum of its diagonal elements,  $\text{tr } A = \sum_{i=1}^n a_{ii}$ .
  3. The identity matrix  $I$  has a unity on the diagonal and zeroes elsewhere.
  4. The transpose of a matrix  $A$ , denoted by  $A'$ , is obtained by interchanging the rows of  $A$  with the columns of  $A$  and vice-versa. An  $m \times n$  matrix  $A$  on transposing results in an  $n \times m$   $A'$ .

**Example 3.5** Determine the transpose of

## Properties

1. The determinant of a square matrix  $A$  equals that of its transpose, i.e.  $\det A = \det A'$ .
  2. The determinant of a product of two square matrices  $A, B$  is equal to the product of their individual determinants, i.e.

$$\det(AB) = (\det A)(\det B)$$

### 3.1.2 Rank

The rank of a matrix is related to the linear independence of the vectors which make up the rows and columns of  $A$ . Let  $A$  be an  $m \times n$  matrix. The  $m$  rows of  $A$ ,  $r^1, r^2, \dots, r^m$  are row vectors, each having  $n$  coordinates. The  $n$  columns of  $A$ ,  $c^1, c^2, \dots, c^n$  are column vectors, each having  $m$  coordinates. The row space of  $A$  is the vector space spanned by the generating set  $r^1, r^2, \dots, r^m$ . The column space is the vector space spanned by the generating set  $c^1, c^2, \dots, c^n$ . The row (column) rank of a matrix is defined as the maximum number of linearly independent row (column) vectors of  $A$ . The row (column) rank of  $A$  is hence equal to the dimension of its row (column) space, see Noble and Daniel (1977) and Lipshutz (1971).

**Theorem 3.1** The row rank of a matrix is equal to its column rank. Their common value is called the rank of the matrix  $A$ .

*Proof.* We refer the reader to any standard book on linear algebra for a proof of this theorem.

**Definition 3.3** Let  $A$  be an  $m \times n$  matrix,  $A: \mathbf{R}^n \rightarrow \mathbf{R}^m$ . The adjoint of  $A$ , denoted by  $A^*$ , is defined as the matrix, which satisfies

$$\langle v, Au \rangle = \langle A^*v, u \rangle \text{ for all } u \in \mathbf{R}^n \text{ and } v \in \mathbf{R}^m \quad (3.3)$$

The definition of the adjoint is dependent on the inner-product defined on the space in which we are interested.

**Example 3.6** In the real inner-product defined in (2.9), show that the adjoint of a real matrix  $A$ , equals its transpose. For two real vectors  $u, v \in \mathbf{R}^n$ , the inner-product is written in the notation of matrix multiplication as

$$\langle v, u \rangle = v' u = \sum_{i=1}^n v_i u_i \quad (3.4)$$

This results in

$$\begin{aligned} \langle v, Au \rangle &= v' Au \\ &= (A'v)' u, \text{ (since } (AB)' = B'A') \\ &= \langle A'v, u \rangle \end{aligned}$$

Comparing this with (3.3), we obtain  $A^* = A'$ . The reader must remember that  $\langle v, Au \rangle$  is defined on  $\mathbf{R}^m$  whereas  $\langle A'v, u \rangle$  is defined on  $\mathbf{R}^n$ .

$A' = A$  for a real symmetric matrix. Here the adjoint of  $A$ ,  $A^*$  equals  $A'$ , which equals  $A$ , i.e.  $A^* = A' = A$ . Such a matrix, more generally an operator, is called a self-adjoint operator. Here the adjoint operator is equal to the original operator. The notion of an operator adjoint is general and can be extended to differential operators. Self-adjoint operators possess many important properties (see Section 3.2).

Most textbooks on linear algebra define an adjoint of a matrix  $A$  as the transpose of the matrix formed by co-factors of  $a_{ij}$ . This is more precisely called the adjugate of  $A$ . The adjoint operator defined here is completely different from the adjoint used in that context. We make special mention of this to avoid any confusion which may exist in the mind of the reader. The adjoint operator which we will be using throughout the text is defined so as to satisfy (3.3) and exists for any ( $m \times n$ ) matrix  $A$  (not necessarily a square matrix). The concept of adjoint operators plays an important role in optimisation (see Luenberger (1966), Burghes (1980)). The role of the adjoint is similar to that of an integrating factor for a system of differential equations.

On using the definition of the complex inner-product (2.11), we would have obtained  $A^* = \bar{A}'$ , where bar over  $A$  denotes complex-conjugate of elements of  $A$ . The conjugate transpose of a complex-matrix is also called its Hermitian, and is denoted as  $A^H$ . Here

$$A^* = A^H = \bar{A}'$$

## 3.2 EIGENVALUES AND EIGENVECTORS

The equation

$$Au = \lambda u \quad (3.5a)$$

where  $\lambda$  is a scalar, is the eigenvalue problem associated with the square matrix  $A$ . This homogeneous

equation admits the trivial solution  $u = 0$ , for all  $\lambda$ . There are a finite number of  $\lambda$ 's for which this equation possesses a nonzero solution. These values of  $\lambda$ 's are called the **eigenvalues**. The corresponding nonzero solution vectors  $u$  are called the **eigenvectors**. *The eigenvector is a nonzero vector by definition.* Equation (3.5a) can be recast as

$$(A - \lambda I)u = 0 \quad (3.5b)$$

This homogeneous equation has a nonzero solution if and only if  $\det(A - \lambda I) = 0$ . For an  $n$ th-order matrix, this condition yields an  $n$ th order polynomial in  $\lambda$  of the form

$$\lambda^n - \lambda^{n-1} (\text{tr } A) + \dots + (-1)^n \det A = 0 \quad (3.6)$$

This polynomial is called the **characteristic equation**. It has  $n$  roots for  $\lambda$ , from the fundamental theorem of algebra. These roots can be real or complex and are the eigenvalues of  $A$ . Every square matrix  $A$  has eigenvalues. For a real matrix, the coefficients of the polynomial are real and the complex eigenvalues always occur as a complex-conjugate pair.

The eigenvector corresponding to each  $\lambda$  is determined by solving for  $u$  in (3.5b). This being a homogeneous equation, the eigenvector can be found only to within a multiplicative constant, i.e.  $cu^1$  is an eigenvector if  $u^1$  is an eigenvector, for every scalar  $c$  (see Noble and Daniel, 1977). A real matrix, as mentioned earlier, can possess complex eigenvalues. The corresponding eigenvectors would possess complex elements. Some coordinates of the eigenvector would be complex numbers. It is therefore desirable to work in a complex vector space even while working with operators which are real matrices. The inner-product one should work in, while dealing with real matrices, is therefore necessarily the complex inner-product, defined as

$$\langle x, y \rangle = \bar{x}' y = \sum_{i=1}^n \bar{x}_i y_i \quad (3.7)$$

The properties of the eigenvalues and eigenvectors associated with the matrix operator on a finite-dimensional space will now be discussed now.

**Theorem 3.2** The eigenvalues of a real matrix  $A$  (i.e. with all its elements as real numbers) are equal to those of its adjoint  $A^*$  ( $= A'$ ).

*Proof.* The eigenvalue problem associated with  $A$  and  $A'$  is

$$Au = \lambda u \quad (3.8a)$$

$$A'v = \eta v \quad (3.8b)$$

The eigenvalues  $\lambda, \eta$  are obtained from the roots of the characteristic equation generated by

$$\det(A - \lambda I) = 0 \quad (3.9a)$$

$$\det(A' - \eta I) = 0 \quad (3.9b)$$

As  $\det B = \det B'$ , (3.9a) implies

$$\det(A - \lambda I)' = 0$$

or

$$\det(A' - \lambda I) = 0$$

Comparing this with (3.9b), it follows that  $A, A'$  generate the same characteristic equation. Thus the eigenvalues  $\lambda$  of  $A$  equal those of its transpose.

**Example 3.7** Determine the eigenvalues of  $A$ ,  $A'$  for

$$(a) A = \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}, \quad (b) A = \begin{bmatrix} 4 & -20 & -10 \\ -2 & 10 & 4 \\ 6 & -30 & -13 \end{bmatrix}$$

(a) The eigenvalues  $\lambda$  of  $A$  are obtained from

$$\det(A - \lambda I) = 0, \text{ i.e., } \det \begin{pmatrix} 2 - \lambda & 2 \\ 1 & 3 - \lambda \end{pmatrix} = 0$$

This yields the characteristic equation

$$\lambda^2 - 5\lambda + 4 = 0$$

The two eigenvalues of  $A$  are  $\lambda_1 = 1$ ,  $\lambda_2 = 4$ . The eigenvalues  $\eta$  of  $A'$  are obtained from

$$\det(A' - \lambda I) = 0, \text{ i.e., } \det \begin{pmatrix} 2 - \eta & 1 \\ 2 & 3 - \eta \end{pmatrix} = 0$$

The characteristic equation here is

$$\eta^2 - 5\eta + 4 = 0$$

The two eigenvalues here are  $\eta_1 = 1 = \lambda_1$ ,  $\eta_2 = 4 = \lambda_2$ .

(b) The eigenvalues  $\lambda$  of  $A$  are obtained from

$$\det(A - \lambda I) = 0, \text{ i.e., } \det \begin{pmatrix} 4 - \lambda & -20 & -10 \\ -2 & 10 - \lambda & 4 \\ 6 & -30 & -13 - \lambda \end{pmatrix} = 0$$

This yields the cubic characteristic equation

$$\lambda^3 - \lambda^2 - 2\lambda = 0$$

The three eigenvalues are  $\lambda_1 = -1$ ,  $\lambda_2 = 0$ ,  $\lambda_3 = 2$ . The cubic characteristic obtained by setting

$$\det(A' - \eta I) = 0$$

or

$$\det \begin{pmatrix} 4 - \eta & -2 & 6 \\ -20 & 10 - \eta & -30 \\ -10 & 4 & -13 - \eta \end{pmatrix} = 0$$

is given by

$$\eta^3 - \eta^2 - 2\eta = 0$$

The eigenvalues of  $A'$  are  $\eta_1 = -1 = \lambda_1$ ,  $\eta_2 = 0 = \lambda_2$ ,  $\eta_3 = 2 = \lambda_3$ .

**Theorem 3.3** The eigenvalues of a real symmetric matrix are real.

*Proof.* Assume the contrary, i.e. an eigenvalue  $\lambda$  of a real symmetric matrix can be complex:  $\lambda = \alpha + i\beta$ . The eigenvector  $u$  corresponding to  $\lambda$  can have complex elements. It is therefore essential to use the definition of the complex inner-product. The eigenvalue problem is

$$Au = \lambda u$$

Taking the inner-product with  $u$  on both sides, we get

$$\langle u, Au \rangle = \langle u, \lambda u \rangle \quad (3.10a)$$

By definition,

$$\langle A^*u, u \rangle = \lambda \langle u, u \rangle$$

$$\langle Au, u \rangle = \lambda \langle u, u \rangle \text{ (as } A = A^*)$$

$$\langle \lambda u, u \rangle = \lambda \langle u, u \rangle$$

or

$$\bar{\lambda} \langle u, u \rangle = \lambda \langle u, u \rangle$$

or

$$(\bar{\lambda} - \lambda) \|u\|^2 = 0 \quad (3.10b)$$

$u$  cannot be the zero vector since it is an eigenvector. Thus the last equation implies  $\lambda = \bar{\lambda}$  or  $\lambda$  is real.

**Remarks:** 1. This property is true for any self-adjoint operator. In Chapter 6 we will extend this theorem to differential operators.

2. This theorem does not imply that a nonsymmetric real matrix will have complex eigenvalues. This is easily verified from Example 3.7. The theorem is a sufficient condition, and it assures us that the eigenvalues are real, whenever the matrix is real and symmetric.

**Definition 3.4** A set of vectors  $\{u^i\}$  is said to be orthogonal if  $\langle u^i, u^j \rangle = 0$  for all  $i \neq j$ . It is said to be orthonormal if  $\langle u^i, u^j \rangle = \delta_{ij}$  for all  $i, j$  where  $\delta_{ij}$  is the Kronecker delta function defined as

$$\delta_{ij} = 1 \quad \text{for } i = j$$

$$= 0 \quad \text{for } i \neq j$$

**Theorem 3.4** The eigen-vectors  $u^i, u^j$  corresponding to two distinct eigenvalues  $\lambda_i, \lambda_j$  are orthogonal for a real symmetric matrix.

*Proof.* Here the eigenvalues are real from the earlier theorem and we can work in the real inner-product (2.9). The eigenvectors  $u^i, u^j$  satisfy

$$Au^i = \lambda_i u^i \quad (3.11a)$$

$$Au^j = \lambda_j u^j \quad (3.11b)$$

Taking the inner-product of (3.11a) with  $u^j$  on the left and of (3.11b) with  $u^i$  on the right and subtracting, we get

$$\langle u^j, Au^i \rangle - \langle Au^j, u^i \rangle = \langle u^j, \lambda_i u^i \rangle - \langle \lambda_j u^j, u^i \rangle$$

or

$$\langle A^*u^j, u^i \rangle - \langle Au^j, u^i \rangle = (\lambda_i - \lambda_j) \langle u^j, u^i \rangle$$

or

$$(\lambda_i - \lambda_j) \langle u^j, u^i \rangle = 0 \quad (3.12)$$

as  $A = A^*$ . Since the eigenvalues were chosen to be distinct,  $\lambda_i - \lambda_j \neq 0$ , and we have  $\langle u^j, u^i \rangle = 0$ , for  $i \neq j$ . This proves the result.

**Example 3.8** Show that the eigenvectors of the matrix  $A$  are orthogonal to each other

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

The cubic characteristic equation obtained from  $\det(A - \lambda I) = 0$  is

$$\lambda^3 - 8\lambda^2 + 19\lambda - 12 = 0$$

The three eigenvalues are  $\lambda_1 = 1$ ,  $\lambda_2 = 3$ ,  $\lambda_3 = 4$ . They are real, as they should be from Theorem 3.3.

The eigenvector  $u^1 (u_1^1, u_2^1, u_3^1)'$  corresponding to  $\lambda_1$  is obtained from

$$Au^1 = \lambda_1 u^1$$

Equating the coordinates, we obtain

$$\begin{aligned} 2u_1^1 - u_2^1 &= 0 \\ -u_1^1 + u_2^1 - u_3^1 &= 0 \\ -u_2^1 + 2u_3^1 &= 0 \end{aligned}$$

Eliminating  $u_2^1$  and using the third equation above, we get

$$2u_1^1 - 2u_3^1 = 0, \quad -u_1^1 + u_3^1 = 0$$

This is one independent equation in the two variables  $u_1^1, u_3^1$ . A nonzero solution can be obtained by assigning  $u_3^1 = 1$  (arbitrarily). This yields  $u_1^1 = 1$ ,  $u_2^1 = 2$ . This arbitrariness in choosing  $u_3^1$  results in the determination of  $u^1$  to within a multiplicative constant. The normalised vector  $u^1$  is  $\frac{1}{\sqrt{6}}(1, 2, 1)'$ . The normalised eigenvector  $u^2$ , corresponding to  $\lambda_2$ , is  $\frac{1}{\sqrt{2}}(1, 0, -1)'$ , and  $u^3$  corresponding to  $\lambda_3$  is  $\frac{1}{\sqrt{3}}(1, -1, 1)'$ .

$$\langle u^1, u^2 \rangle = \frac{1}{\sqrt{12}}[1.1 + 2.0 + (-1).1] = 0$$

Similarly,

$$\langle u^1, u^3 \rangle = 0 = \langle u^2, u^3 \rangle$$

The eigenvectors corresponding to different eigenvalues hence are orthogonal to each other.

**Theorem 3.5** The eigenvector  $u^i$  of a real matrix  $A$  corresponding to  $\lambda_i$  is orthogonal to the complex-conjugate of every eigenvector  $v^j$  of  $A'$  corresponding to an eigenvalue  $\lambda_j$  distinct from  $\lambda_i$ .

*Proof.* By definition  $u^i, v^j$  satisfy

$$Au^i = \lambda_i u^i \quad (3.13a)$$

$$A'v^j = \lambda_j v^j \quad (3.13b)$$

Taking the complex-conjugate of the second problem and remembering that  $A$  is real, we get

$$A'\bar{v}^j = \bar{\lambda}_j \bar{v}^j \quad (3.13c)$$

Taking the inner-product of (3.13a) on the left with  $\bar{v}^j$  and (3.13c) on the right with  $u^i$ , and subtracting

$$\langle \bar{v}^j, Au^i \rangle - \langle A'\bar{v}^j, u^i \rangle = \langle \bar{v}^j, \lambda_i u^i \rangle - \langle \bar{\lambda}_j \bar{v}^j, u^i \rangle$$

or

$$(\lambda_i - \bar{\lambda}_j) \langle \bar{v}^j, u^i \rangle = 0 \quad (3.14)$$

For  $\lambda_i$  distinct from  $\bar{\lambda}_j$ , it follows from  $\langle \bar{v}^j, u^i \rangle = 0$ . Here the inner-product is defined as in (3.7).

**Remark.** This theorem implies that the eigenvector of a matrix  $A$  corresponding to  $\lambda_i$  is orthogonal to the complex-conjugates of all the eigenvectors of its adjoint which correspond to all the other eigenvalues.

Consider two sets of  $n$  vectors each  $\{u^1, u^2, \dots, u^n\}$  and  $\{v^1, v^2, \dots, v^n\}$ . These are said to be **biorthogonal** if  $\langle u^i, v^j \rangle = 0$  for  $i \neq j$  and all  $i, j \in 1, n$ . A complete discussion of the properties of eigenvalues and eigenvectors can be found in (Amundson (1966), Noble and Daniel (1977)).

Let us see how the orthogonality and biorthogonality of the eigenvectors can be exploited in representing an arbitrary vector.

A vector  $x$  can be expressed as a linear combination of a basis  $\{u^i\}$  as

$$x = \sum c_i u^i \quad (3.15)$$

For any matrix  $A$ , which has  $n$  distinct eigenvalues, the eigenvectors of  $A$ , and those of  $A'$  form a biorthogonal set as seen from Theorem 3.5. This can be exploited to obtain the  $c_i$  efficiently if the basis  $\{u^i\}$  is chosen as the eigenvector set of  $A$  in (3.15).

Taking the inner-product of (3.15) on both sides with  $\bar{v}^j$ , and remembering that

$$\langle \bar{v}^j, u^i \rangle = 0 \text{ for } i \neq j$$

we have

$$\langle \bar{v}^j, x \rangle = c_j \langle \bar{v}^j, u^j \rangle$$

This yields

$$c_j = \frac{\langle \bar{v}^j, x \rangle}{\langle \bar{v}^j, u^j \rangle} \quad (3.16a)$$

If the matrix  $A$  is self-adjoint, i.e. real-symmetric, the eigenvectors of  $A$ ,  $u^i$  form an orthogonal set. This enables the  $c_i$  to be determined as

$$c_i = \frac{\langle u^i, x \rangle}{\langle u^i, u^i \rangle} \quad (3.16b)$$

**Example 3.9** Show that Theorem 3.5 holds for the  $2 \times 2$  matrix in Example 3.7. The eigenvector  $u^1(u_1^1, u_2^1)'$  corresponding to  $\lambda_1$  is obtained from

$$u_1^1 + 2u_2^1 = 0, \quad u_1^1 + 2u_2^1 = 0$$

This is one independent equation in two unknowns. Assigning  $u_2^1 = 1$ , an arbitrary nonzero value yields  $u_1^1 = -2$ . The unique normalised eigenvector is  $u^1 = \frac{1}{\sqrt{5}}(-2, 1)'$ .  $u^2$  can be determined as  $u^2 = \frac{1}{\sqrt{2}}(1, 1)'$ .

The eigenvector  $v^1(v_1^1, v_2^1)'$  of  $A'$  corresponding to  $\lambda_1$  is the solution of

$$v_1^1 + v_2^1 = 0, \quad 2v_1^1 + 2v_2^1 = 0$$

The normalised eigenvector  $v^1$  is  $\frac{1}{\sqrt{2}}(-1, 1)'$ . Similarly, we obtain  $v^2 = \frac{1}{\sqrt{5}}(1, 2)'$ . Clearly,

$$\langle u^1, v^2 \rangle = \frac{1}{5}(-2.1 + 1.2) = 0$$

$$\langle u^2, v^1 \rangle = \frac{1}{2}(-1.1 + 1.1) = 0$$

**Example 3.10** Verify that the  $3 \times 3$  matrix  $A$  of Example 3.7 satisfies Theorem 3.5.

Let  $u^1(u_1^1, u_2^1, u_3^1)$  be the eigenvector corresponding to  $\lambda_1$ . The coordinates are found by solving

$$5u_1^1 - 20u_2^1 - 10u_3^1 = 0$$

$$-2u_1^1 + 11u_2^1 + 4u_3^1 = 0$$

$$6u_1^1 - 30u_2^1 - 12u_3^1 = 0$$

Eliminating  $u_1^1$  from the second and third equations and using the first yields  $u_2^1 = 0$  and one independent equation  $u_1^1 = 2u_3^1$ . Assuming  $u_3^1 = 1$ , we get the normalised  $u^1$  as  $\frac{1}{\sqrt{5}}(2, 0, 1)'$ . We can similarly determine  $u^2$  as  $\frac{1}{\sqrt{26}}(5, 1, 0)'$  and  $u^3$  as  $\frac{1}{\sqrt{5}}(0, 1, -2)'$ .

The coordinates  $v_1^1, v_2^1, v_3^1$  of the eigenvector  $v^1$  of  $A'$  corresponding to  $\lambda_1$  are given by

$$5v_1^1 - 2v_2^1 + 6v_3^1 = 0$$

$$-20v_1^1 + 9v_2^1 - 30v_3^1 = 0$$

$$-10v_1^1 + 4v_2^1 - 12v_3^1 = 0$$

This is a system of two independent equations. The last equation is a multiple of the first. Eliminating  $v_1^1$  between the first two equations, we get  $v_2^1 = 6v_3^1$ . Assigning  $v_3^1 = 1$ , we obtain  $v_2^1 = 6$ ,  $v_1^1 = 6/5$ . The normalised eigenvector is  $\frac{1}{31}(-6, 30, 15)'$ . Similarly, we obtain  $v^2 = \frac{1}{\sqrt{21}}(-1, 4, 2)'$  and  $v^3 = \frac{1}{\sqrt{30}}(1, -5, -2)$ . It is easy to verify

$$\langle u^i, v^j \rangle = 0 \quad \text{for } i \neq j$$

**Theorem 3.6** Consider a matrix  $A$  ( $n \times n$ ), whose eigenvalues are all distinct, i.e. no eigenvalue is repeated. The  $n$  eigenvectors of  $A$  form a linearly independent set.

*Proof.* The proof is by contradiction. Assume the contrary, i.e. the set of  $n$  eigenvectors is dependent. Let  $r$  be the maximum number of vectors in the set which are independent. Without any loss of generality we take this to be the first  $r$  eigenvectors,  $(u^1, u^2, \dots, u^r)$ .  $u^{r+1}$  can then be expressed as a linear combination of the first  $r$  vectors

$$u^{r+1} = \sum_{i=1}^r c_i u^i \quad (3.17)$$

Operating on both sides with  $A$ , we obtain

$$\begin{aligned} Au^{r+1} &= \sum_{i=1}^r c_i A u^i \\ \lambda_{r+1} u^{r+1} &= \sum_{i=1}^r c_i \lambda_i u^i \end{aligned} \quad (3.18a)$$

Multiplying (3.17) by  $\lambda_{r+1}$ , we get

$$\lambda_{r+1} u^{r+1} = \sum_{i=1}^r c_i \lambda_{r+1} u^i \quad (3.18b)$$

Subtracting (3.18b) from (3.18a), we obtain

$$\sum_{i=1}^r c_i (\lambda_i - \lambda_{r+1}) u^i = 0 \quad (3.19)$$

Since  $\{u^i\}_{i=1, r}$  is an independent set, and  $\{\lambda_i - \lambda_{r+1}\}$  is nonzero as the eigenvalues are distinct, (3.19) is satisfied only for all  $c_i = 0$ . This implies that  $u^{r+1} = 0$  from (3.18b). This cannot be true as  $u^{r+1}$  is an eigenvector, and must be nonzero. Our assumption that the set of  $n$  eigenvectors was dependent is wrong. They are therefore independent.

The roots of the characteristic equation can be repeated  $m$  times. Such an eigenvalue is said to be algebraically multiple. An algebraically simple eigenvalue corresponds to one which is not repeated, i.e. it occurs only once. Each distinct eigenvalue has at least one eigenvector associated with it. This assures us that the matrix  $A$  has  $n$  eigenvectors in Theorem 3.6. An eigenvalue repeated  $m$  times can have less than  $m$  independent eigenvectors. This will be illustrated in Examples 3.11 and 3.12. The number of eigenvectors associated with an eigenvalue is called its geometric multiplicity. The geometric multiplicity of an eigenvalue is therefore less than or equal to the algebraic multiplicity. An eigenvalue is geometrically simple when both multiplicities are equal.

**Example 3.11** Determine the eigenvalues and eigenvectors of  $A$

$$\begin{bmatrix} 2 & 2 & -6 \\ 2 & -1 & -3 \\ -2 & -1 & 1 \end{bmatrix}$$

The eigenvalues of  $A$  are the roots of the cubic characteristic equation

$$\lambda^3 - 2\lambda^2 - 20\lambda - 24 = 0$$

Therefore,  $\lambda_1 = \lambda_2 = -2$ ,  $\lambda_3 = 6$ . The root  $-2$  is a repeated root. The eigenvalue  $\lambda_3 = 6$  is algebraically simple as opposed to  $\lambda_1 = -2$  which is algebraically multiple.

For  $\lambda_{1,2} = -2$ , the eigenvector  $u^1 (u_1^1, u_2^1, u_3^1)$  is obtained by solving the system of equations.

$$4u_1^1 + 2u_2^1 - 6u_3^1 = 0$$

$$2u_1^1 + u_2^1 - 3u_3^1 = 0$$

$$-2u_1^1 - u_2^1 + 3u_3^1 = 0$$

This is one independent equation in three variables, i.e. the rank of  $(A + 2I)$  is 1. Assigning  $u_3^1 = 0$ ,  $u_1^1 = 1$  yields  $u_2^1 = -2$ . A normalised eigenvector is  $\frac{1}{\sqrt{5}}(1, -2, 0)'$ .

Since the eigenvalue  $-2$  is repeated twice, we can seek a second independent eigenvector corresponding to it. Assigning  $u_3^2 = 1$ ,  $u_1^2 = 1$  yields  $u_2^2 = 1$ .

A second normalised eigenvector corresponding to  $-2$  is  $u^2 \frac{1}{\sqrt{3}}(1, 1, 1)'$ .  $u^1, u^2$  are independent, and span a two-dimensional sub-space in  $R^3$ . All other eigenvectors corresponding to  $\lambda_2 = -2$  can be written as a linear combination  $u^1, u^2$ .

We could have found  $u^2$  such that, in addition to being an independent eigenvector, it would have been orthogonal to  $u^1$ . The coordinates of such a  $u^2$ , denoted by  $u_0^2$ , satisfy

$$2u_{01}^2 + u_{02}^2 - 3u_{03}^2 = 0$$

$$u_{01}^2 - 2u_{02}^2 = 0$$

This yields  $u_0^2 = \frac{1}{\sqrt{70}}(6, 3, 5)$ .  $u^1, u_0^2$  is an orthogonal set which spans the same eigensubspace as spanned by  $u^1, u^2$ .

The eigenvector  $u^3 = \frac{1}{\sqrt{6}}(2, 1, -1)$  along with  $u^1, u^2$  or  $u^1, u_0^2$  forms a linearly independent set of three vectors in  $R^3$ . The algebraic multiplicity of  $\lambda_1$  is 2 as is its geometric multiplicity. This eigenvalue is hence geometrically simple.

**Example 3.12** Determine all eigenvalues of  $A$ .

$$\begin{bmatrix} 7 & 1 & 2 \\ -1 & 7 & 0 \\ 1 & -1 & 6 \end{bmatrix}$$

The three eigenvalues are  $\lambda_1 = 8$ ,  $\lambda_2 = \lambda_3 = 6$ . The eigenvector  $u^1$  is  $\frac{1}{\sqrt{3}}(-1, 1, -1)'$ . The eigenvector corresponding to  $\lambda_2, \lambda_3$ , i.e.  $u^2$  is obtained from

$$u_1^2 + u_2^2 + 2u_3^2 = 0$$

$$-u_1^2 + u_2^2 = 0$$

$$u_1^2 - u_2^2 = 0$$

This is a set of two independent equations in three coordinates. The rank of  $(A - 6I)$  is 2. There is only one independent eigenvector for this eigenvalue. The normalised vector  $u^2$  is

$\frac{1}{\sqrt{3}}(1, 1, -1)'$ . All other eigenvectors are scalar multiples of this vector. The eigenvalue  $\lambda_2 = 6$  has algebraic multiplicity equal to two and geometric multiplicity equal to one here.

Matrix operators whose eigenvalues are not algebraically simple therefore may not have  $n$  independent eigenvectors. The eigenvector set in such cases does not constitute a basis. From now on we assume that all our eigenvalues are algebraically simple. This assures us that the matrix  $A$  has  $n$  independent eigenvectors which constitute a basis in  $\mathbb{R}^n$ . The eigenvectors of a real-symmetric matrix (self-adjoint operator) always generate an orthogonal basis in  $\mathbb{R}^n$ . For a self-adjoint operator with repeated eigenvalues, it can be shown that there always exist  $n$  independent eigenvectors. The general class of matrices which possess  $n$ -independent eigenvectors are called **normal** matrices (see Noble and Daniel, 1977).

The theorems in this chapter will be exploited in obtaining solutions to finite dimensional systems in the next chapter. They, however, have a much wider significance as similar results will be proven for differential operators in Chapter 6.

### 3.3 FREDHOLM ALTERNATIVE (SOLVABILITY CONDITIONS)

We now discuss conditions under which linear algebraic equations have solutions. Let the  $n \times n$  real matrix  $A$  have  $n$  distinct eigenvalues. This assures us that  $A$  has  $n$  independent eigenvectors  $u^i$ . The equation  $(A - \mu I)u = b$  for real  $\mu$  has a unique solution when  $\mu$  is not an eigenvalue of  $A$ , given by

$$u = \sum_{i=1}^n \frac{\langle \bar{v}^i, b \rangle u^i}{\langle \bar{v}^i, u^i \rangle (\lambda_i - \mu)} \quad (3.20)$$

Here  $v^i$  are the eigenvectors of  $A'$  corresponding to  $\lambda_i$ .

When  $\mu$  is an eigenvalue of  $A$ , the equation has a solution if and only if  $\langle v^i, b \rangle = 0$ , for all  $v^i$ , the eigenvectors of  $A'$  corresponding to  $\lambda^i (= \mu)$ . The nonhomogeneous equation under these conditions has an infinity of solutions. These solutions are obtained by adding a constant times  $u^i$ , the eigenvector associated with  $\mu (= \lambda_i)$ , to the particular solution.

We seek the solution

$$u = \sum_{i=1}^n c_i u^i$$

and expand

$$b = \sum_{i=1}^n \beta_i u^i$$

Clearly from (3.16a),

$$\beta_i = \frac{\langle \bar{v}^i, b \rangle}{\langle \bar{v}^i, u^i \rangle} \quad (3.21)$$

Also,  $(A - \mu I)u = b$  implies,

$$\begin{aligned} \sum_{i=1}^n (c_i A u^i - \mu I c_i u^i) &= \sum_{i=1}^n \beta_i u^i \\ \sum_{i=1}^n [(c_i(\lambda_i - \mu) - \beta_i) u^i] &= 0 \end{aligned} \quad (3.22)$$

Since  $u^i$  forms an independent set, (3.22) can only be satisfied if it is a trivial relation, i.e.

$$c_i = \frac{\beta_i}{\lambda_i - \mu} \quad \text{for all } i$$

This implies

$$u = \sum_{i=1}^n \frac{\beta_i u_i}{\lambda_i - \mu}$$

Substituting the expression (3.21) for  $\beta_i$ , we obtain (3.20).

If  $\mu$  is an eigenvalue  $\lambda_i$  of  $A$ , this expression becomes singular. Now  $\mu$  is an eigenvalue of  $A'$  as well, and  $v^i$  satisfies

$$(A' - \mu I)v^i = 0 \quad (3.23a)$$

Also,

$$(A - \mu I)u = b \quad (3.23b)$$

Taking the inner-product of (3.23a) with  $u$  on the left after taking complex-conjugate and of (3.23b) with  $\bar{v}^i$  on the right and subtracting, we get

$$\langle b, \bar{v}^i \rangle = 0 \quad (3.24)$$

for real  $\mu$ . So the equation has a solution only if  $\langle \bar{v}^i, b \rangle = 0$ , for all eigenvectors  $v^i$  of  $A'$  associated with  $\mu$ .

**Remark.** When  $\mu$  is a repeated eigenvalue,  $\langle \bar{v}^i, b \rangle = 0$  must be satisfied by all eigenvectors  $v^i$  of  $A'$  associated with it.

**Corollary.** (i) If  $\mu = 0$ , then  $Au = b$  has a unique solution if and only if zero is not an eigenvalue of  $A$ . This implies  $\det A \neq 0$ , and  $A$  is nonsingular.

(ii) If zero is an eigenvalue of  $A$ ,  $Au = b$  has a solution if and only if  $\langle \bar{v}^i, b \rangle = 0$ , where  $v^i$  is the set of independent eigenvectors of  $A'$  corresponding to the zero eigenvalue. A more formal discussion of the Fredholm alternative can be found in Noble and Daniel (1977) and Stakgold (1979).

We illustrate the applicability of the solvability condition with a few examples.

**Example 3.13** For

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

determine

- (a) whether  $Au = (2, 4, 7)'$  has a solution;
- (b) for what  $a$ , does  $Au = (2, a, 8)'$  possess a solution.

The eigenvalues of  $A$  are  $\lambda_1 = 0$ ,  $\lambda_2 = -4.58$ ,  $\lambda_3 = 4.58$ . The matrix  $A$  is real-symmetric. Its eigenvectors hence form an orthogonal set. The eigenvector set  $v^i$  of  $A'$  and  $u^i$  of  $A$  are identical. The normalised eigenvector,  $u^1$  is  $\frac{1}{\sqrt{3}}(1, 1, 1)'$ . This is the only eigenvector as  $\lambda_1$  is algebraically simple.

(a) Since zero is an eigenvalue of  $A$ , and  $A = A'$ ,

$$\begin{aligned}\langle \bar{v}^1, b \rangle &= \langle u^1, b \rangle \\ &= \frac{13}{\sqrt{3}} \\ &\neq 0\end{aligned}$$

Hence  $Au = b$  has no solution. The system of equations is inconsistent.

(b) The system has a solution if

$$\langle \bar{v}^1, b \rangle = \langle u^1, b \rangle = 0$$

This implies

$$(2 + a + 8)/\sqrt{3} = 0 \text{ or } a = -10$$

The vector  $u^p = (-4/7, 22/7, 0)'$  is a particular solution of  $Au = b$ . The infinity of solutions of the form  $u = u^p + cu^1$ , where  $c$  is an arbitrary constant satisfies  $Au = b$ , as

$$Au = A(u^p + cu^1) = Au^p + cAu^1 = b + 0 = b$$

**Example 3.14** For the matrix  $A$  in Example 3.11, determine if  $(A + 2I)u = b$  has a solution for

(a)  $b = (2, 1, -1)'$

(b)  $b = (1, 1, -1)'$

The matrix  $A$  is nonsymmetric. The eigenvector set  $\{v^i\}$  of  $A'$  is distinct from eigenvector  $\{u^i\}$  of  $A$ .  $\lambda = -2$  is a twice repeated eigenvalue of  $A$ ,  $A'$ .  $v^1 = \frac{1}{\sqrt{5}}(1, 0, 2)'$  and  $v^2 = \frac{1}{\sqrt{2}}(0, 1, 1)'$  are two independent eigenvectors of  $A'$ . They span the eigenspace of  $A'$  corresponding to  $\lambda' = -2$ . We could have chosen these vectors to be orthogonal or chosen any other set of two eigenvectors.

(a) The system has a solution if and only if  $b$  is orthogonal to both  $v^1, v^2$ .

$$\langle \bar{v}^1, b \rangle = 2 + 0 - 2 = 0$$

$$\langle \bar{v}^2, b \rangle = 0 + 1 - 1 = 0$$

The system has an infinity of solutions.

(b)  $\langle \bar{v}^2, b \rangle = 0 + 1 - 1 = 0$

$$\langle \bar{v}^1, b \rangle = 1 + 0 - 2 = -1$$

The system  $(A + 2I)u = b$ , now has no solutions. The system of equations is inconsistent.

It is easy to verify that the above results would not have changed if we had chosen another independent set  $v^1, v^2$  to span the eigenspace of  $A'$ .

### 3.4 RAYLEIGH'S QUOTIENT

An  $n$ th order matrix has  $n$  eigenvalues. In many problems it is not necessary to determine all the eigenvalues explicitly.

While studying transient behaviour, a quantity of interest to the engineer is the “time constant” of a process. The eigenvalues of the matrix  $A$  are a measure of the time constant of a system. The eigenvalues play an important role in determining the transient response of linear systems.

Consider a linear system at steady state. Let the real parts of all the eigenvalues be negative,

and be small in magnitude. Here the system, when subject to a disturbance, takes a sufficiently long time to relax back to the steady state.

When the eigenvalues have negative real parts, with a large magnitude, the system relaxes to the steady state immediately. The time constant of the system here is small.

Theoretically linear systems attain a steady state only as  $t \rightarrow \infty$ . The approach to steady state is exponential and determined usually by fixing a tolerance value which specifies how close the system is to the steady state.

Fixing the tolerance criterion as within 1% of the steady state, we obtain bounds on the time constant ' $\tau$ ' of the process as

$$\min \text{ over all } i \frac{5}{\operatorname{Re}(\lambda_i)} < \tau < \max \text{ over all } i \frac{5}{\operatorname{Re}(\lambda_i)}$$

since  $e^{-5} = .0067 (< 1\%)$ . Here we have used norm of the deviation from the steady state relative to the initial perturbation, i.e.

$$\|x(t) - x^{ss}\| / \|x(0)\|$$

to determine how close the system is to the steady state.

The time-step in the numerical integration of a system of ordinary differential equations is dictated by the lowest eigenvalue of a linear matrix. The spread of eigenvalues indicates the presence of widely different time scales. Hence, it is often of interest to obtain an estimate of the lowest and the largest eigenvalue of a matrix operator. The Rayleigh quotient enables us to obtain estimates of these eigenvalues (see Noble and Daniel, 1977). This concept is valid only for self-adjoint matrix operators as they have real eigenvalues. We are primarily interested in real-symmetric matrices here.

The Rayleigh quotient is a scalar  $\rho(x)$  defined as

$$\rho(x) = \frac{\langle x, Ax \rangle}{\langle x, x \rangle} \quad (3.25)$$

where  $x \in \mathbb{R}^n$ ,  $A$  is a real symmetric  $n \times n$  matrix.

**Theorem 3.7** If the  $n$ -eigenvalues of a real symmetric matrix  $A$  are ordered as

$$\lambda_1 \leq \lambda_2 \leq \lambda_3 \dots \leq \lambda_n$$

then

- (a)  $\lambda_1 = \min \rho(x), \quad x \neq 0 \text{ in } \mathbb{R}^n$
- (b)  $\lambda_n = \max \rho(x), \quad x \neq 0 \text{ in } \mathbb{R}^n$

*Proof.* (a) Any  $x$  can be expressed as a linear combination of the  $n$  eigenvectors  $\{u^i\}$  which form a basis as

$$x = \sum_{i=1}^n c_i u^i$$

Note that we are assured of  $n$  eigenvectors, even if we have repeated eigenvalues, as  $A$  is self-adjoint.

$$\rho(x) = \frac{\langle x, Ax \rangle}{\langle x, x \rangle} = \frac{\left\langle \sum_{i=1}^n c_i u^i, \sum_{i=1}^n c_i \lambda_i u^i \right\rangle}{\left\langle \sum_{i=1}^n c_i u^i, \sum_{i=1}^n c_i u^i \right\rangle}$$

$$= \frac{\sum_{i=1}^n \lambda_i \bar{c}_i c_i}{\sum_{i=1}^n \bar{c}_i c_i} \quad (\text{using the orthogonality of eigenvectors for a self-adjoint operator}) \quad (3.26)$$

Subtracting  $\lambda_1$  from both sides of (3.26), we get

$$\rho - \lambda_1 = \frac{(\lambda_2 - \lambda_1)\bar{c}_2 c_2 + (\lambda_3 - \lambda_1)\bar{c}_3 c_3 + \dots + (\lambda_n - \lambda_1)\bar{c}_n c_n}{\bar{c}_1 c_1 + \bar{c}_2 c_2 + \dots + \bar{c}_n c_n}$$

The ordering of the eigenvalues assures us that  $(\lambda_i - \lambda_1) \geq 0$  for  $i = 2, 3, \dots, n$  or  $\rho - \lambda_1 \geq 0$ . Clearly,

$$\rho = \lambda_1 \quad \text{if } x = c_1 u^1$$

(b) Similarly, we can prove that  $\rho - \lambda_n \leq 0$  or  $\rho \leq \lambda_n$ ,  $\rho = \lambda_n$  when  $x = c_n u^n$ . Choosing different vectors  $x \in \mathbb{R}^n$ , we can obtain bounds on the minimum and maximum eigenvalues by computing  $\rho(x)$ .

**Example 3.15** Obtain an estimate of the eigenvalues of the matrix in Example 3.8.

We choose arbitrary vectors  $x^1 = (1, 1, 1)^T$ ,  $x^2 = (1, 0, 1)^T$  as trial elements in  $\mathbb{R}^3$ .

$$\rho_1(x) = \frac{\langle x^1, Ax^1 \rangle}{\langle x^1, x^1 \rangle} = \frac{\langle (1, 1, 1)^T, (2, 0, 2)^T \rangle}{\langle (1, 1, 1)^T, (1, 1, 1)^T \rangle} = \frac{4}{3} = 1.33$$

$$\rho_2(x) = \frac{\langle x^2, Ax^2 \rangle}{\langle x^2, x^2 \rangle} = \frac{\langle (1, 0, 1)^T, (3, -2, 3)^T \rangle}{\langle (1, 0, 1)^T, (1, 0, 1)^T \rangle} = \frac{6}{2} = 3$$

It follows that the lowest eigenvalue  $\lambda_1 \leq 1.33$  and the largest eigenvalue  $\lambda_3 \geq 3$ . The exact values  $\lambda_1 = 1$  and  $\lambda_3 = 4$ .

**Example 3.16** Use the Rayleigh quotient to determine the eigenvalues of

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

The eigenvectors belong to  $\mathbb{R}^2$  and are of the form  $[x_1 \ x_2]^T$ . Since the eigenvector can only be determined to within a multiplicative constant, we choose our trial vector to be of the form  $(1, a)^T$ . Our objective is to determine  $x$ , i.e. 'a', so that the quotient is maximised or minimised.

$$\rho(x) = \frac{\langle x, Ax \rangle}{\langle x, x \rangle} = \frac{a^2 + 4a + 1}{a^2 + 1}$$

The extremum of  $\rho(x)$  is obtained by setting the derivative of this equation with respect to  $a$  to zero. This yields

$$a^2 - 1 = 0 \text{ or } a = +1, -1$$

Substituting these values in  $\rho(x)$ , we obtain  $\lambda_1 = -1$  and  $\lambda_2 = +3$ . The corresponding eigenvectors are  $u^1 = (1 \ -1)^T$ ,  $u_2 = (1 \ 1)^T$ , (Will this method be always applicable?).

The set of eigenvectors of a matrix  $A$ , occurring in a problem constitute a natural basis. This basis set is generated by the matrix  $A$ . Therefore, it contains some information on the physics of the system. In Chapter 4, we seek solutions to problems in terms of this basis set. The presentation in this chapter is such that the ideas here can be easily extended to partial differential equations, in Chapter 6. All concepts presented in this chapter for matrix operators have a close analogy to those in Chapter 6 for differential operators.

## PROBLEMS

**1.** Discuss whether the following statements are true or false with reason:

- (i) Every  $n \times n$  matrix with real elements has  $n$  real eigenvalues.
- (ii) There is at least one eigenvector associated with every eigenvalue.
- (iii) If an eigenvalue of a matrix is repeated ' $m$ ' times, there are always  $m$  eigenvectors associated with it.
- (iv) A real symmetric matrix has  $n$  real eigenvalues.
- (v) A real symmetric matrix has  $n$  real distinct eigenvalues.
- (vi) Every real symmetric matrix always has an orthonormal set of  $n$  eigenvectors.

**✓ 2.** Determine the eigenvalues and eigenvectors of  $A$ ,  $A'$  and verify the relevant theorems:

$$(i) \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix}, \quad \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$$

$$(ii) \begin{bmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{bmatrix}, \quad \begin{bmatrix} 4 & -20 & -10 \\ -2 & 10 & 4 \\ -6 & -30 & -13 \end{bmatrix}$$

$$(iii) \begin{bmatrix} 7 & 1 & 2 \\ -1 & 7 & 0 \\ 1 & -1 & 6 \end{bmatrix}, \quad \begin{bmatrix} 2 & 2 & -6 \\ 2 & -1 & -3 \\ -2 & -1 & 1 \end{bmatrix}$$

**✓ 3.** Determine an orthonormal set of eigenvectors wherever it is possible:

$$(i) A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

$$(ii) \begin{bmatrix} 7 & -16 & -8 \\ -16 & 7 & 8 \\ -8 & 8 & -5 \end{bmatrix}$$

(iii) 
$$\begin{bmatrix} 2 & 2 & -6 \\ 2 & -1 & -3 \\ -2 & -1 & 1 \end{bmatrix}$$

(iv) 
$$\begin{bmatrix} 5 & 4 & -4 \\ 4 & 5 & 4 \\ -4 & 4 & 5 \end{bmatrix}$$

4. Let  $A = \begin{bmatrix} 2 & 0 & 0 \\ 5 & 4 & 0 \\ 3 & 6 & 1 \end{bmatrix}$

- (i) Obtain the eigenvector set of  $A$ .
- (ii) Find the set biorthogonal to this.
- (iii) Verify the biorthogonality property.
- (iv) Does each of these sets form a basis? Why?

5. A real matrix is skew-symmetric if  $A' = -A$ . For such a matrix prove that the eigenvalues are purely imaginary or zero.

6. Determine the eigenvalues and eigenvectors of the following matrices and verify all the relevant theorems:

(i) 
$$\begin{bmatrix} 5 & 8 \\ -6 & -9 \end{bmatrix}$$

(ii) 
$$\begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}$$

7. Write a computer code to determine the determinant of a general  $n \times n$  matrix.

## REFERENCES

Amundson, N.R., Mathematical Methods in Chemical Engineering, Vol. 1, Prentice-Hall, Englewood Cliffs, New Jersey (1966).

Burghes, D.N. and Graham, A., Introduction to Control Theory including Optimal Control, Ellis Horwood, London (1980).

Kaplan, W., Operational Methods for Linear Systems, Addison-Wesley, Reading, Mass. (1962).

- Kreyszig, E., Advanced Engineering Mathematics, Wiley, New York (1982).
- Lipshutz, Seymour, Schaums Outline of Theory and Problems of State Space and Linear Systems, McGraw-Hill, New York (1971).
- Luenberger, D.G., A generalised maximum principle, in Recent Advances in Optimisation Techniques, Lavi A. and Vogl T.P. (Eds.), Wiley, New York (1966).
- Noble, B. and Daniel, J.W., Applied Linear Algebra, Prentice-Hall, Englewood Cliffs, New Jersey (1977).
- Shilov, G.E., Linear Algebra, translated by Silvermann, R.A., Dover Publications, New York (1977).
- Stakgold, I., Green's Function and Boundary-value Problem, Wiley, New York (1979).

# 4

## Applications to Chemical Engineering Systems

---

In this chapter we develop the method of solution to three classes of equations: (i) linear algebraic equations, (ii) linear homogeneous ordinary differential equations, and (iii) linear nonhomogeneous ordinary differential equations. The differential equations we are concerned with here are initial-value problems (IVP). We discuss examples from various chemical engineering systems which are governed by the above classes of finite-dimensional equations. The method of solution proposed here for linear algebraic systems is extended to solve systems of ordinary differential equations. In the next section we extend the technique to partial differential equations. This is rendered possible as the technique exploits the properties of the operator describing the system in its domain space. The approach thus adopts a natural framework for each system, making it sufficiently general and flexible.

### 4.1 LINEAR ALGEBRAIC EQUATIONS

The general system of linear algebraic equations can be written as

$$Ax = b \quad (4.1)$$

where  $A$  is an  $n \times n$  matrix and  $b, x$  are vectors belonging to  $\mathbf{R}^n$ . For simplicity we further assume that  $A$  has a distinct set of eigenvalues  $\lambda_i$ . Also, we restrict all eigenvalues to be nonzero. The existence of  $n$ -distinct eigenvalues assures us that  $A$  has a linearly independent set of  $n$  eigenvectors  $\{u^1, \dots, u^n\}$  which forms a basis set. This basis set is a *natural* one as it is generated by the matrix  $A$ . The solution  $x$  to (4.1) is sought as a linear combination of this basis in our approach. Writing  $x$  as a linear combination of the  $u^i$ 's, we have

$$x = \sum_{i=1}^n c_i u^i \quad (4.2)$$

To determine  $x$ , we have to find the  $c_i$ 's for all  $i$ . The  $c_i$ 's are to be found such that  $x$  satisfies (4.1). Operating on both sides of (4.2) by the matrix  $A$ , we get

$$Ax = A \sum_{i=1}^n c_i u^i$$

Clearly,  $A$  is a linear operator. So definition (3.1) yields

$$b = \sum_{i=1}^n c_i A u^i$$

Since  $u^i$  is an eigenvector, we have

$$b = \sum_{i=1}^n c_i \lambda_i u^i \quad (4.3)$$

The vector  $b$  can be represented in terms of this basis as

$$b = \sum_{i=1}^n \beta_i u^i \quad (4.4)$$

where  $\beta_i$ 's are scalars.

Substituting this equation in (4.3), we have

$$\sum_{i=1}^n (\beta_i - c_i \lambda_i) u^i = 0 \quad (4.5)$$

The basis set  $\{u^i\}$  is independent. Hence the only relation that can satisfy (4.5) is the trivial relation. This yields

$$c_i = \beta_i / \lambda_i \quad (4.6)$$

The coefficients  $c_i$  in (4.2) have been obtained now in terms of  $\beta_i$ 's in (4.4) and the eigenvalues  $\lambda_i$  of the matrix  $A$ . The properties of the matrix operator  $A$  discussed in Chapter 3 can be exploited to determine the  $\beta_i$ 's elegantly. We consider two cases of the matrix  $A$ :

*Case 1:  $A$  is self-adjoint (real-symmetric).* The eigenvector set  $\{u^i\}$  now forms an orthogonal set, i.e.,

$$\langle u^i, u^j \rangle = 0 \quad \text{if } i \neq j$$

Taking the inner-product of (4.4) on the left with  $u^j$ , we have

$$\langle u^j, \sum_{i=1}^n \beta_i u^i \rangle = \langle u^j, b \rangle$$

Using the orthogonality of the set, we obtain

$$\langle u^j, \beta_j u^j \rangle = \langle u^j, b \rangle$$

or

$$\beta_j = \frac{\langle u^j, b \rangle}{\langle u^j, u^j \rangle} \quad (4.7)$$

Equations (4.6) and (4.7) determine the  $c_i$ 's uniquely. The solution is obtained from (4.2), once the  $c_i$ 's are found.

Should  $A$  be singular, zero is an eigenvalue of  $A$ . From (4.6), one of the  $c_i$ 's becomes indeterminate. If the corresponding  $\beta_j \neq 0$ , there is no solution  $x$  of (4.1) for this case. This is consistent with what we expect. If the corresponding  $\beta_j = 0$ , then we can have an infinite number of solutions (see Fredholm Alternative discussed in Chapter 3).

*Case 2:  $A$  is not a self-adjoint.* The assumption that  $A$  has distinct eigenvalues assures us that  $\{u^i\}$  constitutes a basis in  $\mathbb{R}^n$ . The set  $\{u^i\}$  now is not an orthogonal set. Equation (4.3) is valid for this case also. The  $\beta_i$ 's are obtained elegantly now by exploiting the biorthogonality property

between the eigenvectors  $\{u^i\}$  of  $A$  and the complex-conjugates of the eigenvectors  $\{v^j\}$  of  $A'$ . This is valid only when  $A$  has  $n$  distinct eigenvalues. Taking the inner-product on the left of (4.4) with  $\bar{v}^j$ , we get

$$\langle \bar{v}^j, \sum_{i=1}^n \beta_i u^i \rangle = \langle \bar{v}^j, b \rangle$$

$$\langle \bar{v}^j, \beta_j u^j \rangle = \langle \bar{v}^j, b \rangle$$

or

$$\beta_j = \frac{\langle \bar{v}^j, b \rangle}{\langle \bar{v}^j, u^j \rangle} \quad (4.8)$$

The solution  $x$  is again obtained from (4.2), (4.6) and (4.8). To summarise:

1. Equation (4.1) has a unique solution only if  $A$  is nonsingular. This is a necessary condition as otherwise one of the  $c_i$ 's becomes indeterminate since at least one  $\lambda_i = 0$ .

2. For this method to be applied, it is enough if  $A$  has  $n$  independent eigenvectors  $\{u^i\}$ . This allows us to seek  $x$  as a linear combination of the  $u^i$ 's as in (4.2). For a real symmetric  $A$ , we always have  $n$  orthogonal eigenvectors, even if the eigenvalues are repeated. Therefore, this method is applicable for this case.

3. It is sufficient for a general  $A$  to have  $n$  distinct nonzero eigenvalues. This assures us that we have  $n$  independent eigenvectors and we can exploit the biorthogonality property to obtain the solution as detailed above.

The case where a matrix  $A$  would have repeated eigenvalues needs special consideration. This is a degenerate case and is usually of no interest to the practising engineer. In a physical system the eigenvalues are more likely to be distinct than repeated. We refer the interested reader to Noble and Daniel (1977), for a discussion on the case of repeated eigenvalues, of a matrix. Most physical systems are governed by self-adjoint operators. Here, even if we have repeated eigenvalues, the eigenvector set forms a basis set, which can be chosen to be orthogonal, and the technique discussed here is applicable.

The basis set  $\{u^i\}$  is a natural basis and is generated by the operator  $A$ . Seeking  $x$  the solution in terms of this basis as in (4.2), helps us extend our method to systems of ordinary differential equations and partial differential equations.

## 4.2 FIRST ORDER SYSTEM OF HOMOGENEOUS ORDINARY DIFFERENTIAL EQUATIONS (INITIAL-VALUE PROBLEMS)

The system of linear equations here can be written vectorially as

$$\frac{dx}{dt} = Ax \quad (4.9a)$$

This is a homogeneous system of  $n$  first order equations which are coupled to each other. We restrict  $A$  to be a real constant matrix. To be more precise, its elements are time independent. Such equations occur in the study of the dynamical behaviour of various systems. Here the state of the system characterised by the vector  $x$  is specified at time  $t = 0$ , and the evolution of the system is characterised by studying the variation of  $x$  with time  $t$ . Equation (4.9a) is subject to the initial condition at  $t = 0$ , i.e.

$$x(t = 0) = x^0 \quad (4.9b)$$

Consequently, problem (4.9) is called an initial-value problem (IVP). It is not essential that the independent variable be time 't'. It could as well be a spatial coordinate. The characteristic of an IVP is that all the conditions on the different coordinates of the vector  $x$  are specified at the same point of the independent variable. Again, we assume for convenience that  $A$  has  $n$  distinct eigenvalues. We now allow the possibility that  $A$  can be singular.

The solution vector  $x(t)$  is sought as a linear combination of the  $n$  eigenvectors  $\{u^i\}$  of  $A$  as

$$x(t) = \sum_{i=1}^n c_i(t)u^i \quad (4.10)$$

Since  $A$  is a real constant matrix, the eigenvectors  $u^i$  are time independent. The time dependence of  $x$  in (4.9a) is retained in the coefficients  $c_i$  in (4.10). The evolution of the scalar  $c_i$ 's is determined so that a solution of the form (4.10) satisfies (4.9). This yields

$$\frac{d}{dt} x(t) = Ax$$

or

$$\frac{d}{dt} \sum_{i=1}^n c_i(t)u^i = A \sum_{i=1}^n c_i(t)u^i$$

or

$$\sum_{i=1}^n \left( \frac{d}{dt} c_i(t) - \lambda_i c_i(t) \right) u^i = 0 \quad (4.11)$$

The vector set  $u^i$  is a linearly independent set. Hence (4.11) can hold only if it is the trivial relation. This means that for all time  $t$ , each  $c_i(t)$  satisfies

$$\frac{d}{dt} c_i(t) - \lambda_i c_i(t) = 0$$

or

$$c_i(t) = c_i(0)e^{\lambda_i t}$$

Substituting this in (4.10), we have

$$x(t) = \sum_{i=1}^n c_i(0)e^{\lambda_i t} u^i \quad (4.12a)$$

The constants  $c_i(0)$  are evaluated by using the initial condition (4.9b). This yields

$$x^0 = \sum_{i=1}^n c_i(0)u^i \quad (4.12b)$$

The calculation of the  $c_i(0)$ 's is now similar to the evaluation of the  $\beta_i$ 's in (4.4).

For a self-adjoint matrix  $A$ ,

$$c_i(0) = \frac{\langle u^i, x^0 \rangle}{\langle u^i, u^i \rangle} \quad (4.13a)$$

For a non self-adjoint matrix  $A$ ,

$$c_i(0) = \frac{\langle \bar{v}^i, x^0 \rangle}{\langle \bar{v}^i, u^i \rangle} \quad (4.13b)$$

where the  $v^i$ 's are the eigenvectors of the adjoint operator  $A^*$  or  $A'$ . Having obtained the  $c_i(0)$ 's, we obtain  $x(t)$  from (4.12a). To summarise

1. It is necessary that  $A$  has  $n$  linearly independent eigenvectors, for the above technique to be applicable. Otherwise, we cannot seek  $x$  as in (4.10).
2. If  $A$  has  $n$  distinct eigenvalues, it has  $n$  independent eigenvectors. The requirement of algebraically simple eigenvalues assures us of  $n$  independent eigenvectors.
3. Equation (4.9) has a solution even if  $A$  is singular. If  $A$  is self-adjoint and even if it has repeated eigenvalues, (4.9) can be solved using this technique since it has  $n$  independent eigenvectors (Noble and Daniel, 1977).

### 4.3 NONHOMOGENEOUS FIRST ORDER ORDINARY DIFFERENTIAL EQUATIONS (INITIAL-VALUE PROBLEMS)

The equations of this form can be written vectorially as

$$\frac{dx}{dt} = Ax + b(t) \quad (4.14a)$$

subject to

$$x(t = 0) = x^0 \quad (4.14b)$$

As already mentioned, the operator  $A$  is restricted to be a real constant nonsingular matrix. The vector  $b$  is allowed to have time dependent elements. The presence of the nonzero vector  $b$  renders (4.14a) nonhomogeneous. Such equations arise in studying response of systems to disturbances, as is done in process control. Writing the vectors  $x(t)$ ,  $b(t)$  as a linear combination of the eigenvectors  $u^i$ , we get

$$b(t) = \sum_{i=1}^n \beta_i(t) u^i \quad (4.15a)$$

$$x(t) = \sum_{i=1}^n c_i(t) u^i \quad (4.15b)$$

Once the scalars  $c_i(t)$  are found, the solution  $x(t)$  is determined. We consider two cases of the operator  $A$  as done earlier for determining the  $c_i(t)$ .

*Case 1: The operator  $A$  is self-adjoint.* For the scalars  $\beta_i(t)$ , the orthogonality of the  $u^i$ 's yield

$$\beta_i(t) = \frac{\langle u^i, b(t) \rangle}{\langle u^i, u^i \rangle} \quad (4.16)$$

Substituting (4.15a) and (4.15b) in (4.14a), we get

$$\sum_{i=1}^n \left( \frac{d}{dt} c_i(t) - \lambda_i c_i - \beta_i \right) u^i = 0$$

The linear independence of the  $u^i$ 's, implies that this equation is satisfied only if it is the trivial relation. This implies that for all  $i$ , the  $c_i(t)$  satisfies

$$\frac{d}{dt} c_i(t) - \lambda_i c_i(t) = \beta_i(t) \quad (4.17)$$

This is a linear first order nonhomogeneous equation. Its solution is

$$c_i(t) = c_i(0) e^{\lambda_i t} + e^{\lambda_i t} \int_0^t e^{-\lambda_i \tau} \beta_i(\tau) d\tau \quad (4.18)$$

The constant  $c_i(0)$  here is obtained from the initial condition (4.14b). This yields

$$x^0 = \sum_{i=1}^n c_i(0) u^i \quad (4.19a)$$

where  $c_i(0)$  is obtained from

$$c_i(0) = \frac{\langle u^i, x^0 \rangle}{\langle u^i, u^i \rangle} \quad (4.19b)$$

The solution  $x(t)$  is now obtained from (4.19b), (4.18), and (4.15b).

*Case 2: A is not a self-adjoint operator.* The adjoint  $A^*$  of  $A$  has eigenvectors  $\bar{v}^j$  corresponding to the eigenvalue  $\lambda_j$ . The biorthogonality between the  $\bar{v}^j$  and the  $u^i$ 's, yields, for  $\beta_i(t)$  in (4.15a), the relation

$$\beta_i(t) = \frac{\langle \bar{v}^i, b \rangle}{\langle \bar{v}^i, u^i \rangle} \quad (4.20a)$$

It is easy to verify that (4.17) and (4.18) are still valid. The evaluation of the constant  $c_i(0)$  in (4.19a) is given by

$$c_i(0) = \frac{\langle \bar{v}^i, x^0 \rangle}{\langle \bar{v}^i, u^i \rangle} \quad (4.20b)$$

To summarise, (4.14) can be solved by using the technique of eigenvector expansions. (a) when  $A$  has algebraically simple eigenvalues, for a general  $A$ , and (b) for any self-adjoint matrix  $A$ .

### Example 4.1 The reversible reaction



occurs isothermally in a batch reactor. The forward and reverse reactions are both first order with rate constants  $1s^{-1}$  and  $2s^{-1}$ , respectively. The initial concentrations of  $x_1$ ,  $x_2$  are 2 g mol/cc and 3 g mol/cc. Determine the equilibrium concentration in the reactor. The evolution of the concentrations  $x_1$  and  $x_2$  is given by

$$\frac{dx_1}{dt} = -x_1 + 2x_2, \quad \frac{dx_2}{dt} = x_1 - 2x_2 \quad (4.21)$$

The equilibrium concentrations  $x_{1e}$ ,  $x_{2e}$  of  $x_1$ ,  $x_2$ , are obtained from the steady state solution of (4.21), i.e.,

$$0 = -x_{1e} + 2x_{2e}, \quad 0 = x_{1e} - 2x_{2e} \quad (4.22)$$

This is a system of one independent equation in two unknowns. It cannot be solved for the equilibrium concentrations uniquely as the matrix  $A$  has rank unity. The homogeneous equation (4.22) has a solution of the form  $C \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ . This implies that the concentration ratio of  $x_1$  to  $x_2$  is specified uniquely,

i.e. it is in the ratio 2 : 1, but the actual concentrations are not known. This indeterminacy arises when we solve the system (4.21) as a steady state problem.

The equilibrium solution can also be obtained as the long-time behaviour of the dynamic system (4.21), i.e. the system state as  $t \rightarrow \infty$ . To obtain the dynamic evolution of the system, we solve (4.21) subject to the initial condition at  $t = 0$ ,

$$[x_1, x_2] = [2, 3] = x^0$$

Recasting equation (4.21) in vectorial form (4.9a), we have the matrix

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

Its eigenvalues are  $\lambda_1 = 0$ ,  $\lambda_2 = -3$ . The corresponding eigenvectors are  $u^1 = [2, 1]'$  and  $u^2 = [1, -1]'$ .

The eigenvectors of  $A'$  corresponding to  $\lambda_1, \lambda_2$  are

$$v^1 = [1, 1]' \text{ and } v^2 = [1, -2]'$$

The solution

$$\begin{aligned} x(t) &= \sum_{i=1}^n c_i(0) e^{\lambda_i t} u^i \\ &= c_1(0)u^1 + c_2(0) e^{\lambda_2 t} u^2 \\ &= c_1(0)u^1 + c_2(0) e^{-3t} u^2 \end{aligned}$$

The limiting behaviour as  $t \rightarrow \infty$  is determined only by  $c_1(0)$ , i.e.,

$$\lim_{t \rightarrow \infty} x(t) = c_1(0)u^1$$

where, from (4.13b),

$$c_1(0) = \frac{\langle \bar{v}^1, x^0 \rangle}{\langle \bar{v}^1, u^1 \rangle} = \frac{5}{3}$$

yielding  $x_{1e} = 10/3$ ,  $x_{2e} = 5/3$ .

The apparent indeterminacy in the problem has been resolved. The system  $Ax = 0$  yields the equilibrium state of a closed system. This is governed by principles of thermodynamics. The equilibrium state cannot be revealed by considering the steady state problem. While solving the transient problem, we track the trajectory from the initial state. The initial condition gives a thermodynamic constraint which is used, and the solution is obtained uniquely.

**Example 4.2** A steel ball of volume (1000 cc,  $\rho = 1.25$  g/cc and  $C_p = .8$  cal/g/ $^{\circ}\text{C}$ ) is at a uniform temperature of 100°C. This is dropped into an insulated vessel containing 5000 cc of water at 20°C. Determine the steady state temperature of water and steel ball. Neglect spatial gradients in both the ball and the fluid. The heat transfer rate  $UA$  between the ball and the fluid is 1000 cal/s/ $^{\circ}\text{C}$ . Using the subscript 'b' for ball and 'w' for water, the energy balance equation for each is,

$$\left. \begin{aligned} V_b \rho_b C_{pb} \frac{dT_b}{dt} &= -UA(T_b - T_w) \\ V_w \rho_w C_{pw} \frac{dT_w}{dt} &= +UA(T_b - T_w) \end{aligned} \right\} \quad (4.23)$$

Using the relevant values, we obtain

$$\frac{dT_b}{dt} = -(T_b - T_w), \quad \frac{dT_w}{dt} = .2(T_b - T_w) \quad (4.24)$$

The steady temperature is obtained by setting the time derivatives to zero. This yields

$$T_b = T_w$$

This only tells us that the two temperatures are equal. However, we do not know what the individual steady temperature is. This problem is similar to the previous example where we only know what the ratios of the concentrations are when we solve the steady state system. To obtain the steady temperature, we solve the dynamic system (4.24) and set  $t = +\infty$ . Recasting the equation in vectorial form as (4.9a), we have

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

The eigenvalues are  $\lambda_1 = 0$ ,  $\lambda_2 = -1.2$ . This yields,

$$u^1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad u^2 = \begin{pmatrix} -5 \\ 1 \end{pmatrix}$$

The eigenvectors of  $A'$  corresponding to  $\lambda_1$ ,  $\lambda_2$  are

$$v^1 = \begin{pmatrix} 1 \\ 5 \end{pmatrix}, \quad v^2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

The initial condition is  $\begin{pmatrix} 100 \\ 20 \end{pmatrix}$ . The solution

$$T(t) = \sum_{i=1}^n c_i(0) e^{\lambda_i t} u^i$$

From the initial condition

$$\begin{pmatrix} 100 \\ 20 \end{pmatrix} = c_1(0)u^1 + c_2(0)u^2$$

We have

$$\begin{aligned} c_1(0) &= \frac{\langle (100 \ 20)', (1 \ 5)' \rangle}{\langle (1 \ 1)', (1 \ 5)' \rangle} \\ &= \frac{200}{6} = 33.33 \end{aligned}$$

Therefore,

$$\begin{pmatrix} T_{bss} \\ T_{wss} \end{pmatrix} = \begin{pmatrix} 33.33 \\ 33.33 \end{pmatrix}$$

The steady state temperature of the ball and water at equilibrium is 33.33°C.

In this problem again we have a closed system. Our vessel is insulated and cannot exchange any energy from the surroundings. Consequently, the steady state equations do not tell us the equilibrium temperature. The dynamic equations contain the information from the initial condition. This provides us with the additional relation required to determine the solution uniquely. The heat transfer coefficient term  $UA$  does not affect the steady state. Changing the value of  $UA$  results in the same steady state. The transient behaviour, however, gets affected as  $UA$  effects the nonzero eigenvalue  $\lambda_2$ . The time taken to reach the equilibrium gets changed by changing  $UA$ , and not the equilibrium itself. The steady state value is changed only if we change any of the other parameter values.

**Example 4.3** Consider the series reaction



occurring in a CSTR. We restrict the reactor operation to be isothermal and the reactions to be first order. The feed to the reactor is pure A at a concentration of 1 g mol/cc. Determine the evolution of concentrations of A, B, C in the CSTR to the steady state. The initial concentrations of A, B, C, are 1, 0, 0 g/mol/cc, and the residence time is 4 s<sup>-1</sup>.

$$\left. \begin{aligned} \frac{dC_A}{dt} &= \frac{1}{\tau_{\text{res}}} (C_{Af} - C_A) - k_1 C_A \\ \frac{dC_B}{dt} &= \frac{1}{\tau_{\text{res}}} (C_{Bf} - C_B) + k_1 C_A - k_2 C_B \\ \frac{dC_C}{dt} &= \frac{1}{\tau_{\text{res}}} (C_{Cf} - C_C) + k_2 C_B \end{aligned} \right\} \quad (4.25)$$

The steady state concentration is obtained by setting the time derivative to zero and solving for the unknowns  $C_A$ ,  $C_B$ ,  $C_C$  from

$$\begin{bmatrix} -\frac{1}{\tau_{\text{res}}} - k_1 & 0 & 0 \\ k_1 & -\frac{1}{\tau_{\text{res}}} - k_2 & 0 \\ 0 & k_2 & -\frac{1}{\tau_{\text{res}}} \end{bmatrix} \begin{bmatrix} C_A \\ C_B \\ C_C \end{bmatrix} = \begin{bmatrix} -\frac{C_{Af}}{\tau_{\text{res}}} \\ 0 \\ 0 \end{bmatrix}$$

The matrix  $A$  here is

$$\begin{bmatrix} -2.25 & 0 & 0 \\ 2 & -3.25 & 0 \\ 0 & 3 & -.25 \end{bmatrix}$$

The eigenvalues are  $\lambda_1 = -2.25$ ,  $\lambda_2 = -3.25$ ,  $\lambda_3 = -.25$ .

The corresponding eigenvectors are

$$u^1 = (1 \ 2 \ -3)', \quad u^2 = (0 \ 1 \ -1)', \quad u^3 = (0 \ 0 \ 1)'$$

The eigenvectors of  $A'$  are

$$v^1 = (1 \ 0 \ 0)', \quad v^2 = (-2 \ 1 \ 0)', \quad v^3 = (1 \ 1 \ 1)'$$

The nonhomogeneity  $b = (-1/4 \ 0 \ 0)$ . Seeking  $b = \sum_{i=1}^n \beta_i u^i$  and exploiting the biorthogonality with  $\bar{v}^j$ , we obtain

$$\langle \bar{v}^j, b \rangle = \sum_{i=1}^n \langle \bar{v}^j, u^i \rangle \beta_i$$

or

$$\begin{aligned}\beta_1 &= \frac{\langle \bar{v}^1, b \rangle}{\langle \bar{v}^1, u^1 \rangle} = -\frac{1}{4} \\ \beta_2 &= \frac{1}{2}, \quad \beta_3 = -\frac{1}{4}\end{aligned}$$

Using  $c_i = \beta_i / \lambda_i$  we get  $c_1 = \frac{1}{9}$ ,  $c_2 = -\frac{2}{13}$ ,  $c_3 = 1$ . The steady state solution is

$$\begin{pmatrix} C_{A_{ss}} \\ C_{B_{ss}} \\ C_{C_{ss}} \end{pmatrix} = \frac{1}{9} \begin{pmatrix} 1 \\ 2 \\ -3 \end{pmatrix} - \frac{2}{13} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} + 1 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1/9 \\ 8/117 \\ 32/39 \end{pmatrix}$$

The steady state of a CSTR can be obtained by solving the linear algebraic system. This is in contrast with a batch system. The CSTR is an open system and can exchange mass and energy with the environment.

The transient response of the CSTR given by (4.25) is sought as

$$(C_A, C_B, C_C)' = \sum_{i=1}^n c_i(t) u^i \quad (4.26)$$

From (4.18),

$$c_i(t) = c_i(0) e^{\lambda_i t} + e^{\lambda_i t} \int_0^t e^{-\lambda_i \tau} \beta_i(\tau) d\tau$$

Here,

$$c_1(0) = \frac{\langle \bar{v}^1, x^0 \rangle}{\langle \bar{v}^1, u^1 \rangle} = 1$$

$$c_2(0) = \frac{\langle \bar{v}^2, x^0 \rangle}{\langle \bar{v}^2, u^2 \rangle} = -2$$

$$c_3(0) = \frac{\langle \bar{v}^3, x^0 \rangle}{\langle \bar{v}^3, u^3 \rangle} = 1$$

The  $b$  vector here is  $(-1/4 \ 0 \ 0)$ . Using (4.20a),  $\beta_1 = 1/4$ ,  $\beta_2 = -1/2$ ,  $\beta_3 = 1/4$ .

$$c_1(t) = e^{-9t/4} + \frac{1}{9} (1 - e^{-9t/4})$$

$$c_2(t) = -2e^{-13t/4} - \frac{2}{13} (1 - e^{-13t/4})$$

$$c_3(t) = e^{-t/4} + 1(1 - e^{-t/4})$$

Substituting the above in (4.26), We get

$$C_A(t) = e^{-9t/4} + \frac{1}{9}(1 - e^{-9t/4})$$

$$C_B(t) = 2e^{-9t/4} - 2e^{-13t/4} + \frac{2}{9}(1 - e^{-9t/4}) - \frac{2}{13}(1 - e^{-13t/4})$$

$$C_C(t) = -3e^{-9t/4} + 2e^{-13t/4} + e^{-t/4} - \frac{3}{9}(1 - e^{-9t/4})$$

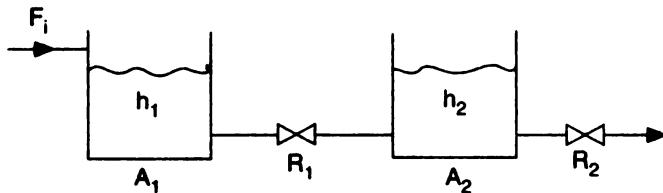
$$+ \frac{2}{13}(1 - e^{-13t/4}) + 1(1 - e^{-t/4})$$

In the limit as  $t \rightarrow \infty$ , we recover the steady state solution  $C_{A_{ss}} = 1/9$ ,  $C_{B_{ss}} = 8/117$ ,  $C_{C_{ss}} = 32/39$ .

The eigenvalues determine the time scale of evolution of the system. Most systems of interest are dissipative in nature, i.e. all eigenvalues are nonpositive. The lowest eigenvalue (or most negative eigenvalue) indicates how fast the system evolves to a steady state. The largest eigenvalue (the one closest to zero) indicates how slowly the system relaxes to steady state.

**Example 4.4** Determine the heights  $h_1$ ,  $h_2$  in the two-tank network shown in Fig. 4.1. The equations are given by

$$\left. \begin{aligned} \frac{dh_1}{dt} &= \frac{F_i}{A_1} - \left( \frac{h_1 - h_2}{R_1 A_1} \right) \\ \frac{dh_2}{dt} &= \frac{h_1 - h_2}{A_2 R_1} - \frac{h_2}{A_2 R_2} \end{aligned} \right\} \quad (4.27)$$



**Fig. 4.1** Two-tank network of Example 4.4: ( $h$ : height in tank,  $R$ : resistance of valve,  $A$ : cross-sectional area,  $F$ : represents flow-rate).

Assume that

$$R_1 A_1 = .5, \quad R_2 A_2 = 0.5, \quad R_1 A_2 = 1, \quad F_i / A_1 = 1$$

This reduces the equations to

$$\begin{aligned} \frac{d}{dt} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} &= \begin{bmatrix} -\frac{1}{R_1 A_1} & \frac{1}{R_1 A_1} \\ \frac{1}{A_2 R_1} & -\frac{1}{R_1 A_2} - \frac{1}{R_2 A_2} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} F_i / A_1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} -2 & 2 \\ 1 & -3 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} F_i / A_1 \\ 0 \end{bmatrix} \end{aligned}$$

The steady state is obtained from

$$\begin{bmatrix} -2 & 2 \\ 1 & -3 \end{bmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

The eigenvalues of  $A$  are

$$\lambda_1 = -1, \lambda_2 = -4$$

The eigenvectors of  $A$  corresponding to this are

$$u^1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad u^2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

and of  $A'$  are

$$v^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad v^2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

Seek

$$\begin{pmatrix} h_1^{ss} \\ h_2^{ss} \end{pmatrix} = \sum_{i=1}^2 c_i u^i$$

$$c_1 = \frac{\beta_1}{\lambda_1} = \frac{\langle \bar{v}^1, b \rangle}{\lambda_1 \langle \bar{v}^1, u^1 \rangle} = \frac{1}{3}$$

$$c_2 = \frac{\beta_2}{\lambda_2} = -\frac{1}{4} \frac{\langle \bar{v}^2, b \rangle}{\langle \bar{v}^2, u^2 \rangle} = \frac{1}{12}$$

$$\begin{pmatrix} h_1^{ss} \\ h_2^{ss} \end{pmatrix} = \frac{1}{3} \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \frac{1}{12} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 3/4 \\ 1/4 \end{bmatrix}$$

If  $F/A_1$  were to change to 4, the system would evolve to another steady state given by  $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$ . The evolution occurs from the initial condition determined by  $\begin{bmatrix} 3/4 \\ 1/4 \end{bmatrix}$ . The approach of the system to steady state because of this disturbance, i.e. change in  $F/A_1$ , is treated in process control using Laplace transform. Here we employ our method of solution to study the system behaviour in the time domain directly. The resulting system of equations is

$$\frac{d}{dt} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} -2 & 2 \\ 1 & -3 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} 4 \\ 0 \end{bmatrix}$$

The coefficients  $\beta_i$  from (4.20a) are given by

$$\beta_1 = \frac{4}{3}, \quad \beta_2 = \frac{4}{3}$$

where

$$[4, \ 0]' = \sum_{i=1}^2 \beta_i u^i$$

Seeking,

$$\begin{bmatrix} h_1(t) \\ h_2(t) \end{bmatrix} = \sum_{i=1}^2 c_i(t) u^i$$

we have, from (4.18), the relations

$$c_1(t) = c_1(0) e^{\lambda_1 t} + e^{\lambda_1 t} \int_0^t e^{-\lambda_1 \tau} \beta_1(\tau) d\tau$$

$$c_1(0) = \frac{\langle \bar{v}^1, h^0 \rangle}{\langle \bar{v}^1, u^1 \rangle} = \frac{1}{3}$$

$$c_2(0) = \frac{\langle \bar{v}^2, h^0 \rangle}{\langle \bar{v}^2, u^2 \rangle} = \frac{1}{12}$$

These yield

$$c_1(t) = \frac{1}{3} e^{-t} + \frac{4}{3}(1 - e^{-t})$$

$$c_2(t) = \frac{1}{12} e^{-4t} + \frac{4}{12}(1 - e^{-4t})$$

and

$$h_1(t) = \frac{2}{3} e^{-t} + \frac{1}{12} e^{-4t} + 3 - \frac{8}{3} e^{-t} - \frac{4}{12} e^{-4t}$$

$$h_2(t) = \frac{1}{3} e^{-t} + \frac{1}{12} e^{-4t} + 1 - \frac{4}{3} e^{-t} + \frac{1}{3} e^{-4t}$$

The eigenvalues  $\lambda_1, \lambda_2$  represent how fast the variables approach the new state. The perturbation dampens out here and the system is “stable” as both the eigenvalues are negative. If the eigenvalues are very negative, the system approaches the steady state very quickly. The closer the eigenvalue is to zero, the more the time taken to approach the new steady state.

#### 4.4 GEOMETRIC BASIS OF THE METHOD

The method of solution presented here is based on working in a vector space where the basis is generated by the eigenvectors. This is in contrast to working in the standard orthogonal basis. We will see the advantages of this representation geometrically. Consider the solution of (4.1).

The coordinates of vector  $b$  in the standard orthogonal basis of  $\mathbb{R}^n$ , viz.

$$e^1 = (1, 0, 0, \dots, 0)$$

$$e^2 = (0, 1, 0, \dots, 0)$$

$\vdots$

$$e^n = (0, 0, 0, \dots, 1)$$

represent the coefficients in the linear combination

$$b = \sum_{i=1}^n b_i e^i \tag{4.28a}$$

In terms of the eigenvector basis,  $\{u^i\}$ ,  $b$  is represented as

$$b = \sum_{i=1}^n \beta_i u^i \quad (4.28b)$$

Let  $P$  be the matrix whose columns represent the eigenvectors of  $A$ , then, clearly, from (4.28a) and (4.28b),

$$\begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = P \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} \quad (4.29)$$

So the coordinates of any vector in the standard basis  $e^i$  given by  $x$  is related to its coordinates in the eigenvector basis  $x'$  by

$$x = Px' \quad (4.30)$$

Representing both  $x$ ,  $b$  in (4.1) in the eigenvector basis, we obtain

$$APx' = Pb'$$

or

$$P^{-1}APx' = b' \quad (4.31a)$$

From the similarity transformation in linear algebra when  $A$  has  $n$  distinct eigenvalues,  $P^{-1}AP = D$ , where  $D$  is a diagonal matrix having the eigenvalues  $\lambda_i$  on the diagonal. This reduces (4.31a) to

$$Dx' = b' \quad (4.31b)$$

yielding, for each coordinate  $x'_i$ , the equation

$$x'_i = \frac{b'_i}{\lambda_i} \quad (4.32)$$

The similarity transformation, i.e. the representation in the eigenbasis, enables us to decouple the system, of equations (4.1) (see Noble and Daniel (1977) for more details). The solution  $x$  in terms of the standard basis  $\{e^i\}$  comes from using (4.30) and (4.32).

$$x_j = \sum_{i=1}^n u_j^i b'_i / \lambda_i$$

$$x = \sum_{i=1}^n \frac{b'_i}{\lambda_i} u^i$$

Remembering  $b'_i = \beta_i$ , we recover (4.6).

Our method of solving coupled ordinary differential equations is also based on decoupling the equations by using the eigenvector basis  $u^i$  instead of the regular classical basis  $e^i$ . We demonstrate the method for linear homogeneous equations (4.9a) and (4.9b).

It is left as an exercise to the reader to verify the validity of this statement to the nonhomogeneous system of equations (4.14a) and (4.14b).

Representing  $x$  and  $x^0$  in (4.9a) and (4.9b) in terms of the eigenvector basis  $\{u^i\}$ , we obtain

$$x(t = 0) = x^0 \quad (4.33a)$$

where

$$x^0 = Px' \quad (4.33b)$$

$$\frac{d}{dt} Px' = APx' \quad (4.34)$$

As  $A$  is a constant matrix, so is  $P$  and we can recast (4.34) as

$$\frac{dx'}{dt} = P^{-1}APx', \quad \frac{dx'}{dt} = Dx' \quad (4.35)$$

Equations (4.35) are decoupled once again and each coordinate is of the form  $x'_i = x_i(0)e^{\lambda_i t}$ .

In terms of the standard basis, the solution is

$$x = Px' = \sum_{i=1}^n x_i(0) e^{\lambda_i t} u^i \quad (4.36)$$

Clearly,  $x_i(0)$  equals  $c_i(0)$  in the earlier notation, i.e. the representation of  $x_0$  in the eigenbasis. Equations (4.36) and (4.12a) are identical as expected.

When  $A$  has repeated eigenvalues, it is not possible to reduce the matrix  $A$  to diagonal matrix  $D$  by a similarity transformation. Instead, we obtain a Jordan canonical form (see Noble and Daniel, 1977) and the representation now is not very elegant. The complete decoupling of the equations is not possible anymore. A matrix  $A$  is defective when it does not possess an independent set of  $n$  eigenvectors. This can only happen when  $A$  has repeated eigenvalues. The concept of the generalised eigenvector then comes into the picture. However, we now lose the properties of orthogonality, biorthogonality, etc., and the representation in terms of eigenbasis is not appealing anymore. For a general matrix, with repeated eigenvalues and a complete set of eigenvectors, we cannot use the biorthogonality property as was seen in Chapter 3.

The other decompositions of a matrix like the Q-R decomposition and singular-value decomposition arise when we represent  $x$  and  $b$  in (4.1) using two different bases. The interested reader is referred to Noble and Daniel, 1977, for a detailed discussion of these concepts.

## 4.5 IMPLICATIONS IN PROCESS CONTROL

In process control applications we are frequently interested in the dynamic behaviour of second and higher order systems. An  $n$ th order system is described by an  $n$ th-order differential equation of the form

$$a_0 \frac{d^n x_1}{dt^n} + a_1 \frac{d^{n-1} x_1}{dt^{n-1}} + \dots + a_n x_1 = f(t) \quad (4.37a)$$

This can be recast into a system of  $n$  first order equations by defining

$$\begin{aligned} \frac{dx_1}{dt} &= x_2 \\ \frac{dx_2}{dt} &= \frac{d^2 x_1}{dt^2} = x_3 \\ \frac{dx_3}{dt} &= \frac{d^2 x_2}{dt^2} = \frac{d^3 x_1}{dt^3} = x_4 \\ &\vdots \\ \frac{dx_{n-1}}{dt} &= x_n \end{aligned}$$

$$\frac{dx_n}{dt} = f(t) - a_n x_1 - a_{n-1} x_2 - \dots - a_1 x_n$$

or, in vectorial form,

$$\frac{dx}{dt} = Ax + b \quad (4.37b)$$

where

$$x = [x_1, \dots, x_n]', A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & \dots & \dots \\ \vdots & \vdots & \vdots & 0 & 1 \\ -a_n & -a_{n-1} & \dots & -a_2 & -a_1 \end{bmatrix}, b = [0, 0, \dots, f(t)]'$$

Equations (4.37b) is called the state-space representation of (4.37a) in control theory. This representation is preferred in analysing systems since it allows us to extend methods for single-input single-output (SISO) systems to multiple input multiple output (MIMO) systems. This extension is possible only because of the general principles which we have introduced in Chapter 2, where we extend the geometric concepts using the algebraic representation.

The dynamic behaviour of the system (4.37) subject to perturbations is studied in control theory (see Stephanopoulos, 1984). The disturbances are located in  $f(t)$ . These problems can also be solved by the method of solution presented here, since (4.37a) can be represented in vectorial form as (4.37b).

In process control the transient response of a system to disturbances is investigated. As we are usually interested in the behaviour near a steady state, the systems can be represented here by linearised equations (even though the actual interactions are nonlinear, see Chapter 11). This provides the motivation for studying the transient response of linear systems as already discussed.

Consider a second order system modelled by

$$\dot{x} = Ax + b$$

where  $x, b \in \mathbb{R}^2$  and  $A$  is a  $2 \times 2$  matrix.

The response of this system to a step input can be obtained in the frequency domain or time domain. In process control, one works usually in the frequency domain using Laplace transforms and then uses an inverse transform to obtain the system response in time domain. The methods discussed so far in this chapter allow us to analyse system response directly in time domain.

The response of second order systems can be underdamped or overdamped. In underdamping, when there is a step input imposed on the system, the response is oscillatory in the approach to the new steady state. In overdamping, the system approaches the new steady state monotonically. This essential difference is determined by the nature of the eigenvalues of  $A$ . When the eigenvalues of  $A$  are both real, negative, and unequal, we have an overdamped response. When the eigenvalues of  $A$  are complex-conjugates, with negative real parts, the response is underdamped, and when they are real, negative and equal, the response is critically damped. The characteristic variable in frequency domain in process control is identical to the eigenvalue of the matrix  $A$ .

#### 4.6 NON SELF-ADJOINT SYSTEMS (A SPECIAL APPROACH)

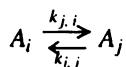
Many chemical engineering systems are governed by non self-adjoint operators. It is possible to

construct the solution to these systems using the biorthogonality property between the eigenvector sets of the original and the adjoint systems.

In some cases it is possible to render the system self-adjoint by defining a new inner-product. The operator is non self-adjoint in the standard inner-product which generates the Euclidean metric. However, since it is self-adjoint in another inner-product, it allows us to establish the properties of self-adjoint systems, i.e. real eigenvalues, orthogonality of eigenvectors (in the new inner-product), etc.

We illustrate the idea on a model problem of a batch reactor sustaining a system of reversible reactions. This problem has been investigated in detail by Wei and Prater (1962). The problem was cast in an operator theoretic framework by Ramkrishna and Amundson (1985). The problem involves estimating the reaction rate constants of each reaction from batch reactor data. In chemical engineering the parameter estimation is usually accomplished by statistical techniques like nonlinear regression. Since the system behaviour is sensitive to the parameters, these techniques are prone to give erroneous results. We illustrate a method which is based on the principles discussed in this section.

Consider an isothermal batch reactor sustaining a system of reversible monomolecular elementary reactions. Each reaction is of the form



Here we assume that each chemical species  $A_i$  can get converted to every other chemical  $A_j$  species and vice-versa. Let  $x_j$  denote the mole fraction of species  $A_j$ . The evolution of  $x_i$  is determined by,

$$\frac{dx_i}{dt} = k_{i,1} x_1 + k_{i,2} x_2 + \dots - \sum_{j \neq i}^N k_{j,i} x_i + k_{i,N} x_N \quad (4.38)$$

Here  $\sum_{j,i}^N$  denotes  $\sum_{\substack{j=1 \\ j \neq i}}^N$ , where  $N$  denotes the total number of species in the batch system. The evolution of the system can be written compactly as

$$\frac{dx}{dt} = Kx \quad (4.39)$$

where

$$x = [x_1, x_2, \dots, x_n]'$$

$$K = \begin{bmatrix} -\sum_{j=1}^N k_{j,1} & k_{1,2} & k_{1,3} \dots k_{1,N} \\ k_{2,1} & -\sum_{j=2}^N k_{j,2} & k_{2,3} \dots k_{2,N} \\ k_{3,1} & \dots & k_{3,N} \\ \vdots & \vdots & \vdots \\ k_{N,1} & \dots & -\sum_{j=N}^N k_{j,N} \end{bmatrix}$$

Clearly, for  $k_{i,j} = k_{j,i}$  the  $K$  matrix is real symmetric and consequently self-adjoint. We would like to consider the more general case of  $k_{i,j} \neq k_{j,i}$ . The treatment for the symmetric case can then be analysed as a special case.

Clearly, the rows of the  $K$ -matrix are linearly dependent, since they add up to give the zero vector. This implies  $\text{rank}(K) < N$ , and 0 is an eigenvalue of the matrix. We assume that the rank ( $K$ ) =  $n - 1$ , i.e. 0 is an algebraically simple eigenvalue. The corresponding eigenvector is the equilibrium composition of the batch system. We also assume in what follows that the other eigenvalues are algebraically simple.

The system, i.e. the operator  $K$ , can be rendered self-adjoint by defining a new inner-product. This is rendered possible since the reaction system satisfies the principle of microscopic reversibility or detailed balancing. At equilibrium this implies

$$k_{i,j} u_i^0 = k_{j,i} u_j^0 \quad (4.40a)$$

where  $u_i^0$  represents the mole fraction of the  $i$ th component at equilibrium. This results in

$$\frac{k_{i,j}}{u_i^0} = \frac{k_{j,i}}{u_j^0} \quad (4.40b)$$

Let us define the inner-product

$$\langle x, y \rangle = \sum_{i=1}^N \frac{\tilde{x}_i y_i}{u_i^0} \quad (4.41)$$

This is a valid definition of the inner-product since  $u_i^0 > 0$  for all  $i$ , since we assume every species can form every other species.

We define the  $(i, j)$ th element of  $K$  matrix for clarity as

$$k_{ij} = k_{i,j} \quad \text{for } i \neq j$$

$$k_{ii} = - \sum_{j, i}^N k_{j,i}$$

Clearly,

$$\begin{aligned} \langle x, Ky \rangle &= \left\langle x, \sum_{j=1}^N k_{ij} y_j \right\rangle = \sum_{i=1}^N \sum_{j=1}^N \frac{\tilde{x}_i k_{ij} y_j}{u_i^0} \\ &= \sum_{i=1}^N \sum_{j=1}^N \frac{\tilde{x}_j k_{ji} y_i}{u_j^0} \quad (\text{interchanging the indices } i \text{ with } j) \\ &= \sum_{i=1}^N \sum_{j=1}^N \frac{\tilde{x}_j k_{ij} y_i}{u_i^0} \quad (\text{from the principle of microscopic reversibility}) \\ &= \langle Kx, y \rangle \quad (\text{since } K \text{ has only real elements}) \end{aligned} \quad \left. \right\} \quad (4.42)$$

So,  $K$  is self-adjoint in the inner-product (4.41).

Since  $K$  is self-adjoint in this inner-product, we can work in it and prove that its eigenvalues are real. Let  $u$  be the eigenvector of  $K$  corresponding to eigenvalue  $\lambda_1$ . Then,

$$\langle Ku, u \rangle = \langle u, Ku \rangle$$

or

$$\langle \lambda u, u \rangle = \langle u, \lambda u \rangle$$

or

$$(\bar{\lambda} - \lambda) \langle u, u \rangle = 0 \quad (4.43)$$

Since  $\langle u, u \rangle > 0$  for  $u \neq 0$ ,  $\lambda = \bar{\lambda}$ , or the eigenvalues are real.

We have already established that one of the eigenvalues is zero. We next show that the nonzero eigenvalues are negative. Here again,

$$\begin{aligned}
 \langle Ku, u \rangle &= \left\langle \sum_j k_{ij} u_j, u \right\rangle \\
 &= \sum_i \sum_j \frac{k_{ij} u_j u_i}{u_i^0} \\
 &= \sum_{\substack{i, j \\ i \neq j}} \frac{k_{ij} u_i u_j}{u_i^0} + \sum_{i=1}^N k_{ii} \frac{u_i^2}{u_i^0} \\
 &= \sum_{\substack{i, j \\ i \neq j}} k_{i,j} \frac{u_i u_j}{u_i^0} - \sum_{\substack{i, j \\ i \neq j}} k_{j,i} \frac{u_i^2}{u_i^0} \\
 &= \sum_{\substack{i, j \\ i \neq j}} \sqrt{\frac{k_{i,j}}{u_j^0}} u_j \sqrt{\frac{k_{j,i}}{u_i^0}} u_i - \sum_{\substack{i, j \\ i \neq j}} \left( \sqrt{\frac{k_{i,j}}{u_j^0}} u_j \right)^2
 \end{aligned}$$

Interchanging indices  $i$  with  $j$ , we obtain

$$\langle Ku, u \rangle = \sum_{\substack{i, j \\ i \neq j}} \sqrt{\frac{k_{j,i}}{u_i^0}} u_i \sqrt{\frac{k_{i,j}}{u_j^0}} u_j - \sum_{\substack{i, j \\ i \neq j}} \left( \sqrt{\frac{k_{j,i}}{u_i^0}} u_i \right)^2$$

Adding these expressions, we get

$$\begin{aligned}
 \langle Ku, u \rangle &= -\frac{1}{2} \sum_{\substack{i, j \\ i \neq j}} \left( \sqrt{\frac{k_{j,i}}{u_i^0}} u_i - \sqrt{\frac{k_{i,j}}{u_j^0}} u_j \right)^2 \\
 &\leq 0
 \end{aligned} \tag{4.44}$$

This establishes that all the eigenvalues are negative. So the system converges to an equilibrium value starting from any initial condition. The approach to this equilibrium is monotonic. In particular, the approach is not oscillatory.

The orthogonality of the eigenvectors can also be established in a straightforward manner. This orthogonality is with respect to the inner-product (4.41). Thus,

$$\langle Ku^i, u^j \rangle = \langle u^i, Ku^j \rangle$$

$$\langle \lambda_i u^i, u^j \rangle = \langle u^i, \lambda_j u^j \rangle$$

or  $(\lambda_i - \lambda_j) \langle u^i, u^j \rangle = 0$  (since all  $\lambda$ 's are real;  $\tilde{\lambda}_i = \lambda_i$ ). For  $\lambda_i \neq \lambda_j$ , clearly we have

$$\langle u^i, u^j \rangle = 0 \tag{4.45}$$

Our objective is to determine the various rate constants  $k_{i,j}$ . We now explain how this can be done using the concept of eigenvector expansions. We define the vector

$$k^n = \{k_1^n, k_2^n, \dots, k_n^n\}''$$

where

$$k_n^n = - \sum_{\substack{j=1 \\ j \neq n}}^N k_{j,n} \quad k_i^n = k_{i,n} \quad (4.46)$$

Clearly,

$$\langle k^n, u^0 \rangle = \sum_{i=1}^N \frac{k_i^n u_i^0}{u_i^0} = \sum_{i=1}^N k_i^n = 0 \quad (4.47)$$

Let us represent the rate constant vector  $k^n$  in terms of the eigenvectors  $u^j$ , which forms an orthogonal basis. Now,

$$k^n = \sum_{j=1}^{N-1} \langle k^n, u^j \rangle u^j \quad (4.48)$$

Obviously,

$$\langle k^n, u^j \rangle = \sum_{i=1}^N \frac{k_i^n u_i^j}{u_i^0} = \sum_{i=1}^N \frac{k_i^n u_i^j}{u_n^0} = \frac{\lambda_j u_n^j}{u_n^0}$$

Substituting in (4.48), we get.

$$k^n = \sum_{j=1}^{N-1} \frac{\lambda_j u_n^j}{u_n^0} u^j \quad (4.49)$$

$$k_{i,n} = k_i^n = \sum_{j=1}^{N-1} \frac{\lambda_j u_n^j}{u_n^0} u_i^j \quad (4.50)$$

So once the spectral properties, i.e. the eigenvalues and the eigenvectors of  $K$  are determined, the rate constants  $k_{i,j}$  can be obtained from (4.50). The problem of determining the kinetics now reduces to determining the eigenvalues and eigenvectors from batch reactor experiments.

The solution of (4.39) subject to initial condition  $x(0) = x_0$  is obtained as

$$x(t) = \sum_{i=0}^{N-1} c_i(0) e^{\lambda_i t} u^i \quad (4.51)$$

where

$$c_i(0) = \frac{\langle x(0), u^i \rangle}{\langle u^i, u^i \rangle}$$

Here the inner-product used is the one defined in (4.41).

An arbitrary initial composition vector  $x(0)$  will have components along each of  $u^i$ 's, so that  $c_i(0) \neq 0$  for any  $i$ . If we could choose our initial composition to be such that it is of the form  $u^0 + u^j$ , then we would have

$$c_i(0) = 0 \text{ for } i \neq j$$

$$c_j(0) \neq 0 \text{ for } i = j$$

This follows since  $\{u^i\}$  forms an orthogonal basis in  $\mathbb{R}^n$ , with respect to the inner-product (4.41). The experiments have to be conducted such that the initial condition lies along these eigendirections. The system will then converge to the equilibrium composition and the rate of approach is an estimate of the eigenvalue of the system.

For an initial condition which has a component along  $u^j$  and  $u^0$  only, we have

$$x(t) = u^0 + \langle x_0, u^j \rangle e^{\lambda_j t} u^j \quad (4.52)$$

So,

$$\frac{x_i - u_i^0}{x_k - u_k^0} = \frac{u_i^j}{u_k^j} = b_{i,k}^j \quad (4.53a)$$

The  $b_{i,k}^j$ 's can be used to determine the direction of  $u^j$ . This along with the normalisation relation

$$\|u^j\| = 1 \quad (4.53b)$$

specifies  $u^j$  uniquely.

An estimate of  $u^j$  is obtained by following the evolution of the system from an arbitrary (initial condition) close to the equilibrium value. This point is our  $x(t)$ . In general, this vector will contain components  $u^0, u^1, u^2$ . To determine  $u^1$ , we construct an estimate of  $u^1$ , using (4.53). This vector is extended in a direction till one of the  $x_i$ 's equals zero (say,  $x_k$ ). Here the other  $x_i$ 's will be nonzero. Using this as our next initial condition we determine the reaction path. When the system is close to the equilibrium point, the reaction path is again estimated by a straight line and this is extended backwards using

$$x_i = u_i^0 - \frac{u_k^0}{u_k^j} u_i^j \quad (4.54a)$$

The estimate for the new initial conditions is obtained by using

$$x(0) = u^0 - \frac{u_k^0}{u_k^j} u^j \quad (4.54b)$$

The eigenvalue  $\lambda_1$  is obtained after this iterative scheme converges on the eigenvector, as slope of  $\ln(x_i - u_i^0)$  vs.  $t$  curve.

The estimates of the other eigenvectors are obtained from (4.53) iteratively.

Repeating this iterative procedure, we converge on all the eigenvalues and eigenvectors.

We illustrate the method discussed so far with the help of two examples:

1. Consider the two-component system



in a batch reactor.

The evolution of the mole fractions is governed by  $\dot{x} = Kx$ , where

$$K = \begin{bmatrix} -k_{2,1} & k_{1,2} \\ k_{2,1} & -k_{1,2} \end{bmatrix}, \quad x = [x_1, x_2]$$

Clearly,  $\lambda_0 = 0$  is an eigenvalue. The corresponding eigenvector is  $u^0 = (u_1^0, u_2^0)$ . This can be determined experimentally. Normalising this vector in the inner-product defined by (4.41), we obtain

$$\|u^0\| = \langle u^0, u^0 \rangle^{1/2} = (u_1^0 + u_2^0)^{1/2} = 1$$

This equation follows since the sum of mole fractions must add up to unity.

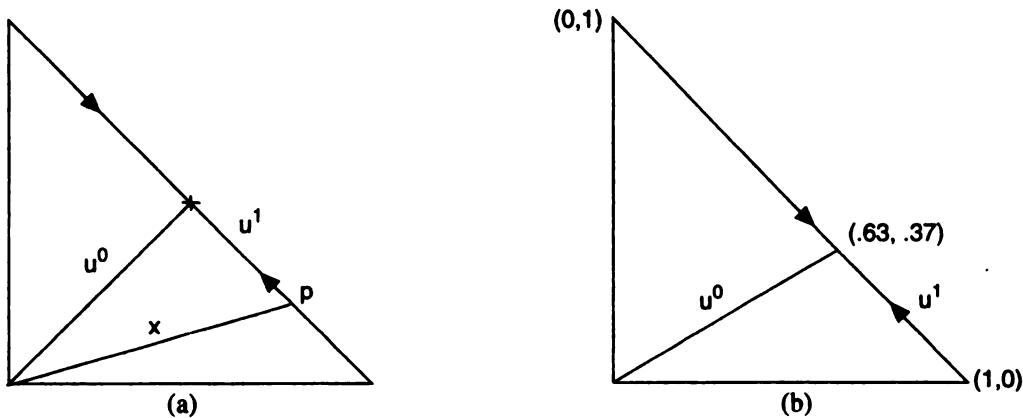
The mole fractions  $x_1, x_2$  at any instant of time must also add up to unity. The state of a system can be followed elegantly in  $x_1 - x_2$  plane. Tracking the system evolution is therefore equivalent to moving along  $x_1 + x_2 = 1$ . Time here is a parameter along the line. The reaction path is a straight line. The system evolves along this line till it reaches equilibrium.

Clearly  $u^0$ , satisfies

$$-k_{2,1}u_1^0 + k_{1,2}u_2^0 = 0$$

When  $k_{2,1} = k_{1,2}$ ,  $u_1^0 = u_2^0$ . The eigenvector  $u^0$  is the  $45^\circ$ -line, as shown in Fig. 4.2(a). Any composition  $x(P)$  can be resolved as a linear combination of  $u^0, u^1$ . So as shown,  $u^1$  is along  $x_1 + x_2 = 1$ . Clearly, for this case  $K$  is real symmetric and  $u^0, u^1$  are orthogonal in the Euclidean inner-product.

For  $k_{2,1} \neq k_{1,2}$ . The eigenvector  $u^0$  is as shown in Fig. 4.2(b). The two vectors  $u^0$  and  $u^1$  are not perpendicular anymore.



**Fig. 4.2** Evolution of a composition in phase plane: (a)  $k_{2,1} = k_{1,2}$ ; (b)  $k_{2,1} \neq k_{1,2}$ .

$$x(t) = u^0 + \langle x(0), u^1 \rangle e^{\lambda_1 t} u^1$$

Clearly,

$$\frac{x_1(t) - u_1^0}{x_2(t) - u_2^0} = \frac{u_1^1}{u_2^1}$$

This along with the normalisation condition  $\|u^1\| = 1$ , uniquely determines  $u^1$ .  $\lambda_1$  is obtained as slope of  $\ln(x_1(t) - u_1^0)$  vs.  $t$  curve.

**Example 4.5** Consider the batch reactor data of the two-component system measured as a function of time shown in Table 4.1. The first eigenvalue  $\lambda_0 = 0$  and its corresponding eigenvector is [.63, .37].

Choosing  $t = 0.35$ , we get

$$\frac{u_1^1}{u_2^1} = \frac{.59 - .63}{.41 - .37} = -1$$

This uniquely determines the direction of the eigenvector  $u^1$  which, along with the normalisation condition, yields

$$u^1 = (.48 - .48)'$$

$\lambda_1$  is obtained by plotting  $\ln(x_1(t) - u_1^0)$  vs. 't'. This yields  $\lambda_1 = -3.16$ .

Clearly, from (4.50) the rate constants are

$$k_{1,2} = 2.0, \quad k_{2,1} = 1.16$$

**Table 4.1** Evolution of the Mole Fractions of the Batch Reactor in Example 4

$t$	$y_1$	$y_2$	$t$	$y_1$	$y_2$
0	0	1.0000	0	1.0000	0
0.0781	0.1385	0.8615	0.0781	0.9197	0.0803
0.1433	0.2304	0.7696	0.1508	0.8609	0.1391
0.2119	0.3089	0.6911	0.2272	0.8120	0.1880
0.2834	0.3744	0.6256	0.3071	0.7720	0.2280
0.3579	0.4287	0.5713	0.3909	0.7396	0.2604
0.4359	0.4733	0.5267	0.4790	0.7137	0.2863
0.5176	0.5096	0.4904	0.5718	0.6932	0.3068
0.6033	0.5389	0.4611	0.6698	0.6771	0.3229
0.6936	0.5622	0.4378	0.7736	0.6648	0.3352
0.7887	0.5806	0.4194	0.8840	0.6554	0.3446
0.8893	0.5948	0.4052	1.0016	0.6484	0.3516
0.9960	0.6057	0.3943	1.1275	0.6433	0.3567
1.1096	0.6139	0.3861	1.2630	0.6397	0.3603
1.2309	0.6200	0.3800	1.4093	0.6372	0.3628
1.3611	0.6243	0.3757	1.5684	0.6355	0.3645
1.5013	0.6274	0.3726	1.7425	0.6344	0.3656
1.6533	0.6295	0.3705	1.9344	0.6337	0.3663
1.8190	0.6309	0.3691	2.1478	0.6333	0.3667
2.0009	0.6318	0.3682	2.3879	0.6331	0.3669
2.2022	0.6323	0.3677	2.6612	0.6330	0.3670
2.4272	0.6326	0.3674	2.9774	0.6329	0.3671
2.6816	0.6328	0.3672	3.3504	0.6329	0.3671
2.9732	0.6329	0.3671	3.8018	0.6329	0.3671
3.3135	0.6329	0.3671			

The procedure here is noniterative, since the reaction path is a straight line for this two-component system, see Fig. 4.2. All reaction paths here necessarily lie along  $u^1$ , which we obtain by subtracting the equilibrium composition  $u^0$ , from any point on the trajectory as shown in Fig. 4.2. Hence it is enough to conduct only one run for this binary system.

**PROBLEMS**

1. Solve  $Ax = b$ , where

$$(a) A = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$(b) A = \begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$(c) A = \begin{bmatrix} -1 & 0 & 1 \\ 3 & 0 & -4 \\ 3 & 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

2. Solve  $\dot{x} = Ax + b$

$$(a) A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 3 \\ 0 & 1 & 0 \end{bmatrix}, \quad b = [0 \quad 0 \quad 0]', \quad x^0 = [2 \quad 0 \quad 2]'$$

$$(b) A = \begin{bmatrix} 5 & 8 \\ -6 & -9 \end{bmatrix}, \quad b = [1 \quad 1]', \quad x^0 = [0 \quad 0]', \quad x^0 = [0 \quad 0]'$$

$$(c) A = \begin{bmatrix} -1 & 0 & -5 \\ 1 & 2 & -1 \\ 1 & 1 & 1 \end{bmatrix}, \quad b = [0 \quad 0 \quad 0]', \quad x_0 = [1 \quad 2 \quad 0]'$$

3. An isothermal batch reactor sustains the following network of reactions (see Fig. 4.3).

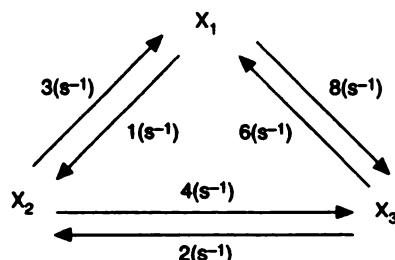


Fig. 4.3 Reaction network of Problem 3.

The initial concentration of  $X_1$  is 3 units and  $X_2, X_3$  is 0. Determine the equilibrium concentration of  $X_1, X_2, X_3$ .

4. Solve  $\dot{x} = Ax$ , where

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

with (i)  $x_0 = [0, 0, 0]'$

(ii)  $x_0 = [1, 0, 0]'$

5. Solve  $\dot{x}(t) = Ax$ , where

$$A = \begin{bmatrix} -0.4 & 0.5 \\ -.16 & 0.2 \end{bmatrix}, \quad x(t=0) = \begin{bmatrix} 100 \\ 1000 \end{bmatrix}$$

6. Consider  $\frac{dx}{dt} = Ax$ , where

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

$$x(0) = [2 \ 1 \ 4]'$$

(a) Find the steady state  $[x_1 \ x_2 \ x_3]$ , setting the time derivative to zero.

(b) Find the time dependence of  $x_1(t), x_2(t), x_3(t)$  and then obtain the steady state of the system.

7. Solve  $Ax = b$ , where

$$A = \begin{bmatrix} 3 & 0 & 0 \\ 5 & 4 & 0 \\ 3 & 6 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}$$

8. Solve  $\dot{x} = Ax + b$ , where

$$b = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \text{ with } x(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Solve for

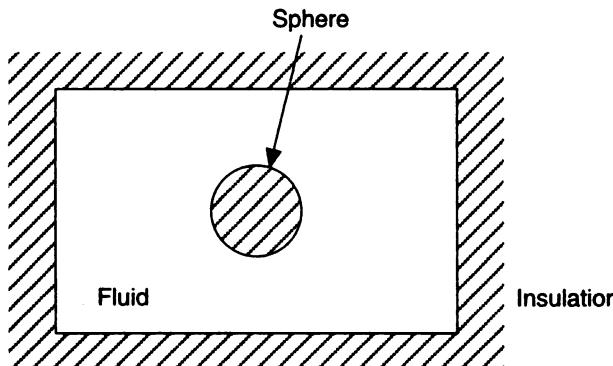
$$(a) A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

$$(b) A = \begin{pmatrix} -4 & 0 \\ 0 & -2 \end{pmatrix}$$

**9.** Re-work Problem 5 for the initial conditions

$$(a) \begin{bmatrix} 2 \\ 1 \\ 6 \end{bmatrix}, \quad (b) \begin{bmatrix} 4 \\ 9 \\ 5 \end{bmatrix}$$

**10.** Consider a steel sphere at 100°C of radius 3 cm. This is dipped into 250 ml of water at 30°C. This is in a vessel which is completely insulated, see Fig. 4.4. What is the equilibrium temperature inside the vessel?



**Fig. 4.4.** The system studied in Problem 10.

Solve (a) as a steady state problem; (b) as a transient problem. (See Example 4.2 for physical properties of steel.)

**11.** Solve  $Ax = b$ , for the following matrices:

$$(a) A = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 3 & -1 \\ 0 & -1 & 2 \end{bmatrix}$$

$$(b) A = \begin{bmatrix} 2 & 2 & -6 \\ 2 & -1 & -3 \\ -2 & -1 & 1 \end{bmatrix}$$

$$(c) A = \begin{bmatrix} 7 & -16 & -8 \\ -16 & 7 & 8 \\ -8 & -8 & -5 \end{bmatrix}$$

$$(d) A = \begin{bmatrix} 5 & 4 & -4 \\ 4 & 5 & 4 \\ -4 & 4 & 5 \end{bmatrix}$$

Solve  $Ax = b$ , where  $b = (1 \ 1 \ 1)'$ . Seek the solution in terms of a linear combination of the eigenbasis. Use orthogonality, or biorthogonality wherever possible and explain when this method is not applicable.

**12.** A continuous stirred tank reactor sustains a first-order endothermic reaction. To maintain the temperature in the reactor, heat is supplied at  $Q$  cal/s. The temperature and concentration profiles in the reactor are governed by

$$V_p C_p \frac{dT}{dt} = q_p C_p (T_{in} - T) + Q + (-\Delta H) V k C$$

$$V \frac{dC}{dt} = q(C_{in} - C) - V k C$$

For a particular system, these linear equations take the form

$$\frac{dx_1}{dt} = -x_1 + x_2 + a + q$$

$$\frac{dx_2}{dt} = -bx_2 + b$$

Obtain the steady state for the above system when  $a = 1$ ,  $b = 2$ ,  $q = 3$ . At  $t = 0$  the heating  $q$  is abruptly shut off. Determine the variation of  $x_1$ ,  $x_2$  with time. This problem is similar to studying the response of a system to a step-change.

**13.** A continuously stirred tank is cooled by circulating cold water at  $T_{cin}$  through the cooling coil. The energy balance equations modelling the system are:

$$\frac{dT}{dt} = \frac{q}{V}(T_{in} - T) - \frac{UA}{V_p C_p} (T - T_c)$$

$$\frac{dT_c}{dt} = \frac{q_c}{V_c}(T_{cin} - T_c) + \frac{UA}{V_c \rho_c C_{pc}} (T - T_c)$$

Determine the steady state of the above system when it is modelled by,

$$\frac{dx_1}{dt} = -2x_1 + x_2 + 1$$

$$\frac{dx_2}{dt} = x_1 - (1 + \alpha)x_2 + 2$$

where  $\alpha = 3$ .

At  $t = 0$  there is a failure in the coolant pump. This causes  $q_c$  or  $\alpha$  to be reduced to zero. What is the trajectory traced by the system?

**14.** A CSTR sustains the series reaction



The dynamic equations governing the system are

$$\frac{dC_A}{dt} = \frac{1}{\tau} (C_{A,in} - C_A) - k_1 C_A$$

$$\frac{dC_B}{dt} = \frac{1}{\tau} C_B + k_1 C_A - k_2 C_B$$

$$\frac{dC_C}{dt} = -\frac{1}{\tau} C_C + k_2 C_B$$

Discuss the evolution of the system to a steady state given that the initial state of the system is  $(0, 0, 0)'$ . Take

$$C_{A,in} = 1, \tau = 1, k_1 = 1, k_2 = 1$$

The frequency response of a system is obtained in process control by varying a parameter periodically. Determine how the system behaves when  $C_{A,in}$  is  $(1 + \sin 2t)$ .

**15.** The dynamics of a system is given by

$$\dot{x} = Ax + b$$

with

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

The system is at a steady state with  $\alpha = 0$ . Calculate the trajectory of the system when  $\alpha$  is changed to 3. What is the state the system tends to as  $t \rightarrow \infty$ ?

**16.** Consider the system,

$$\dot{x} = Ax + b$$

with

$$A = \begin{bmatrix} 3 & 9 \\ a & 4 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ \alpha \end{bmatrix}$$

(a) Obtain the steady state of the system for  $a = 2, \alpha = 3$ . Determine the response of the system when  $\alpha$  is changed to 4. Is this overdamped or underdamped?

(b) Obtain the steady state of the system for  $a = 4, \alpha = 3$ .

## REFERENCES

Noble, B. and Daniel, J.W., Applied Linear Algebra, Prentice-Hall (1977).

Ramkrishna, D. and Amundson, N.R., Linear operator methods in chemical engineering with applications to transport and chemical reaction systems, Prentice-Hall, Englewood Cliffs, New Jersey (1985).

Stephanopoulos, G., Chemical Process Control: An introduction to theory and practice, Prentice-Hall of India, New Delhi (1984).

Wei, J. and Prater, C.C., The Structure and Analysis of Complex Reaction Systems: Advances in catalysis, 13, 5, Academic Press, New York (1962).

# 5

# Partial Differential Equations

---

In Chapters 1–4, we dealt with finite dimensional spaces  $\mathbf{R}^n$ . The elements in the space were vectors, each with a finite number of coordinates, and the operators were matrices. In this part we will be discussing partial differential equations which arise in modelling many chemical engineering systems. The vector space of interest now is infinite dimensional and the elements in it are functions. Differential operators in this space transform and map these elements to other elements. The presentation of various concepts in Chapters 5–9 is along the same lines as in Chapters 1–4. This illustrates and helps bring out the applicability of the concepts already discussed.

This chapter deals with introductory concepts of linear partial differential equations. It includes the classification of these equations, the associated boundary conditions, and discusses the principles of linearity and superposition.

## 5.1 FUNDAMENTAL CONCEPTS: A REVIEW

In Chapter 1 we described some physical situations in which partial differential equations arise. Now we formally define concepts which are necessary while working with these equations.

**Definition 5.1** The order of a differential equation is the order of the highest derivative occurring in it, see Kersten (1969), Kreyszig (1982). For example,

$$\frac{du}{dt} + \left(\frac{du}{dt}\right)^2 = u^2 \text{ is a first order nonlinear equation}$$

$$\frac{\partial u}{\partial t} = \frac{\partial^3 u}{\partial x^3} \text{ is a third order linear equation}$$

**Definition 5.2** The degree of a differential equation is the power to which the highest order derivative has been raised, see Kersten (1969), and Kreyszig (1982).

$$\frac{\partial u}{\partial t} = \left(\frac{\partial^2 u}{\partial x^2}\right)^3 + \left(\frac{\partial^3 u}{\partial x^3}\right)^n$$

is of  $n$ th degree.

The differential equations which arise in modelling a problem are solved subject to boundary conditions. The differential equation specifies the relationship which the dependent variables must satisfy in the region or domain of interest. The boundary conditions are the constraints the variable must satisfy on the boundary of the domain or region. A solution to a problem is sought such that

the equation and the boundary conditions are satisfied in the domain and on the boundary respectively. A problem is said to be well-posed when the solution is uniquely determined and it is a sufficiently smooth and differentiable function of the independent variables. We will be primarily concerned with second order linear partial differential equations in this section.

**Example 5.1** Consider the two-dimensional (in space) steady heat conduction problem defined on a rectangular region  $0 \leq x \leq a$ ,  $0 \leq y \leq b$  (see Fig. 5.1). Assuming that there are no sources or sinks in the domain, the temperature is governed by

$$\left( \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} \right) = 0 \quad (5.1)$$

in the region  $0 < x < a$ ,  $0 < y < b$

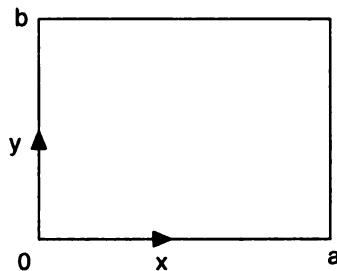


Fig. 5.1 Rectangular domain of Example 5.1.

The boundary conditions are specified on the surfaces of the problem, i.e.  $x = 0$ ,  $x = a$ ,  $y = 0$ ,  $y = b$ . Typically, these conditions specify the value of the function or the first derivative or their combination and are of the form

$$u(x = 0, y) = f(y) \quad (5.2a)$$

$$\frac{\partial u}{\partial x}(x = a, y) = 0 \quad (5.2b)$$

$$\frac{\partial u}{\partial y}(x, y = 0) - hu(x, y = 0) = g(x) \quad (5.2c)$$

$$u(x, y = b) = h(x) \quad (5.2d)$$

The boundary conditions have to be consistent with each other in order for a problem to be well-posed. This means that at the common points, i.e. the vertices of the rectangle, the boundary conditions must not violate each other. So at  $x = 0$ ,  $y = b$ , from (5.2a) and (5.2d), we have

$$u(x = 0, y = b) = f(b) = h(0)$$

If  $f(b) \neq h(0)$ , the problem is ill-posed. Under these conditions there would be sharp gradients near this corner if the solution were to try and satisfy both these conditions simultaneously. Similarly, for well-posedness the consistency condition at  $x = 0$ ,  $y = 0$  yields

$$\frac{\partial u}{\partial y}(x = 0, y = 0) - hu(x = 0, y = 0) = g(0)$$

or

$$\frac{df}{dy} (y = 0) - hf(0) = g(0)$$

**Example 5.2**

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} \quad (5.3)$$

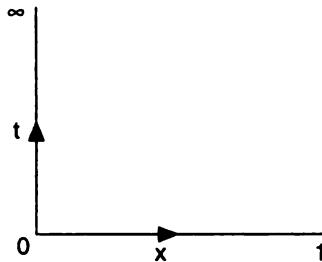
subject to

$$T(x, t = 0) = \sin \pi x \quad (5.4a)$$

$$T(x = 0, t) = 0 \quad (5.4b)$$

$$T(x = 1, t) = 1 \quad (5.4c)$$

Equations (5.3) and (5.4) govern the evolution of temperature in space ( $x$ ) and time ( $t$ ) in a one-dimensional slab. This problem is defined in the region  $0 < x < 1$ ,  $0 < t < \infty$ , (Fig 5.2). This is an ill-posed problem as the initial condition, at  $t = 0$  violates the boundary condition at  $x = 1$ . This results in a ‘singularity’ at  $x = 1$ , which manifests itself in sharp temperature gradients near this surface for small  $t$ , i.e. at the initial stages.



**Fig. 5.2** Space-time domain of Example 5.2.

**Example 5.3**

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \quad \text{in } 0 < x < 1, \quad 0 < y < 1 \quad (5.5a)$$

This is the steady-state problem of Example 5.1. Let this be subject to the boundary conditions

$$T(x = 0, y) = \sin \pi y \quad (5.5b)$$

$$T(x = 1, y) = 0 \quad (5.5c)$$

$$T(x, y = 0) = 0 \quad (5.5d)$$

$$T(x, y = 1) = x(1 - x) \quad (5.5e)$$

This is a well-posed problem since at the common points of the different boundaries both the boundary conditions are satisfied simultaneously.

A linear equation can be broadly classified as follows:

(i) **Homogeneous equation.** The differential equation does not contain any term independent of the dependent variable or its derivatives. In particular, all terms in the equation contain the dependent variable. Two examples of a homogeneous equation are

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

**(ii) Nonhomogeneous equation.** The equation contains terms which are independent of the dependent variable. These terms are the nonhomogeneities present and they represent sources or sinks in a problem.

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \sin x$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \sin x \cos y$$

The terms  $\sin x$  and  $\sin x \cos y$  render these equations nonhomogeneous.

The boundary conditions of a problem can be similarly homogeneous or nonhomogeneous, depending on whether they contain terms independent of  $u$ , see Weinberger (1965) and Berg (1964). The condition

$$\frac{\partial u}{\partial x}(x=1) + u(x=1) = 0$$

is a homogeneous boundary condition as it contains no term independent of  $u$ . The relation

$$u(x=0) = f(y)$$

is an example of a nonhomogeneous condition as  $f(y)$  is independent of  $u$ .

The partial differential equations which we come across in chemical engineering frequently are second order equations. These arise typically in reaction-diffusion systems, heat-transfer, fluid-flow problems, see Aris (1975) and Bird et al. (1960).

## 5.2 CLASSIFICATION OF SECOND ORDER PARTIAL DIFFERENTIAL EQUATIONS

Consider a second order partial differential equation in  $n$  independent variables  $x_1, x_2, \dots, x_n$ . This can be written as

$$\sum_{i=1}^n \sum_{j=1}^n a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} = f\left(\frac{\partial u}{\partial x_i}, \dots, \frac{\partial u}{\partial x_n}, x_1, \dots, x_n, u\right) \quad (5.6)$$

Equation (5.6) has been written such that all second order terms occur on the left-hand side. The coefficients  $a_{ij}$  are assumed to be independent of  $u$  and its derivatives. They can however be functions of  $x_i$ . The  $x_i$ 's in this section denote the various coordinates. Thus we can identify  $x_1, x_2, x_3$  with  $x, y, z$ .

The  $a_{ij}$ 's in (5.6) can always be written so that they satisfy

$$a_{ij} = a_{ji} \quad \text{for } i \neq j$$

This is rendered possible as

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

Thus the  $a_{ij}$ 's are the elements of a real symmetric matrix. The matrix  $A$  it follows has all real eigenvalues. Equation (5.6) is classified on the basis of the nature of these eigenvalues, as:

- (a) **Elliptic**, if all its eigenvalues are of the same sign, i.e. all positive or all negative;
- (b) **Hyperbolic**, if some eigenvalues are positive and some negative; and (c) **Parabolic**, if at least one eigenvalue is zero.

The classification is therefore dependent only on the second order terms (see Weinberger, 1965). It is global if the  $a_{ij}$ 's are independent of the  $x_i$ 's. Should the  $a_{ij}$ 's be a function of the  $x_i$ 's, the nature of the equation changes with the  $x_i$ 's and our classification would be local. The second order terms in an equation are generated by diffusion in mass transfer and conduction in energy transfer. Such problems arise when diffusion contributes to overall transport process. Elliptic problems usually arise while studying steady state behaviour of diffusive systems, and parabolic problems typically arise while analysing the transient behaviour of diffusive systems. It must be pointed out that certain steady state problems can be parabolic (see Example 5.4 (iii), and Problem 6 (ii) at the end of the chapter).

#### **Example 5.4** Classify the equations

$$(i) \quad \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

$$(ii) \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial u}{\partial x}$$

$$(iii) \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \frac{\partial u}{\partial z}$$

$$(iv) \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + 4 \frac{\partial^2 u}{\partial x \partial y} = \sin x$$

(i) There are two independent variables  $x, t$ . Let  $x_1 = x$  and  $x_2 = t$ . We have  $a_{11} = 1$ ,  $a_{12} = a_{21} = a_{22} = 0$ . The eigenvalues of  $A$  are 1, 0. The equation is parabolic. The first order derivative with respect to time indicates the presence of the variable  $t$ . Although there are no second order derivatives with respect to  $t$ , the first order time derivative indicates  $A$  is a  $2 \times 2$  matrix and is responsible for contributing the zero eigenvalue.

(ii) Identify  $x$  with  $x_1$  and  $y$  with  $x_2$ . The mixed derivative terms are absent. This yields  $a_{11} = a_{22} = 1$  and  $a_{12} = a_{21} = 0$ . The eigenvalues of  $A$  are 1, 1, and the equation is elliptic.

(iii) The equation has three independent variables  $x (= x_1)$ ,  $y (= x_2)$ ,  $z (= x_3)$ . The only nonzero elements of  $A$  are  $a_{11} = a_{22} = 1$ . The eigenvalues of  $A$  are 1, 1, 0, and the equation is parabolic. (This is a steady state problem where convection is important in the  $z$ -direction and conduction in the  $x$ -,  $y$ -direction.)

(iv) We rewrite the equation as

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + 2 \frac{\partial^2 u}{\partial x \partial y} + 2 \frac{\partial^2 u}{\partial y \partial x} = \sin x$$

with  $x_1 = x$ ,  $x_2 = y$ . This assures us that  $a_{ij} = a_{ji}$ . We have  $a_{11} = a_{22} = 1$ , and  $a_{12} = a_{21} = 2$ . The eigenvalues of  $A$  are -1, 3, and the equation is hyperbolic.

**Example 5.5** Determine the nature of

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + 2x \frac{\partial^2 u}{\partial x \partial y} = \sin x + \frac{\partial u}{\partial x}$$

Let  $x_1 = x$ ,  $x_2 = y$ . Clearly,  $a_{11} = a_{22} = 1$ . It appears that  $a_{21} = 2x_1$  and  $a_{12} = 0$ . We use

$$2x_1 \frac{\partial^2 u}{\partial x_2 \partial x_1} = x_1 \frac{\partial^2 u}{\partial x_2 \partial x_1} + x_1 \frac{\partial^2 u}{\partial x_1 \partial x_2}$$

and obtain  $a_{21} = a_{12} = x_1$ . The matrix  $A$  is now symmetric and its eigenvalues which are real depend on  $x_1$ .  $\lambda_1 = 1 - x_1$  and  $\lambda_2 = 1 + x_1$ . The classification is therefore local and changes with  $x_1$ . For  $x_1 > 1$ ,  $\lambda_1 < 0$ ,  $\lambda_2 > 0$ , and the equation is hyperbolic. For  $x_1 < -1$ ,  $\lambda_1 > 0$ ,  $\lambda_2 < 0$ , and the equation is hyperbolic. For  $x_1 = \pm 1$ , the equation is parabolic, and for  $-1 < x_1 < 1$ , the equation is elliptic.

### 5.2.1 Boundary Conditions

While dealing with second order partial differential equations, the boundary conditions specify the variable or its derivative or a combination of the two on the surface of the domain of interest. Four types of boundary conditions occur frequently. These are now discussed.

**(i) Dirichlet condition.** A boundary condition is said to be a Dirichlet condition when the value of the dependent variable is specified on a surface. If the value of the function is uniformly zero, we have a homogeneous Dirichlet condition; else it is nonhomogeneous. The general Dirichlet condition is of the form

$$u = f \text{ on the surface} \quad (5.7a)$$

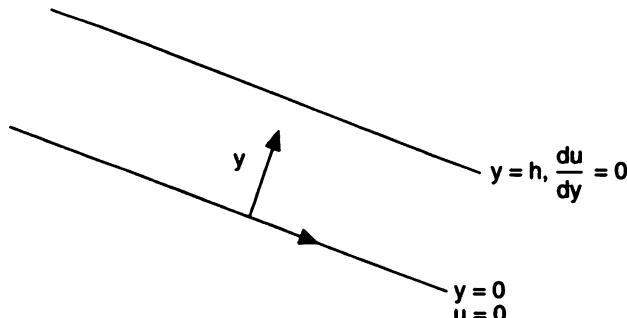
Consider a fluid flowing over a stationary flat plate. If the surface of the plate is at  $y = 0$ , the no slip condition yields the homogeneous Dirichlet boundary condition for the velocity in the  $x$ -direction, viz.  $u$ . That is

$$u(y = 0) = 0$$

Consider a sphere of radius  $R$  immersed in a well-stirred fluid at a temperature  $T_a$ . The agitation in the fluid lowers the resistance to heat transfer in the fluid. The temperature of the surface of the sphere can be taken to be equal to  $T_a$ . This yields the nonhomogeneous Dirichlet condition

$$T(r = R) = T_a$$

A similar condition arises for the concentration on the surface of a catalytic pellet in a fluidised catalytic cracker.



**Fig. 5.3** Fluid flowing down an inclined plane, no slip condition at  $y = 0$ , no shear stress at  $y = h$ .

For the low mass transfer resistance due to the agitation, the surface concentration  $C$  in the spherical catalyst is

$$C(r = R) = C_0$$

where  $C_0$  is the concentration in the bulk fluid.

**(ii) Neumann condition.** A boundary condition where we specify the outward normal derivative on a surface is called the Neumann condition. This condition is of the form

$$-n \cdot \nabla u = f \quad (5.7b)$$

where  $n$  is the outward normal unit vector on the surface, and  $f$  is the flux away from surface.

Let the flux of heat through the curved surface of a solid cylinder be  $q$ . Heat flow occurs inside the cylinder due to conduction. This results in the nonhomogeneous Neumann condition

$$-k \frac{\partial T}{\partial r} \Big|_{r=R} = q$$

In a general problem,  $q$  does not have to be a constant and can vary with the angular position  $\theta$  as well as the axial distance  $z$  on the surface  $r = R$ .

The heat flux across an insulated surface is zero. This yields the general homogeneous Neumann condition

$$n \cdot \nabla T = 0$$

Consider a fluid flowing as a film of constant thickness (fully developed flow) down an inclined plane (Fig. 5.3). The vanishing of the shear stress on the free surface of the film yields the homogeneous Neumann condition

$$\frac{\partial u}{\partial y} = 0 \quad \text{at } y = h$$

The zero mass flux across an impermeable surface at  $x = 0$  yields the homogeneous Neumann condition for the concentration of a species  $i$  as

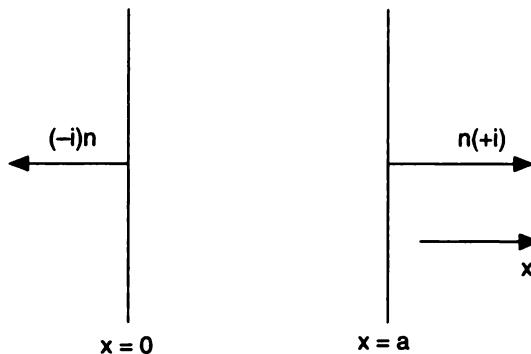
$$\frac{\partial C_i}{\partial x} (x = 0) = 0$$

**(iii) Robin (or mixed) boundary condition.** In many problems a combination of the outward normal derivative and the variable is specified on a surface. Here the external resistance to transfer is comparable with the internal resistance. Such a condition is called the Robin or the mixed boundary condition. It is of the form

$$n \cdot \nabla u + hu = hu_0 \quad (5.7c)$$

We recover the Neumann condition from this in the limit  $h \rightarrow 0$ , and the Dirichlet condition in the limit  $h \rightarrow \infty$ .

Consider a rectangular slab ( $0 < x < a$ ). Let the two faces at  $x = 0, a$  lose heat to the ambient at  $T_a$  by Newton's law of cooling, i.e. the rate of heat transfer is proportional to the temperature difference ( $T_s - T_a$ ). Heat is transferred to the surface in the slab by conduction. At  $x = a$ ,  $n$  is in the positive  $x$ -direction (Fig. 5.4).



**Fig. 5.4** The direction of the outward normal at two surfaces of body  $0 < x < a$ .

The gradient vector is  $\left( i \frac{\partial T}{\partial x} + j \frac{\partial T}{\partial y} + k \frac{\partial T}{\partial z} \right)$ . Substituting in (5.7c), we get

$$k \frac{\partial T}{\partial x}(x = a) + hT(x = a) = hT_a$$

At  $x = 0$ ,  $n$  is in the negative  $x$ -direction. Substituting the gradient vector again in (5.7c) yields

$$-k \frac{\partial T}{\partial x} + hT = hT_a \quad \text{at } x = 0$$

The presence of  $T_a$  in these equations renders them nonhomogeneous. This is a Robin or a mixed condition as we specify a mixture or a combination of the outward normal derivative and the variable at  $x = a$  or  $x = 0$ .

A similar condition arises when we describe the concentration at a surface when the external resistance to mass transfer is not negligible. The boundary condition for concentration at the surface  $C_s$  of a spherical catalyst pellet is

$$-D \frac{\partial C}{\partial r}(r = R) = k_m(C(r = R) - C_a)$$

In transport phenomena we encounter these boundary conditions, depending on the nature of the physical system.

When the external resistance to transfer is low as compared to the internal resistance, we have Dirichlet boundary conditions. This is the usual situation that prevails when we have well-stirred conditions. The external transfer coefficients are very high and the dependent variable at the surface attains the ambient value.

At the other extreme, when the external resistance to the transfer is high as compared to the internal resistance, we have Neumann conditions. Here the external transport coefficients are very low.

When the external and internal resistances to transfer are comparable, we have the mixed or Robin boundary conditions.

Consider the heat transfer through a sphere placed in a fluid at a temperature  $T_a$ . Assuming heat transfer to the fluid occurs only due to convection, the general boundary condition is

$$k \frac{\partial T}{\partial r} + h(T - T_a) = 0$$

Rendering this equation dimensionless, using  $T/T_a = T^*$  and  $r^* = r/D$ , where  $D$  is the diameter of the sphere, we have

$$\frac{\partial T^*}{\partial r^*} + \text{Bi} (T^* - 1) = 0$$

Here, the Biot number  $\text{Bi} = hD/k$ , where  $h$  is the heat transfer coefficient and  $k$  is the thermal conductivity of the solid.

As  $h \rightarrow \infty$ ,  $\text{Bi} \rightarrow \infty$ , and we have the Dirichlet condition  $T^* = 1$ . In this limit the spatial gradients inside the sphere play an important role and cannot be neglected.

As  $h \rightarrow 0$ ,  $\text{Bi} \rightarrow 0$ . The external resistance is high and the external transfer is the rate determining step. This limit corresponds to large  $k$  (relative to  $hD$ ). Here we can neglect the spatial gradients inside the sphere and treat the temperature in a sphere to be uniform.

**(iv) Cauchy conditions.** These boundary conditions occur usually in solving hyperbolic equations. Here both the value of the dependent variable and the value of its derivative are specified individually at the surface. For example, at  $t = 0$

$$u = 0 \quad (5.8a)$$

$$\frac{\partial u}{\partial t} = 0 \quad (5.8b)$$

This forms a homogeneous set of Cauchy conditions.

We have seen in Chapter 3 that functions map real variables to real variables, and matrices map vectors to vectors. In this section we will be discussing differential operators which map functions from their domain space to other functions in their range space. The elements in the spaces could be functions of one or several variables.

### Example 5.6

$$g(x) = \frac{d}{dx} f(x)$$

can be written as  $g = Lf$ .  $L$  represents the derivative operator  $d/dx$  operating on the function  $f$ . We may further restrict  $f$  to be defined in a closed interval  $[a, b]$ .

### Example 5.7

$$v(x, t) = \left( \frac{\partial}{\partial t} - \frac{\partial^2}{\partial x^2} \right) u(x, t)$$

can also be written as  $v(x, t) = Lu(x, t)$ . The operator  $L$  takes the first derivative with respect to 't' and from this subtracts the second derivative with respect to 'x'. These operations result in the function  $v(x, t)$ .

## 5.3 LINEARITY AND SUPERPOSITION

A **linear operator**, as we have already seen, is one that satisfies

$$(i) Lu = 0 \quad \text{if } u = 0 \quad (5.9a)$$

$$(ii) L(c_1u^1 + c_2u^2) = c_1Lu^1 + c_2Lu^2 \quad (5.9b)$$

where  $u$ ,  $u^1$ ,  $u^2$  are elements on which  $L$  operates and  $c_1$ ,  $c_2$  are scalars.

Introducing the notion of the operator  $L$  enables us to recast the differential equations in a form similar to linear algebraic systems

$$Ax = b$$

Here the matrix  $A$ , transforming the vector  $x$  to the vector  $b$ , is similar to  $L$  which maps  $u(x, t)$  in Example 5.7 to  $v(x, t)$ . A good introduction for an operator theoretic approach to differential systems can be found in Kaplan (1962).

Clearly, the operators in the above examples are linear. In this section we will be interested in problems governed by second order partial differential equations which frequently arise in modelling engineering systems. The general linear problem here can be written as

$$Lu = f \quad \text{in } V \quad (5.10a)$$

$$u(t = 0) = u_0 \quad \text{in } V \quad (5.10b)$$

$$B_1u = g \quad \text{on } S_D \quad (5.10c)$$

$$B_2u = h \quad \text{on } S_R \quad (5.10d)$$

The Problem (5.10) is defined in a general region  $V$ , whose boundary (surface) is  $S$ . The differential equation is valid in the interior of  $V$  and the boundary conditions are valid on the boundary  $S$ . The boundary  $S$  is subdivided into two parts:  $S_D$ ,  $S_R$ .  $S_D(S_R)$  is the portion of the boundary where Dirichlet (Robin) conditions are specified. Neumann conditions are taken to be a special case of the Robin conditions. In Example 5.1, for instance,  $V$  is the rectangular region defined by  $0 < x < a$ ,  $0 < y < b$ .  $S_D$  consists of two line segments:

$$(i) x = 0, 0 \leq y \leq b; (ii) y = b, 0 \leq x \leq a$$

Similarly,  $S_R$  is made up of the following line segments:

$$(i) x = a, 0 \leq y \leq b; (ii) y = 0, 0 \leq x \leq a.$$

In the linear problem (5.10),  $L$  is a linear differential operator, and  $B_1$ ,  $B_2$  are linear boundary operators. All three satisfy (5.9a) and (5.9b). A partial differential equation is said to be linear only when the equation is linear, the boundary conditions are linear, and the domain  $V$  and its boundary  $S$  are independent of the dependent variable. Violation of any one of these conditions renders the system nonlinear.

### 5.3.1 Superposition

The general problem (5.10) has four sources of nonhomogeneities present, viz.  $f$ ,  $g$ ,  $h$  and  $u_0$ . The linearity of  $L$ ,  $B_1$  and  $B_2$  allows us to construct the solution  $u$ , using the principle of superposition. We decompose the original problem into sub-problems, such that each sub-problem has only one nonhomogeneity. The actual solution is obtained by adding (superposing) the solutions of the individual sub-problems (see Weinberger (1965)). This idea is illustrated with a two-point boundary-value problem (ordinary differential equation).

**Example 5.8** Consider the equation

$$\frac{d^2u}{dx^2} = x \quad \text{in } 0 < x < 1 \quad (5.11)$$

subject to

$$u(0) = 1, \quad u(1) = 2$$

The domain of the problem  $V$  is the open interval  $(0, 1)$  and the boundary  $S$  consists of the end points of this interval, i.e.  $x = 0$ , and  $x = 1$ . This equation can be solved explicitly by integrating twice. The two constants of integration are obtained by imposing the boundary conditions. This yields the solution as

$$u(x) = \frac{x^3}{6} + \frac{5x}{6} + 1 \quad (5.12)$$

We now apply the principle of superposition on this problem. The problem has three nonhomogeneities,  $x$  in the domain and 1, 2 at the two boundary points. We seek the solution  $u(x)$  as the sum of the solutions  $u_1(x) + u_2(x) + u_3(x)$ .

Each sub-problem  $u_i$  considers the effect of only one nonhomogeneity. Thus we take

$$\frac{d^2 u_2(x)}{dx^2} = 0, \quad u_2(0) = 1, \quad u_2(1) = 0 \quad (5.13a)$$

$$\frac{d^2 u_2(x)}{dx^2} = 0, \quad u_2(0) = 0, \quad u_2(1) = 0 \quad (5.13b)$$

$$\frac{d^2 u_3(x)}{dx^2} = 0, \quad u_3(0) = 0, \quad u_3(1) = 2 \quad (5.13c)$$

The problem (5.13a) for  $u_1$  has homogeneous boundary conditions at  $x = 0$  and  $x = 1$  and a nonhomogeneity in the differential equation. The problem for  $u_2$  (5.13b), is a homogeneous equation and has one nonhomogeneity on the boundary. The same is true for problem (5.13c) which governs  $u_3$ .

The solutions to the above equations yield

$$u_1(x) = \frac{x^3 - x}{6}, \quad u_2(x) = 1 - x, \quad u_3(x) = 2x$$

Clearly,  $u(x) = u_1(x) + u_2(x) + u_3(x)$  satisfies problem (5.11).

This example is over simplistic. It enables us to understand how the principle of superposition is employed. This principle forms the basis of solving linear partial differential equations. Before discussing the applications of this principle, we discuss some important features of these systems. The typical second order linear partial differential equation we come across in engineering applications is of the form

$$\nabla^2 u = a \frac{\partial u}{\partial t} + b \frac{\partial^2 u}{\partial t^2} + cu + f(x_1, x_2, x_3, t) \quad (5.14a)$$

subject to appropriate boundary conditions and initial conditions. For convenience and ease of physical interpretation, the independent variable time has been given a distinct identity  $t$  instead of denoting it generically as  $x_4$ . The Laplacian contains derivatives with respect to spatial coordinates  $x_1, x_2, x_3$  only.

The partial differential equation (5.14a) is solved in a three-dimensional region  $V$  which is bounded or unbounded. The boundary of  $V$  is denoted as  $S$ . On the spatial surface  $S$  we have boundary conditions of the form

$$\alpha(s, t) n \cdot \nabla u + \beta(s, t)u = h(s, t) \quad (5.14b)$$

where  $n$  is the outward normal direction to  $S$  and  $s$  represents the spatial coordinates along  $S$ . Clearly,

- (a) for  $b = 0, a = 0$  (5.14a) is elliptic;
- (b) for  $a \neq 0, b = 0$  (5.14a) is parabolic;
- (c) for  $b > 0$ , (5.14a) is hyperbolic;
- (d) for  $f = 0$ , (5.14a) is homogeneous;
- (e) for  $\alpha \neq 0, \beta \neq 0$ , the boundary condition (5.14b) is Robin;
- (f) for  $\alpha = 0$ , the boundary condition (5.14b) is Dirichlet; and
- (g) for  $\beta = 0$ , the boundary condition (5.14b) is Neumann.

For the remaining part of this text we restrict ourselves to the case where  $\alpha, \beta, h$  are time independent.

A boundary is closed if it completely surrounds the region of interest  $V$  and boundary conditions are specified everywhere on it, even if a part of the boundary goes to infinity. A boundary is open if it goes to infinity and no boundary condition is imposed along the part at infinity.

In addition to the conditions for well-posedness of a given equation described earlier, the conditions for a unique stable solution to exist for the different equations are now enumerated.

(i) **Elliptic equations** of the form (5.14a) need Dirichlet, Neumann or Robin conditions on a closed boundary. Neumann conditions cannot be described everywhere on the boundary.

(ii) **Parabolic equations** need Dirichlet conditions on an open boundary. The solution is unique and stable only in the positive direction of this coordinate.

(iii) **Hyperbolic equations** need Cauchy conditions on an open boundary, see Berg (1964), Kaplan (1962).

We assume these results in this text. The formal basis of these criteria is beyond the scope of this study. But the reader will be able to intuitively understand these in this section.

The domain of an elliptic equation is of the form shown in Fig. 5.5(a). We restrict ourselves to U-shaped domains for hyperbolic equations as shown in Fig. 5.5(b). Two conditions have to be specified at the base of the U. Therefore, we need Cauchy conditions at the base of the U and Dirichlet, Neumann, or Robin conditions along the sides.

For parabolic equations we again have U-shaped domains. We need Dirichlet, Neumann, or Robin conditions along the sides and Dirichlet conditions at the base of the U (Fig. 5.5c). These are needed everywhere on the open boundary as long as the boundary is open in the direction of increasing time. Going backwards in time leads to irregularities in the solution. In these problems we can predict the future of a system but cannot reconstruct the past.

All linear partial differential equations of the form (5.14) can be decomposed by the method of superposition into one of the following four sub-problems, which contains one nonhomogeneity.

(i) **Elliptic Boundary Condition (EBC):** The equation is elliptic, homogeneous, but the boundary conditions are nonhomogeneous.

(ii) **Parabolic Initial Condition (PIC):** The equation is homogeneous parabolic and is subject to homogeneous boundary conditions, but a nonzero initial condition.

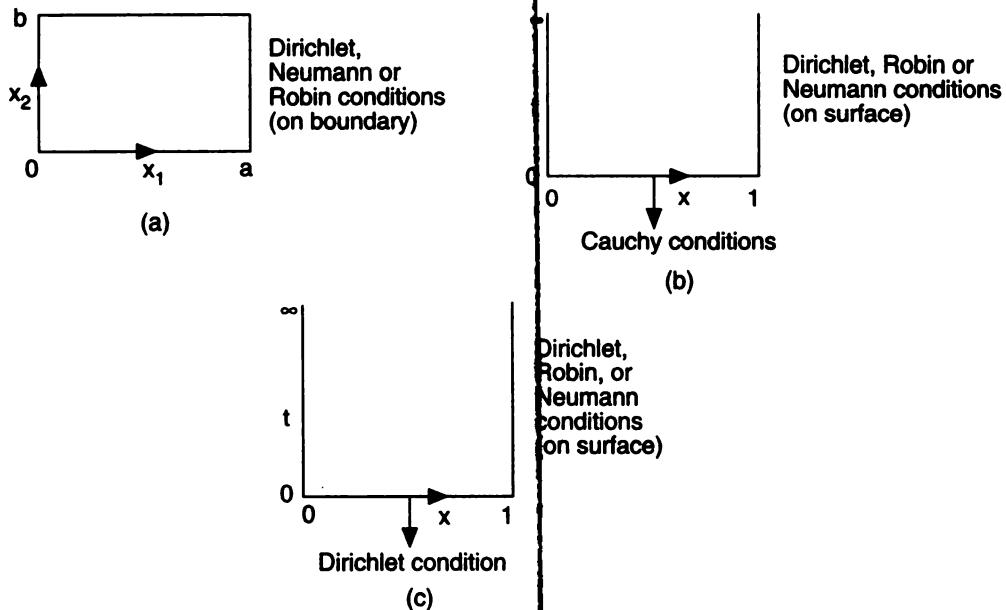


Fig. 5.5 Visualisation of domain and specifying the boundary conditions of: (a) elliptic; (b) hyperbolic; (c) parabolic systems.

(iii) *Hyperbolic Initial Condition (HIC):* The hyperbolic equation is homogeneous. It has homogeneous boundary conditions but a nonzero initial condition.

(iv) *Nonhomogeneous Equation (NHE):* The equation (elliptic, parabolic or hyperbolic) is nonhomogeneous and has homogeneous boundary conditions and zero initial conditions.

We now see how these four problems arise in the solution of elliptic and parabolic partial differential equations.

(i) *Elliptic problems.* The general steady state problem is of the form

$$Lu = f(x_1, x_2, x_3) \quad \text{in } V \quad (5.15a)$$

subject to

$$u = u_D \quad \text{on } S_D \quad (5.15b)$$

$$n \cdot \nabla u + hu = hu_R \quad \text{on } S_R \quad (5.15c)$$

The operator  $L$  is typically the Laplacian, denoted by  $\nabla^2$  or  $\Delta$ . This problem has three sources of nonhomogeneities  $f(x_1, x_2, x_3)$ ,  $u_D$  and  $u_R$ . We seek the solution  $u$  to this problem as

$$u = u_1 + u_2 + u_3$$

such that  $u_1$  satisfies

$$Lu_1 = f(x_1, x_2, x_3) \quad \text{in } V \quad (5.16a)$$

$$n \cdot \nabla u_1 + hu_1 = 0 \quad \text{on } S_R \quad (5.16b)$$

$$u_1 = 0 \quad \text{on } S_D \quad (5.16c)$$

$u_2$ ,  $u_3$  are the solutions of

$$Lu_2 = 0 \quad \text{in } V \quad (5.17a)$$

subject to

$$n \cdot \nabla u_2 + hu_2 = hu_R \quad \text{on } S_R \quad (5.17b)$$

$$u_2 = 0 \quad \text{on } S_D \quad (5.17c)$$

and

$$Lu_3 = 0 \quad \text{in } V \quad (5.18a)$$

$$n \cdot \nabla u_3 + hu_3 = 0 \quad \text{on } S_R \quad (5.18b)$$

$$u_3 = u_D \quad \text{on } S_D \quad (5.18c)$$

The problem defining  $u_1$  is called the nonhomogeneous equation (NHE) problem. The equation, elliptic in this case, is nonhomogeneous and it satisfies homogeneous boundary conditions on  $S$ . The problems defining  $u_2, u_3$  are called the elliptic boundary condition (EBC) problem. The equation is homogeneous, but there is a nonhomogeneity occurring in the boundary conditions here.

(ii) *Parabolic problems.* The general parabolic problem is of the form

$$Lu = \frac{\partial u}{\partial t} + f(x_1, x_2, x_3) \quad \text{in } V \quad (5.19a)$$

This is subject to the initial condition

$$u(t = 0) = u_0(x_1, x_2, x_3) \quad \text{in } V \quad (5.19b)$$

and boundary conditions

$$n \cdot \nabla u + hu = hu_R \quad \text{on } S_R \quad (5.19c)$$

$$u = u_D \quad \text{on } S_D \quad (5.19d)$$

The operator  $L$  here is again typically of the form  $\nabla^2$

The solution  $u$  is sought as

$$u(x) = u_1(x) + u_2(x) + u_3(x) + u_4(x)$$

since (5.19a)–(5.19d) has four nonhomogeneities. Here  $u_1$  is the solution of

$$Lu_1 = f(x_1, x_2, x_3) \quad \text{in } V \quad (5.20a)$$

subject to

$$u_1(t = 0) = 0 \quad \text{in } V \quad (5.20b)$$

$$n \cdot \nabla u_1 + hu_1 = 0 \quad \text{on } S_R \quad (5.20c)$$

$$u_1 = 0 \quad \text{on } S_D \quad (5.20d)$$

$u_2$  is obtained from

$$Lu_2 = \frac{\partial u_2}{\partial t} \quad \text{in } V \quad (5.21a)$$

subject to

$$u_2(t = 0) = u_0 \quad \text{in } V \quad (5.21b)$$

$$n \cdot \nabla u_2 + hu_2 = 0 \quad \text{on } S_R \quad (5.21c)$$

$$u_2 = 0 \quad \text{on } S_D \quad (5.21d)$$

Similarly,  $u_3, u_4$  are the solutions of

$$Lu_3 = \frac{\partial u_3}{\partial t} \quad \text{in } V \quad (5.22a)$$

subject to

$$u_3(t=0) = 0 \quad \text{in } V \quad (5.22\text{b})$$

$$n \cdot \nabla u_3 + hu_3 = hu_R \quad \text{on } S_R \quad (5.22\text{c})$$

$$u_3 = 0 \quad \text{on } S_D \quad (5.22\text{d})$$

and

$$Lu_4 = \frac{\partial u_4}{\partial t} \quad \text{in } V \quad (5.23\text{a})$$

subject to

$$u_4(t=0) = 0 \quad \text{in } V \quad (5.23\text{b})$$

$$n \cdot \nabla u_4 + hu_4 = 0 \quad \text{on } S_R \quad (5.23\text{c})$$

$$u_4 = u_D \quad \text{on } S_D \quad (5.23\text{d})$$

Equations (5.20a)–(5.20d) for  $u_1$  is called the nonhomogeneous equation (NHE) problem, and for  $u_2$  (5.21) is a parabolic initial condition (PIC) problem. The problems for  $u_3, u_4$  do not conform to any of the four classes defined earlier. To achieve this they are further split into a time independent (steady state) part and a time dependent (transient) part as

$$u_3(x, t) = v_3(x) + w_3(x, t) \quad (5.24\text{a})$$

$$u_4(x, t) = v_4(x) + w_4(x, t) \quad (5.24\text{b})$$

Substituting (5.24a) in (5.22a)–(5.22d), we have

$$Lv_3(x) + Lw_3 = \frac{\partial w_3}{\partial t} \quad (5.25\text{a})$$

subject to

$$w_3(t=0) = -v_3(x) \quad (5.25\text{b})$$

$$n \cdot \nabla v_3 + n \cdot \nabla w_3 + hv_3 + hw_3 = hu_R \quad \text{on } S_R \quad (5.25\text{c})$$

$$v_3(x) + w_3(x, t) = 0 \quad \text{on } S_D \quad (5.25\text{d})$$

We seek  $v_3, w_3$  such that

$$Lv_3 = 0 \quad \text{in } V \quad (5.26\text{a})$$

subject to

$$n \cdot \nabla v_3 + hv_3 = hu_R \quad \text{on } S_R \quad (5.26\text{b})$$

$$v_3 = 0 \quad \text{on } S_D \quad (5.26\text{c})$$

and

$$Lw_3 = \frac{\partial w_3}{\partial t} \quad \text{in } V \quad (5.27\text{a})$$

subject to

$$w_3(t=0) = -v_3(x) \quad \text{in } V \quad (5.27\text{b})$$

$$n \cdot \nabla w_3 + hw_3 = 0 \quad \text{on } S_R \quad (5.27\text{c})$$

$$w_3 = 0 \quad \text{on } S_D \quad (5.27\text{d})$$

The equation governing  $v_3$  is an EBC and that for  $w_3$  is a PIC. Although  $u_3$  does not fall into one of the four categories described earlier  $v_3, w_3$  do. Clearly, (5.25a) does not imply (5.26a) and (5.27a). The solution  $u_3$  that we obtain using the systems (5.26) and (5.27) is a particular solution

to (5.22). We will prove in Chapter 9 that these equations have a unique solution. Hence  $u_3$  obtained from (5.24a), where  $v_3, w_3$  are solutions of (5.26), (5.27) is the only solution. Similarly,  $v_4, w_4$  in (5.24b) are obtained from

$$Lv_4 = 0 \text{ in } V \quad (5.28a)$$

subject to

$$n \cdot \nabla v_4 + hv_4 = 0 \quad \text{on } S_R \quad (5.28b)$$

$$v_4 = u_D \quad \text{on } S_D \quad (5.28c)$$

and

$$Lw_4 = \frac{\partial w_4}{\partial t} \quad \text{in } V \quad (5.29a)$$

subject to

$$w_4(t=0) = -v_4(x) \quad \text{in } V \quad (5.29b)$$

$$n \cdot \nabla w_4 + hw_4 = 0 \quad \text{on } S_R \quad (5.29c)$$

$$w_4 = 0 \quad \text{on } S_D \quad (5.29d)$$

The problems defining  $v_4$  is an EBC and  $w_4$  is a PIC.

### Example 5.9 Decompose the parabolic problem

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}, \quad 0 < x < 1,$$

subject to  $T(x=0) = 0, T(x=1) = T_0, T(t=0) = 0$  into a PIC and an EBC. We seek the solution to the transient heat conduction equation as

$$T(x, t) = T_{ss}(x) + T^*(x, t),$$

$T_{ss}(x)$  is obtained as the solution to

$$\frac{d^2 T_{ss}}{dx^2} = 0$$

subject to

$$T_{ss}(x=0) = 0, \quad T_{ss}(x=1) = T_0$$

and  $T^*(x, t)$  as the solution to

$$\frac{\partial T^*}{\partial t} = \frac{\partial^2 T^*}{\partial x^2},$$

subject to

$$T^*(t=0) = -T_{ss}, \quad T^*(x=0) = 0, \quad T^*(x=1) = 0$$

The problem for  $T_{ss}(x)$  is an EBC and for  $T^*(x, t)$  is a PIC. This decomposition is equivalent to physically obtaining the transient solution  $T(x, t)$  as a steady state part  $T_{ss}(x)$  and as a perturbation from the steady state  $T^*(x, t)$  as done in transport phenomena.

Hyperbolic equations are amenable to the same treatment as parabolic and elliptic systems. They are less frequently encountered by chemical engineers and we leave it to the student to apply the techniques discussed here to hyperbolic equations.

The remainder of this section will be devoted to solving the four classes of problems using two analytical techniques: (a) **separation of variables**, and (b) **Green's function**. We will illustrate the former primarily in the context of solving EBCs, PICs, and HICs, and the latter for solving NHEs. It must be remembered that Green's function can be used to solve EBCs, PICs, and HICs and the separation of variables for solving NHEs. The separation of variables technique is based on seeking the solution as a series expansion in terms of eigenfunctions. This is very similar to the approach adopted in Chapter 4 where the solution vector was expanded in terms of eigenvectors. The Green's function, on the other hand, is the inverse of the differential operator and is analogous to the matrix inverse  $A^{-1}$  in finite dimensional spaces.

## PROBLEMS

- 1. Show that the most general hyperbolic problem**

$$Lu = f(x, y, z)$$

subject to

$$u(t = 0) = u_0, \quad \frac{\partial u}{\partial t} (t = 0) = u_1$$

$$u = u_D \quad \text{on } \partial V_D$$

$$n \cdot \nabla u + hu = hu_R \quad \text{on } \partial V_R$$

can be written as the sum of a nonhomogeneous equation, four HICs, and two EBC's, where

$$L = \left( \frac{\partial^2}{\partial t^2} - \nabla^2 \right)$$

- 2. Apply the principle of superposition to write**

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}$$

subject to

$$T(t = 0, x) = 1 + \sin(\pi x), \quad T(t, x = 0) = 1, \quad T(t, x = 1) = 0$$

as a combination of an EBC and a PIC. What is the physical significance of each of the problems.

- 3. Classify:**

$$(i) \frac{\partial u}{\partial t} + \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2} + \sin x$$

$$(ii) \frac{\partial u}{\partial x} + \frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial z^2} + \frac{\partial u}{\partial y} = 0$$

**4.** Classify the following equations as parabolic, elliptic or hyperbolic:

$$(i) \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

$$(ii) \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

$$(iii) \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

**5.** Describe the boundary conditions at the following surfaces:

(a) The surface is insulated.

(b) Steam at atmospheric pressure is condensing on the surface.

(c) Fluid is flowing on a flat plate, whose interface is exposed to the atmosphere.

(d) The normal and tangential velocity components at a rigid wall  $x = 0$ .

**6.** Discuss how you would apply linearity and superposition to solve

$$(i) \frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}, \quad 0 < x < 1, \quad 0 < y < 1, \quad t > 0$$

subject to

$$T(x, y, 0) = \sin \pi x \sin \pi y$$

$$T(0, y, t) = 0, \quad T(1, y, t) = y(1 - y)$$

$$T(x, 0, t) = x(1 - x), \quad T(x, 1, t) = 0$$

$$(ii) \frac{\partial C}{\partial z} = \frac{\partial^2 C}{\partial y^2} \quad 0 < y < 1, \quad z > 0$$

$$C(y, 0) = C_0 y, \quad \frac{\partial C}{\partial y}(0, z) = 0, \quad C(1, z) = C_0$$

Classify these equations.

**7.** The potential function  $\phi$ , and the stream function  $\psi$  are defined for a two-dimensional irrotational flow-field. They satisfy  $\nabla^2 \phi = 0$ ,  $\nabla^2 \psi = 0$ . Yet the solutions ' $\phi$ ', and ' $\psi$ ' are different. Explain qualitatively how this can happen.

**8.** The second order partial differential equation in two independent variables  $x_1, x_2$ , viz.

$$A \frac{\partial^2 u}{\partial x_1^2} + 2B \frac{\partial^2 u}{\partial x_1 \partial x_2} + C \frac{\partial^2 u}{\partial x_2^2} = f(u, \dots)$$

is obtained in terms of  $(B^2 - AC)$ . Prove that this is consistent with the classification discussed in this chapter in terms of the eigenvalues.

**REFERENCES**

- Aris, R., *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Clarendon Press, Oxford (1975).
- Berg, P.W. and McGregor, J.L., *Elementary Partial Differential Equations*, Holden Day, San Francisco (1964).
- Bird, R.B., Stewart, W.E. and Lightfoot, E.N., *Transport Phenomena*, Wiley, New York (International Edition), (1960).
- Fahien, R., *Transport Phenomena*, McGraw-Hill, New York (1980).
- Kaplan, W., *Operational Methods for Linear Systems*, Addison-Wesley, Reading, Mass. (1962).
- Kersten, R.D., *Engineering Differential Systems*, McGraw-Hill, New York (1969).
- Kreyszig, E., *Advanced Engineering Mathematics*, Wiley, New York (1982).
- Weinberger, H.F., *A First Course in Partial Differential Equations: With complex variables and transform methods*, Wiley, New York (1965).

# 6

## Sturm-Louiville Theory

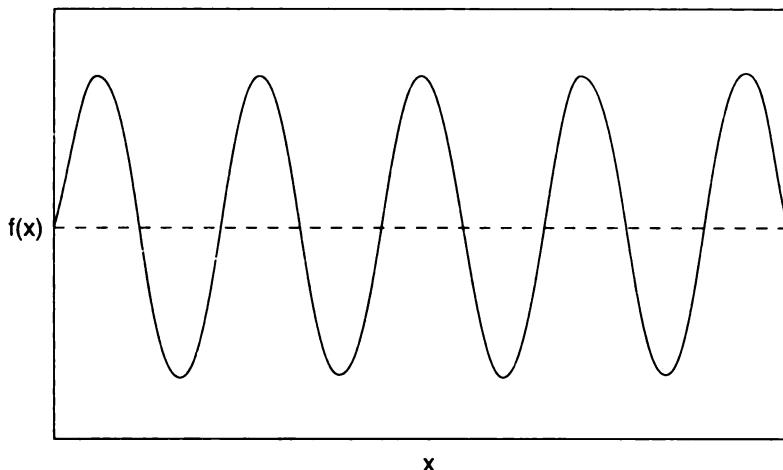
---

In this chapter we introduce the basic concepts required for solving partial differential equations. The concepts of metric, norm and inner-product defined on  $\mathbb{R}^n$  in Chapter 3 are extended to infinite dimensional spaces. The concept of a complete space is defined. The notion of the adjoint operator for a differential operator and the properties of eigenvalues and eigenfunctions of self-adjoint systems are established. The close similarity in the results obtained here and those in Chapter 3 is brought out. The most frequently encountered eigenvalue problems in the different coordinate systems are discussed in detail. We also see how the method of solution presented in Chapter 4 generates the concept of a Fourier series. This is discussed as a general concept and is not restricted to the Fourier series of periodic functions. We will discuss the applications of this in detail in Chapter 7 when we solve partial differential equations by the method of separation of variables. We conclude this chapter by defining the Rayleigh's quotient for a differential operator. We begin by seeing how two-point boundary value problems automatically give rise to the concept of an infinite dimensional space.

### 6.1 INFINITE DIMENSIONAL SPACES

Differential operators transform, i.e. map functions of one or more variables to other functions as in Examples 5.6 and 5.7. The functions that the operator acts on are defined on a domain. For the sake of concreteness, we consider the functions to be defined on the one-dimensional closed interval  $[a, b]$ . Functions defined on  $[a, b]$  can be thought of as infinite dimensional vectors. Any function on this interval can be represented approximately by a vector whose coordinates, i.e. elements, are the function values at a predetermined number of points (say, 10 points). Such a representation is only approximate and is non-unique. More than one function can assume the same values at these 10 points (Fig. 6.1). Increasing the number of points from 10 to 20 makes the approximation more accurate. However, even now there can be many functions which are represented by the same 20-dimensional vector. As we increase  $n$  to infinity (assuming that the points are dense in  $[a, b]$  and the function is continuous), we have an infinite dimensional vector and the function is represented accurately and uniquely. This is only to be expected as the function is defined on an infinity of points in  $[a, b]$ . A function defined on  $[a, b]$  can be hence viewed as an infinite dimensional vector. The coordinates of this vector represent the function values in the interval  $[a, b]$ .

The intuitive concept of dimension introduced here is an extension of its definition on a vector space. In Chapter 2 we identified the dimension of a vector space as being the number of coordinates of the vectors in that space. It must be recalled that the formal definition of dimension is based on the concept of linear independence. Hence to prove that a space is infinite dimensional



**Fig. 6.1** Graph of two functions which are equal to each other at a finite number (10) of points.

it is essential to establish that the basis contains an infinite number of elements. We will not establish this formally in this text.

In the earlier section we defined the vector space  $\mathbb{R}^n$  as the space containing vectors with  $n$  real coordinates closed under the operations of vector addition and scalar multiplication.

We would now like to formally define the relevant space we will be working on in this section. The set of functions which are continuous in the open interval  $(a, b)$  is denoted by  $C(a, b)$ . This space is closed under the operations of addition and scalar multiplication. Similarly,  $C^2(a, b)$  represents the set of smooth functions which are twice differentiable and have continuous second derivatives in the open interval  $(a, b)$ . It would appear that the space of our interest should be  $C^2(a, b)$  as most of our differential operators are second order. Therefore, elements in the domain space of such operators must be at least twice differentiable. This space, however, is not complete (see Stakgold, 1968). Before we understand the concept of completeness of a space, we introduce the definitions of metric, norm and inner-product in an infinite-dimensional space.

### 6.1.1 Metric, Norm and Inner-product in an Infinite Dimensional Space

In Chapter 2 we introduced these concepts for a general finite dimensional vector space. The definitions introduced were for an  $n$ -dimensional space and we now extend these to infinite dimensional spaces. For specificity we restrict ourselves to functions of one variable ‘ $x$ ’ defined on the interval  $[a, b]$ . We restrict ourselves to  $C(a, b)$ .

**(i) Metric.** Let  $f(x), g(x)$  be two elements in  $C(a, b)$ . The metric or distance between these is denoted as  $d(f(x), g(x))$ . A metric can be defined in many ways as seen in  $\mathbb{R}^n$ . Each of these must satisfy axioms D1–D4 in Chapter 2.

Three possible candidates for a metric are:

$$d_2(f(x), g(x)) = \left( \int_a^b (f(x) - g(x))^2 dx \right)^{1/2} \quad (6.1a)$$

$$d_1(f(x), g(x)) = \int_a^b |f(x) - g(x)| dx \quad (6.1b)$$

$$d_\infty(f(x), g(x)) = \max_{x \in [a,b]} |f(x) - g(x)| \quad (6.1c)$$

These three metrics satisfy axioms D1–D4 (Chapter 2) and are valid candidates for a metric. These are analogous to definitions on a vector space 2.6(a)–2.6(c). As already seen, the metrics need not be defined on a linear space, because axioms D1–D4 do not involve scalar multiplication or vector addition, see Stakgold (1968), Naylor and Sell (1971).

**(ii) Norm.** The norm or length of an element is defined in  $C(a, b)$  so as to satisfy axioms N1-N4 in Chapter 2. Three possible candidates for the norm are:

$$(i) \|f(x)\|_2 = \left( \int_a^b f^2(x) dx \right)^{1/2} \quad (6.2a)$$

$$(ii) \|f(x)\|_1 = \int_a^b |f(x)| dx \quad (6.2b)$$

$$(iii) \|f(x)\|_\infty = \max_{x \in [a,b]} |f(x)| \quad (6.2c)$$

Every norm has to be necessarily defined on a linear space or a vector space (see Chapter 2) as the axioms N1–N4 in Chapter 2 involve concepts of scalar multiplication and vector addition. Every linear space on which a norm is defined is automatically a metric space, as a metric is defined on it. The metric generated by the norm is defined as

$$d(f(x), g(x)) = \|f(x) - g(x)\| \quad (6.3)$$

**(iii) Inner-product.** The inner-product is a measure of the angle between two elements in a function space. A valid candidate for the inner-product of  $f(x)$  with  $g(x)$  defined on the space of real-valued functions is

$$\langle f(x), g(x) \rangle = \int_a^b f(x)g(x) dx \quad (6.4a)$$

In the space of complex-valued functions, this definition is extended as

$$\langle f(x), g(x) \rangle = \int_a^b \bar{f}(x)g(x) dx \quad (6.4b)$$

It can be verified that these definitions satisfy, respectively, axioms I1–I4 and IC1–IC4 given in Chapter 2. Besides, they are an extension of the inner-product defined there on  $\mathbb{R}^n$  or  $\mathbb{C}^n$ . The definition of an inner-product on a linear space generates a norm and a metric from

$$\begin{aligned} d_2(f(x), g(x)) &= \|f(x) - g(x)\|_2 \\ &= \langle f(x) - g(x), f(x) - g(x) \rangle^{1/2} \end{aligned} \quad (6.5)$$

Therefore, an inner-product space is automatically a normed linear space and a metric space. Extending the definition given in Chapter 2, two functions  $f(x)$ ,  $g(x)$  are said to be orthogonal when

$$\langle f(x), g(x) \rangle = 0 \quad (6.6)$$

### 6.1.2 Completeness

The concept of metric in a vector space introduces the notion of convergence. Consider a sequence of functions  $\{u_i(x)\}$  defined in the interval  $[a, b]$ . The sequence  $u_i(x)$  is said to converge to an element  $u^*(x)$  if

$$d(u_n(x), u^*(x)) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

In finite dimensional vector spaces  $\mathbf{R}^n$ , each element (coordinate) of the vector was a real number. Sequences of vectors therefore converged (if at all) to a vector whose elements were real numbers. Hence the converged vector also belongs to  $\mathbf{R}^n$ . This is not the case for infinite dimensional spaces.

A sequence of elements  $\{u_i(x)\}$  is said to be Cauchy when  $d(u_m, u_n) < \epsilon$  for  $m, n > P$ , where  $P$  is a large integer. If Cauchy sequences exist on a set or space such that they converge to  $u^*(x)$ , and if  $u^*(x)$  does not belong to the original set or space, we call the set or space incomplete. If all Cauchy sequences in a space converge to elements in that space, we have a complete space.

This idea can be clearly understood by considering the set of rational numbers. This is an incomplete set. Sequences of rational numbers exist which converge to irrational numbers. This set is made complete by including the set of irrational numbers. This yields the set of real numbers  $\mathbf{R}$ , which is complete. The set of irrational numbers is a hypothetical set which is added to the rational number set to enable every possible sequence of rational numbers to converge to an element of that set. This construction is the process of completion of the set of rational numbers.

As an example, consider the sequence generated by the iterative scheme

$$x_{n+1} = 4 - \frac{1}{x_n}$$

Starting from the initial point  $x_0 = 1$ , we generate the sequence of rational numbers (3, 11/3, 41/11, 153/41...). This sequence of rational numbers converges as  $n \rightarrow \infty$ , to the irrational number  $2 + \sqrt{3}$ , which is a root of the quadratic  $x^2 - 4x + 1 = 0$ . As a second example, consider the series  $S_n = \sum_{k=1}^n \frac{1}{k!}$ . It generates elements  $\{S_i\}$  which are rational numbers for different  $n$ . The series converges to the irrational number 'e' as  $n \rightarrow \infty$ . We now illustrate by means of an example why the space of twice differentiable continuous functions  $C^2(a, b)$  is not a complete space.

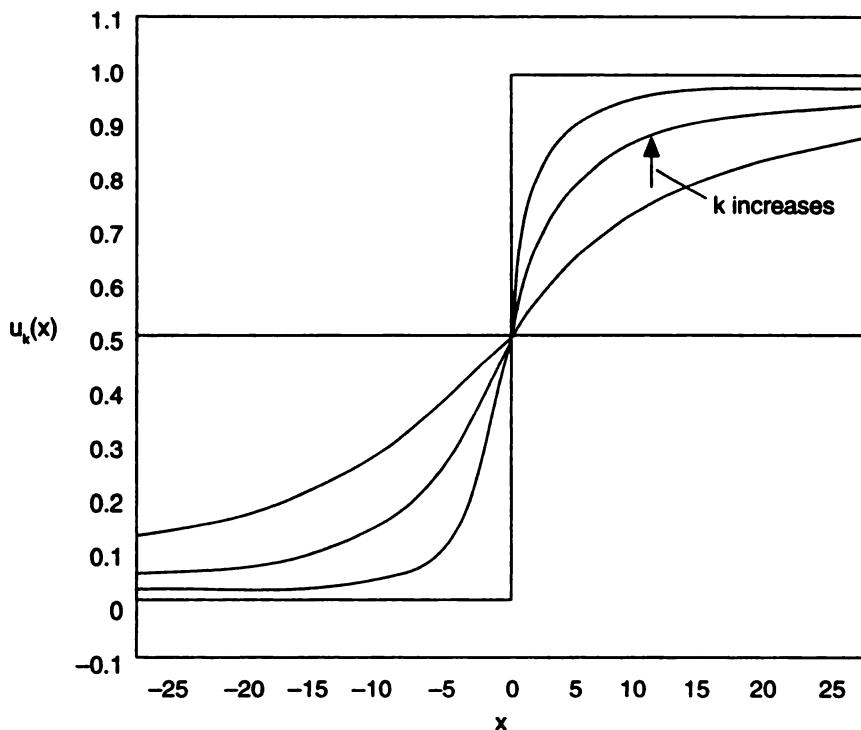
Consider the sequence of functions

$$u_k(x) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1} kx$$

defined in the interval  $-\infty < x < \infty$ , for all integer values  $k$ . The range of the functions  $u_k(x)$  is  $(0, 1)$ , see Fig. 6.2. As  $k \rightarrow \infty$ , the sequence of continuous functions converges to the discontinuous function

$$\begin{aligned} u^*(x) &= 0, & -\infty < x < 0 \\ &= 1, & 0 < x < \infty \end{aligned}$$

Thus the space of functions  $C^2(a, b)$  is not complete in the  $d_2$  metric. The differential operators we normally encounter in chemical engineering and with which we will be primarily concerned in this book are second order differential operators. The functions  $u(x)$  must therefore possess continuous second derivatives and belong to  $C^2(a, b)$ . To overcome the limitation of the incompleteness of  $C^2(a, b)$ , we will be concerned primarily with the space  $L^2(a, b)$ . This is a hypothetical space which consists of all functions in  $C^2(a, b)$  and other functions included to yield a complete space.



**Fig. 6.2** The plot of the sequence  $u_k(x)$  generated for various  $k$ .

This assures us that sequences and series in this space converge to elements in this space in a suitable metric.

The notion of completeness is an important concept in infinite dimensional spaces. In Chapter 7 we construct a solution as a linear combination of an infinite number of eigenfunctions. When we work in  $L^2(a, b)$ , we are assured that the sum, i.e. the infinite series representing the solution, will converge to an element in that space.

The integral in (6.1a) is defined now in the Lebesgue sense; hence the notation  $L^2$ .<sup>†</sup> The superscript 2 denotes that the metric we are working with is the  $d_2$  metric. The construction of  $L^2(a, b)$  is analogous to the construction of the real line  $\mathbb{R}$  from the set of rational numbers and the role played by the set of rational numbers is analogous to the role of  $C^2(a, b)$ . A good introduction to the theory of Lebesgue integration can be found in Ramkrishna (1985).

A complete normed linear space is called a **Banach space**. A complete inner-product space is called an **Hilbert space**. The completeness of these spaces here is in the natural metric, i.e. the one generated by the norm and the inner-product respectively (see Stakgold, 1968).

---

<sup>†</sup>The integral  $\int_a^b f(x) dx$  can be defined in a Riemannian sense or in a Lebesgue sense. For an engineering student the area under the curve  $f(x)$  is obtained, in the Riemannian sense, by dividing the  $x$ -axis into small divisions  $dx$  and in the Lebesgue sense, by dividing the  $y$ -axis into small divisions  $dy$ . The Lebesgue integration enlarges the class of "integrable" functions and includes those that are integrable in the Riemannian sense.

## 6.2 EIGENVALUE PROBLEMS

A common technique used to solve partial differential equations is the method of separation of variables (see Kreyszig, 1982). The basis of the method and its application are discussed in detail in Chapter 7. Here, the solution which is a function of many independent variables is sought as the product of different functions, such that each function depends only on one independent variable. For example, the solution  $u(x, y, t)$  is sought in the form

$$u(x, y, t) = X(x)Y(y)T(t)$$

where  $X$  is independent of  $y, t$ , and  $Y$  is independent of  $x, t$ . These functions  $X(x)$ ,  $Y(y)$ , etc. of a single variable are determined as the solutions of ordinary differential equations of the form

$$a_0(x) \frac{d^2 u}{dx^2} + a_1(x) \frac{du}{dx} + a_2(x)u = -\lambda a_3(x)u \quad \text{in } a < x < b \quad (6.7a)$$

The homogeneous equation (6.7a) is subject to appropriate boundary conditions. For the sake of concreteness and simplicity, we take these to be homogeneous Dirichlet conditions. For the two-point boundary value problem, this implies

$$u(x = a) = 0, \quad u(x = b) = 0 \quad (6.7b)$$

Relation (6.7a) is a linear homogeneous system. Here,  $\lambda$  is a scalar and  $a_0(x), a_1(x), a_2(x), a_3(x)$  are real functions of  $x$ . It is a two-point boundary value problem (BVP) as the conditions (6.7b) on  $u$  are specified at two distinct points  $x = a, x = b$ . This is in contrast with the initial value problems (IVP) we came across in Chapter 4 where all conditions were specified at one point of the independent variable.

Problems of the form of (6.7) occur frequently (see Garabedian, 1964). The functions  $a_i(x)$  take on specific forms depending on the coordinate system being used. In rectangular cartesian coordinates, equation (6.7a) is typically of the form

$$\frac{d^2 u}{dx^2} + \lambda u = 0 \quad (6.8a)$$

This corresponds to the specific case (6.7a), where  $a_0(x) = 1, a_3(x) = 1$  and  $a_1(x) = a_2(x) = 0$ . In cylindrical coordinates we encounter equations of the form

$$\frac{d^2 u}{dr^2} + \frac{1}{r} \frac{du}{dr} + \lambda u = 0 \quad (6.8b)$$

Comparing this equations with (6.7a), we identify  $r$  with  $x, a_0 = 1, a_3 = 1, a_1 = 1/r$  and  $a_2 = 0$ . In spherical coordinates the typical ordinary differential equations are of the form

$$\frac{d^2 u}{dr^2} + \frac{2}{r} \frac{du}{dr} + \lambda u = 0 \quad (6.8c)$$

Here again,  $r = x, a_0 = 1, a_1 = 2/r, a_2 = 0$  and  $a_3 = 1$ . The origin of the different equations (6.8a)–(6.8c) will become clear in Chapter 7. Equation (6.7a) is a linear homogeneous equation subject to linear homogeneous boundary conditions (6.7b). It admits  $u = 0$ , the trivial solution for all  $\lambda$ . It can admit nontrivial solutions for discrete values of  $\lambda$ . These values of  $\lambda$  are the eigenvalues and the corresponding nonzero solutions are the eigenfunctions of (6.7). The system (6.7) is analogous to (3.2) and is said to define an eigenvalue problem. It is not necessary for an eigenvalue problem

to have Dirichlet boundary conditions. What is necessary is that boundary conditions and the equation should be linear and homogeneous. This will permit the trivial solution for all  $\lambda$  and a nontrivial solution for a discrete set of  $\lambda$ .

The similarity of (6.7), (6.8) with (3.2) can best be seen by recasting them in a more compact and general form as

$$Lu = -\lambda a_3(x)u \quad (6.9a)$$

subject to

$$B_1u = 0, \quad B_2u = 0, \quad (6.9b)$$

Here  $L$  represents a linear differential operator and  $B_1, B_2$  represent linear boundary operators. Comparing (6.9a) with (6.7a), we identify  $L$  with

$$\left( a_0 \frac{d^2}{dx^2} + a_1 \frac{d}{dx} + a_2 \right)$$

and the boundary operators with the Dirichlet boundary conditions. Equation (6.9a) is structurally similar to the matrix eigenvalue problem (3.5) given by

$$Au = \lambda u$$

The introduction of the notion of the operator has enabled us to establish this similarity. We can now determine properties of the eigenvalues and eigenfunctions of the differential operator  $L$  proceeding along the same lines as we did for the matrix  $A$  in Chapter 3. To do this we need to define and determine the adjoint operator.

### 6.2.1 Adjoint Operators

We have already related the properties of the matrix to its adjoint. To achieve a similar objective, we now define the adjoint operator  $L^*$  associated with the original operator  $L$ .  $L^*$  operates on the adjoint function  $v$ . We associate this operator with the adjoint boundary conditions denoted as  $B_1^*$ ,  $B_2^*$  (see Kaplan, 1962).

Using Definition 3.3,  $L^*$  is defined such that it satisfies

$$\langle v, Lu \rangle = \langle L^*v, u \rangle \quad (6.10)$$

The adjoint operator  $L^*$  depends upon the definition of the inner-product used. Let us define

$$\langle u, v \rangle = \int_a^b u(x) v(x) dx \quad (6.11)$$

For the operator  $L$  defined in (6.7), this implies

$$\langle v, Lu \rangle = \int_a^b v(x) (a_0(x)u''(x) + a_1(x)u'(x) + a_2(x)u(x)) dx \quad (6.12)$$

Here the prime denotes the derivative with respect to  $x$ . To obtain  $L^*$  from (6.10), we need to write the term inside the integral in (6.12) as a product of  $u$  with some operator acting on  $v$ . Hence the derivatives on  $u$  have to be removed. We achieve this by carrying out an integration by parts. This yields

$$\begin{aligned}
\langle v, Lu \rangle &= v(x)a_0(x)u'(x) \Big|_a^b - \int_a^b (v(x)a_0(x))' u'(x) dx + a_1(x)v(x)u(x) \Big|_a^b \\
&\quad - \int_a^b (a_1(x)v(x))' u(x) dx + \int_a^b a_2(x)v(x)u(x) dx \\
&= [v(x)a_0(x)u'(x) - (v(x)a_0(x))' u(x) + a_1(x)v(x)u(x)]_a^b \\
&\quad + \int_a^b [(a_0(x)v(x))'' - (a_1(x)v(x))' + a_2(x)v(x)] u(x) dx \\
&= \langle L^*v, u \rangle + J(u, v)
\end{aligned} \tag{6.13}$$

where

$$\begin{aligned}
L^*v &= (a_0(x)v(x))'' - (a_1(x)v(x))' + (a_2(x)v(x)) \\
&= a_0(x) \frac{d^2}{dx^2}v(x) + (2a'_0(x) - a_1(x)) \frac{dv}{dx} + (a''_0(x) - a'_1(x) + a_2(x))v
\end{aligned} \tag{6.14a}$$

and the bilinear concomitant  $J(u, v)$  containing the boundary terms is defined as

$$J(u, v) = (v(x)a_0(x)u'(x) - (v(x)a_0(x))' u(x) + a_1(x)v(x)u(x)) \Big|_a^b \tag{6.14b}$$

To reduce (6.13) to (6.10), we have to set the bilinear concomitant to zero. This yields the adjoint boundary conditions  $B_1^*$ ,  $B_2^*$  on  $v$ . To summarise, the adjoint  $L^*$  for  $L$  corresponding to (6.1) is given by (6.14a), and the adjoint boundary conditions  $B^*$  are obtained from (6.14b) by setting

$$J(u, v) = 0 \tag{6.15}$$

An operator  $L$  is said to be self-adjoint if  $L^* = L$ ,  $B^* = B$ . This definition is similar to the definitions for the matrix operator, where for self-adjointness we needed  $A^* = A$ . The differential operator  $L$  is always associated with a boundary operator  $B$ . For self-adjointness it is necessary for both  $L$  and  $B$  to be equal to their respective adjoints. (Some other authors define  $L$  to be self-adjoint when  $L = L^*$  without imposing  $B = B^*$ .)

Consider the specific case of  $L$  in (6.7), where  $a_1(x) = a'_0(x)$ . This implies that  $a'_1(x) = a''_0(x)$ . Under these conditions, (6.14a) yields

$$L^*v = \left( a_0(x) \frac{d^2}{dx^2} + a_1(x) \frac{d}{dx} + a_2(x) \right) v = Lv \tag{6.16}$$

This does not imply that  $L$  is self-adjoint as it is likely that  $B^* \neq B$ . Before we discuss the determination of the adjoint conditions  $B^*$ , we deal with an important transformation. This allows us to transform any  $L$  defined as in (6.7a) to a modified form  $L_M$ , where  $L_M = L_M^*$ . From (6.16), we note that  $L = L^*$  when  $L$  is of the form

$$L = \left[ \frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u \right] \tag{6.17}$$

Hence we rewrite equation (6.7a) as

$$\frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x)u + \lambda \omega(x)u = 0 \tag{6.18}$$

Expanding the first term in (6.18), using the product rule of differentiation, and comparing coefficients of (6.18) with (6.7a), we obtain

$$\left. \begin{aligned} p(x) &= \exp \left[ \int \frac{a_1(x) dx}{a_0(x)} \right] \\ q(x) &= p(x) \frac{a_2(x)}{a_0(x)} \\ w(x) &= p(x) \frac{a_3(x)}{a_0(x)} \end{aligned} \right\} \quad (6.19)$$

The eigenvalue problem (6.18) is called a Sturm-Louiville problem, see Weinberger (1965); Courant and Hilbert (1975). The theory which describes the properties of self-adjoint systems of the form (6.18) is called the Sturm-Louiville theory. The transformations (6.19) allow us to recast (6.7), where  $a_0(x) \neq 0$  in the form (6.17). The linear operator occurring in (6.18) is such that  $L = L^*$ . This follows from (6.16). We use the definition of the adjoint operator to demonstrate this again and obtain the bilinear concomitant for the operator in (6.18).

$$\begin{aligned} \langle v, Lu \rangle &= \int_a^b v(x) (p(x)u'(x))' + v(x)q(x)u(x) \\ &= (v(x)p(x)u'(x) - v'(x)p(x)u(x)) \Big|_a^b + \int_a^b u(x) ((p(x)v'(x))' + q(x)v(x)) dx \end{aligned} \quad (6.20)$$

Clearly,  $L^* v = \frac{d}{dx} \left[ p(x) \frac{du}{dx} \right] + q(x)v$ , and hence  $L^* = L$  as expected. The bilinear concomitant  $J(u, v)$  is

$$J(u, v) = [v(x)p(x)u'(x) - v'(x)p(x)u(x)] \Big|_a^b \quad (6.21)$$

The adjoint boundary conditions  $B^*$  are obtained from

$$J(u, v) = 0 \quad (6.22)$$

We introduce the conditions  $Bu = 0$  in (6.22) and obtain the boundary conditions on  $v$  as  $B^*v = 0$ . The criterion for self-adjointness now reduces to checking only for  $B = B^*$ , as  $L = L^*$  by the transformation (6.19).

The adjoint  $B^*$  is always determined assuming the boundary conditions are homogeneous. If the original boundary conditions are nonhomogeneous, we render them homogeneous and then determine  $B^*$ . Consider the specific case where the boundary conditions on  $u$  are Dirichlet, i.e.,

$$B_1 u = u(a) = 0, \quad B_2 u = u(b) = 0 \quad (6.23a)$$

Inserting (6.23a) in (6.21), we get

$$v(b) p(b) u'(b) - v(a) p(a) u'(a) = 0$$

This is satisfied if and only if (assuming  $p(b), p(a)$  are not equal to zero)

$$B_1^* v = v(a) = 0, \quad B_2^* v = v(b) = 0 \quad (6.23b)$$

Thus operator (6.7) subject to (6.23a) is self-adjoint.

**Example 6.1** Determine  $L^*$ ,  $B^*$  for

$$Lu = \frac{d^2u}{dx^2} - \frac{du}{dx} \quad \text{in } 0 < x < 1$$

subject to  $u(0) = 0$ ,  $u(1) = 2$ .

Since the boundary conditions are nonhomogeneous, we render them homogeneous for the evaluation of  $L^*$ ,  $B^*$ . Thus we have, for  $B$ ,

$$u(0) = u(1) = 0$$

$$\begin{aligned} \langle v, Lu \rangle &= \int_0^1 (v(x)u''(x) - v(x)u'(x)) dx \\ &= [v(x)u'(x) - u(x)v'(x) - u(x)v(x)] \Big|_0^1 + \int_0^1 u(x) [v''(x) + v'(x)] dx \end{aligned}$$

These yield

$$L^* = \left( \frac{d^2}{dx^2} + \frac{d}{dx} \right)$$

Using the boundary conditions on  $u$ , the bilinear concomitant vanishes when

$$v(1)u'(1) - v(0)u'(0) = 0$$

This results in Dirichlet conditions on  $v$  at  $x = 0$  and  $x = 1$ . Thus  $B^*$  is

$$v(0) = v(1) = 0$$

In this problem,  $L \neq L^*$ , but  $B = B^*$ . Consequently, the operator is not self-adjoint.

**Example 6.2** Convert the operator  $L$  in the earlier example to self-adjoint form. Is the system self-adjoint?

$$Lu = \frac{d^2u}{dx^2} - \frac{du}{dx}$$

We identify

$$a_0(x) = 1, a_1(x) = -1, a_2(x) = 0$$

Using (6.19), we obtain

$$p(x) = e^{-x}, \quad q(x) = 0$$

The operator  $Lu$  is modified to  $L_Mu$ , where  $L_Mu = \frac{d}{dx} \left( e^{-x} \frac{du}{dx} \right)$ . This renders  $L_M = L_M^*$  and  $B^* = B$ . The system in Example 6.1, which was not self-adjoint, has been rendered self-adjoint by the transformation (6.19).

**Example 6.3** Determine the adjoint operator and boundary conditions for

$$Lu = \frac{d^2u}{dx^2} \quad \text{in } 0 < x < 1$$

subject to  $u(0) = 2u'(1)$ ,  $u(1) = 0$ .

The boundary conditions are homogeneous. From (6.14a), with  $a_0(x) = 1$ ,  $a_1(x) = a_2(x) = 0$ , we have

$$L^* v = \frac{d^2 v}{dx^2}$$

Thus,  $L^* = L$ .

From (6.14b),

$$v(x) u'(x) - v'(x) u(x) \Big|_0^1 = 0$$

$$v(1)u'(1) - v'(1)u(1) - v(0)u'(0) + v'(0)u(0) = 0$$

Substituting from  $Bu = 0$ , we get

$$v(1)u'(1) - v(0)u'(0) + 2v'(0) = 0$$

The adjoint conditions  $B^*v$  are given by

$$v(0) = 0, \quad v(1) + 2v'(0) = 0$$

Clearly,  $B^* \neq B$ , and the operator is again not self-adjoint. We cannot render this system self-adjoint using (6.19). The transformation in (6.19) can be used to render only  $L = L^*$ , and not  $B = B^*$ . In Chapter 3 we established many theorems on the eigenvalues and eigenvectors of a matrix operator. The Sturm-Louiville theory establishes analogous results for linear differential operators. The theorems in Chapter 3 and those in this chapter have a close correspondence. This has been made possible only because we have generalised the various concepts, and put them under one umbrella using the notion of the linear operator and vector spaces. The eigenvalue problem we consider is of the form

$$\frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + q(x) u = -\lambda \omega(x) u, \quad 0 < x < 1 \quad (6.24a)$$

subject to

$$u(0) = 0 \quad (6.24b)$$

$$u(1) = 0 \quad (6.24c)$$

In what follows, we assume that  $p(x) \geq 0$ ,  $q(x) \leq 0$ ,  $\omega(x) > 0$ , and they are all real. These conditions are satisfied while solving problems in the three natural coordinate systems—rectangular, cylindrical, spherical, as we shall discuss later.

For the sake of simplicity and concreteness we consider only Dirichlet boundary conditions. Analogous results can be established for other boundary operators as long as the system is self-adjoint, see Weinberger (1965), Courant and Hilbert (1975).

For the present we discuss the general properties of equations of the form (6.24). The discussion is along general lines. Consequently, the emphasis is not on specific equations, where  $p(x)$  etc. take on specific forms.

**Theorem 6.1** The eigenvalues of a self-adjoint operator are real.

*Proof.* The eigenvalue problem (6.24) can be recast as

$$Lu = -\lambda \omega(x) u \quad (6.25)$$

where  $L$  is defined in (6.17). Taking the inner-product of (6.25) with  $u$ ,

$$\langle u, Lu \rangle = -\langle u, \lambda \omega(x) u \rangle$$

From the definition of adjoint this yields

$$\langle L^*u, u \rangle = -\langle u, \lambda\omega(x)u \rangle$$

As  $L$  is self-adjoint,  $L^* = L$ , and we have

$$\langle Lu, u \rangle = -\langle u, \lambda\omega(x)u \rangle$$

or

$$-\langle \lambda\omega(x)u, u \rangle + \langle u, \lambda\omega(x)u \rangle = 0 \quad (6.26)$$

The inner-product one has to use to establish this result is the complex inner-product, i.e. (6.4b), as we have to prove that  $\lambda$  is real. Since  $\lambda$  is a scalar,

$$\langle \lambda\omega(x)u, u \rangle = \bar{\lambda} \langle \omega(x)u, u \rangle$$

$$\langle u, \lambda\omega(x)u \rangle = \lambda \langle u(x), \omega(x)u \rangle$$

Equation (6.26) now reduces to

$$(\lambda - \bar{\lambda}) \langle \omega(x)u, u \rangle = 0 \quad (6.27)$$

For real  $\omega(x) > 0$ , and from the definition of the inner-product  $\langle \omega(x)u, u \rangle > 0$ , (6.27) can only be satisfied if  $\lambda = \bar{\lambda}$  or  $\lambda$  is real.

**Theorem 6.2** The eigenvalues of the self-adjoint operator in (6.24) are non-negative.

*Proof.* Let  $u$  be an eigenfunction of  $L$ , corresponding to  $\lambda$ . Then

$$Lu = -\lambda\omega(x)u, \quad \langle u, Lu \rangle = -\langle u, \lambda\omega(x)u \rangle$$

or

$$\int_0^1 u \left( \frac{d}{dx} \left( p(x) \frac{du}{dx} \right) + qu \right) dx = -\lambda \int_0^1 \omega(x)u^2(x) dx$$

or

$$up(x) \frac{du}{dx} \Big|_0^1 - \int p(x) \left( \frac{du}{dx} \right)^2 + \int_0^1 qu^2 = -\lambda \int \omega(x)u^2(x) dx$$

Imposing the boundary conditions

$$u(0) = 0 = u(1)$$

we obtain

$$\lambda = \frac{\int_0^1 (p(x)u'^2 - qu^2) dx}{\int_0^1 \omega u^2 dx}$$

For  $p \geq 0$ ,  $q \leq 0$ ,  $\omega > 0$ , clearly,  $\lambda$  is always non-negative. It can be zero if a constant  $u$  can be an admissible eigenfunction (with  $q(x) = 0$ ). Before establishing the next theorem we generalise the notion of orthogonality between two functions defined in (6.6). Two functions  $u(x)$ ,  $v(x)$  are said to be orthogonal with respect to the weighting function  $\omega(x) > 0$  in  $(a, b)$  when

$$\langle u(x), \omega(x) v(x) \rangle = 0$$

or

$$\int_a^b \bar{u}(x)\omega(x)v(x) dx = 0$$

**Theorem 6.3** The eigenfunctions  $u_m, u_n$  corresponding to two distinct eigenvalues  $\lambda_m, \lambda_n$  of a self-adjoint operator are orthogonal with respect to the weighting function  $\omega(x)$ .<sup>†</sup>

*Proof.* The eigenvalue problems for  $\lambda_m, \lambda_n$  are

$$Lu_m = -\lambda_m \omega(x)u_m \quad (6.28a)$$

$$Lu_n = -\lambda_n \omega(x)u_n \quad (6.28b)$$

$$\langle u_n, Lu_m \rangle = -\langle u_n, \lambda_m \omega(x)u_m \rangle \quad (6.28c)$$

$$\langle Lu_n, u_m \rangle = -\langle \lambda_n \omega(x)u_n, u_m \rangle \quad (6.28d)$$

Subtracting (6.28c) and (6.28d), since  $L^* = L$ , and the boundary conditions on  $u_n, u_m$  are homogeneous (as we have eigenvalue problems), we get

$$0 = (\lambda_n - \lambda_m) \langle u_n, \omega(x)u_m \rangle \quad (6.29)$$

We restrict ourselves to the real inner-product here as we are assured of real eigenvalues from theorem 6.1. Clearly, for distinct  $\lambda_n, \lambda_m, \lambda_m - \lambda_n \neq 0$ . And we have

$$\langle u_n, \omega(x)u_m \rangle = 0 \quad \text{for } n \neq m$$

**Example 6.4** Verify the validity of Theorems 6.1–6.3 for

$$\frac{d^2u}{dx^2} = -\lambda u, \quad 0 < x < 1$$

subject to

$$u(0) = u(1) = 0$$

Let the eigenvalues  $\lambda$  be complex. We work with the complex inner-product

$$\langle u, v \rangle = \int_0^1 \bar{u}(x)v(x) dx$$

This yields

$$\begin{aligned} \langle u, Lu \rangle &= \int_0^1 \bar{u}(x) \frac{d^2u}{dx^2} dx \\ &= \int_0^1 \frac{d^2\bar{u}}{dx^2} u \quad (\text{integrating by parts and employing the boundary conditions}) \\ &= \langle Lu, u \rangle \end{aligned}$$

Hence

$$\langle u, \lambda u \rangle - \langle \lambda u, u \rangle = 0$$

or

$$(\lambda - \bar{\lambda}) \langle u, u \rangle = 0$$

as  $\langle u, u \rangle > 0$  for  $u \neq 0$ ,  $\lambda - \bar{\lambda} = 0$  or  $\lambda$  is real.

<sup>†</sup>For non self-adjoint systems, a biorthogonality relationship exists between the eigenfunctions set of the original system and that of the adjoint system. This is similar to what we saw in Chapter 3. We do not discuss this here as classical problems in diffusive systems; what we are primarily concerned with are self-adjoint systems.

The eigenvalues  $\lambda$  have to be non-negative. We consider three separate cases  $\lambda < 0$ ,  $\lambda = 0$ ,  $\lambda > 0$  to prove this.

Let  $\lambda < 0$ , say,  $\lambda = -\mu^2$ . The equation becomes

$$\text{Case 1: } \frac{d^2u}{dx^2} = \mu^2 u$$

The solutions to the above equation are

$$u(x) = Ae^{\mu x} + Be^{-\mu x}$$

The boundary conditions yield

$$A + B = 0, \quad Ae^{-\mu} + Be^{-\mu} = 0$$

This has only the trivial solution  $A = B = 0$ . Consequently, for  $\lambda < 0$ , we have no nontrivial solution for  $u(x)$ .

*Case 2:* If  $\lambda = 0$ ,  $u(x) = A + Bx$ . Imposing the boundary conditions further yields  $A = B = 0$ . Hence  $\lambda = 0$  is not an eigenvalue, as this also admits only the trivial solution  $u = 0$ .

*Case 3:* If  $\lambda > 0$ , let  $\lambda = \mu^2$ . The solution  $u(x)$  now is

$$u(x) = A \sin \mu x + B \cos \mu x$$

$u(0) = 0$  implies  $B = 0$ . The condition at  $x = 1$  yields  $A \sin \mu = 0$ . If we allow  $A$  to be zero we have again the trivial solution. Hence we look for  $\mu$  such that  $\sin \mu = 0$ . So,  $\mu = n\pi$  and the eigenvalues  $\lambda_n$  are

$$\lambda_n = n^2\pi^2 \text{ for } n = 1, 2, \dots, \infty$$

The system has a countable infinity of eigenvalues, as it belongs to an infinite-dimensional space, as opposed to matrices in Chapter 3 which had a finite number. The corresponding eigenfunctions are

$$u_n = A_n \sin(n\pi x)$$

Clearly, the eigenvalues are all positive here. Consider two eigenfunctions  $u_m(x)$  and  $u_n(x)$  corresponding to two distinct eigenvalues  $\lambda_m$ ,  $\lambda_n$ . Here  $\omega(x) = 1$ .

$$\begin{aligned} \langle u_m, \omega(x)u_n \rangle &= A_m A_n \int_0^1 \sin(m\pi x) \sin(n\pi x) dx \\ &= 0 \text{ for } m \neq n. \end{aligned}$$

This proves the orthogonality relation between two eigenfunctions corresponding to distinct eigenvalues (see Weinberger, 1965).

**Example 6.5** Find the eigenvalues and eigenfunctions of

$$\frac{d^2u}{dx^2} - \frac{du}{dx} + \lambda u = 0, \quad 0 < x < 1$$

subject to  $u(0) = u(1) = 0$ .

We have seen in Example 6.2 that the operator  $L$  subject to  $B$  can be converted to self-adjoint form. Hence we seek only those  $\lambda$ 's which are real. The eigenvalues can be negative, zero, or positive.

(i) Let  $\lambda < 0$ , i.e.  $\lambda = -\mu^2$ . We have

$$u'' - u' = \mu^2 u$$

Since this equation is a linear equation with constant coefficients, we seek  $u = e^{mx}$ .  $m$  is a root of the characteristic equation

$$m^2 - m - \mu^2 = 0$$

or

$$m_{1,2} = \frac{1 \pm \sqrt{(1 + 4\mu^2)}}{2}$$

This yields

$$u(x) = A \exp [(1 + \sqrt{(1 + 4\mu^2)})x/2] + B \exp [(1 - \sqrt{(1 + 4\mu^2)})x/2]$$

The boundary conditions yield

$$A + B = 0$$

$$A \exp [(1 + \sqrt{(1 + 4\mu^2)})/2] + B \exp [(1 - \sqrt{(1 + 4\mu^2)})/2] = 0$$

$$A = B = 0$$

Here again there is no negative eigenvalue. If  $\lambda = 0$ , we obtain  $u = Ae^x + B$ . Imposing the boundary conditions again, we get  $A = B = 0$ .  $\lambda = 0$  yields only the trivial solution and hence is not an eigenvalue. For  $\lambda > 0$ , we set  $\lambda = \mu^2$ . This yields

$$u(x) = A \exp [(1 + \sqrt{(1 - 4\mu^2)})x/2] + B \exp [(1 - \sqrt{(1 - 4\mu^2)})x/2]$$

The boundary conditions yield

$$A + B = 0$$

$$e^{i/2} (A \exp [\sqrt{(1 - 4\mu^2)/2}] + B \exp [-\sqrt{(1 - 4\mu^2)/2}]) = 0$$

If  $4\mu^2 \leq 1$ , then again  $A = B = 0$ , and we have only the trivial solution. For notational convenience, define

$$\sqrt{(1 - 4\mu^2)/2} = i\alpha \text{ when } \mu^2 > 1/4, \quad i = \sqrt{-1}.$$

Then we have, for a nontrivial solution,

$$(e^{i\alpha} - e^{-i\alpha}) = 0$$

or

$$2i \sin \alpha = 0, \quad \text{or} \quad \alpha_n = n\pi$$

The eigenvalues are, therefore,

$$\lambda_n = \frac{1}{4} + n^2\pi^2 \quad \text{for } n = 1, 2, \dots, \infty$$

The case  $4\mu^2 = 1$  has to be treated separately as this corresponds to repeated roots. It can be verified that this yields again the trivial solution. The eigenfunctions corresponding to the eigenvalues  $\lambda_n$  are

$$u_n(x) = A_n e^{x/2} \sin (n\pi x)$$

**Example 6.6** Is Theorem 6.3 valid for Example 6.5.

The eigenfunctions corresponding to two distinct eigenvalues  $\lambda_m, \lambda_n$  are

$$u_m(x) = A_m e^{x/2} \sin (m\pi x)$$

$$u_n(x) = A_n e^{x/2} \sin (n\pi x)$$

$$\langle u_m(x), u_n(x) \rangle = A_m A_n \int_0^1 e^x \sin (m\pi x) \sin (n\pi x) dx \neq 0$$

The operator  $L$  in Example 6.5 is not self-adjoint and hence the two eigenfunctions are not orthogonal to each other.

We write the eigenvalue problem in a self-adjoint form (Example 6.3) as

$$\frac{d}{dx} \left( e^{-x} \frac{du}{dx} \right) + \lambda e^{-x} u = 0$$

subject to

$$u(0) = 0, \quad u(1) = 0$$

This equation and boundary conditions are identical to the original eigenvalue problem in Example 6.4. Hence its eigenvalues and eigenfunctions have to be the same. Moreover, this is in self-adjoint form (i.e. similar to 6.17). The eigenfunctions  $u_m(x)$ ,  $u_n(x)$  are now orthogonal with respect to the weighting functioning  $e^{-x}$ .

$$\langle u_m(x), e^{-x} u_n(x) \rangle = \int_0^1 \sin(m\pi x) \sin(n\pi x) dx = 0$$

The transformation into self-adjoint form yields this important property of orthogonality, between the eigenfunctions. This can be exploited to construct solutions elegantly to partial differential equations. This is similar to the situation prevailing in finite dimensional spaces where real-symmetric (self-adjoint) matrices have an orthogonal set of eigenvectors. We chose this as our basis to represent vectors. We have already seen in Chapter 3 why it is desirable to have an orthogonal basis and not just an independent basis. Fortunately in most applications we have self-adjoint systems. These generate a natural orthogonal eigenbasis, which we like to work with. If our system is not self-adjoint, we try to render it self-adjoint, as in the Wei-Prater problem discussed in Section 4.6 or by using the transformation in (6.17).

### 6.3 CLASSICAL EIGENVALUE PROBLEMS

So far we established theorems for a general linear operator  $L$  of the form given by (6.17) subject to homogeneous linear boundary conditions  $Bu = 0$ . The theorems discuss the properties of the eigenvalues and eigenfunctions of these operators. The results are valid for the restrictions placed on  $p(x)$ ,  $q(x)$ , and  $\omega(x)$ , i.e.  $p(x) \geq 0$ ,  $q(x) \leq 0$  and  $\omega(x) > 0$  in the interval of interest. We now apply the theorems to specific linear operators. These operators and the corresponding eigenvalue problems arise while solving partial differential equations in curvilinear coordinates—cylindrical and spherical. The classical problem in rectangular cartesian coordinates has been discussed in Example 6.4. We will see the origin of these specific eigenvalue problems (ordinary differential equations) in Chapter 7. They occur frequently in modelling different situations and deserve the special attention which we now devote to them. The arbitrary functions  $p(x)$ ,  $q(x)$ ,  $\omega(x)$  in (6.17) now take specific forms as we are now going to discuss specific operators.

#### 6.3.1 Cylindrical Coordinates

**Eigenvalue problems in  $r$ -direction.** The commonly encountered eigenvalue problem in the  $r$ -direction in cylindrical coordinates is of the form

$$\frac{d^2 u}{dr^2} + \frac{1}{r} \frac{du}{dr} + \left( \lambda^2 - \frac{\mu^2}{r^2} \right) u = 0 \quad \text{in } 0 < r < R \quad (6.30a)$$

subject to the boundary conditions

$$\left. \begin{array}{l} u(r = R) = 0 \\ u(r = 0) \text{ is bounded} \end{array} \right\} \quad (6.30b)$$

Comparing with (6.7a), we identify the independent variable  $x$  with  $r$  and now have  $a_0(r) = 1$ ,  $a_1(r) = 1/r$ ,  $a_2(r) = -\mu^2/r^2$ ,  $a_3(r) = 1$ . The above equation occurs for a fixed value of  $\mu$ . It is an eigenvalue problem as the trivial solution exists for all  $\lambda$ . Eigenfunctions or nonzero solutions are sought to the above equation for a fixed  $\mu$ . The values of  $\lambda$  for which the eigenfunctions exist are called eigenvalues and depend on the value of  $\mu$ . We now proceed to determine the eigenvalues  $\lambda$  and eigenfunctions  $u$ . We adopt the same approach as used by Weinberger (1965). Substituting  $s = \lambda r$ , (6.30) is transformed into

$$\frac{d}{ds} \left( s \frac{du}{ds} \right) - \frac{\mu^2 u}{s} + su = 0 \quad (6.31a)$$

The parameter  $\lambda$  is now eliminated from the equation and it appears that this equation is not an eigenvalue problem. This is incorrect as  $\lambda$  now occurs in the boundary condition

$$u(s = \lambda R) = 0, \quad u(s = 0) = \text{bounded} \quad (6.31b)$$

The form of the boundary condition at  $r = 0$  arises since the equation is singular at  $r = 0$  as  $p(0) = 0$  and the coefficient multiplying the second derivative vanishes at  $s = 0$ . Equation (6.31a) is a linear equation with variable coefficients. The solutions are obtained in the form of a power series using the Frobenius method. This method of solving equations is discussed in detail in Piskunov (1981) and Ince (1956). Here we briefly describe the method and show how it is applied to this problem. We seek the solution  $u(s)$  to (6.31) in the Frobenius method in the form of a power series as

$$u(s) = s^\alpha \sum_{i=0}^{\infty} a_i s^i$$

Substituting this form in (6.31a) and collecting all terms raised to the same power of  $s$ , we have

$$s^{\alpha-1} \left[ (\alpha^2 - \mu^2)a_0 + ((\alpha+1)^2 - \mu^2)a_1 s + \sum_{i=2}^{\infty} ((\alpha+i)^2 - \mu^2)a_i s^i \right] = 0 \quad (6.32)$$

This equation holds for  $0 < s < \lambda R$  if all the coefficients vanish with  $a_0 \neq 0$ . Clearly, this implies  $a_1 = 0$ , and  $\alpha = \pm \mu$ , if the first two terms are to vanish. The recurrence relation between the various coefficients arises by imposing the condition that the terms inside the summation sign in (6.32) vanish, yielding thereby

$$a_i = \left( -\frac{a_{i-2}}{(\alpha+i)^2 - \mu^2} \right) \quad (6.33)$$

This recursion relation enables us to determine all the coefficients  $a_i$ . Setting  $\alpha = \mu$  and  $a_0 = 2^{-\mu}/\mu!$  we obtain the solution

$$u_1(s) = \sum_{k=0}^{\infty} \frac{(-1)^k (s/2)^{\mu+2k}}{k! (\mu+k)!} \quad (6.34)$$

When  $\mu$  is an integer  $n$ , the above series is denoted more compactly as  $J_n(s)$ . For noninteger  $\mu$  the factorial in (6.34) is defined in terms of the Gamma function ( $\Gamma$ ) see Kreyszig (1982). This is called

the Bessel function of the first kind of order  $n$ , see Weinberger (1965) and Kreyszig (1982). This is a solution to (6.31a) for  $\mu = n$ . Being a second order equation, the linear homogeneous equation has two independent solutions. The second solution  $u_2(s)$  can be obtained using  $\alpha = -\mu$ . For integer  $\mu (= n)$ , this yields a linearly dependent solution on  $J_n(s)$  as

$$J_{-n}(s) = (-1)^n J_n(s) \quad (6.35)$$

For integer  $n$ , the other independent solution is obtained using the method of variation of parameters. This solution is denoted by  $Y_n(s)$  and is called the Bessel function of the second kind of order  $n (= \mu)$ ,

$$Y_0(x) = \frac{2}{\pi} \left[ J_0(x) \left( \ln \frac{x}{2} + \gamma \right) + \sum_{m=1}^{\infty} \frac{(-1)^{m-1} h_m x^{2m}}{2^{2m} (m!)^2} \right]$$

where  $\gamma = .57721 \dots$  is called the Euler constant, and  $h_m = 1 + \frac{1}{2} \dots + \frac{1}{m}$ .  $Y_n(x)$  is defined for  $n \geq 1$  as

$$Y_n(x) = \lim_{v \rightarrow n} \frac{1}{\sin v\pi} [J_v(x) \cos v\pi - J_{-v}(x)]$$

The general solution to (6.31) for an integer  $\mu (= n)$  is given by a linear combination of these two solutions

$$u(s) = c_1 J_n(s) + c_2 Y_n(s) \quad (6.36a)$$

For a noninteger  $\mu$  in (6.31a), say  $\gamma$ , the two solutions to the linear homogeneous equation in series form arise for  $\alpha = \pm \gamma$ . These are denoted as  $J_\gamma(s)$  and  $J_{-\gamma}(s)$  and are linearly independent. The solution  $u(s)$  for this case is now given as

$$u(s) = c_1 J_\gamma(s) + c_2 J_{-\gamma}(s) \quad (6.36b)$$

The student coming across Bessel's function for the first time (or even the second time) is likely to feel intimidated at this stage. He must remember that these are names which are given to specific power series expansions. In this sense they are similar to the trigonometric functions  $\sin x$ ,  $\cos x$  which have the following unique power-series expansions:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \dots$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots$$

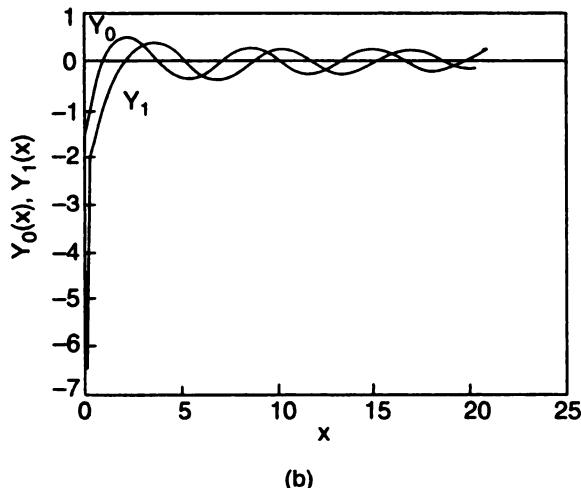
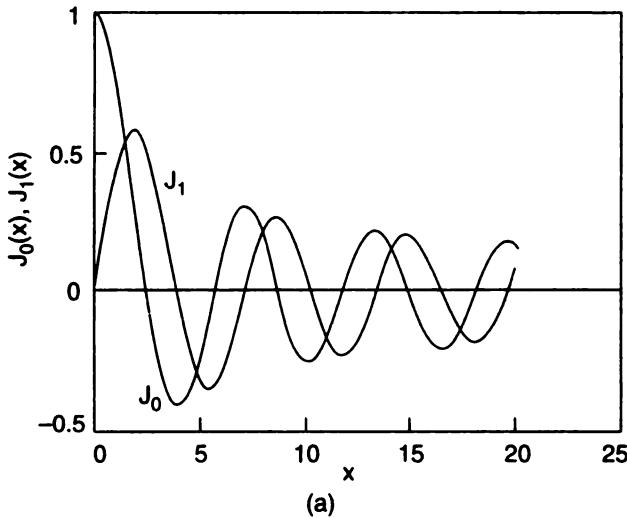
It may be worthwhile to recall the initial discomfort a student might have faced with trigonometry. Familiarity with trigonometry makes the student comfortable with the power series  $\sin x$ ,  $\cos x$ , etc. A student will be at ease with different Bessel's functions  $J_n(s)$ ,  $Y_n(s)$  after he has encountered them a few times. This will also make him conversant with the different properties needed in using them. A detailed listing of these properties can be found in Abramowitz and Stegun (1965).

The eigenvalues are determined by imposing the boundary conditions. In many problems  $\mu$  is an integer. For integer  $n$ ,  $Y_n(s)$  is unbounded at  $s = 0$ . Since we seek  $u(s)$  to be bounded, this implies  $c_2 = 0$  in (6.36a) as otherwise the solution will be unbounded. The constant  $c_1$  cannot be zero as this will yield the trivial solution. We satisfy the other boundary condition at  $s = \lambda R$ , for a nonzero  $c_1$ . This yields the eigenvalues  $\lambda$  as the roots of  $J_n(\lambda R) = 0$ .

The eigenvalues are obtained from the zeroes of  $J_n(x) = 0$ . The Bessel function  $J_n(x)$  has an infinite number of positive real roots as can be seen from Fig. 6.3. Consequently, the equation has an infinite number of eigenvalues. This is consistent with the fact that  $L$  operates on an infinite dimensional space. We identify the eigenvalues as  $\lambda_{n,m}$ , where the first subscript specifies the parameter  $\mu$  and the second subscript the  $m$ th eigenvalue. Equation (6.30a) is not in self-adjoint form. It can be recast in the self-adjoint form as

$$\frac{d}{dr} \left( r \frac{du}{dr} \right) - \frac{\mu^2}{r} u + \lambda^2 r u = 0$$

subject to  $u(r = R) = 0$ ,  $u(r = 0)$  is bounded.



**Fig. 6.3** Bessel functions: (a)  $J_0(x)$ ,  $J_1(x)$ ; (b)  $Y_0(x)$ ,  $Y_1(x)$ .  $Y_0$ ,  $Y_1$  are unbounded at  $x = 0$ .

Comparing this with (6.18), we have  $x = r$ ,  $p(r) = r$ ,  $q(r) = -\frac{\mu^2}{r}$ ,  $\omega(r) = r$ . Clearly,  $p(r) \geq 0$ ,  $q(r) < 0$ ,  $\omega(r) > 0$  in  $0 < r < R$ . The theorems established earlier are therefore applicable to this system. The eigenvalues are real and countably infinite. This is easily verified from the fact that  $J_n(s)$  has an infinite number of real zeroes. The eigenfunctions  $J_n(\lambda_{n,m}r)$  and  $J_n(\lambda_{n,p}r)$  corresponding to two distinct eigenvalues  $\lambda_{n,m}$  and  $\lambda_{n,p}$  are orthogonal with respect to the weighting function  $\omega(r) = r$ . Thus,

$$\int_0^R J_n(\lambda_{n,m}r) J_n(\lambda_{n,p}r) r dr = 0 \quad \text{for } m \neq p \quad (6.37a)$$

The following normalisation relation for  $m = p$  comes from identities of Bessel's function

$$\int_0^R J_n^2(\lambda_m r) r dr = \frac{R^2}{2} J_{n+1}^2(\lambda_m R) \quad (6.37b)$$

Figures (6.3a) and (6.3b) depict the various Bessel functions as a function of  $r$ . The different properties of these functions are now summarised.

**Property 1:**  $J_n(s)$  has an infinite number of zeroes and is a bounded function.

**Property 2:**  $Y_n(s)$  is unbounded at  $s = 0$ .

**Property 3:**  $\frac{d}{ds}(s^m J_m(s)) = s^m J_{m-1}(s)$ .

**Property 4:**  $\frac{d}{ds}(s^{-m} J_m(s)) = -s^{-m} J_{m+1}(s)$ .

These properties are often used in determining elegant analytical solutions to problems in cylindrical coordinates. These are similar to the identities in trigonometry and calculus. More identities involving the integrals of Bessel's functions can be found in Luke (1962).

A second class of Bessel's equations arise while solving problems in cylindrical coordinates which have nonhomogeneous boundary conditions in the  $r$ -direction. These equations are of the form

$$\frac{d^2 u}{dr^2} + \frac{1}{r} \frac{du}{dr} - \frac{\mu^2}{r^2} u - \alpha^2 u = 0 \quad (6.38a)$$

This equation can be rewritten as

$$\frac{d}{dr} \left( \frac{r du}{dr} \right) - \frac{\mu^2}{r} u - \alpha^2 r u = 0 \quad (6.38b)$$

In this, both  $\mu$ , and  $\alpha$  are fixed parameters. Although this equation is homogeneous, it does not admit the trivial solution  $u = 0$  because when it occurs it is associated with nonhomogeneous boundary conditions. Equations (6.38) do not arise as an eigenvalue problem. The equation does not satisfy the conditions of the Sturm-Liouville formulation either, as  $q(r) > 0$ . Besides, there are no free parameters to be determined. The solution to this linear homogeneous equation is determined

by using the method of Frobenius which has already been discussed. We seek solutions to (6.38) for a fixed  $\mu$  and  $\alpha$ . For integer  $\mu$ , i.e.  $\mu = n$ , the solution is of the form

$$u(r) = c_1 I_n(\alpha r) + c_2 K_n(\alpha r) \quad (6.39)$$

where  $I_n$  is called the modified Bessel's function of the first kind of order  $n$  and  $K_n$  is called the modified Bessel function of the second kind of order  $n$ . For noninteger  $\mu$ , say  $\mu = \gamma$ , we have the solution to (6.38) of the form

$$u(r) = c_1 I_\gamma(\alpha r) + c_2 I_{-\gamma}(\alpha r)$$

The expressions for these functions are

$$I_\gamma(x) = \sum_{m=0}^{\infty} \frac{x^{2m+\gamma}}{2^{2m+\gamma} m! \Gamma(m+\gamma+1)}$$

$$K_\gamma(x) = \frac{\pi}{2 \sin \gamma \pi} [I_{-\gamma}(x) - I_\gamma(x)]$$

$$K_n(x) = \lim_{\gamma \rightarrow n} K_\gamma(x)$$

A detailed discussion of the properties of Bessel's functions can be found in Sneddon (1956). The modified Bessel's functions  $I_n(r)$ ,  $K_n(r)$  possess, the following important properties (see Abramowitz and Stegun, 1965).

- (i)  $I_n(r)$  is bounded at  $r = 0$
- (ii)  $I_{-n}(r)$ ,  $K_n(r)$  are unbounded at  $r = 0$ .

These are depicted in Fig. 6.4.

**Eigenvalue problems in  $\theta$ -direction.** The eigenvalue problem in the  $\theta$ -direction in cylindrical coordinates takes the form

$$\frac{d^2 u}{d\theta^2} + \lambda^2 u = 0 \quad \text{for } -\pi < \theta < \pi \quad (6.40a)$$

The boundary conditions here are mathematical statements of the periodicity conditions of the function  $u$  (see Weinberger, 1965). The periodicity of  $u$  and its derivative imply

$$\left. \begin{array}{l} u(\pi) = u(-\pi) \\ u'(\pi) = u'(-\pi) \end{array} \right\} \quad (6.40b)$$

From the differential equation, the continuity of  $u(u')$  at  $\pi$  implies the continuity of all even (odd) derivatives of  $u$  at  $\theta = \pm \pi$ . So we have

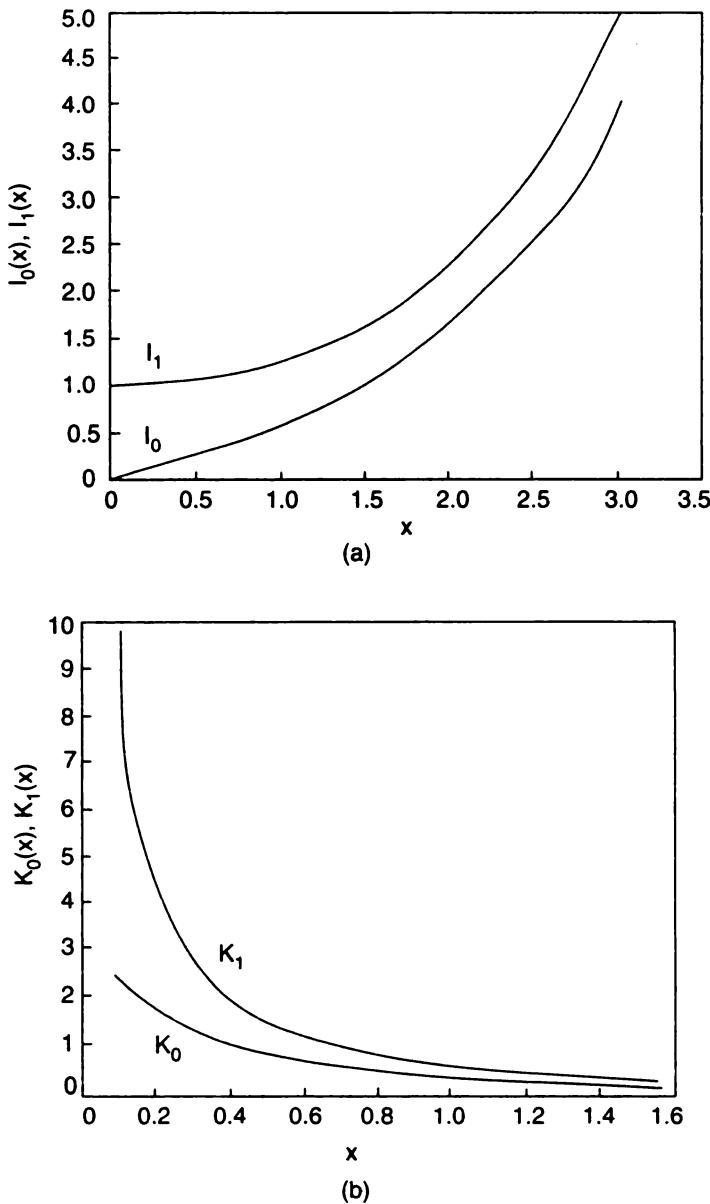
$$u''(\pi) = u''(-\pi), \dots u'''(\pi) = u'''(-\pi), \dots$$

To determine the eigenvalues we consider three distinct cases.

*Case 1:*  $\lambda^2 < 0$ .

*Case 2:*  $\lambda^2 = 0$ .

*Case 3:*  $\lambda^2 > 0$ .



**Fig. 6.4** Bessel functions: (a)  $I_0(x)$ ,  $I_1(x)$ ; (b)  $K_0(x)$ ,  $K_1(x)$ .

It is easy to establish that  $\lambda^2 < 0$ , yields only the trivial solution.

Case 2 yields condition

$$u = A\theta + B$$

which implies,

$$A\pi + B = A(-\pi) + B \text{ or } A = 0$$

The other condition does not yield any new result. Hence  $\lambda^2 = 0$  is an eigenvalue. The eigenfunction corresponding to it is the constant eigenfunction,  $u = B$ .

Case 3 yields

$$u = A \sin \lambda\theta + B \cos \lambda\theta \quad (6.41)$$

Imposing the boundary conditions, we obtain the eigenvalues  $\lambda$  as the roots of

$$\lambda \sin \lambda\pi = 0$$

This yields as eigenvalues  $\lambda = n^2$ , where

$$n = 1, 2, 3, \dots \quad (6.42a)$$

The corresponding eigenfunctions are

$$u = A \sin n\theta + B \cos n\theta \quad (6.42b)$$

For  $n = 1, 2, \dots$ , we have two independent eigenfunctions for each  $\lambda$ . These eigenvalues are therefore called double eigenvalues. We do not consider  $n = 0$ , here, as this reduces to case 2. We can elegantly combine cases 2 and 3, to yield the eigenvalues  $\lambda = n^2$ , where

$$n = 0, 1, 2, \dots, \infty \quad (6.43a)$$

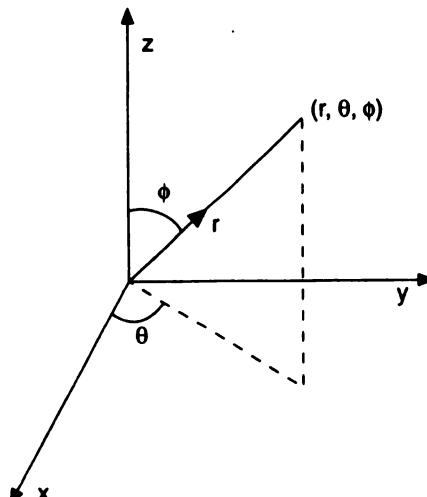
with the corresponding eigenfunctions

$$u_n = A \sin n\theta + B \cos n\theta \quad (6.43b)$$

The eigenfunctions generated by a self-adjoint operator subject to periodic boundary conditions are therefore the sine and cosine trigonometric functions. This forms the basis of the Fourier series expansion of periodic functions in terms of these trigonometric functions (see Churchill, 1963).

### 6.3.2 Spherical Coordinate Systems

In this coordinate system the three independent variables are  $r, \theta, \phi$  (Fig. 6.5). The  $r$ -direction is usually bounded in most applications and nonhomogeneous boundary conditions prevail in this



**Fig. 6.5**  $r, \theta, \phi$  in the spherical coordinate system,  $-\pi < \theta < \pi, 0 < \phi < \pi$ .

direction. Hence we do not encounter eigenvalue problems in this direction in the frequently encountered circumstances. In the  $\phi$ -direction the domain extends from  $(-\pi, \pi)$ . The eigenvalue problem in this direction here is identical to the problem in the  $\theta$ -direction in cylindrical coordinates. We obtain the periodic eigenfunctions  $\sin m\phi, \cos m\phi$  in this direction, where  $m = 0, 1, 2, \dots$

The  $\theta$ -direction in the spherical coordinate system extends from  $(0, \pi)$ . In solving problems in the spherical coordinate system by separation of variables, the eigenvalue problem originating in the  $\theta$ -direction is of the form

$$(\sin(\theta)T')' - \frac{m^2 T}{\sin \theta} + \lambda \sin(\theta)T = 0 \quad (6.44a)$$

where  $m = 0, 1, 2, \dots$  The domain of the definition of this problem is  $0 < \theta < \pi$ . The equation is singular at these two end points, as here the coefficient multiplying the highest order derivative vanishes. The boundary conditions we impose, are  $\theta, \theta'$ , be bounded at the two ends ( $\theta = 0, \pi$ ). This renders (6.44a) an eigenvalue problem with two singular end points. Substitution of

$$t = \cos \theta, \quad T(\theta) = P(\cos \theta) = P(t)$$

transforms (6.44a) into

$$\frac{d}{dt} \left[ (1 - t^2) \frac{dP}{dt} \right] - \frac{m^2}{1 - t^2} P + \lambda P = 0, \quad -1 < t < 1 \quad (6.44b)$$

The eigenvalues  $\lambda$  now are those values of  $\lambda$  for which we have nonzero solutions such that  $P$  and  $(1 - t^2)^{1/2} P'(t)$  remain bounded at  $t = \pm 1$ . To obtain the eigenvalues and eigenfunctions of (6.44b), we consider the cases  $m = 0$  and  $m = 1, 2, \dots$  separately.

*Case  $m = 0$ :* The Equation (6.44b) now reduces to

$$\frac{d}{dt} \left[ (1 - t^2) \frac{dP}{dt} \right] + \lambda P = 0 \quad (6.45)$$

Clearly,  $\lambda = 0$  is an eigenvalue as it admits  $P = \text{constant}$  as a possible candidate for the eigenfunction. We seek the solution  $P(t)$  in the neighbourhood of  $t = 1$ , as a power series in  $(t - 1)$ , see Weinberger (1965).

$$P(t) = (t - 1)^\alpha \sum_{k=0}^{\infty} c_k (t - 1)^k \quad (6.46)$$

Substituting (6.46) in (6.45), we obtain a power series which is to be identically zero, i.e.

$$-c_0 2\alpha^2 (t - 1)^{\alpha-1} - \sum_{k=0}^{\infty} \left\{ 2(k + \alpha + 1)^2 c_{k+1} + (-\lambda + (k + \alpha + 1)(k + \alpha)) c_k \right\} (t - 1)^{k+\alpha} = 0 \quad (6.47)$$

If  $c_0$  is zero, we get only the trivial solution. For  $c_0 \neq 0$ , we must have  $\alpha = 0$ , for the first term to vanish. This is a double root. We obtain the recursion formula from (6.47) as

$$c_{k+1} = \frac{-[k(k + 1) - \lambda] c_k}{2(k + 1)^2}$$

This yields

$$c_k = \frac{[k(k - 1) - \lambda][k(k - 2) - \lambda] \dots [2 - \lambda][-\lambda](-1)^k c_0}{2^k (k!)^2}$$

Applying the ratio test (see Kreyszig, 1982), we see that the series is convergent for

$$|t - 1| < 2 \text{ or } -1 < t < 1$$

When  $(k + 1) > \lambda$ , we have

$$-\frac{c_{k+1}}{c_k} > \frac{1}{2} \frac{k-1}{k+1}$$

The function therefore approaches  $\pm\infty$  as  $t \rightarrow -1$ . The only exception occurs if the series terminates, i.e. for some  $n$ ,  $c_{n+1} = c_{n+2} = \dots = 0$ . This occurs if

$$\lambda = n(n + 1)$$

A bounded function can be obtained in  $[-1, 1]$  as the solution to (6.45) only if  $\lambda = n(n + 1)$ , for  $n = 0, 1, 2, \dots$ . These are the eigenvalues. Setting  $c_0 = 1$ , we obtain  $P_n(t)$  corresponding to  $\lambda = n(n + 1)$  as

$$P_n(t) = \sum_{k=0}^n \frac{(n+k)!}{(n-k)!(k!)^2} \frac{(t-1)^k}{2^k} \quad (6.48)$$

$P_n(t)$  is an  $n$ th order polynomial in  $t$  and is called the Legendre polynomial.

The first few polynomials are

$$P_0(t) = 1, \quad P_1(t) = t, \quad P_2(t) = \frac{3}{2}t^2 - \frac{1}{2} \dots$$

$\alpha = 0$  is a double root of (6.47) when the first coefficient has to vanish. The second independent solution to (6.45) can be obtained by the method of variation of parameters. This is denoted as  $Q_n(t)$  and is of no interest to us as it is unbounded at  $t = \pm 1$ .

*Case  $m \neq 0$ :* Let  $m$  be a positive integer. Differentiating (6.45) with  $m = 0, m$  times with respect to  $t$ , we get

$$(1 - t^2)P^{[m+2]} - 2(m+1)t P^{[m+1]} + [\lambda - m(m+1)]P^{[m]} = 0 \quad (6.49)$$

Here  $P^{[m]}$  represents  $m$ th derivative of  $P$  with respect to  $t$ .

Defining,

$$Q(t) = (1 - t^2)^{m/2} P^{[m]}(t)$$

the above equation becomes

$$[(1 - t^2)Q']' - \frac{m^2}{1 - t^2} Q + \lambda Q = 0 \quad (6.50)$$

Thus from any solution  $P(t)$  of (6.45), we can obtain a solution  $(1 - t^2)^{m/2} P^{[m]}(t)$  of (6.50). It can be proven that for a given  $m$ , the eigenvalues  $\lambda$  of (6.50) are of the form  $n(n + 1)$ , where  $n = m, m + 1, \dots$  etc. (see Weinberger, 1965 and Sneddon, 1956). In particular,  $n(n + 1)$  is not an eigenvalue for  $n < m$ .

The corresponding eigenfunctions

$$\begin{aligned} P_n^m(t) &= (1 - t^2)^{m/2} \frac{d^m}{dt^m} [P_n(t)] \\ &= \frac{1}{2^n n!} (1 - t^2)^{m/2} \frac{d^{n+m}}{dt^{m+n}} (t^2 - 1)^n \end{aligned} \quad (6.51)$$

are called the associated Legendre function. The eigenfunction  $P_n^m(t)$  corresponds to the eigenvalue  $\lambda_{m,n} = n(n+1)$ , where  $n \geq m$ . The second independent solution is  $Q_n^m(t)$  and is unbounded at  $t = \pm 1$ .

## 6.4 FOURIER SERIES

So far we have seen the eigenfunctions and eigenvalues which frequently arise in different coordinate systems. The boundary conditions considered were of the Dirichlet type for simplicity. The eigenvalues and eigenfunctions would be different if either the differential equation or the boundary condition were to be different.

We now revert to some generalities and discuss how an arbitrary function can be represented as a linear combination of eigenfunctions. This is analogous to the representation of an arbitrary vector as a linear combination of independent eigenvectors in Chapter 3. In determining the solutions of linear algebraic equations we represented the unknown solution vector in terms of a basis set. This set was chosen to be the eigenvector set which was natural to the system. In this approach we seek

$$u = \sum_{i=1}^N c_i u^i$$

In many situations where the vector  $u$  depended on time we wrote

$$u = \sum_{i=1}^N c_i(t) u^i$$

The summation index  $N$  represents the dimension of the vector space. In a finite dimensional space the summation is over a finite number of terms. In an infinite dimensional space we have an infinite number of eigenvalues and eigenfunctions. Here we may want to represent an arbitrary function  $u(x)$  in terms of the eigenfunctions  $u_i(x)$  in the interval  $[a, b]$ . We write

$$u(x) = \sum_{i=1}^{\infty} c_i u_i(x) \quad (6.52a)$$

In writing (6.52a) we have tacitly assumed that the eigenfunctions form a complete set. We will define the completeness of a set shortly. This notion of a complete set is not to be confused with the completeness of a space which we saw earlier when we spoke of  $L^2(a, b)$ . There is a practical difficulty in representing  $u(x)$  as in (6.52a) where the summation consists of a summation over an infinite number of terms. Any computation or evaluation can be done only over a finite number of terms  $N$ . The partial sum in (6.52a) over  $N$  terms is denoted by  $s_N(x)$ . This  $s_N(x)$  is only an approximate representation of the function  $u(x)$ . Hence

$$u(x) \approx s_N(x) = \sum_{i=1}^N c_i u_i(x) \quad (6.52b)$$

Our aim is to obtain  $c_i$  such that the partial sum  $s_N(x)$  is the best possible representation of the function  $u(x)$ . In particular, we would like to minimise the error in the representation of  $u(x)$  by  $s_N(x)$  for a fixed  $N$ . Ideally, we would like to choose  $c_i$  so that we have pointwise convergence. Here for every  $x$  in  $(a, b)$  we would like  $s_N(x)$  to be near  $u(x)$  as  $N \rightarrow \infty$ , i.e.

$$\lim_{N \rightarrow \infty} |s_N(x) - f(x)| < \epsilon \quad \forall x \text{ in } (a, b) \quad (6.53)$$

It is, however, much easier to aspire for convergence in the mean or in the sense of least squares. This implies

$$\lim_{N \rightarrow \infty} \int_a^b (u(x) - s_N(x))^2 \omega(x) dx = 0 \quad (6.54)$$

or

$$d_2(u(x), s_N(x)) \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

The convergence in the mean (6.54) suggests that  $s_N(x)$  approximates  $u(x)$  very well in almost the whole interval  $(a, b)$  except over a set, i.e. interval of measure (length) zero. The metric defined here is the one generated by the inner-product in which the eigenfunctions are orthogonal with respect to the weight function  $\omega(x)$ . That is,

$$\langle u(x), v(x) \rangle = \int_a^b \bar{u}(x)v(x)\omega(x) dx$$

Consider now the set of eigenfunctions  $u_i(x)$  which are orthogonal with respect to the weighting function  $\omega(x)$ , i.e.,

$$\int_a^b \bar{u}_m(x)u_n(x)\omega(x) dx = 0 \quad \text{for } m \neq n$$

We now determine the  $c_i$  in (6.52b) which will yield the best approximation to the arbitrary function  $u(x)$  for a fixed  $N$ . The best  $c_i$  will result in the minimum value of the metric  $d_2(u(x), s_N(x))$  for a fixed  $N$ .

$$\begin{aligned} \int_a^b (u(x) - s_N(x))^2 \omega(x) dx &= \int_a^b \left( u(x) - \sum_{i=1}^N c_i u_i(x) \right)^2 \omega(x) dx \\ &= \int_a^b u^2(x)\omega(x) dx - 2 \sum_{i=1}^N c_i \int_a^b u(x)u_i(x)\omega(x) dx + \sum_{i=1}^N c_i^2 \int_a^b u_i^2(x)\omega(x) dx \\ &= \int_a^b u^2(x)\omega(x) dx - \sum_{i=1}^N \frac{\left( \int_a^b u(x)u_i(x)\omega(x) dx \right)^2}{\int_a^b u_i^2(x)\omega(x) dx} \\ &\quad + \sum_{i=1}^N \left( c_i - \frac{\int_a^b u(x)u_i(x)\omega(x) dx}{\int_a^b u_i^2(x)\omega(x) dx} \right)^2 \int_a^b u_i^2(x)\omega(x) dx \end{aligned}$$

The terms have been rearranged so that  $c_i$  now occurs only in the last term on the right. This is a sum of squares with positive coefficients. The least value this term can attain is zero. This occurs when

$$c_i = \frac{\int_a^b u(x)u_i(x)\omega(x) dx}{\int_a^b u_i^2(x)\omega(x) dx} \quad (6.55)$$

This choice of  $c_i$  minimises the deviation of  $s_N(x)$  from  $f(x)$  in the sense of least squares. For any other choice of  $c_i$  and the same  $N$ , the error over the interval  $(a, b)$  will be more than this minimum value. The coefficients  $c_i$  defined above are called **Fourier coefficients** and with this choice (6.52) is called a **Fourier series**.

An added advantage of the Fourier coefficients defined above is that the  $i$ th coefficient  $c_i$  is independent of the  $j$ th coefficient  $c_j$ . To obtain  $s_N(x)$  it is necessary to determine the coefficients  $c_i$  where  $i \in 1, N$ . For a better approximation we may want to increase  $N$  to  $P$ , where  $P > N$ . To obtain  $s_P(x)$  we only have to compute the coefficients  $c_{N+1}, \dots, c_P$  and use the already determined values of the coefficients  $c_1, \dots, c_N$ . We can do this as the coefficients  $c_{N+1}, \dots, c_P$  are independent of  $c_1, \dots, c_N$ . The choice of Fourier coefficients yields

$$\int_a^b (u(x) - s_N(x))^2 \omega(x) dx = \int_a^b u^2(x) \omega(x) dx - \frac{\sum_{i=1}^N \left( \int_a^b u(x) u_i(x) \omega(x) dx \right)^2}{\int_a^b u_i^2(x) \omega(x) dx}$$

Since the integral on the left is a non-negative quantity

$$\sum_{i=1}^N \frac{\left( \int_a^b u(x) u_i(x) \omega(x) dx \right)^2}{\int_a^b u_i^2(x) \omega(x) dx} \leq \int_a^b u^2(x) \omega(x) dx \quad (6.56)$$

This is called **Bessel's inequality** (see Weinberger, 1965). The sum on the left is a sum of positive terms and is nondecreasing as  $N$  increases. For finite  $\int_a^b u^2(x) \omega(x) dx$ , it is bounded from above. Hence the sum on the left converges as  $N \rightarrow \infty$ , and we have, using (6.55), the inequality

$$\sum_{i=1}^{\infty} c_i^2 \int_a^b u_i^2(x) \omega(x) dx \leq \int_a^b u^2(x) \omega(x) dx$$

If  $s_N(x)$  converges to  $u(x)$  in the mean as  $N \rightarrow \infty$ , we have

$$\sum_{i=1}^{\infty} c_i^2 \int_a^b u_i^2(x) \omega(x) dx = \int_a^b u^2(x) \omega(x) dx \quad (6.57)$$

This equation is called **Parseval's equation** (see Weinberger, 1965).

The evaluation of  $s_N(x)$  using the Fourier coefficients  $c_i$  enables us to approximate a function in an interval  $(a, b)$ . Being an infinite dimensional space we can now only hope to increase the accuracy of the approximation by taking on larger values of  $N$ . This is similar to employing the method of finite differences in determining the numerical solution of an equation. Here again we discretise an infinite dimensional system and approximate it over a discrete number of points, say  $N$ . The accuracy is improved in the finite difference scheme by increasing  $N$ , the number of grid points. This is analogous to increasing  $N$  in  $s_N(x)$  to get a better representation of  $f(x)$ .

Till now we have considered  $u_i(x)$  in defining the  $s_N(x)$  to be a set of infinite orthogonal eigenfunctions. Such a representation of an arbitrary function in terms of an infinite set is not possible when the infinite set of functions is arbitrary. The set of functions  $u_i(x)$  must be complete

so that we can represent a function as a linear combination of the set. A set of functions  $\{u_i(x)\}$  is said to be complete (Weinberger, 1965), if for every function  $f(x)$  in  $[a, b]$  such that

$$\int_a^b f^2(x) \omega(x) dx \text{ is finite for } \omega(x) > 0$$

we have

$$\lim_{N \rightarrow \infty} \int_a^b (f(x) - s_N(x))^2 \omega(x) dx = 0$$

where  $s_N(x)$  is the partial sum  $\sum_{i=1}^N c_i u_i(x)$ .

It is quite difficult to establish the property of completeness of a set of functions. This property is analogous to that of linear independence of a set in a finite dimensional space. If a set of  $n$  vectors is linearly dependent in  $R^n$ , we cannot use it as a basis. Similarly, if a set of functions is incomplete, it cannot be a basis in  $(a, b)$ . The maximal set of linearly independent elements in a space is a complete set. We state the following theorem without proof.

**Theorem 6.4** The eigenfunctions of a self-adjoint operator form a complete set.

Fortunately in most applications we encounter self-adjoint systems. This theorem allows us to use the eigenfunctions of the operator as a basis in representing an arbitrary function. The classical Fourier series is used in representing periodic functions in the interval  $-\pi, \pi$ . The trigonometric functions  $\sin n\theta, \cos n\theta$  are the eigenfunctions of the system with periodic conditions see (6.43). Other forms of series representations arise when we use eigenfunctions generated by the naturally occurring system. The Fourier coefficients are more generically called finite Fourier transform as here the domain of the problem  $[a, b]$  is finite. We will see more of this in Chapter 7.

## 6.5 RAYLEIGH'S QUOTIENT

In Chapter 3 we were able to establish for a matrix operator upper bounds on the lowest eigenvalue  $\lambda_1$  and lower bounds on the largest eigenvalue  $\lambda_n$ , where the eigenvalues are arranged so that

$$\lambda_1 \leq \lambda_2 \leq \lambda_3, \dots, \leq \lambda_n$$

We restricted our discussions to self-adjoint operators, which assured us that all eigenvalues would be real. The  $n$ th order matrix has  $n$  eigenvalues and it is possible to have lower and upper bounds on these eigenvalues. For a second order differential operator we have an infinite number of eigenvalues. Besides, as already seen, for self-adjoint systems, these eigenvalues are non-negative. When arranged in an ascending order the largest eigenvalue approaches infinity. Here clearly zero is a lower bound on the eigenvalues. We would like to improve our estimate of the lowest eigenvalue by obtaining an upper bound for it. We extend the definition of Rayleigh's quotient from Chapter 3 now to differential operators.

Consider a self-adjoint operator,  $L$ . Then

$$Lu = -\lambda \omega(x)u$$

subject to

$$u(0) = 0, u(1) = 0$$

(6.58)

Here  $L$  is of the form as in (6.17). Let  $f(x)$  be an arbitrary piecewise differentiable function which satisfies the boundary conditions of  $u$ , i.e.  $f(0) = 0, f(1) = 0$ . We represent

$$f(x) = \sum_{i=1}^N c_i u_i(x)$$

where  $u_i(x)$  are the eigenfunctions of (6.58). Consider the Fourier series representation of

$$\frac{1}{\omega(x)} [(p(x)f'(x))' + q(x)f(x)]$$

Its Fourier coefficients  $b_i$  are given by

$$b_i = \frac{\int_a^b [(p(x)f'(x))' + q(x)f(x)]u_i(x) dx}{\int_a^b \omega(x)u_i^2(x) dx} \quad (6.59)$$

Integrating by parts and using the conditions on  $f$  and  $u_i$  at the end points, we obtain

$$b_i = -\lambda_i c_i \quad (6.60)$$

where  $c_i$  are the Fourier coefficients of  $f(x)$ . From Parseval's equation, we have

$$\left. \begin{aligned} \int_0^1 f^2(x)\omega(x) dx &= \sum_{i=1}^{\infty} c_i^2 \int_0^1 u_i^2(x)\omega(x) dx \\ \int_0^1 [p(x)f'^2(x) + q(x)f^2(x)] dx &= - \int_0^1 f(x) ((pf')' + qf) dx \\ &= - \int_0^1 \sum_{i=1}^{\infty} c_i u_i(x) \sum_{j=1}^{\infty} b_j u_j(x)\omega(x) dx \end{aligned} \right\} \quad (6.61)$$

Using the orthogonality of the eigenfunctions, Parseval's equation and (6.60), the right-hand side becomes

$$\text{R.H.S.} = \sum_{i=1}^{\infty} \lambda_i c_i^2 \int_0^1 \omega(x)u_i^2(x) dx$$

Let the  $\lambda$ 's be arranged such that

$$\lambda_1 < \lambda_2 < \lambda_3 < \dots < \lambda_n \dots$$

Replacing all the eigenvalues in the summation sign by  $\lambda_1$  the lowest eigenvalue, we get

$$\int_0^1 (p(x)f'^2(x) - q(x)f^2(x)) dx \geq \lambda_1 \sum_{i=1}^{\infty} c_i^2 \int_0^1 \omega(x)u_i^2(x) dx$$

Using Parseval's equation we obtain

$$\lambda_1 \leq \frac{\int_0^1 (p(x)f'^2(x) - q(x)f^2(x))}{\int_0^1 \omega(x)f^2(x) dx}$$

For  $\omega(x) = 1$ , we can represent the above inequality as

$$\lambda_1 \leq -\frac{\langle f, Lf \rangle}{\langle f, f \rangle}$$

This is exactly similar to Rayleigh quotient we had defined in Chapter 3. It is possible to obtain estimates or bounds on the other eigenvalues as well using this method (see Weinberger, 1965).

**Example 6.7** Obtain upper and lower bounds on the lowest eigenvalue of

$$\frac{d^2u}{dx^2} + \lambda^2 u = 0$$

subject to  $u(0) = u(1) = 0$ .

It is easy to prove that all the eigenvalues of this system are positive. Hence a lower bound on the lowest eigenvalue  $\lambda_1^2$  is zero, i.e.,

$$0 \leq \lambda_1^2$$

To obtain an upper bound on  $\lambda_1^2$ , consider a function that satisfies the boundary conditions. The simplest function to do this is

$$f(x) = x(1-x)$$

Clearly,

$$f'(x) = (1-2x), \quad p(x) = 1$$

Hence,

$$\lambda_1^2 \leq \frac{\int_0^1 (1-2x)^2 dx}{\int_0^1 x^2(1-x)^2 dx} \leq 10$$

The least eigenvalue  $\lambda_1^2$  we know is  $\pi^2$ . Hence the function  $f(x)$  chosen provides a very good upper bound for the eigenvalue  $\lambda_1^2$ . We can try to improve this estimate by considering different functions  $f(x)$ .

## PROBLEMS

1. Consider

$$\frac{d^2u}{dx^2} = -\lambda u$$

subject to

$$u'(0) = u'(1) = 0$$

Why does the proof of  $\lambda$ 's being positive fail here?

2.  $(1+x)^2 \frac{d^2u}{dx^2} = -\lambda u$  in  $0 < x < 1$

subject to

$$u(0) = 0, \quad u(1) = 0,$$

Verify that its eigenvalues are positive, and the eigenfunctions corresponding to two distinct eigenvalues are orthogonal.

3.  $(1+x)^3 \frac{\partial}{\partial x} \left( \frac{1}{1+x} \frac{\partial u}{\partial x} \right) - \lambda u = 0$

subject to

$$u(0) = u(1) = 0$$

Compute the eigenvalues and eigenfunctions of these equations.

**4. Consider**

$$Lu = \frac{d^2u}{dx^2} - du/dx, \quad u(0) = 0, \quad u(1) = 0$$

Find eigenfunctions of  $L$ .

Find eigenfunctions of  $L^*$ .

Convert to self-adjoint form and find eigenfunctions of  $L$ .

**5. Repeat Problem 4 for**

$$\frac{d^2u}{dx^2} = \lambda u$$

subject to

$$u(0) = 2u'(1), \quad u(1) = 0$$

**6. Find the eigenvalues and eigenfunctions of**

(a)  $\frac{d^2u}{dx^2} + \lambda u = 0$

subject to

$$u(1) = u(0), \quad u'(1) = u'(0)$$

(b)  $\frac{d}{dx} \left( (1+x)^2 \frac{du}{dx} \right) + \lambda u = 0$

subject to

$$u(0) = u(1) = 0$$

**7. Find the adjoint operator and boundary conditions for**

$$Lu = \frac{d^2u}{dx^2} + \frac{du}{dx}$$

subject to

$$u(0) = 2u'(1) + 3, \quad u(1) = 0$$

**8. Consider**

$$Lu = \frac{d^2u}{dx^2} - \frac{du}{dx}$$

subject to

$$u(0) = 0, \quad u(1) = 0$$

- (a) Find the eigenfunctions  $u_m$  and the eigenvalues of  $L$ .
- (b) Find  $L^*$ ,  $B^*$ .
- (c) Find the eigenfunctions  $v_n$  and the eigenvalues of  $L^*$ .
- (d) Prove the biorthogonality relationship between  $u_m$ ,  $v_n$ .
- (e) Convert to Sturm-Louiville form, obtain eigenfunctions, and the orthogonality relationship.

**9. Find an estimate of the first zero of  $J_1(x) = 0$ .**

**10. Find the adjoint operators  $L^*$ ,  $B^*$  when**

$$Lu = \frac{d^2u}{dx^2}$$

subject to

$$u(0) = u'(1), \quad u(1) = u'(0)$$

## REFERENCES

- Abramowitz, M. and Stegun, I.A. (Eds.), Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, Dover, New York (1965).
- Churchill, R.V., Fourier Series and Boundary Value Problems, McGraw-Hill, New York (1963).
- Courant, R. and Hilbert, D., Methods of Mathematical Physics, Wiley Eastern, New Delhi (1975).
- Garabedian, P.R., Partial Differential Equations, Wiley, New York (1964).
- Ince, E.L., Ordinary Differential Equations, Dover, New York (1956).
- Kaplan, W., Operational Methods for Linear Systems, Addison Wesley, Reading, Mass. (1962).
- Kreyszig, E., Advanced Engineering Mathematics, Wiley, New York (1982).
- Luke, Y.L., Integrals of Bessel Functions, McGraw-Hill, New York (1962).
- Naylor, A.W. and Sell Ha., Linear Operator Theory in Engineering and Science, Holt, Reinhart and Winston, New York (1971).
- Piskunov, N., Differential and Integral Calculus, Mir Publishers, Moscow (1981).

- Ramkrishna, D. and Amundson, N.R., Linear Operator Methods in Chemical Engineering: With applications to transport and chemical reaction systems, Prentice-Hall, Englewood Cliffs, New Jersey (1985).
- Sneddon, I.N., Special Functions of Mathematical Physics and Chemistry, Oliver & Boyd, London (1956).
- Stakgold, I., Boundary-value Problems of Mathematical Physics, Macmillan, New York (1968).
- Weinberger, H.F., A First Course in Partial Differential Equations: With complex variables and transform methods, Wiley, New York (1965).

# 7

## Separation of Variables and Fourier Transforms

---

In this chapter we discuss the methods of solving some of the different classes of second order linear partial differential equations. The equations we consider here are homogeneous and are subject to nonhomogeneous initial conditions and boundary conditions. The method described here is an extension of the one discussed in Chapter 4 for finite dimensional systems. In that chapter we expanded the unknown solution in terms of the eigenvectors of the matrix operator occurring in the problem. In this chapter we use the eigenfunctions of the naturally occurring differential operator to construct the solution. The two problems of main interest to us are: (i) the parabolic initial condition problem, and (ii) the elliptic boundary condition problem. In this chapter the method is applied to solve problems in the different coordinate systems—rectangular (cartesian), cylindrical and spherical coordinates. These problems are typical and occur in a wide range of engineering applications (see Berg (1964) and Kersten (1969)).

We also see how the method of separation of variables gives rise to the concept of a Fourier series and, more generally, a finite Fourier transform for problems in spatially bounded domains. The extension of this idea to spatially unbounded domains is the basis for the Fourier transform. These transforms are used to reduce partial differential equations to simpler ordinary differential equations which can be solved using first principles. The solution to the original system is then obtained by taking the inverse transform.

### 7.1 RECTANGULAR CARTESIAN COORDINATES

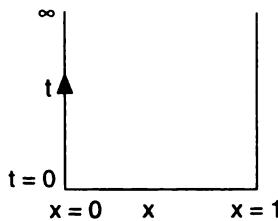
**Example 7.1** Consider the problem of heat conduction in a one-dimensional slab ( $0 < x < 1$ ), see Fig. 7.1. The surface at  $x = 0$  is insulated and that at  $x = 1$  is losing heat to the environment at a rate proportional to the temperature difference between the surface and the ambient (see Carslaw and Jaeger, 1959). The transient temperature profile is governed by

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (7.1a)$$

subject to

$$u(t = 0) = u_0(x) \quad (7.1b)$$

$$\frac{\partial u}{\partial x}(x = 0) = 0 \quad (7.1c)$$



**Fig. 7.1** Domain of one-dimensional heat conduction problem.

$$\frac{\partial u}{\partial x}(x=1) + u(x=1) = 0 \quad (7.1d)$$

The above equation has a single nonhomogeneity, i.e., the nonzero initial condition. The system of equations (7.1) describes the evolution of the temperature profile  $u$  in space and time. From physical considerations we expect the temperature to decay to zero at all points in space as  $t \rightarrow \infty$ , since there is no heat generation term in the domain  $(0, 1)$  and the slab will lose heat and attain the ambient value  $u = 0$  until all existing temperature gradients are wiped out. To determine the transient behaviour, we seek the solution in separable form as

$$u(x, t) = T(t)X(x) \quad (7.2)$$

Substituting this in (7.1a), we obtain

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} \quad (7.3)$$

Since the left hand side is only a function of  $t$  and the right-hand side is only a function of  $x$ , they can be equal only if both sides are identically equal to the constant. For convenience we choose this constant to be  $-\lambda^2$  and get

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} = -\lambda^2 \quad (7.4)$$

Equation (7.4) yields two ordinary differential equations

$$\frac{T'(t)}{T(t)} = -\lambda^2 \quad (7.5a)$$

$$\frac{X''(x)}{X(x)} = -\lambda^2 \quad (7.5b)$$

Equation (7.5b) is subject to the boundary conditions

$$X'(0) = 0 \quad (7.5c)$$

$$X'(1) + X(1) = 0 \quad (7.5d)$$

The constant  $\lambda^2$  is evaluated such that the system (7.5b)–(7.5d) is an eigenvalue problem. When  $\lambda^2$  is not an eigenvalue, we obtain the trivial solution  $X(x) = 0$ .

In choosing the constant in (7.4) as  $-\lambda^2$ , we have used our intuition that the eigenvalues in the  $x$ -direction will be positive or non-negative. This yields the negative sign in  $-\lambda^2$ . The squared constant indicates the sign of the constant and gets rid of cumbersome square-root signs in the solution.

The eigenvalue problem in the  $x$ -direction has a solution of the form

$$X(x) = A \sin(\lambda x) + B \cos(\lambda x) \quad (7.6)$$

Condition (7.5c) implies  $A\lambda = 0$ .  $\lambda$  cannot be zero as  $\lambda = 0$  is not an eigenvalue. Hence  $A = 0$ . The second boundary condition (7.5d) yields the eigenvalues as roots of

$$\lambda = \cot \lambda \quad (7.7a)$$

This is a transcendental equation. It has an infinite number of roots. These correspond to the intersections of the  $45^\circ$ -line with the curves  $\cot \lambda$  as shown in Fig. 7.2. We have to evaluate the roots numerically. Let them be ordered as  $\lambda_1, \lambda_2, \lambda_3, \dots$ :

$$\lambda_1 < \lambda_2 < \lambda_3 \dots$$

The corresponding eigenfunctions are

$$X_n(x) = B_n \cos \lambda_n x \quad (7.7b)$$

Since  $\cot \lambda$  is an odd function of  $\lambda$ , if  $\lambda_1$  is a root, so is  $-\lambda_1$ . We do not consider these roots since (i) the eigenvalue is the square of the root and  $(\lambda_1)^2 = (-\lambda_1)^2$ , and (ii) no new linear independent eigenfunction arises by considering negative  $\lambda_j$ 's.

Using the  $\lambda_i$ 's from (7.7a) in (7.4), we obtain

$$T_n(t) = D_n \exp(-\lambda_n^2 t)$$

where  $D_n$  is an arbitrary constant which is yet to be determined and corresponds to  $T_n(0)$ .

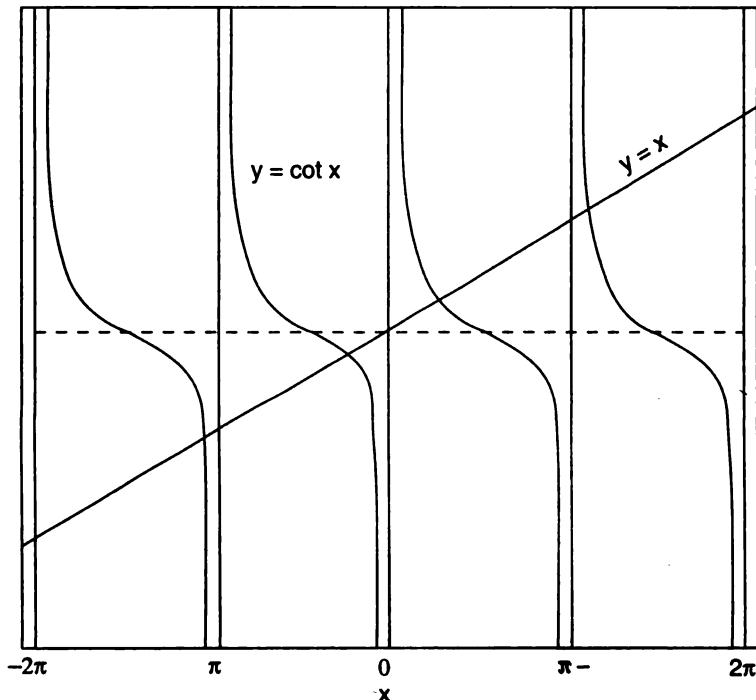


Fig. 7.2 Roots of  $\cot x = x$ .

A solution of the form  $X_n(x)T_n(t)$  will therefore satisfy the homogeneous equation (7.1a) and the homogeneous conditions (7.1c) and (7.1d). Since these equations are homogeneous, a linear combination of the  $X_n(x)T_n(t)$  will also satisfy (7.1a)–(7.1d). We now only have to see how we can ensure that the nonhomogeneous initial condition (7.1b) can also be satisfied.

Substituting the expressions for  $X_n(x)$ ,  $T_n(t)$  and using linear superposition, we obtain

$$u(x, t) = \sum_{n=1}^{\infty} C_n \exp(-\lambda_n^2 t) \cos \lambda_n x \quad (7.8)$$

where the constant  $C_n = B_n D_n$ . The constants  $C_n$  in the eigenfunction expansion (7.8) are determined by using the initial condition, yielding thereby

$$u_0(x) = \sum_{n=1}^{\infty} C_n \cos \lambda_n x \quad (7.9a)$$

The Fourier coefficients  $C_n$  are determined by exploiting the orthogonality of the  $\cos \lambda_n x$ , i.e.,

$$\int_0^1 \cos \lambda_n x \cos \lambda_m x dx = 0 \quad \text{for } n \neq m$$

This yields

$$C_n = \frac{\int_0^1 u_0(x) \cos \lambda_n x dx}{\int_0^1 \cos^2 \lambda_n x dx} \quad (7.9b)$$

It follows from (7.8) that as  $t \rightarrow \infty$ ,  $u(x, t) \rightarrow 0$  for all  $x \in [0, 1]$  as  $\lambda_n^2 > 0$  for all  $n \geq 1$ . This is consistent with what we had discussed earlier, based on purely physical reasoning.

In this example the problem has only the nonzero initial condition. The equation and the boundary conditions are homogeneous. Consequently, the need to split the original problem using linearity and superposition does not arise.

We now consider a multidimensional problem with several different nonhomogeneities.

**Example 7.2** The two-dimensional parabolic problem of heat conduction in a rectangular slab is given by

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < x < a, \quad 0 < y < b \quad (7.10a)$$

This problem is subject to the initial condition

$$u(t = 0) = f(x, y) \quad (7.10b)$$

For convenience, we take all the boundary conditions to be Dirichlet

$$u(x = 0) = g(y) \quad (7.10c)$$

$$u(x = a) = 0 \quad (7.10d)$$

$$u(y = 0) = 0 \quad (7.10e)$$

$$u(y = b) = h(x) \quad (7.10f)$$

The domain of this system is shown in Fig. 7.3. There are three sources of nonhomogeneity

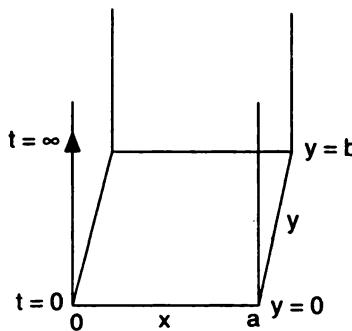


Fig. 7.3 Domain of system (7.10).

in (7.10), i.e.  $f(x, y)$ ,  $g(y)$  and  $h(x)$ . Consequently, we seek the solution  $u$  as a linear combination of three parts:

$$u(x, y, t) = u_1(x, y, t) + u_2(x, y, t) + u_3(x, y, t)$$

Each  $u_i$  is defined in the above such that it considers only one nonhomogeneity at a time. Using the subscript to denote derivative with respect to a variable, we write the problems defining the different  $u_i$ 's as

$$\left. \begin{array}{l} u_{1t} = u_{1xx} + u_{1yy} \\ u_1(t=0) = f(x, y) \\ u_1(0, y, t) = u_1(a, y, t) = 0 \\ u_1(x, 0, t) = u_1(x, b, t) = 0 \end{array} \right\} \quad (7.11b)$$

$$\left. \begin{array}{l} u_{2t} = u_{2xx} + u_{2yy} \\ u_2(x, y, 0) = 0, \quad u_2(0, y, t) = g(y), \quad u_2(a, y, t) = 0 \\ u_2(x, 0, t) = 0, \quad u_2(x, b, t) = 0 \end{array} \right\} \quad (7.11c)$$

$$\left. \begin{array}{l} u_{3t} = u_{3xx} + u_{3yy} \\ u_3(t=0) = 0, \quad u_3(0, y, t) = 0, \quad u_3(a, y, t) = 0 \\ u_3(x, 0, t) = 0, \quad u_3(x, b, t) = h(x) \end{array} \right\} \quad (7.11d)$$

The problem (7.11b) is a PIC problem, whereas (7.11c) and (7.11d) do not conform either to an EBC or a PIC. Equations (7.11c) and (7.11d) are further decomposed into an EBC and a PIC as discussed in Chapter 5. We write

$$u_2(x, y, t) = v_2(x, y) + w_2(x, y, t) \quad (7.12a)$$

where  $v_2(x, y)$  satisfies

$$\left. \begin{array}{l} v_{2xx} + v_{2yy} = 0 \\ v_2(0, y) = g(y), \quad v_2(a, y) = 0, \quad v_2(x, 0) = 0, \quad v_2(x, b) = 0 \end{array} \right\} \quad (7.12b)$$

and  $w_2(x, y, t)$  satisfies

subject to

$$\left. \begin{array}{l} w_{2t} = w_{2xx} + w_{2yy} \\ w_2(t=0) = -v_2(x, y) \\ w_2(0, y) = w_2(a, y) = 0 \\ w_2(x, 0) = w_2(x, b) = 0 \end{array} \right\} \quad (7.12c)$$

$w_2$  satisfies a PIC as in the case of  $u_1$ , and  $v_2$  satisfies an EBC. Problem (7.11d) can be similarly decomposed into an EBC and a PIC. These are problems in two spatial directions. To avoid being repetitious, we solve for  $u_1, v_2$  as a typical PIC and an EBC in this problem. We leave it to the student to proceed, and solve the Problem (7.10) in its entirety as an exercise by determining  $w_2, v_3, w_3$ .

*PIC of  $u_1$ .*  $u_1$  is dependent on two space variables  $x, y$  and time  $t$ . The solution  $u_1$  is sought as

$$u_1(x, y, t) = X(x) \cdot Y(y) \cdot T(t)$$

Substituting this expression in (7.11), we obtain

$$\frac{T'}{T} - \frac{X''}{X} = \frac{Y''}{Y}$$

Arguing as earlier, the two sides have to be equal to a constant, say  $\mu$ .

$$\frac{T'}{T} - \frac{X''}{X} = \frac{Y''}{Y} = \mu$$

$X(x), Y(y)$  satisfy homogeneous Dirichlet conditions. The constant has the physical significance of eigenvalues. In a typical problem we have to determine the eigenvalues—whether they are positive, negative or zero. Here both the boundary conditions are Dirichlet (and of course homogeneous). For convenience, the constant  $\mu$  is chosen as  $-\lambda^2$ . This choice is dictated by the fact that we need an eigenvalue problem in the  $y$ -direction, where the boundary conditions are homogeneous. Otherwise, we will have only the trivial solution  $Y(y) = 0$  which will yield  $u_1 = 0$ . This yields

$$Y'' + \lambda^2 Y = 0$$

subject to

$$Y(0) = 0, \quad Y(b) = 0 \quad (7.13)$$

The eigenvalues from Example 6.4 are  $\lambda_n^2 = n^2\pi^2/b^2$ , and the corresponding eigenfunctions are

$$Y_n(y) = A_n \sin(n\pi y/b) \text{ for } n = 1, 2, 3, \dots, \infty$$

The  $\lambda_n^2$  have now been determined uniquely. Substituting in (7.12) and rearranging, we get

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} - \lambda_n^2 = \gamma \quad (7.14)$$

Arguing as we did earlier, both these sides can be equal only if they are identically equal to a constant. The initial condition on  $T$  is nonzero and  $X$  is subject to homogeneous boundary conditions

in the  $x$ -direction. To obtain a nonzero value for  $X(x)$ , we need an eigenvalue problem in the  $x$ -direction as well.  $X(x)$  is obtained from

$$X''(x) - (\lambda_n^2 + \gamma) X = 0,$$

subject to

$$X(0) = X(a) = 0$$

For convenience, and to reduce to standard form, define

$$\alpha^2 = -(\lambda_n^2 + \gamma)$$

We now obtain

$$X''(x) + \alpha^2 X = 0$$

subject to

$$X(0) = X(a) = 0 \quad (7.15)$$

This is similar to (7.13) and admits the eigenvalues

$$\alpha_m^2 = m^2 \pi^2 / a^2$$

and the eigenfunctions

$$X_m(x) = B_m \sin(m\pi x/a), \quad m = 1, 2, 3, \dots$$

The index  $m$  has been chosen distinct from the index used in defining  $Y(y)$ . This is necessary because we associate with each integer value  $n$  denoting an eigenvalue in the  $y$ -direction an infinite number of eigenfunctions and eigenvalues in the  $x$ -direction represented by  $m$ . The time evolution is governed by the constant  $\gamma$  (see 7.14). This is defined as

$$\gamma_{m,n} = -\lambda_n^2 - \alpha_m^2 = -\left(\frac{n^2}{b^2} + \frac{m^2}{a^2}\right)\pi^2 \quad (7.16a)$$

The double subscript on  $\gamma$  indicates that there is a double infinity of the  $\gamma$ 's present. The time dependence is governed by

$$T_{m,n}(t) = C_{m,n} e^{\gamma_{m,n} t} \quad (7.16b)$$

The constant  $C_{m,n}$  depends on both the indices  $m$  and  $n$ . The solution  $u_1$  is sought by superposing the different eigenfunctions in  $x$  and  $y$  directions, which satisfy the homogeneous equation and boundary conditions. This yields

$$u_1(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} C_{m,n} e^{\gamma_{m,n} t} B_m \sin(m\pi x/a) \sin(n\pi y/b) A_n \quad (7.17a)$$

Defining a new constant,  $D_{m,n}$ , which depends on both the subscripts  $m, n$ , as

$$D_{m,n} = C_{m,n} B_m A_n$$

we get

$$u_1(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} D_{m,n} \exp\left[-\left(\frac{m^2 \pi^2}{a^2} + \frac{n^2 \pi^2}{b^2}\right)t\right] \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b} \quad (7.17b)$$

The coefficients  $D_{m,n}$  are obtained from the initial condition of  $u_1$ . Equation (6.4a) is extended to this two-dimensional problem by defining the inner-product as

$$\langle f(x, y), g(x, y) \rangle = \int_0^a dx \int_0^b dy f(x, y)g(x, y)$$

Here,  $f(x, y)$ ,  $g(x, y)$  are real-valued functions. The orthogonality of the sines implies

$$D_{m,n} = \frac{\int_0^a \int_0^b u_0(x, y) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} dx dy}{\int_0^a \int_0^b \sin^2 \frac{m\pi x}{a} \sin^2 \frac{n\pi y}{b} dx dy} \quad (7.17c)$$

The coefficients of the eigenfunctions expansion in (7.17b) are determined in a way similar to the eigenvector expansions in Chapter 4. This has been formalised by generalising the definition of the inner-product, introducing the notion of an operator, and exploiting the orthogonality property between the eigenfunctions of the operator.

In this example, the solution is sought as a summation over two indices. This accounts for the eigenvalue problems which occur in the two directions. This has been made possible only because we have used linearity and superposition to decompose a problem such that only one nonhomogeneity is considered at a time. (The number of summation signs is one less than the number of independent variables in a problem.)

The solution of EBC of  $v_2$ , is sought as  $v_2(x, y) = X(x) Y(y)$ . Substituting in these equations, we obtain, as earlier,

$$\frac{Y''(y)}{Y(y)} = \frac{-X''(x)}{X(x)} = \mu \quad (7.18)$$

The boundary conditions in the  $y$ -direction are homogeneous, whereas in the  $x$ -direction they are not. The constant  $\mu$  has to be chosen such that we obtain an eigenvalue problem in the  $y$ -direction; otherwise the resulting solution will be the trivial solution  $Y(y) = 0$ . The boundary conditions in the  $y$ -direction are homogeneous and **Dirichlet**. The choice  $\mu = \lambda^2$  yields only the trivial solution  $Y(y) = 0$ , and for  $\mu = -\lambda^2$ , we have the eigenfunctions  $Y_n(y) = A_n \sin(n\pi y/b)$ , corresponding to the eigenvalues  $\lambda_n^2 = (n^2\pi^2/b^2)$ .

Substituting this value in (7.18), we have

$$X''(x) - (n^2\pi^2/b^2) X(x) = 0 \quad (7.19)$$

subject to  $X(a) = 0$ . The boundary condition at  $x = 0$  cannot be used now as  $X(x)$  is independent of  $y$ . The solution to (7.19) is of the form

$$X_n(x) = A_n \sinh(n\pi x/b) + B_n \cosh(n\pi x/b) \quad (7.20a)$$

A more convenient choice of the independent functions enables us to determine the constants elegantly. This choice is guided by the boundary conditions:

$$X_n(x) = C_n \sinh \frac{n\pi(x-a)}{b} + D_n \cosh \left[ \frac{n\pi}{b} (x-a) \right] \quad (7.20b)$$

Now  $X_n(a) = 0$ , yields  $D_n = 0$ .

The solution  $v_2(x, y)$  now is of the form

$$v_2(x, y) = \sum_{n=1}^{\infty} E_n \sinh \left[ \frac{n\pi}{b} (x-a) \right] \sin \frac{n\pi y}{b}$$

where  $E_n = C_n A_n$ . The constant  $E_n$  is obtained by imposing the boundary condition on  $v_2(x = 0, y)$ . Invoking the orthogonality property again, we have

$$E_n = \frac{\int_0^b g(y) \sin \frac{n\pi y}{b} dy}{\sinh\left(-\frac{n\pi a}{b}\right) \int_0^b \sin^2 \frac{n\pi y}{b} dy}$$

We leave it as an exercise to the reader to obtain the above solution using (7.20a).

## 7.2 CYLINDRICAL COORDINATES

The heat conduction equation in an infinitely long cylinder is modelled by the one-dimensional equation

$$\frac{\partial u}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) \quad (7.21a)$$

In this equation  $u$  is independent of  $z$ , because of the assumption of infinite length. This allows us to neglect the edge effects. In the bulk of the cylinder,  $u$  does not vary with axial position for a fixed  $t$ ,  $r$ ,  $\theta$ , and so (7.21a) is a good model here. The  $\theta$  dependence in (7.21a) is neglected. This is the assumption of axi-symmetry. It is justified only if the initial temperature profile, and any other sources of nonhomogeneities are  $\theta$ -symmetric (see Weinberger, 1965).

**Example 7.3** Determine the solution to equation (7.21a) subject to the initial condition

$$u(t=0) = u_0(r) \quad (7.21b)$$

and the boundary conditions

$$u(r=R_1) = 0 \quad (7.21c)$$

$$u(r=0) = \text{bounded} \quad (7.21d)$$

The only nonhomogeneity occurs in the initial condition of this PIC. It is therefore not necessary to decompose this problem into sub-problems. The function  $u(r, t)$  is obtained as a product of two functions

$$u(r, t) = R(r) T(t) \quad (7.22a)$$

Substituting in (7.21a), we get

$$\frac{T'(t)}{T(t)} = \frac{1}{rR} \left( \frac{d}{dr} r \frac{dR(r)}{dr} \right) = \mu \quad (7.22b)$$

The boundary conditions are homogeneous in the  $r$ -direction. The constant  $\mu$  is chosen as  $-\lambda^2$  to obtain an eigenvalue problem in the  $r$ -direction.  $R(r)$  is governed by

$$\frac{d}{dr} \left( r \frac{dR}{dr} \right) + \lambda^2 r R = 0 \quad (7.23a)$$

This ordinary differential equation is subject to

$$R(R_1) = 0 \quad (7.23b)$$

$$R(0) = \text{bdd} \quad (7.23c)$$

Equations (7.23a)–(7.23c) constitute the Sturm-Louville problem with  $x = r$ ,  $p(x) = r$ , and  $\omega(x) = r$ . The solutions to (7.23a) are

$$R(r) = C_1 J_0(\lambda r) + C_2 Y_0(\lambda r) \quad (7.24)$$

The condition at  $r = 0$  implies  $C_2 = 0$ , as  $Y_0(0)$  is unbounded. The eigenvalues  $\lambda_n^2$  are obtained from (7.23b) as roots of  $J_0(\lambda R_1) = 0$ . This equation has an infinite number of zeroes  $\lambda$ . The eigenvalues  $\lambda_n^2$  have to be obtained numerically as in Example 7.1.

The corresponding eigenfunction is  $R_n(r) = C_{1,n}J_0(\lambda_n r)$ . Substituting this  $\lambda_n$  in (7.22b),  $T_n(t)$  is obtained as

$$T_n(t) = T_n(0) \exp(-\lambda_n^2 t)$$

The solution  $u(r, t)$  is represented by using superposition as

$$u(r, t) = \sum_{n=1}^{\infty} C_n \exp(-\lambda_n^2 t) (J_0(\lambda_n r)) \quad (7.25)$$

with  $C_n = T_n(0) C_{1,n}$ .

Once again from the initial condition,  $C_n$  is obtained as

$$u_0(r) = \sum_{n=1}^{\infty} C_n J_0(\lambda_n r)$$

The coefficients  $C_n$  are determined using the orthogonality property of  $J_0(\lambda_n r)$  and  $J_0(\lambda_m r)$ . This orthogonality is only with respect to the weighting function  $r$ . This yields

$$C_n = \frac{\int_0^{R_1} u_0(r) r J_0(\lambda_0 r) dr}{\int_0^{R_1} r J_0^2(\lambda_n r) dr}$$

In particular, it must be emphasised that

$$C_n \neq \frac{\int_0^{R_1} u_0(r) J_0(\lambda_n r) dr}{\int_0^{R_1} J_0^2(\lambda_n r) dr}$$

**Example 7.4** Solve the transient heat conduction equation in two dimensions in polar coordinates.

$$\frac{\partial u}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} \quad (7.26a)$$

subject to the initial condition

$$u(r, \theta, t = 0) = u_0(r, \theta) \quad (7.26b)$$

and the boundary conditions

$$u(1, \theta, t) = 0 \quad (7.26c)$$

$$u(0, \theta, t) = bdd \quad (7.26d)$$

$$u(r, \pi, t) = u(r, -\pi, t) \quad (7.26e)$$

$$\frac{\partial u}{\partial \theta}(r, \pi, t) = \frac{\partial u}{\partial \theta}(r, -\pi, t) \quad (7.26f)$$

Equations (7.26) constitute a more general form of (7.21), where the variation of  $u$  with respect to  $\theta$  is incorporated. This occurs because the initial profile in (7.26b) is dependent on  $\theta$ , in contrast

to (7.21b). The conditions (7.26e) and (7.26f) are periodicity conditions in the  $\theta$ -direction which the solution must satisfy at all  $r, t$ .  $u$  is sought as a product of three functions

$$u = R(r) \cdot \theta(\theta) \cdot T(t)$$

Substituting in (7.26a) and rearranging, we obtain

$$r^2 \left[ \frac{T'(t)}{T(t)} - \frac{1}{rR} \frac{d}{dr} \left( r \frac{dR}{dr} \right) \right] = \frac{\theta''}{\theta} = \mu \quad (7.27)$$

The boundary conditions in the  $\theta$ -direction (7.26e) and (7.26f) are linear and homogeneous. The constant  $\mu$  is chosen so that we obtain an eigenvalue problem in the  $\theta$ -direction where we have homogeneous boundary conditions. This leads us to choose  $\mu = -\alpha^2$ . From Chapter 6 this yields the eigenvalues  $\mu_n = n^2$ , where  $n = 0, 1, 2, \dots$ , and  $\theta_n = A_n \cos n\theta + B_n \sin n\theta$ .

The boundary conditions in the  $r$ -direction are also homogeneous. So we need an eigenvalue problem in this direction also. From (7.27), on rearranging, we obtain

$$\frac{T'(t)}{T(t)} = \frac{1}{rR} \frac{d}{dr} \left( r \frac{dR}{dr} \right) - \frac{n^2}{r^2} = -\lambda^2 \quad (7.28)$$

The choice of the separation constant as  $-\lambda^2$  reduces the  $R(r)$  equation to the standard Bessel function equation

$$\frac{d}{dr} \left( r \frac{dR}{dr} \right) - \frac{n^2}{r} R + \lambda^2 r R = 0 \quad (7.29a)$$

$$R(1) = 0, \quad R(0) = \text{bdd} \quad (7.29b,c)$$

The solution  $R(r)$ , for each  $n$  is of the form

$$R(r) = C_1 J_n(\lambda r) + C_2 Y_n(\lambda r)$$

The boundedness condition (7.29c) implies  $C_2 = 0$  as  $Y_n(0)$  is unbounded at  $r = 0$ . The separation constant  $\lambda$  is obtained from (7.29b) as the zeroes of

$$J_n(\lambda_{m,n}) = 0 \quad (7.30)$$

For each  $n$ , i.e. the eigenvalue in the  $\theta$ -direction, (7.30) has an infinite number of zeroes. This is represented by the index  $m$  which goes from 1 to  $\infty$  for every  $n$ . The eigenfunctions in the  $r$ -direction now are

$$R_{m,n}(r) = C_{m,n} J_n(\lambda_{m,n} r)$$

By  $\lambda_{m,n}$  we mean the  $m$ th zero of the Bessel function of the first kind of order  $n$ .

Clearly, from (7.28)

$$T_{m,n}(t) = T_{m,n}(0) \exp(-\lambda_{m,n}^2 t)$$

The solution  $u(r, \theta, t)$  follows from linearity and superposition as

$$u(r, \theta, t) = \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} \exp(-\lambda_{m,n}^2 t) (D_{m,n} \sin n\theta + E_{m,n} \cos n\theta) J_n(\lambda_{m,n} r) \quad (7.31)$$

where

$$D_{m,n} = T_{m,n}(0) B_n C_{m,n}$$

$$E_{m,n} = T_{m,n}(0) A_n C_{m,n}$$

Remembering that the Bessel function of the first kind are orthogonal to each other only with respect to the weighting function  $r$  (see Weinberger, 1965) we obtain the coefficients as

$$D_{m,n} = \frac{\int_0^1 dr \int_{-\pi}^{\pi} d\theta u_0(r, \theta) r J_n(\lambda_{m,n} r) \sin n\theta}{\int_0^1 dr \int_{-\pi}^{\pi} d\theta r J_n^2(\lambda_{m,n} r) \sin^2 n\theta}$$

$$E_{m,n} = \frac{\int_0^1 dr \int_{-\pi}^{\pi} d\theta u_0(r, \theta) r J_n(\lambda_{m,n} r) \cos n\theta}{\int_0^1 dr \int_{-\pi}^{\pi} d\theta r J_n^2(\lambda_{m,n} r) \cos^2 n\theta}$$

when  $u_0$  is independent of  $\theta$ , it follows that

$$D_{m,n} = 0 \text{ for } n = 0, 1, 2, \dots, \infty$$

$$E_{m,n} = 0 \text{ for } n = 1, 2, \dots, \infty$$

only  $E_{m,0} \neq 0$ . This reduces (7.31) to

$$u(r, t) = \sum_{m=1}^{\infty} \exp(-\lambda_{m,0}^2 t) J_0(\lambda_{m,0} r) E_{m,0}$$

This expression is identical to (7.25), where, to begin with, we considered the problem to be one-dimensional in space. It is therefore advantageous to use the physical insight into the model to simplify it to the maximum extent possible. This will reduce the algebraic complexity of the solution procedure. In our next example we see a situation where the modified Bessel's function of the first and second kind arise.

**Example 7.5** Consider a cylindrical pellet of finite length  $0 < z < 1$ . The temperature profile on the curved surface is maintained at  $f(z)$ . The temperature at  $z = 0$ ,  $z = 1$  is assumed to be zero.

The steady-state temperature profile for this problem is governed by the equation

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2} = 0 \quad (7.32a)$$

subject to

$$u(z = 0) = u(z = 1) = 0 \quad (7.32b)$$

$$u(r = 0) = \text{bounded} \quad (7.32c)$$

$$u(r = 1) = f(z) \quad (7.32d)$$

For the problem to be well defined,  $f(0) = f(1) = 0$ . This will ensure that  $u(r = 1, z = 0)$  and  $u(r = 1, z = 1)$ , as obtained from (7.32b)–(7.32d), are all equal. As  $f(z)$  is independent of  $\theta$  in (7.32d), we expect the solution to be  $\theta$ -independent. Imposition of this condition of axisymmetry simplifies (7.32a) as

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{\partial^2 u}{\partial z^2} = 0 \quad (7.33a)$$

We seek  $u = R(r)Z(z)$ . Substituting this in (7.33a), we get

$$\frac{1}{r} \frac{1}{R} \frac{d}{dr} \left( r \frac{dR}{dr} \right) = -\frac{Z''}{Z} = \mu \quad (7.33b)$$

Clearly,  $R(0)$  must be bounded and  $Z$  satisfies homogeneous Dirichlet boundary conditions at  $z = 0, 1$ . We choose  $\mu = \lambda^2$ , to obtain a standard eigenvalue problem in the  $z$ -direction (as we have homogeneous conditions in this direction).

The eigenvalues clearly are  $\lambda_n^2 = n^2\pi^2$ , and the corresponding eigenfunctions are  $Z_n = A_n \sin(n\pi z)$ . Using this value of  $\mu_n$  in (7.33b),  $R(r)$  is obtained from

$$\frac{d}{dr} \left( r \frac{dR}{dr} \right) - \lambda_n^2 r R = 0 \quad (7.34)$$

This is not an eigenvalue problem, since (a)  $R$  does not satisfy the homogeneous conditions at  $R = 1$ , and (b) the weighting function  $w(r) < 0$ . Besides, there are no free parameters or eigenvalues to be found. The solution to (7.34) as seen in Chapter 6 is

$$R_n(r) = C_{1n} I_0(n\pi r) + C_2 K_0(n\pi r)$$

The boundedness of  $R_n(0)$  implies  $C_2 = 0$  as  $K_0(0)$  is unbounded. The solution  $u(r, z)$  is now obtained as

$$u(r, z) = \sum_{n=1}^{\infty} C_n I_0(n\pi r) \sin(n\pi z)$$

where  $C_n = C_{1n} A_n$ . The coefficient  $C_n$  is obtained from (7.32d) as

$$C_n = \frac{\int_0^1 f(z) \sin(n\pi z) dz}{I_0(n\pi) \int_0^1 \sin^2(n\pi z) dz}$$

### 7.3 SPHERICAL COORDINATE SYSTEMS

The Laplacian operator in spherical coordinate systems is given by

$$\nabla^2 u = \left( \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2} \right) \quad (7.35)$$

In a spherical region of radius  $R$ , these coordinates vary from  $0 < r < R$ ,  $0 < \theta < \pi$ ,  $-\pi < \phi < \pi$ . The method of separation of variables can be used to solve homogeneous equation problems here as well. We now discuss a few examples.

#### Example 7.6

$$\nabla^2 u = 0 \quad \text{in } 0 < r < 1$$

with  $u(r = 1) = f(\theta)$ .

The nonhomogeneity  $f(\theta)$  is independent of  $\phi$ . Hence we expect our solution  $u$  to be independent of  $\phi$ . This is called  $\phi$ -symmetry or axisymmetry. We now solve the two-dimensional problem

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial u}{\partial \theta} \right) = 0 \quad (7.36a)$$

Seeking  $u(r, \theta) = R(r) \cdot T(\theta)$ , we have

$$\frac{1}{R} \frac{d}{dr} \left( r^2 \frac{dR}{dr} \right) = - \frac{1}{T \sin \theta} \frac{d}{d\theta} \left( \sin \theta \frac{dT}{d\theta} \right) = \lambda \quad (7.36b)$$

The boundary condition in the  $r$ -direction is nonhomogeneous. The conditions in the  $\theta$ -direction are not yet specified. Denoting the separation constant as  $\lambda$ , we have, in the  $\theta$ -direction, the equations

$$\frac{d}{d\theta} \left( \sin \theta \frac{dT}{d\theta} \right) + \lambda \sin(\theta) T = 0 \quad (7.37a)$$

We impose the boundary conditions in the  $\theta$ -direction as

$$T(0) = \text{bounded} \quad (7.37b)$$

$$T(\pi) = \text{bounded} \quad (7.37c)$$

$$T'(0) = \text{bounded} \quad (7.37d)$$

$$T'(\pi) = \text{bounded} \quad (7.37e)$$

These conditions arise because the equation is singular at its end-points. Here,  $p(0) = p(\pi) = 0$ , where  $p(\theta)$  is  $\sin \theta$ , see (6.17). Besides  $p(\theta) > 0$ ,  $\omega(\theta) > 0$  in  $0 < \theta < \pi$ , and so we can exploit the results of Sturm-Louiville theory in Chapter 6.

Substituting  $t = \cos \theta$ , we have

$$\frac{d}{dt} \left( (1 - t^2) \frac{dP(t)}{dt} \right) + \lambda P(t) = 0, \quad -1 < t < 1 \quad (7.38)$$

where  $P(t) = T(\theta)$ . From the results in Chapter 6, the solution is

$$P(t) = c_1 P_n(t) + c_2 Q_n(t)$$

where  $P_n$  is the Legendre polynomial of degree  $n$  and  $Q_n$  is the Legendre function of degree  $n$ .

The eigenvalues are  $\lambda_n = n(n + 1)$ ,  $n = 0, 1, 2, \dots$ ,  $Q_n$  is unbounded at  $t = \pm 1$ . Consequently,  $c_2$  is zero.

In the  $r$ -direction from (7.36b), we now have

$$\frac{d}{dr} \left( r^2 \frac{dR_n}{dr} \right) - \lambda_n R_n = 0 \quad (7.39a)$$

$$r^2 \frac{d^2 R_n}{dr^2} + 2r \frac{dR_n}{dr} - n(n + 1) R_n = 0 \quad (7.39b)$$

This equation is called Euler's equation. It admits solutions of the form

$$R(r) = r^\alpha$$

Substituting this form in (7.39b), we obtain  $\alpha = n$ , or  $-(n + 1)$ . Now,

$$R_n(r) = c_{1n} r^n + c_{2n} r^{-n-1}$$

To obtain a solution bounded at  $r = 0$ , we set  $c_{2n} = 0$ . This yields

$$u_n(r, \theta) = c_{1n} r^n P_n(\cos \theta) \quad (7.40a)$$

This solution satisfies the boundary conditions at  $r = 0$ , those in the  $\theta$ -direction (7.37b) and (7.37e) as well as the homogeneous equation (7.36a) for  $n = 0, 1, 2, \dots$ . Hence the superposition of all solutions for different  $n$  also satisfies these conditions. So we have

$$u(r, \theta) = \sum_{n=0}^{\infty} c_{1n} r^n P_n(\cos \theta) \quad (7.40b)$$

The boundary condition at  $r = 1$  implies

$$f(\theta) = \sum_{n=0}^{\infty} c_{1n} P_n(\cos \theta) \quad (7.41)$$

From the orthogonality of the eigenfunctions in the  $\theta$ -direction, we have

$$c_{1n} = \frac{\int_0^\pi f(\theta) P_n(\cos \theta) \sin \theta d\theta}{\int_0^\pi P_n^2(\cos \theta) \sin \theta d\theta} \quad (7.42)$$

**Example 7.7** Re-work the above example when the nonhomogeneity  $f$  is dependent on  $\phi$ ,

$$u(1, \theta, \phi) = f(\theta, \phi)$$

In this problem we cannot seek an axi-symmetric solution. Seeking  $u$  as

$$u(r, \theta, \phi) = R(r) \Theta(\theta) \phi(\phi)$$

we have

$$\left( \frac{(r^2 R')'}{R} + \frac{(\sin(\theta) \Theta')'}{\Theta \sin \theta} \right) \sin^2 \theta = -\frac{\phi''}{\phi} = \mu^2 \quad (7.43)$$

The boundary conditions in the  $\phi$ -direction emanate from the periodicity conditions. This is similar to what we had in the  $\theta$ -direction of cylindrical coordinates.

$$\phi(\pi) = \phi(-\pi), \phi'(\pi) = \phi'(-\pi)$$

The choice of the separation constant is such that we first obtain an eigenvalue problem in the  $\phi$ -direction. Here the eigenvalues are

$$\mu^2 = m^2, m = 0, 1, 2, \dots, \infty$$

$$\phi_m(\phi) = A_m \sin m\phi + B_m \cos m\phi$$

Separating (7.43) in the  $r$ -direction, it follows that

$$-\frac{(r^2 R')'}{R} = \frac{(\sin(\theta) \Theta')'}{\Theta \sin \theta} - \frac{m^2}{\sin^2 \theta} = -\lambda^2 \quad (7.44)$$

In the  $\theta$ -direction we need an eigenvalue problem. This is because we have homogeneous boundary conditions in this direction. This leads to the choice of the separation constant as in (7.44), resulting in

$$(\sin(\theta) \Theta')' - \frac{m^2 \Theta}{\sin \theta} + \lambda^2 \sin(\theta) \Theta = 0 \quad (7.45)$$

Substituting  $t = \cos \theta$ ,  $P(t) = \Theta(\theta)$ , we obtain

$$\frac{d}{dt} \left( (1 - t^2) \frac{dP}{dt} \right) - \frac{m^2}{1 - t^2} P + \lambda^2 P = 0$$

The solution to this equation is

$$P(\cos \theta) = c_1 P_n^m (\cos \theta) + c_2 Q_n^m (\cos \theta)$$

as we have seen in Chapter 6. The corresponding eigenvalues  $\lambda^2$  are of the form

$$\lambda^2 = n(n + 1), \quad n = m, m + 1, \dots$$

As  $Q_n^m (\cos \theta)$  is unbounded at  $\theta = 0, \pi$ , we have  $c_2 = 0$ .

The solutions in the  $r$ -direction can be again obtained from Euler's equation as

$$R_{m,n}(r) = c_{1m,n} r^n + c_{2m,n} r^{-(n+1)} \quad (7.46)$$

From the boundedness of the solution at  $r = 0$ , we have  $c_{2m,n} = 0$ . The solution is obtained from superposition, see Weinberger (1965), Kreyszig (1982) as

$$u(r, \theta, \phi) = \sum_{n=m}^{\infty} \sum_{m=0}^{\infty} r^n P_n^m (\cos \theta) [c_{m,n} \sin m\phi + D_{m,n} \cos m\phi]$$

The constants are evaluated as

$$C_{m,n} = \frac{\int_{-\pi}^{\pi} d\theta \int_0^{\pi} d\phi f(\theta, \phi) P_n^m (\cos \theta) \sin m\phi \sin \theta}{\int_{-\pi}^{\pi} d\theta \int_0^{\pi} d\phi P_n^{m^2} (\cos \theta) \sin \theta \sin^2 m\phi} \quad (7.47a)$$

$$D_{m,n} = \frac{\int_{-\pi}^{\pi} d\theta \int_0^{\pi} d\phi f(\theta, \phi) P_n^m (\cos \theta) \cos m\phi \sin \theta}{\int_{-\pi}^{\pi} d\theta \int_0^{\pi} d\phi P_n^{m^2} (\cos \theta) \sin \theta \cos^2 m\phi} \quad (7.47b)$$

## 7.4 FOURIER SERIES AND FINITE FOURIER TRANSFORMS

In the method of separation of variables, as we have just seen, we represent the solution in terms of an infinite series. This series is generated by the eigenfunctions of the operator occurring in the problem. Such a series is called a *Fourier series*. The constant coefficients mentioned in Chapter 6 for functions of a single variable are called the Fourier coefficients. This is a more general version of the classical Fourier series which is used to represent periodic functions. We will see how this generates the concept of the Fourier transform. This transform enables us to reduce a partial differential equation in a bounded domain to an infinite system of ordinary differential equations. The Fourier transform can also be used and, in fact, is normally used in solving problems in unbounded domains. We will first discuss the transform for finite spatial domains and then extend it to spatially unbounded (infinite) domains (see Weinberger (1965) and Churchill (1963)).

Consider a continuous function  $f(x)$  defined in the interval  $(-\pi, \pi)$ . Let the function be periodic with period  $2\pi$ . This implies

$$f(\pi) = f(-\pi) \quad (7.48a)$$

$$f'(\pi) = f'(-\pi) \quad (7.48b)$$

The Fourier series of  $f(x)$  is

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \quad (7.49a)$$

Here we are representing  $f(x)$  in terms of a complete set, i.e. the eigenfunction set of (6.40). This is a complete set since these are eigenfunctions of a self-adjoint operator. If a function is defined only in  $(-\pi, \pi)$ , it can always be extended as a periodic function over the entire interval  $(-\infty, \infty)$ . The series (7.49a) would be an approximate representation of such a function. Using

$$\cos nx = (e^{inx} + e^{-inx})/2, \quad \sin nx = (e^{inx} - e^{-inx})/2i$$

we rewrite (7.49a) as

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{-inx} \quad (7.49b)$$

where

$$c_n = \begin{cases} (a_n + ib_n)/2 & \text{for } n \geq 0 \\ (a_{-n} - ib_{-n})/2 & \text{for } n \leq 0 \end{cases} \quad (7.49c)$$

The coefficients  $a_n$ ,  $b_n$ ,  $c_n$  are obtained from

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx \quad (7.50a)$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx \quad (7.50b)$$

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{inx} dx \quad (7.50c)$$

The above representation is similar to seeking a vector  $u \in \mathbb{R}^n$  in terms of a basis set as in Chapter 4. In (7.49a) we have represented  $f(x)$  in  $(-\pi, \pi)$  in terms of the basis set  $(\sin x, \sin 2x, \dots, 1, \cos x, \cos 2x, \dots, \cos nx)$ . This is the set of eigenfunctions of (6.40), see (6.42). The coefficients  $a_n$ ,  $b_n$  are the components of  $f(x)$  in terms of this basis.

Consider now the case where  $f(x)$  is an odd function. Here  $f(-x) = -f(x)$ . Clearly, for such a function

$$a_n = 0 \quad \text{for } n = 0, 1, 2\dots$$

$$b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin(nx) dx$$

$$= \frac{2}{\pi} \int_0^{\pi} f(x) \sin(nx) dx$$

The representation (7.49a) is now reduced to

$$f(x) = \sum_{n=1}^{\infty} b_n \sin nx \quad (7.51)$$

This is called a **Fourier sine series**. Such a series exists for  $f(x)$  defined on  $(0, \pi)$ . Here the function  $f(x)$  is extended in  $(-\pi, 0)$  such that  $f(x)$  is odd. The set  $(\sin x, \sin 2x, \dots, \sin nx)$  is now the basis set. It is a complete set, i.e., every series of the form (7.51) converges to  $f(x)$  in the mean as  $n \rightarrow \infty$  for  $f(x)$  in  $L^2(0, \pi)$ . This basis set is the set of eigenfunctions of a self-adjoint operator as seen earlier and so we expect it to form a complete set.

When  $f(x)$  is an even function

$$f(-x) = f(x)$$

Here

$$a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos(nx) dx$$

$$b_n = 0$$

The Fourier series now reduces to

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nx \quad (7.52)$$

This is a **Fourier cosine series** (Churchill, 1963). Such a series is again defined for functions on  $(0, \pi)$ . Here the function is extended on  $(-\pi, 0)$  such that it is an even function. The basis set here is  $(1, \cos x, \cos 2x, \dots, \cos nx)$ . This is again the set of eigenfunctions of a self-adjoint operator and is therefore a complete set.

It is this property of completeness which allows us to represent any function as a linear combination of that set and assures us of convergence in the mean as  $n \rightarrow \infty$ . As explained earlier, this is analogous to the linear independence of a finite number of vectors in  $\mathbb{R}^n$ . When a function is defined on  $(0, \pi)$ , we are only interested in representing the function in this range. It is a matter of convenience as to whether we extend the function as an odd function or an even function and get the appropriate representation. Choosing a sine series or a cosine series is analogous to choosing between different bases to represent a function.

In separation of variables we were interested in the solution of

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (7.53)$$

subject to

$$u(x = \pi) = 0, \quad u(x = 0) = 0$$

$$u(t = 0) = u_0(x)$$

The solution can be represented as

$$u(x, t) = \sum_{n=1}^{\infty} T_n(t) \sin nx \quad (7.54)$$

Similarly, the solution to

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = 0 \quad (7.55)$$

can be written as

$$u(r, \theta) = \frac{a_0(r)}{2} + \sum_{n=1}^{\infty} [a_n(r) \cos n\theta + b_n(r) \sin n\theta] \quad (7.56)$$

The coefficients  $T_n(t)$ ,  $a_n(r)$ ,  $b_n(r)$  are called finite Fourier transform and are governed by ordinary differential equations as they are functions of a single variable. These form a generalisation of the Fourier coefficients of a function of a single variable. Extending the definition of the Fourier coefficients of a function of one variable to (7.54), we obtain  $T_n(t)$  as

$$T_n(t) = \frac{2}{\pi} \int_0^\pi u(x, t) \sin(nx) dx \quad (7.57a)$$

$T_n(t)$  is a function of  $n$  and  $t$  and is called the finite sine transform of  $u(x, t)$ . This is analogous to the  $c_n(t)$  defined in Chapter 4 while solving IVPs. The only restriction we have imposed on  $u(x, t)$  is that it should be continuous. If  $\partial^2 u / \partial x^2$  is also continuous (i.e.,  $u$  is twice differentiable with respect to  $x$  and the second derivative is continuous), its finite Fourier sine transform is defined as

$$\frac{2}{\pi} \int_0^\pi \frac{\partial^2 u}{\partial x^2} \sin(nx) dx = -n^2 T_n(t) \quad (7.57b)$$

This result is obtained from an integration by parts twice and using the boundary conditions on  $u$  at  $x = 0$  and  $\pi$ .

The finite Fourier sine transform of  $\partial u / \partial t$  is

$$\frac{2}{\pi} \int_0^\pi \frac{\partial u}{\partial t} \sin(nx) dx = \frac{d}{dt} T_n(t) \quad (7.57c)$$

This is obtained by interchanging the order of integration and differentiation. Equations (7.54) and (7.57a) are said to constitute a transform pair. The ordinary differential equation governing  $T_n(t)$  is obtained by taking the finite sine transform of (7.53). This yields

$$\frac{dT_n(t)}{dt} + n^2 T_n(t) = 0, \quad n = 1, 2, \dots, \infty \quad (7.58)$$

which in turn yields

$$T_n(t) = e^{-n^2 t} T_n(0)$$

Substituting in (7.54), we get

$$u(x, t) = \sum_{n=1}^{\infty} T_n(0) e^{-n^2 t} \sin nx$$

where

$$T_n(0) = \frac{\int_0^\pi u_0(x) \sin nx}{\int_0^\pi \sin^2 nx}$$

The partial differential equation (7.53) has been now converted to an infinite system of ordinary differential equations (7.58). These ordinary differential equations are decoupled, since our basis set is orthogonal. We have an infinite system of equations as our space is now infinite dimensional. This infinite system is countably infinite since we have a spatially bounded region.

The coefficients  $\{a_n(r), b_n(r)\}$  in (7.56) are obtained from

$$a_n(r) = \frac{1}{\pi} \int_{-\pi}^{\pi} u(r, \theta) \cos(n\theta) d\theta \quad (7.59a)$$

$$b_n(r) = \frac{1}{\pi} \int_{-\pi}^{\pi} u(r, \theta) \sin(n\theta) d\theta \quad (7.59b)$$

The set  $\{a_n(r), b_n(r)\}$  is called the finite Fourier transform of  $u(r, \theta)$ . These coefficients are again governed by ordinary differential equations. Taking the Fourier transform on both sides of (7.55), we obtain

$$\frac{d^2 a_n}{dr^2} + \frac{1}{r} \frac{da_n}{dr} - \frac{n^2}{r^2} a_n = 0 \quad (7.60a)$$

$$\frac{d^2 b_n}{dr^2} + \frac{1}{r} \frac{db_n}{dr} - \frac{n^2}{r^2} b_n = 0 \quad (7.60b)$$

for  $n = 1, 2, 3, \dots$

The solution to this infinite system of ordinary differential equations is

$$a_n = r^m = b_n$$

where

$$m(m - 1) + m - n^2 = 0$$

This yields  $m = n$  or  $m = -(n + 1)$ . Since we are interested only in bounded solutions, we argue as earlier and obtain

$$a_n(r) = r^n C_n, \quad b_n(r) = r^n D_n$$

Substituting in (7.50), we obtain

$$u(r, \theta) = \frac{C_0}{2} + \sum_{n=1}^{\infty} C_n r^n \cos n\theta + D_n r^n \sin n\theta$$

Thus we see that finite Fourier transforms are just a generalisation of the Fourier coefficients (refer Churchill, 1963). Whenever the homogeneous equation can be solved using the method of separation of variables, we can use these finite Fourier transforms to reduce the partial differential equation to an infinite system of homogeneous ordinary differential equations. This method can be used while solving nonhomogeneous equation problems as well, see Chapter 4. We illustrate this using an example.

### Example 7.8 Solve

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2} + f(x, t), \quad 0 < x < \pi$$

subject to

$$T(x, 0) = 0, \quad T(x, t) = 0, \quad T(\pi, t) = 0$$

We seek

$$T(x, t) = \sum_{n=1}^{\infty} T_n(t) \sin nx$$

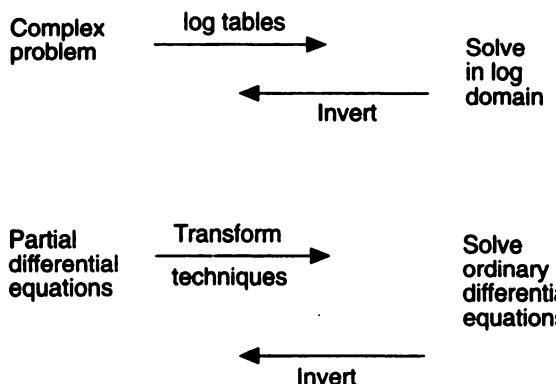
and write

$$f(x, t) = \sum_{n=1}^{\infty} f_n(t) \sin nx$$

where  $f_n$  is the finite Fourier transform of  $f(x, t)$ . The choice of the basis set as  $\sin nx$  is because this is the set of eigenfunctions of our problem. Taking the Fourier transform of the differential equation, we obtain the infinite system of ordinary differential equations

$$\frac{dT_n(t)}{dt} + n^2 T_n(t) = f_n(t), \quad n = 1, 2, 3\dots$$

This is analogous to (4.17) which arose while solving (4.14a). Here  $n^2$  plays the part of  $\lambda_i$ , the eigenvalues of the operator  $\partial^2/\partial x^2$ ,  $f_n$  plays the role of  $\beta_i$ , and  $f(x, t)$  plays the role of the nonhomogeneity  $b$  in (4.10a).



**Fig. 7.4** Analogy between transform techniques and log tables.

The infinite set of ordinary differential equations which arise here are decoupled as the basis set chosen for the Fourier transform is orthogonal. Consequently,  $T_n(t)$  is independent of  $T_m(t)$ , where  $n \neq m$ .

We have not addressed the issue of convergence of the infinite series in (7.54) and (7.56). In this text, we are going to assume this property. The interested reader is referred to Weinberger (1965) for a formal proof. It must be remembered that representations of this kind are possible only when we are assured of the convergence of the series. This is guaranteed when the basis set is complete and the space we are working in is complete.

In the numerical solution of partial differential equations, using techniques like orthogonal collocation and weighted residuals, the solution is sought as a series expansion. The set of functions used as a basis here is chosen such that the boundary conditions are satisfied and they are orthogonal. Such sets must possess the important property of completeness. Only then can we justify a representation in the form of a series. Theoretically the solution is obtained only when we take an infinite sum. The truncation of the series to a finite series tantamounts to considering only a few (modes) terms in the series for the solution. The convergence of the series permits this approximation. Taking into account more terms in the series generates a more accurate solution. Numerical methods are used primarily to solve nonlinear equations. The resulting system of ordinary differential equations

which arise using finite Fourier transforms are coupled and nonlinear. The interested reader is referred to Gupta (1995) and Finlayson (1972) for more details.

## 7.5 FOURIER TRANSFORMS UNBOUNDED DOMAINS

The method of separation of variables leads us naturally to the concepts of Fourier series and finite Fourier transform. The series representation of the solution exists only for problems in finite domains. The extension of this idea to solve problems in infinite or unbounded domains generates the notion of the Fourier transform.

The eigenvalues of the one-dimensional elliptic operator  $d^2/dx^2$  in  $(0, L)$  subject to Dirichlet conditions are of the form

$$\lambda_n^2 = \frac{n^2\pi^2}{L^2}, \quad n = 1, 2, 3, \dots$$

The difference between two successive values is

$$\Delta\lambda_n^2 = |\lambda_{n+1}^2 - \lambda_n^2| = (2n + 1)\pi^2/L^2$$

For a finite domain, the eigenvalues form a discrete countably infinite set. As the domain size is increased, i.e.,  $L \rightarrow \infty$ , the distance between successive eigenvalues decreases, and the eigenvalues become more and more dense on the real line. The spectrum of an operator consists of the set of eigenvalues it generates. For a finite domain problem, the operator has a discrete spectrum. In the direction of an unbounded domain, the spectrum is continuous. In the limit  $L \rightarrow \infty$ , the series representation (7.54), (7.56) we expect will be replaced by an integral representation.

From equations (7.49b), (7.50c) for an  $f(x)$  in  $(-L, L)$

$$f(x) = \sum_{n=-\infty}^{\infty} \frac{1}{2L} \left( \int_{-L}^L f(x) e^{inx/L} dx \right) e^{-inx/L} \quad (7.61)$$

Defining a new continuous variable  $\omega_n = n\pi/L$ , we have

$$(\omega_{n+1} - \omega_n) = \pi/L \text{ or } \Delta\omega_n/\pi = 1/L$$

This yields

$$f(x) = \sum_{n=-\infty}^{\infty} \frac{\Delta\omega_n}{2\pi} \int_{-L}^L f(x) \exp(i\omega_n x) dx \exp(-i\omega_n x) \quad (7.62a)$$

Taking  $\lim L \rightarrow \infty$ , we keep  $\omega_n$  constant and obtain

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\omega_n \int_{-\infty}^{\infty} (f(x) \exp(i\omega_n x) dx) \exp(-i\omega_n x) \quad (7.62b)$$

Here,  $\omega_n$  is a continuous variable and is called the transform variable, it plays the role of an eigenvalue.  $A(\omega_n)$  defined below plays the role of  $c_n$ , the Fourier coefficient.

$$A(\omega_n) = \int_{-\infty}^{\infty} f(x) \exp(i\omega_n x) dx \quad (7.62c)$$

Since (7.62b) and (7.62c) involve integrals from  $(-\infty, \infty)$ , we need to impose restrictions on  $f(x)$

to assure us that the integrals exist. A function is said to be absolutely integrable when the improper integral

$$\int_{-\infty}^{\infty} |f(x)| dx$$

converges. The Fourier coefficients  $c_n^{(L)}$  of such a function defined on  $(-L, L)$  satisfy

$$\begin{aligned} |c_n^{(L)}| &= \frac{1}{2L} \left| \int_{-L}^L f(x) e^{inx/L} dx \right| \\ &\leq \frac{1}{2L} \int_{-L}^L |f(x)| dx \end{aligned} \quad (7.63)$$

As  $L \rightarrow \infty$ , clearly the coefficients  $|c_n^{(L)}| \rightarrow 0$  for each fixed  $n$ . The limit of  $2Lc_n^{(L)}$  for a fixed  $n$  is given by

$$\begin{aligned} \lim_{L \rightarrow \infty} 2Lc_n^{(L)} &= \lim_{L \rightarrow \infty} \int_{-L}^L f(x) e^{inx/L} dx \\ &= \int_{-\infty}^{\infty} f(x) dx \quad \text{for a fixed } n \\ &= \text{constant} \end{aligned} \quad (7.64)$$

For a fixed  $n$ ,  $2Lc_n^{(L)}$  is a constant as  $L \rightarrow \infty$ . Defining the continuous variable  $\omega_n$  as  $n\pi/L$ , the limiting function  $\hat{f}(\omega_n)$  for a fixed  $\omega_n$  is defined as

$$\hat{f}(\omega_n) = \lim_{L \rightarrow \infty} 2Lc_{n\pi/L}^{(L)} = \int_{-\infty}^{\infty} f(x) e^{i\omega_n x} dx \quad (7.65)$$

Here  $n$  is not fixed as  $L \rightarrow \infty$ , it varies so as to keep  $\omega$  constant. This helps us see how the Fourier transforms is a generalisation of the Fourier coefficients.

The continuous function  $\hat{f}(\omega)$  plays the role of the discrete coefficients  $c_n^{(L)}$ . It is called the Fourier transform of  $f(x)$ . Analogous to the coefficients, each function has a unique Fourier transform. This exists for all  $f(x)$  which are absolutely integrable.

In order to be useful, the Fourier transform  $\hat{f}(\omega)$  must have the same basic properties as the Fourier coefficients or the finite transform. The transform of the second derivative of  $f(x)$  must be obtained by multiplying  $\hat{f}(\omega)$  by  $-\omega^2$  (which plays the role of the eigenvalue). Consider  $f(x)$  such that  $\hat{f}(\omega)$  exists and

$$\lim_{\substack{x \rightarrow \infty \\ x \rightarrow -\infty}} f(x) = \lim_{\substack{x \rightarrow \infty \\ x \rightarrow -\infty}} f'(x) = 0 \quad (7.66)$$

These conditions are required when we carry out an integration by parts and they arise from physical considerations. Integrating by parts twice and using (7.66), we have

$$\lim_{\substack{L_1 \rightarrow \infty \\ L_2 \rightarrow \infty}} \int_{-L_1}^{L_2} f''(x) e^{i\omega x} dx = -\omega^2 \hat{f}(\omega) \quad (7.67)$$

The coefficients  $c_n^{(L)}$  determine  $f(x)$  uniquely by (7.49). The following theorem describes how to represent  $f(x)$  in terms of its transform  $\hat{f}(\omega)$ .

**Inversion theorem.** If  $f(x)$  is square integrable, i.e.  $\int_{-\infty}^{\infty} |f(x)|^2 dx$  converges, then the Fourier transform of its Fourier transform is  $2\pi f(-x)$ . This implies

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-i\omega x} d\omega \quad (7.68)$$

This can also be obtained from (7.62b), (7.62c) and (7.65).

The restriction of  $f(x)$  to be square integrable is a severe constraint. It is possible to obtain a transform for  $f(x)$  and to recover  $f(x)$  from  $\hat{f}(\omega)$  even when  $f(x)$  is absolutely integrable. A detailed discussion of these aspects and a formal proof of the inversion theorem can be found in Weinberger.

### 7.5.1 Fourier sine and cosine Transform

For  $f(x)$  defined on  $(0, L)$ , we saw how the Fourier series gets modified to a Fourier sine series, and a Fourier cosine series when  $f(x)$  is odd and even respectively. We proceed analogously for an  $f(x)$  defined on  $(0, \infty)$  and see that the Fourier transform gets modified to a Fourier sine transform and a Fourier cosine transform.

Let  $f(x)$  be defined on  $[0, \infty]$ . Extending it as an odd function in  $[-\infty, 0]$ , we have

$$\int_{-\infty}^{\infty} f(x) \cos(\omega x) dx = 0$$

and the Fourier transform (7.65) now becomes

$$\begin{aligned} \hat{f}(\omega) &= i \int_{-\infty}^{\infty} f(x) \sin(\omega x) dx \\ &= 2i \int_0^{\infty} f(x) \sin(\omega x) dx \end{aligned} \quad (7.69)$$

If  $f(x)$  is defined only on  $(0, \infty)$ , we define a sine transform as

$$\hat{f}_s(\omega) = \int_0^{\infty} f(x) \sin(\omega x) dx \quad (7.70)$$

From (7.69) and (7.70)

$$\hat{f}(\omega) = 2i \hat{f}_s(\omega) \quad (7.71)$$

The inversion theorem now yields

$$\begin{aligned} f(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{-i\omega x} d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\omega x} 2i \hat{f}_s(\omega) d\omega \\ &= \frac{2}{\pi} \int_0^{\infty} \sin(\omega x) \hat{f}_s(\omega) d\omega \end{aligned} \quad (7.72)$$

remembering that  $\hat{f}_s(\omega)$  is also an odd function of  $\omega$ .

The Fourier cosine transform is defined analogously as

$$\hat{f}_c(\omega) = \int_0^\infty f(x) \cos(\omega x) dx \quad (7.73a)$$

for  $f(x) \in [0, \infty]$ . Extending  $f(x)$  as an even function on  $[-\infty, 0]$ , we have

$$\hat{f}(\omega) = 2\hat{f}_c(\omega) \quad (7.73b)$$

From the inversion theorem, and remembering  $\hat{f}_c(\omega)$  is also an even function, we obtain

$$f(x) = \frac{2}{\pi} \int_0^\infty \cos(\omega x) \hat{f}_c(\omega) d\omega \quad (7.73c)$$

The evaluation of the Fourier transform involves integrations in the complex plane. This is normally done using standard methods from complex variables. However, there are tables of Fourier transforms available, which contain the transforms of a wide range of functions. The inverse transform can also be obtained from these tables by a suitable change of the variables. For our purposes we will use these tables while solving problems.

**Example 7.9** Consider the parabolic heat conduction problem in an infinite domain

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{for } -\infty < x < \infty, t > 0$$

subject to

$$u(x, 0) = e^{-x^2}$$

Observe that we have chosen the initial temperature profile such that it tends to zero, as  $x \rightarrow \pm \infty$ . The solution  $u(x, t)$  represents the evolution of temperature with space and time in an infinite domain, with no sources or sinks. It is reasonable to expect  $u(x, t) \rightarrow 0$  and  $\frac{\partial u}{\partial x}(x, t) \rightarrow 0$  as  $x \rightarrow \pm \infty$  for  $t > 0$ .

Since the problem is in an infinite domain, we take the Fourier transform in the  $x$ -direction, and obtain

$$\frac{d}{dt} \hat{u}(\omega, t) + \omega^2 \hat{u}(\omega, t) = 0$$

where

$$\hat{u}(\omega, t) = \int_{-\infty}^{\infty} u(x, t) e^{i\omega x} dx$$

Integrating this first order equation, we obtain

$$\hat{u}(\omega, t) = \hat{u}(\omega, 0) e^{-\omega^2 t}$$

Here

$$\begin{aligned} \hat{u}(\omega, 0) &= \int_{-\infty}^{\infty} u(x, 0) e^{i\omega x} dx \\ &= \sqrt{\pi} e^{-\omega^2/4} \end{aligned}$$

So,

$$\hat{u}(\omega, t) = \sqrt{\pi} e^{-\omega^2/4} e^{-\omega^2 t}$$

Taking the inverse transform, we obtain

$$u(x, t) = 1/(2\sqrt{\pi}) \int_{-\infty}^{\infty} e^{-\omega^2/4} e^{-\omega^2 t} e^{-i\omega x} d\omega$$

The evaluation of this integral needs concepts from complex variables. We refer the interested reader to Weinberger (1963) and Churchill (1960) for details.

**Example 7.10** Consider the steady-state heat conduction problem in an infinite strip

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \text{ for } 0 < y < 1, -\infty < x < \infty$$

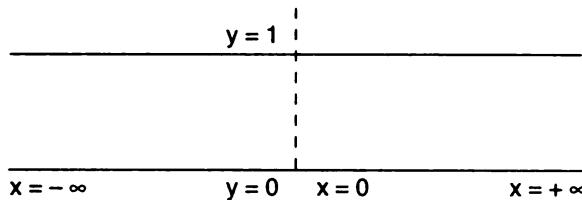
subject to

$$u(x, 0) = e^{-2bx}, u(x, 1) = 0$$

$$u(x, y) \rightarrow 0 \text{ as } x \rightarrow \pm \infty$$

Note that  $u(x, 0) \rightarrow 0$  as  $x \rightarrow \pm \infty$ .

The nonhomogeneity here is present in the boundary condition at  $y = 0$  (see Fig. 7.5). In a spatially bounded system we would have sought the solution in terms of eigenfunctions in the



**Fig. 7.5** Domain of system in Example 7.10.

$x$ -direction (along which we have homogeneous conditions). As an analogy to this we take the Fourier transform in the  $x$ -direction, to obtain

$$\frac{d^2 \hat{u}(\omega, y)}{dy^2} - \omega^2 \hat{u}(\omega, y) = 0$$

where

$$\hat{u}(\omega, y) = \int_{-\infty}^{\infty} e^{i\omega x} u(x, y) dx$$

This ordinary differential equation in  $y$  is subject to the conditions

$$\hat{u}(\omega, 1) = 0$$

$$\hat{u}(\omega, 0) = \frac{4}{\omega^2 + 4}$$

Here we have used

$$\int_{-\infty}^{\infty} e^{-2|x|} e^{i\omega x} = \frac{4}{\omega^2 + 4}$$

The solution to  $\hat{u}(\omega, y)$  is

$$\hat{u}(\omega, y) = A(\omega) \sinh \omega(1 - y)$$

This satisfies the differential equation as well as the condition at  $y = 1$ . Using the boundary condition at  $y = 0$ , we obtain

$$\hat{u}(\omega, y) = \frac{\sinh \omega(1 - y)}{\sinh \omega} \left( \frac{4}{\omega^2 + 4} \right)$$

Once again the inverse transform yields the solution in terms of an integral

$$u(x, y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{u}(\omega, y) e^{-i\omega x} d\omega$$

**Example 7.11** Solve the heat conduction problem

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

subject to

$$\begin{aligned} u(x, 0) &= 0, \quad u(x, 1) = 0, \quad u(0, y) = y(1 - y) \\ u(x, y) &\rightarrow 0 \text{ as } x \rightarrow \infty \end{aligned}$$

This problem is defined on a semi-infinite domain  $0 < x < \infty$ . At first sight it would appear to be amenable to solution using Fourier sine or cosine transforms. The nonhomogeneity however occurs in the boundary condition at  $x = 0$ . Consequently we need to use eigenfunction expansion in the  $y$ -direction (or the finite Fourier sine transform, since we have homogeneous Dirichlet conditions). Thus we seek

$$u(x, y) = \sum_{n=1}^{\infty} u_n(x) \sin(n\pi y)$$

where

$$u_n(x) = 2 \int_0^1 u(x) \sin(n\pi y) dy$$

Taking the finite Fourier sine transform in the  $y$ -direction we have

$$\frac{d^2 u_n}{dx^2} - n^2 \pi^2 u_n = 0$$

or

$$u_n(x) = A_n e^{n\pi x} + B_n e^{-n\pi x}$$

The boundedness of  $u_n(x)$  as  $x \rightarrow \infty$  implies  $A_n = 0$ . This yields

$$u(x, y) = \sum_{n=1}^{\infty} B_n e^{-n\pi x} \sin(n\pi y)$$

From the boundary condition at  $x = 0$

$$y(1 - y) = \sum_{n=1}^{\infty} B_n \sin(n\pi y)$$

where

$$\begin{aligned} B_n &= 2 \int_0^1 y(1-y) \sin(n\pi y) dy \\ &= \frac{8}{n^3 \pi^3} \quad \text{for odd } n \\ &= 0 \quad \text{for even } n \end{aligned}$$

This yields

$$u(x, y) = \sum_{n=1,3,5}^{\infty} \frac{8}{n^3 \pi^3} e^{-n\pi x} \sin(n\pi y)$$

## 7.6 LAPLACE TRANSFORM

The Fourier transform is useful in solving partial differential equations on a spatially unbounded domain. This enables us to eliminate spatial derivatives in the direction in which the domain is unbounded. This reduces the problem to an ordinary differential equation or an algebraic equation. It may be necessary to take repeated Fourier transforms for this. A parabolic partial differential equation is reduced to a first order ordinary differential equation with time as the independent variable in this approach (see Weinberger, 1985).

The Laplace transform allows us to integrate the time derivative in a parabolic problem. It reduces a system of ordinary differential equations to an algebraic system and a parabolic partial differential equation to an elliptic boundary value problem. This transform is applicable only for initial-value problems (IVP). The dynamic simulation of most engineering systems is modelled as an IVP. Our interest in IVP lies in determining the evolution of a system from a given initial state. Such functions are not defined for  $t < 0$ . They can be conveniently assumed to be zero for  $t < 0$ .

Consider a function  $f(x)$  which vanishes for  $x < 0$ . Let  $e^{-s_1 x} f(x)$  be absolutely integrable. Then  $e^{-sx} f(x)$  is also absolutely integrable for  $s > s_1$ .

The Fourier transform of  $f(x)$  with the variable 'is' is defined as the Laplace transform  $L(f(x))$  of the function with  $s$

$$L(f(x)) = \hat{f}(is) = \int_0^{\infty} e^{-sx} f(x) dx \quad (7.74)$$

since  $f(x) = 0$  for  $x < 0$ .

The Laplace transform of the first derivative of  $f$  can be obtained by an integration by parts as

$$L(f'(x)) = \int_0^{\infty} e^{-sx} f'(x) dx = sL(f(x)) - f(0) \quad (7.75)$$

To use the Laplace transform it is, therefore, necessary to know  $f(x)$  at  $x = 0$ . The Laplace transform of the higher order derivatives of  $f(x)$  needs conditions on the derivatives of  $f(x)$  at  $x = 0$ . Consequently, this transform can only be used for IVPs, where all the conditions are specified at  $x = 0$ . In contrast, boundary-value problems specify conditions at both ends of a domain.

The Laplace transform (7.74) was defined in terms of the Fourier transform. It would appear, therefore, that the inversion theorem can be applied to invert the Laplace transform. This extension

is nontrivial and involves the use of contour integrals in the complex plane. We do not introduce any of these concepts here and refer the interested reader to Churchill (1960), Weinberger (1965) for a detailed discussion on these lines. For our purposes we will use the transform tables for determining the Laplace transform and its inverse. The tables contain the transform and the inverse of the most frequently encountered functions.

**Example 7.12** Solve the problem of Example 7.9 using Laplace transforms. Taking the Laplace transforms yields

$$\int_{-\infty}^{\infty} e^{-st} \frac{\partial u}{\partial t} dt - \int_0^{\infty} \frac{\partial^2 u}{\partial x^2} e^{-st} dt = 0$$

$$s\hat{u} - \frac{\partial^2 \hat{u}}{\partial x^2} = e^{-x^2}$$

where

$$\hat{u}(x, s) = \int_0^{\infty} u(x, t) e^{-st} dt$$

This problem is defined in the infinite interval  $-\infty < x < \infty$ . Taking the Fourier transform with respect to  $x$  yields

$$\omega^2 u^* + su^*(\omega, s) = \sqrt{\pi} e^{-\omega^2/4}$$

where

$$u^*(\omega, s) = \int_{-\infty}^{\infty} \hat{u}(x, s) e^{i\omega x} dx$$

$$= \sqrt{\pi} \frac{e^{-\omega^2/4}}{\omega^2 + s}$$

Taking the inverse Laplace transform, we obtain

$$\tilde{u}(\omega, t) = \sqrt{\pi} e^{-\omega^2/4} e^{-\omega^2 t}$$

This is identical to the result obtained in Example 7.9 and the solution  $u(x, t)$  is obtained using an inverse Fourier transform as explained earlier.

The Fourier transform is normally used when solving a problem in  $(-\infty, \infty)$ . The Fourier sine or cosine transforms are used while solving a problem from  $[0, \infty]$ . The Fourier transform can be used to solve a problem which contains odd and even order derivatives. (Explain why?) The Fourier sine and cosine transforms can, however, be used only for problems whose operators have even order derivatives in the direction of the transform.

These transforms are useful in reducing a partial differential equation in an unbounded region to an ordinary differential equation. Here we end up with only a single ordinary differential equation in contrast to infinite number of equations in a spatially bounded region. The transforms used to reduce partial differential equations to ordinary differential equations to solving these equations and inverting them is analogous to using of logarithmic tables (see Fig. 7.4). Taking the logarithm facilitates solving a problem in the log-domain. The solution is found by using the inverse logarithm (transform). So far we have seen partial differential equations with only one dependent variable. The techniques presented here can be extended to systems where there are many dependent variables. We refer the interested reader to Ramkrishna and Amundson (1985) for such systems.

### PROBLEMS

**1. Solve**

$$\nabla^2 u = 0 \quad \text{in } 0 < x < \pi, 0 < y < \pi, 0 < z < \pi$$

subject to

$$u = 0 \quad \text{for } x = 0, x = \pi, y = 0, y = \pi, z = \pi$$

$$u(x, y, z = 0) = \sin x \sin^3 y$$

$$2. \nabla^2 u = 0 \text{ in } 0 < r < 1, 0 < z < \pi, -\pi < \theta < \pi$$

subject to

$$u(r, \theta, 0) = u(r, \theta, \pi) = 0$$

$$u(1, \theta, z) = z(\pi - z) \cos^2 \theta$$

$$3. \nabla^2 u = 0 \quad \text{in } -\pi < \theta < \pi, 0 < r < 1, 0 < z < 1$$

subject to

$$\frac{\partial u}{\partial r} (1, z) = \sin^2 \pi z$$

$$u(r, \theta, 0) = 0, u(r, \theta, 1) = 0$$

$$4. \nabla^2 u = 0 \quad \text{in } 0 < \theta < \pi, 0 < z < \pi, 0 < r < 1$$

subject to

$$u(r, \theta, 0) = u(r, \theta, \pi) = 0$$

$$u(r, 0, z) = u(r, \pi, z) = 0,$$

$$u(1, \theta, z) = z(\pi - z)$$

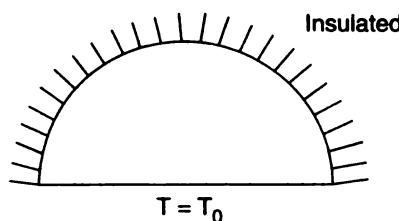
$$5. \nabla^2 u - u = 0 \quad \text{in } 0 < x < \pi, 0 < z < 1$$

subject to

$$u(x = 0) = 0, \quad \frac{\partial u}{\partial x} (x = \pi) = 0$$

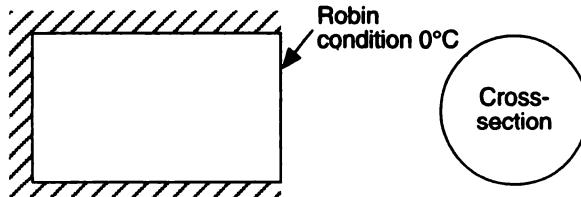
$$u(z = 1) = 0, \quad \frac{\partial u}{\partial z} (z = 0) = 2x - \pi$$

6. Obtain the steady state temperature distribution in a hemisphere whose temperature at the flat surface is a constant  $T_0$  and whose curved surface is insulated (Fig. 7.6).



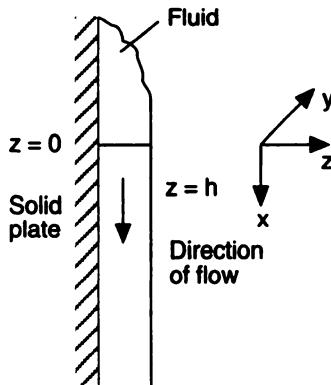
**Fig. 7.6** The physical system of Problem 6. Heat conduction in a hemisphere.

7. A finite cylinder of radius  $R$  and length  $L$  is at a uniform initial temperature  $T_0$ . The curved surface is insulated and the cylinder loses heat to the ambient at  $0^\circ\text{C}$  at the flat surface ( $z = L$ ). The other flat surface is insulated. Formulate the problem and solve for the temperature profile (Fig. 7.7).



**Fig. 7.7** The physical system of Problem 7. Heat conduction across an insulated cylinder.

8. In an absorption unit, liquid containing pure A is flowing down a vertical plate as shown in Fig. 7.8 with a mean velocity  $v_0$ . The thickness of the film is  $h$  and gas B is absorbed by this fluid. The mass transfer coefficient is  $k_m$  at this surface and concentration of B in gas is  $C_{B0}$ .



**Fig. 7.8** Fluid flow down a vertical flat plate.

Determine the concentration of B in the liquid film at steady state (neglect diffusion in the  $x$ -direction and convection in  $z$ -direction in liquid). Assume the profile to be independent of  $y$ .

9. A finite cylinder of radius  $R_1$  and length  $L$  has its curved surface insulated. There are no sources or sinks in the cylinder. The flat surface at  $z = 0$  is maintained at  $T_0$ , and the upper flat surface at  $z = L$  is maintained at 0. Find the temperature profile in the cylinder.

10. Consider the transient one-dimensional heat conduction equation

$$\frac{\partial T}{\partial t} = \frac{\partial^2 T}{\partial x^2}, \quad t > 0, 0 < x < 1$$

Let both ends of the slab be insulated. The initial temperature of the slab is given by  $\cos \pi x$ . Find the steady state temperature in the slab.

- (a) Model the problem as an elliptic problem.
- (b) Solve the parabolic problem.

Discuss the result obtained.

**11. Solve**

$$(a) \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \text{in } 0 < x < \pi, 0 < y < 1$$

subject to

$$u(y=0) = \sin^3 x, \quad u(y=1) = \sin^3 x$$

$$u(x=\pi) = 0, \quad u(x=0) = \sin \pi y$$

$$(b) \nabla^2 u(x, y) = 0 \quad \text{in } 0 < x < \pi, 0 < y < \pi$$

subject to

$$u(y=\pi) = 0, \quad u(y=0) = x^2$$

$$\frac{\partial u}{\partial x}(x=0) = 0, \quad \frac{\partial u}{\partial x}(x=\pi) = 0$$

$$(c) \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \text{in } 0 < x < \pi, 0 < y < 1$$

subject to

$$u(0, y) = u(x, 0) = u(x, 1) = 0$$

$$u(\pi, y) = \sin y$$

$$(d) \nabla^2 C(r, \theta) = 0 \quad \text{in the cylinder } 0 < r < 1, -\pi < \theta < \pi$$

subject to

$$C(r=1) = \sin^2 \theta$$

$$(e) \nabla^2 T(r, \theta) = 0 \quad \text{in } 0 < r < 1, 0 < \theta < \pi/2$$

subject to

$$T(1, \theta) = \theta, \quad T(r, 0) = 0, \quad T(r, \pi/2) = 0$$

**12.** Consider a fluid at rest in a pipe of radius  $R$ . A pressure gradient is applied across the fluid. The equation for the velocity profile is described by

$$\frac{\partial v}{\partial t} = 4 + \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial v}{\partial r} \right)$$

(a) Find the steady state velocity profile. What boundary conditions will you choose, and why?

$$(b) \text{ Write } v(r, t) = v_{ss}(r) + v'(r, t)$$

Find  $v'(r, t)$ . Is this an EBC or a PIC?

$$(c) \text{ Find } v(r, t).$$

**13.** A first order chemical reaction is occurring in a cylindrical pellet (assumed to be infinitely long). The pellet has no reactant initially. It is dipped into a well-stirred vessel containing pure A, at pressure  $P$  and temperature  $T$ . Find  $C(t, r, \theta)$  in pellet. State all assumptions and justify them.

**14.** A thin film of thickness  $\delta$  is flowing down a vertical flat plate. Assume the velocity across the film is a constant at  $\bar{u}$ . Water is entering at temperature  $0^\circ\text{C}$ . The plate is made of a poor conducting material. The free surface of the film is exposed to an atmosphere of temperature  $T_0$ . Find the temperature profile in the film as a function  $x, y$ . State all assumptions with justification. Is this problem well posed? Why? ( $y$ -coordinate normal to plate,  $x$  = coordinate along plate.)

**15.** Can Example 7.4 be solved by first seeking the eigenvalues in the  $r$ -direction and then in the  $\theta$ -direction. Explain.

**16.** Solve  $u_{xx} + u_{yy} + u_x = 0$  in  $0 < x < \pi, 0 < y < \pi$

subject to

$$u(y = 0) = 0, \quad u(y = \pi) = 0$$

$$u(x = 0) = 0, \quad u(x = \pi) = \sin y$$

**17.** Determine the concentration profile in an infinite slab sustaining a first order reaction at steady state. The lower face is impermeable. The governing equation is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - u = 0$$

subject to

$$\frac{\partial u}{\partial y}(x, 0) = 0, \quad u(x, 1) = e^{-x^2}$$

$$u(x, y) \rightarrow 0 \quad \text{as } x \rightarrow \pm \infty$$

## REFERENCES

- Berg, P.W. and McGregor, J.L., Elementary Partial Differential Equations, Holden Day, San Francisco (1964).
- Carslaw, H.S. and Jaeger, J.C., Conduction of Heat in Solids, Clarendon Press, Oxford (1959).
- Churchill, R.V., Complex Variables and Applications, McGraw-Hill, New York (1960).
- \_\_\_\_\_, Fourier Series and Boundary Value Problems, McGraw-Hill, New York (1963).
- Finlayson, A., The Method of Weighted Residuals and Variational Principles with Applications in Fluid Mechanics, Heat and Mass Transfer, Academic Press, New York (1972).
- Gupta, S.K., Numerical Methods for Engineers, Wiley Eastern, New Delhi (1995).
- Holman, J.P., Heat Transfer, McGraw-Hill, New York (1972).
- Kaplan, W., Operational Methods for Linear Systems, Addison-Wesley, Reading, Mass. (1962).
- Kersten, R.D., Engineering Differential Systems, McGraw-Hill, New York (1969).

- Kreyszig, E., *Advanced Engineering Mathematics*, Wiley, New York (1982).
- Ramakrishna, D. and Amundson, N.R., *Linear Operator Methods in Chemical Engineering: With applications to transport and chemical reaction systems*, Prentice-Hall, Englewood Cliffs, New Jersey (1985).
- Weinberger, H.F., *A First Course in Partial Differential Equations: With complex variables and transform methods*, Wiley, New York (1965).

# 8

## Green's Functions

---

---

We have seen in Chapter 5 that the principle of linearity and superposition can be used to decompose any parabolic and elliptic problem into three basic problems: (a) the parabolic initial condition (PIC) problem, (b) the elliptic boundary condition (EBC) problem, and (c) the nonhomogeneous equation (NHE) problem. In Chapter 7 we saw how the first two classes of problems can be solved using separation of variables. The nonhomogeneous equation problem can also be solved using this method. However, in this text we do not emphasise this approach for this class of problems.

In this chapter we present a different method of solving the nonhomogeneous equation problem. Here the nonhomogeneity occurs in the differential equation and not in the boundary condition. The methodology involves constructing the Green's function for the operator and associated boundary conditions and seeking the solutions in terms of the Green's function (see Stakgold, 1968, 1979). The concept of the Green's function brings out another analogy between finite dimensional spaces and infinite dimensional spaces. It is the equivalent of the matrix inverse for a differential operator. The numerical technique of boundary element method is based on this concept. It is an efficient technique and can be used in solving many problems elegantly. Unlike the finite element method which discretises the entire domain of the problem, this discretises only the boundary over which the problem is defined. Consequently, this method is computationally superior. A detailed discussion of this numerical method is available in Brebbia and Dominguez (1989).

Consider the general linear partial differential equation written in operator form as

$$\left. \begin{array}{l} Lu = f(x, y, z) \quad \text{in } V \\ \text{subject to} \\ Bu = g(x, y, z) \quad \text{on } S \end{array} \right\} \quad (8.1a)$$

Here  $V$  is the domain of the problem and  $S$  is its boundary on which the boundary conditions are specified.

The nonhomogeneity in the boundary can be eliminated using superposition as discussed in Chapter 7. We need to determine how to solve the nonhomogeneous equation arising from (8.1a).

$$\left. \begin{array}{l} Lu_1 = f(x, y, z) \quad \text{in } V \\ \text{subject to} \\ Bu_1 = 0 \quad \text{on } S \end{array} \right\} \quad (8.1b)$$

The nonhomogeneous matrix equation

$$Ax = b \quad (8.1c)$$

can be solved by expanding  $x$  in terms of the eigenvectors of  $A$  as in Chapter 4. This is analogous

to the method of separation of variables discussed in Chapter 7. Another method of solving (8.1c) is to compute the inverse operator  $A^{-1}$  and determine  $x$  as  $A^{-1}b$ . This particularly is a useful method when we seek  $x$  for different nonhomogeneities  $b$ . The inverse  $A^{-1}$  has to be computed only once and stored, and the different solutions are obtained by a simple matrix multiplication. We would like to derive an analogous method to solve for  $u$  as  $u = L^{-1}f$ . This inverse operator plays the part of the matrix inverse and is known as the Green's function of the differential operator  $L$ . Once this is determined, we can compute the solution for different values of  $f$  in (8.1b) elegantly as long as the operator  $L$  and the boundary conditions  $B$  remain unchanged. In this chapter we discuss: (a) Methods of determining Green's functions for different operators, and (b) obtaining the solution  $u$  to the nonhomogeneous equation in terms of Green's function.

Once again our discussions will be concerned with ordinary differential equations and elliptic and parabolic partial differential equations.

## 8.1 ORDINARY DIFFERENTIAL EQUATIONS

We illustrate the concept and motivation behind using the Green's function approach for an ordinary differential equation and later on extend it to partial differential equations. Consider the two-point boundary value problem defined in  $0 < x < 1$ :

$$Lu = f(x) \quad 0 < x < 1 \quad (8.2a)$$

subject to

$$B_1 u = h \quad (8.2b)$$

$$B_2 u = k \quad (8.2c)$$

For the above system we can determine the associated adjoint operators as  $L^*$ ,  $B_1^*$ ,  $B_2^*$ . The causal Green's function of this problem is defined as the distribution  $g(x/x_0)$  due to a concentrated unit point source at  $x_0$  instead of  $f(x)$ , the distributed source for the original system. The adjoint Green's function is defined as the distribution  $g^*(x/x_1)$  due to a unit point source at  $x_1$  for the adjoint system. The unit point source at  $x_0$  is represented by the Dirac delta function  $\delta(x - x_0)$ . The Green's functions are defined by

$$Lg(x/x_0) = -\delta(x - x_0) \quad (8.3a)$$

subject to

$$B_1 g(x/x_0) = 0 \quad (8.3b)$$

$$B_2 g(x/x_0) = 0 \quad (8.3c)$$

$$L^* g^*(x/x_1) = -\delta(x - x_1) \quad (8.4a)$$

subject to

$$B_1^* g^*(x/x_1) = 0 \quad (8.4b)$$

$$B_2^* g^*(x/x_1) = 0 \quad (8.4c)$$

Both Green's functions are defined with homogeneous boundary conditions, as we are interested only in the effect of a unit source function and not any nonhomogeneities in the boundary. The two variables  $x$ ,  $x_0$  in  $g(x/x_0)$  are used to represent the distribution as a function of  $x$  when the source is located at  $x_0$  (see Stakgold, 1979).

**Dirac delta function.** The Dirac delta function defined in a one-dimensional region denoted by

$\delta(x - x_0)$  is a mathematical idealisation of a unit impulse. It has the following properties:

1. The graphical property:

$$\delta(x - x_0) = \begin{cases} 0 & \text{for } x \neq x_0 \\ \infty & \text{for } x = x_0 \end{cases} \quad (8.5a)$$

2. The sifting property:

$$\int_{x_0-\epsilon}^{x_0+\epsilon} f(x)\delta(x-x_0) dx = f(x_0) \quad (8.5b)$$

Applying (8.5b) with  $f(x) = 1$ , we obtain  $\int_{-\infty}^{\infty} \delta(x - x_0) dx = 1$ . Thus, the function is a unit impulse.

We would like to obtain  $u$  defined by (8.2) in terms of the Green's function. Since it is only possible to relate the function defined by the original system  $L, B$  to the function defined by the adjoint system  $L^*, B^*$ , we will

- (i) first relate  $g(x/x_0)$  to  $g^*(x/x_1)$ ,
- (ii) then relate  $u(x)$  to  $g^*(x/x_1)$ ,
- (iii) use (i) and (ii) to relate  $u(x)$  to  $g(x/x_0)$ ,
- (iv) determine  $g(x/x_0)$  and use (iii) to obtain  $u(x)$ .

Taking the inner-product of (8.3a) with  $g^*(x/x_1)$  and (8.4a) with  $g(x/x_0)$ , and subtracting, we obtain

$$g(x_1/x_0) = g^*(x_0/x_1) \quad (8.6)$$

Taking the inner-product of (8.4a) with  $u(x)$  and (8.2a) with  $g^*(x/x_1)$ , and subtracting, we get

$$u(x_1) = - \int_0^1 g^*(x/x_1)f(x) dx + J(u, g^*(x/x_1)) \Big|_{x=0}^{x=1} \quad (8.7a)$$

The adjoint boundary conditions  $B^*$  are determined by setting the bilinear concomitant to zero when the original boundary conditions  $B$  are homogeneous. The bilinear concomitant in (8.7a) here is nonzero as the boundary conditions on  $u$  are nonhomogeneous (see 8.2b, c), refer Kaplan (1962) and Stakgold (1968).

Using the relation between  $g(x/x_0)$  and  $g^*(x_0/x)$ , viz. (8.6), we have

$$u(x_1) = - \int_0^1 g(x_1/x)f(x) dx + J(u, g(x_1/x)) \Big|_{x=0}^{x=1} \quad (8.7b)$$

Hence although  $u$  can only be related to  $g^*$ , we have used (8.6) to relate  $u(x)$  to  $g(x_1/x)$ . Consequently, even for a non self-adjoint system, it is not necessary to determine  $g^*(x/x_1)$  at all although  $u$  can only be related to  $g^*(x/x_1)$ . We can work with the causal Green's function  $g(x/x_0)$  itself.

For a self-adjoint system,  $L = L^*$ ,  $B = B^*$ . Comparing (8.3a) and (8.4a), we obtain for a self-adjoint system the relation

$$g(x/x_1) = g^*(x/x_1) \quad (8.8a)$$

Using (8.6) and (8.8a), we have

$$g(x/x_1) = g(x_1/x) \quad (8.8b)$$

Any Green's function that satisfies (8.8b) is said to be symmetric in its arguments  $x, x_1$ . The Green's function of a self-adjoint operator is symmetric. If  $u$  satisfies homogeneous boundary conditions, then the bilinear concomitant in (8.7) vanishes and we have, after changing  $x_1$  with  $x$  and  $x$  with  $x_0$ ,

$$u(x) = - \int_0^1 g(x/x_0) f(x_0) dx_0 \quad (8.9)$$

### 8.1.1 Construction of Green's Function for Ordinary Differential Equations (Boundary-value problems)

So far we have obtained  $u(x)$  in terms of  $g(x/x_0)$ . We can use this relation to determine  $u(x)$  only if we know  $g(x/x_0)$ . This is the fourth step listed a little earlier. We explain how to obtain  $g(x/x_0)$  using a specific example (see Stakgold, 1979).

Consider the linear operator  $L = d^2/dx^2$  subject to Dirichlet conditions at 0, 1. The causal Green's function is defined by

$$\frac{d^2}{dx^2} g(x/x_0) = -\delta(x - x_0) \quad \text{in } 0 < x, x_0 < 1 \quad (8.10a)$$

subject to

$$g(0/x_0) = 0 \quad (8.10b)$$

$$g(1/x_0) = 0 \quad (8.10c)$$

Since the Dirac delta function vanishes for  $x \neq x_0$ , the problem (8.10a) can be split into two parts

$$\frac{d^2}{dx^2} g_1(x/x_0) = 0, \quad 0 \leq x < x_0 \quad (8.11)$$

$$\frac{d^2}{dx^2} g_2(x/x_0) = 0 \quad \text{in } x_0 < x \leq 1 \quad (8.12)$$

This is equivalent to defining the Green's function in two parts as

$$g(x/x_0) = \begin{cases} g_1(x/x_0), & 0 < x < x_0 \\ g_2(x/x_0), & x_0 < x < 1 \end{cases} \quad (8.13)$$

$g_1$  must satisfy the boundary condition at the left-end point ( $x = 0$ )

$$g_1(0/x_0) = 0 \quad (8.14a)$$

Similarly,  $g_2$  satisfies

$$g_2(1/x_0) = 0 \quad (8.14b)$$

To determine  $g_1(x/x_0)$ ,  $g_2(x/x_0)$ , we need two more conditions. These arise from the continuity of the Green's function at  $x_0$ , which we expect from the physical considerations

$$g_1(x_0^-/x_0) = g_2(x_0^+/x_0) \quad (8.14c)$$

Integrating (8.10a) from  $x_0 - \varepsilon$  to  $x_0 + \varepsilon$  and letting  $\varepsilon \rightarrow 0$ , we obtain

$$\int_{x_0-\varepsilon}^{x_0+\varepsilon} \frac{d^2 g}{dx^2} dx = \int_{x_0-\varepsilon}^{x_0+\varepsilon} -\delta(x - x_0) dx$$

$$\left. \frac{dg_2}{dx} \right|_{x_0+\epsilon} - \left. \frac{dg_1}{dx} \right|_{x_0-\epsilon} = -1 \quad (8.14d)$$

This condition tells us that the derivative of the Green's function is discontinuous at  $x = x_0$ . We now have four conditions (8.14a)–(8.14d), and we can solve for  $g(x/x_0)$ . Equation (8.11) implies

$$g_1(x/x_0) = Ax + B \quad (8.15a)$$

Similarly,

$$g_2(x/x_0) = Cx + D \quad (8.15b)$$

The four constants  $A, B, C, D$  are obtained from the boundary conditions (8.14a)–(8.14d). Then  $g_1(0/x_0) = 0$  implies

$$B = 0 \quad (8.16a)$$

Again, the condition  $g_2(1/x_0) = 0$  implies

$$C = -D \quad (8.16b)$$

Thus we have

$$g_1(x/x_0) = Ax \quad (8.17a)$$

$$g_2(x/x_0) = (1 - x)D \quad (8.17b)$$

The two constants  $A, D$  are obtained by imposing the condition of continuity of the function and jump discontinuity in the derivative at  $x = x_0$  (8.14c) and (8.14d). From the continuity of  $g$  we have

$$Ax_0 = (1 - x_0)D \quad (8.18a)$$

Using (8.14d) yields

$$D = x_0, A = 1 - x_0 \quad (8.18b)$$

The Green's function is

$$g(x/x_0) = \begin{cases} x(1 - x_0), & 0 \leq x \leq x_0 \\ x_0(1 - x), & x_0 \leq x \leq 1 \end{cases} \quad (8.19)$$

The operator

$$Lu = \frac{d^2u}{dx^2}$$

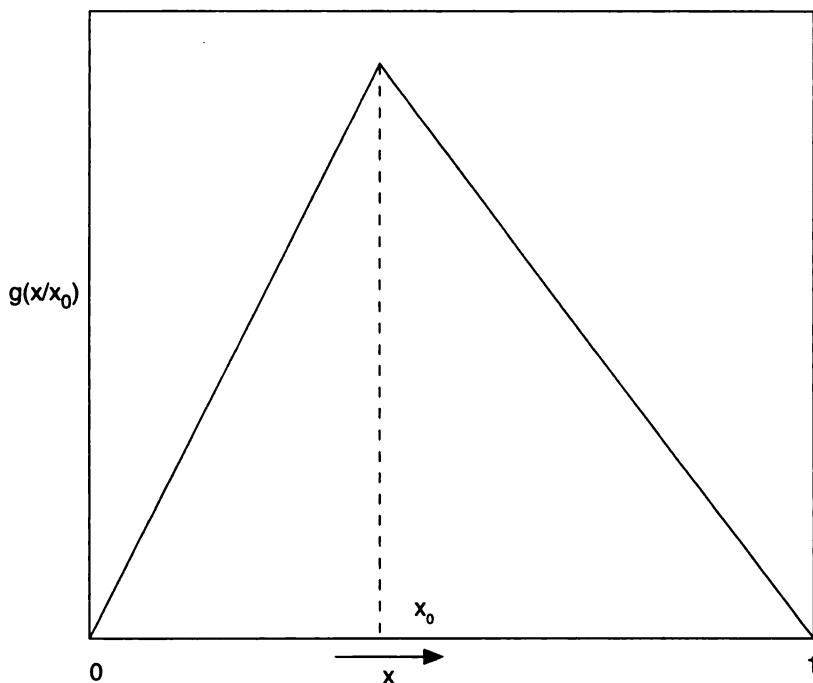
subject to Dirichlet boundary conditions  $u(0) = u(1) = 0$  is a self-adjoint operator. Hence  $g(x/x_0)$  must be symmetric or  $g(x/x_0) = g(x_0/x)$ . By interchanging  $x, x_0$  in (8.19), we get

$$g(x_0/x) = \begin{cases} x_0(1 - x), & x_0 \leq x \\ x(1 - x_0), & x \leq x_0 \end{cases} \quad (8.20)$$

This is clearly the same as  $g(x/x_0)$  in (8.19), see Fig. 8.1.

**Example 8.1** Solve the completely nonhomogeneous problem

$$\frac{d^2u}{dx^2} = \sin x,$$



**Fig. 8.1** Green's function or temperature distribution of a unit source at  $x = x_0$  (see (8.19))

subject to

$$u(0) = 1, \quad u(1) = 2 \quad (8.21)$$

Using (8.7b) and remembering that  $g^*(x/x_0) = g(x/x_0)$ , for the given self-adjoint operator we have

$$u(x_0) = g(1/x_0) u'(1) - u(1) g'(1/x_0) - g(0/x_0) u'(0) + u(0) g''(0/x_0) - \int_0^1 g(x/x_0) \sin x \, dx \quad (8.22a)$$

where  $g(x/x_0)$  is defined in (8.19). Using the boundary condition on  $g(x/x_0)$  and  $u(x)$ , we have

$$u(x_0) = g'(0/x_0) - 2g'(1/x_0) - \int_0^{x_0} g(x/x_0) \sin x \, dx - \int_{x_0}^1 g(x/x_0) \sin x \, dx \quad (8.22b)$$

We have already determined  $g(x/x_0)$ . From this we have

$$g'(x=0) = (1-x_0), \quad g'(x=1) = -x_0$$

In the first integral, clearly  $x$  ranges from  $(0, x_0)$  as  $0 < x < x_0$ . Hence the value of  $g(x/x_0)$  is  $g_1(x/x_0)$ , which is  $x(1-x_0)$ . In the second integral,  $x$  lies in  $[x_0, 1]$ . Here,  $g(x/x_0)$  takes the value  $x_0(1-x)$ . Thus we have

$$u(x_0) = (1-x_0) + 2x_0 - \int_0^{x_0} \sin(x) x(1-x_0) \, dx - \int_{x_0}^1 \sin(x) x_0 (1-x) \, dx \quad (8.23)$$

Since we integrate over  $x$ , we can treat  $x_0$  as a constant and take the terms containing  $x_0$  outside the integral sign. Replacing  $x_0$  by  $x$  after we perform the integrations, we get the solution  $u(x)$  to (8.21) as

$$u(x) = 1 + x - \sin x + x \sin 1$$

This example illustrates how the solution  $u(x)$  can be obtained using the Green's function approach.

**Example 8.2** Consider the two-point Green's function problem with Neumann boundary conditions. Here

$$\frac{d^2g}{dx^2}(x/x_0) = -\delta(x - x_0) \quad \text{in } 0 < x, x_0 < 1 \quad (8.24)$$

subject to

$$g'(0/x_0) = 0 = g'(1/x_0)$$

Following the approach described above we have

$$g(x/x_0) = \begin{cases} Ax + B, & 0 \leq x \leq x_0 \\ Cx + D, & x_0 \leq x \leq 1 \end{cases} \quad (8.25)$$

Imposing the boundary conditions at  $x = 0, x = 1$ , we have

$$g(x/x_0) = \begin{cases} B, & 0 \leq x \leq x_0 \\ D, & x_0 \leq x \leq 1 \end{cases} \quad (8.26)$$

From the continuity of  $g(x/x_0)$  at  $x = x_0$ , we have  $B = D$ . The jump discontinuity yields the equation  $0 = -1$ . The constants  $B, D$  cannot be uniquely determined. The Green's function or the inverse operator cannot be determined for this problem. This anomalous situation occurs because the completely homogeneous problem

$$\frac{d^2u}{dx^2} = 0$$

subject to

$$u'(0) = u'(1) = 0 \quad (8.27)$$

admits a nonzero solution,  $u = A$  (a constant). This becomes clear when we consider the analogous problem in finite dimensional space. The homogeneous system of linear algebraic equations  $Ax = 0$  has a nontrivial solution when the rank of  $A$  is less than the order of  $A$ , i.e., when  $\det A = 0$ . This implies the inverse operator  $A^{-1}$  does not exist for this situation. For the purely homogeneous Neumann Problem (8.27), we have a similar situation. It has a nonzero solution and the inverse, in this case the Green's function, does not exist.

**Example 8.3** Solve the following problem using Green's function

$$\frac{d^2T}{dx^2} - \frac{dT}{dx} = x$$

subject to

$$T(x = 0) = 1, \quad \frac{dT}{dx}(x = 1) = 2$$

The differential system here is not self-adjoint as seen in Chapter 6.

The causal Green's function is defined as

$$\frac{d^2 g}{dx^2}(x/x_0) - \frac{dg}{dx}(x/x_0) = -\delta(x - x_0)$$

subject to homogeneous conditions

$$g(0/x_0) = 0, \quad \frac{dg}{dx}(1/x_0) = 0$$

Defining

$$g(x/x_0) = \begin{cases} g_1(x/x_0), & 0 < x < x_0 \\ g_2(x/x_0), & x_0 < x < 1 \end{cases}$$

$g_1(x/x_0)$  is the solution of

$$\frac{d^2 g_1}{dx^2}(x/x_0) - \frac{dg_1}{dx}(x/x_0) = 0$$

or

$$\frac{dg_1}{dx}(x/x_0) - g_1(x/x_0) = A \text{ (a constant)}$$

which yields

$$g_1(x/x_0) = Be^x - A$$

Similarly,  $g_2(x/x_0)$  can be found as

$$g_2(x/x_0) = De^x - C$$

The constants  $A, B, C, D$  are obtained as earlier from the boundary conditions

$$g_1(0/x_0) = 0$$

$$\frac{dg_2}{dx}(1/x_0) = 0$$

$$g_1(x_0^-/x_0) = g_2(x_0^+/x_0)$$

$$\frac{dg_2}{dx}(x_0^+/x_0) - \frac{dg_1}{dx}(x_0^-/x_0) = -1$$

This yields

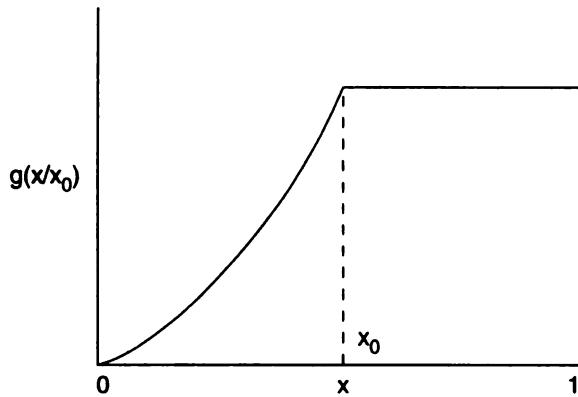
$$D = 0, \quad A = B = e^{-x_0}, \quad C = e^{-x_0} - 1$$

The Green's function here is

$$g(x/x_0) = \begin{cases} e^{-x_0}(e^x - 1), & 0 < x < x_0 \\ 1 - e^{-x_0}, & x_0 < x < 1 \end{cases}$$

and is plotted in Fig. 8.2. The adjoint Green's function is defined as the solution of the adjoint problem

$$\frac{d^2 g^*}{dx^2}(x/x_1) + \frac{dg^*}{dx}(x/x_1) = -\delta(x - x_1)$$



**Fig. 8.2** Green's function or temperature distribution of a unit source at  $x = x_0$  (see Example 8.3).

subject to

$$g^*(0/x_0) = 0$$

$$\frac{dg^*}{dx} (1/x_1) + g^*(1/x_1) = 0$$

The corresponding Green's function is

$$g^*(x/x_1) = \begin{cases} 1 - e^{-x}, & x < x_1 \\ (e^{x_1} - 1) e^{-x}, & x_1 < x < 1 \end{cases}$$

Clearly,

$$\begin{aligned} g^*(x_0/x) &= \begin{cases} 1 - e^{-x_0}, & x_0 < x < 1 \\ (e^x - 1) e^{-x_0}, & 0 < x < x_0 \end{cases} \\ &= g(x/x_0) \end{aligned}$$

Taking the inner-product of the equation governing  $T$  with  $g^*(x/x_1)$  and the equation governing  $g^*(x/x_1)$  with  $T$  and subtracting, we get

$$T(x_1) = - \int_0^1 g^*(x/x_1)x \, dx + g^*(1/x_1)T'(1) + g^*(0/x_1)T(0)$$

Using (8.8a), we obtain

$$T(x_1) = - \int_0^1 g(x_1/x)x \, dx + g(x_1/1)T'(1) + g'(x_1/0)T(0)$$

Inserting the values of the Green's function as calculated and the boundary conditions on  $T$ , we get

$$\begin{aligned} T(x) &= - \int_0^x (1 - e^{-x_1})x_1 \, dx_1 - \int_x^1 e^{-x_1}(e^x - 1)x_1 \, dx_1 + 2 \cdot e^{-1}(e^x - 1) + 1 \cdot 1 \\ &= 4e^{x-1} - 4e^{-1} + 1 - x - \frac{x^2}{2} \end{aligned}$$

**Example 8.4** Solve for  $u(r)$ , using the Green's functions

$$\frac{1}{r} \left( \frac{d}{dr} r \frac{du}{dr} \right) = r^2 \quad \text{in } 0 < r < 1 \quad (8.28)$$

subject to

$u(r = 0)$  is bounded

$u(r = 1) = 1$

We determine the causal Green's function  $g(r/r_0)$  from

$$\frac{1}{r} \frac{d}{dr} \left( r \frac{dg}{dr} (r/r_0) \right) = - \frac{\delta(r - r_0)}{2\pi r} \quad \text{in } 0 < r < 1 \quad (8.29)$$

(Later we will see why the unit source function is defined like this.)

Define

$$g_1(r/r_0) = g(r/r_0) \quad \text{for } 0 < r < r_0$$

$$g_2(r/r_0) = g(r/r_0) \quad \text{for } r_0 < r < 1$$

The problems defining  $g_1(r/r_0)$  and  $g_2(r/r_0)$  are

$$\frac{1}{r} \frac{d}{dr} \left( r \frac{dg_1}{dr} (r/r_0) \right) = 0$$

subject to

$$\left. \begin{array}{l} g_1(0/r_0) = \text{bounded} \\ g_1(r_0/r_0) = g_2(r_0/r_0) \end{array} \right\}$$

$$\frac{1}{r} \frac{d}{dr} \left( r \frac{dg_2}{dr} (r/r_0) \right) = 0, \quad g_2(1/r_0) = 0$$

Integrating both sides of (8.29) from  $(r_0 - \varepsilon)$  to  $(r_0 + \varepsilon)$ , we obtain

$$\frac{dg_2}{dr} (r_0/r_0) - \frac{dg_1}{dr} (r_0/r_0) = - \frac{1}{2\pi r_0}$$

This yields,

$$g_1(r/r_0) = C_1 \ln r + C_2$$

$$g_2(r/r_0) = C_3 \ln r + C_4$$

Using the boundary conditions, we obtain  $C_1 = C_4 = 0$  and  $C_3 = -1/2\pi$  and  $C_2 = -\ln r_0/2\pi$ , we get the Green's function

$$g(r/r_0) = \begin{cases} -\frac{1}{2\pi} \ln r_0, & 0 < r < r_0 \\ -\frac{1}{2\pi} \ln r, & r_0 < r < 1 \end{cases} \quad (8.30)$$

It is easy to verify that subject to the Dirichlet boundary conditions the operator  $L = \frac{1}{r} \frac{d}{dr} \left( r \frac{d}{dr} \right)$  is self-adjoint. Hence

$$g^*(r/r_0) = g(r/r_0)$$

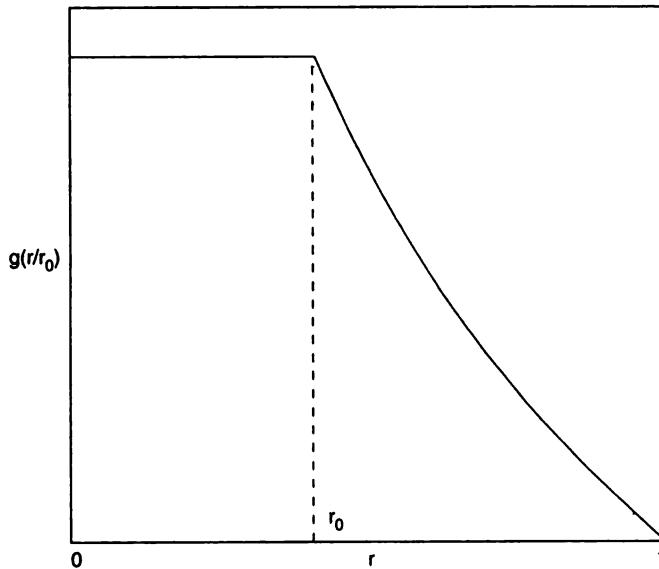
As  $g^*(r/r_0) = g(r_0/r)$ , we have

$$g(r/r_0) = g(r_0/r)$$

The Green's function is symmetric with respect to the arguments. This can be readily verified by interchanging  $r$  with  $r_0$ .

$$\begin{aligned} g(r_0/r) &= -\frac{\ln r}{2\pi}, \quad r_0 < r \\ &= -\frac{\ln r_0}{2\pi}, \quad r < r_0 \end{aligned} \quad (8.31)$$

This is identical to (8.30). This function is depicted in Fig. 8.3.



**Fig. 8.3** Green's function  $g(r/r_0)$  of Example 8.4.

To obtain  $u$ , take the inner-product of (8.28) with  $g(r/r_0)$  and (8.29) with  $u(r)$ . Remembering that an elemental area is  $2\pi r dr$ ,

$$\int_0^1 \left( \frac{1}{r} \frac{d}{dr} \left( r \frac{du}{dr} \right) g(r/r_0) - u(r) \frac{1}{r} \frac{d}{dr} \left( r \frac{dg}{dr} (r/r_0) \right) \right) 2\pi r dr$$

$$= \int_0^1 r^2 g(r/r_0) 2\pi r dr + \int_0^1 \frac{u}{r} \frac{\delta(r - r_0)}{2\pi} 2\pi r dr$$

or

$$2\pi \left[ g(r/r_0) r \frac{du}{dr} - u(r) r \frac{dg}{dr} (r/r_0) \right] \Big|_0^1 = u(r_0) + \int_0^{r_0} r^2 \frac{(-\ln r_0)}{2\pi} \cdot 2\pi r dr + \int_{r_0}^1 \frac{r^2 (-\ln r)}{2\pi} 2\pi r dr$$

yielding

$$-2\pi u(l) \frac{dg}{dr} (1/r_0) + \int_0^{r_0} r^3 dr \ln r_0 + \int_{r_0}^l r^3 \ln r dr = u(r_0)$$

or

$$u(r_0) = \frac{15}{16} + \frac{r_0^4}{16}$$

or

$$u(r) = \frac{15}{16} + \frac{r^4}{16}$$

### 8.1.2 Physical Significance of the Green's Function $g(x/x_0)$

The causal Green's function is defined in (8.3). The Dirac delta function represents a unit source term localised at the point  $x = x_0$ . The strength is taken to be unity since the integral over the neighbourhood of  $x_0$  is unity due to the sifting property. It is a point source of infinite strength. Depending on the sign of the term, we have a unit source or unit sink. As written in (8.10), it is a source term. For the sake of concreteness, let us assume it to be a source of heat. The Green's function  $g(x/x_0)$  can be viewed as representing the temperature distribution corresponding to this source of heat. The diffusion term or the second derivative term in (8.11) represents conduction carrying away the heat generated at  $x = x_0$ . As there are no source or sinks at  $x \neq x_0$ , the temperature profile or  $g(x/x_0)$  is a linear function of  $x$  as seen in (8.20), see Fig. 8.1.

Consider for the moment a nonhomogeneous ordinary differential equation (8.2a) subject to homogeneous boundary conditions (i.e. with  $h = k = 0$  in (8.2b) and (8.2c)). Then the solution  $u$  given by (8.9) represents the temperature distribution in  $0 < x < 1$ . When the source term is of unit strength and located at  $x_0$ , the temperature distribution is given by  $g(x/x_0)$ . For a source at  $x_0$  of strength  $f(x_0)$ , the temperature distribution will be  $f(x_0)$  times  $g(x/x_0)$  as our equation is linear. Clearly, our source term is  $(-f(x))$  and is distributed in  $0 < x < 1$ . By superposing the effect of the source term as  $x_0$  varies continuously from 0 to 1, we get

$$u(x) = - \int_0^1 f(x_0) g(x_0/x) dx_0$$

which is the same as (8.9) with  $x$  replaced by  $x_0$  and vice-versa.

## 8.2 GREEN'S FUNCTION FOR PARTIAL DIFFERENTIAL EQUATIONS

We have so far discussed Green's function for ordinary differential equations. We will now see how we can extend these ideas to partial differential equations. We are interested essentially in two basic kinds of nonhomogeneous problems: the nonhomogeneous parabolic problem and the nonhomogeneous elliptic problem.

### 8.2.1 Elliptic Equations

Consider the general boundary value problem

$$\left. \begin{aligned} Lu(x) &= f(x) \text{ in } V \\ Bu(x) &= h(x) \text{ on } S \end{aligned} \right\} \quad (8.32a)$$

subject to

Here  $x$  represents all the independent variables,  $x_1, x_2, \dots, x_n$ . The causal Green's function problem for this operator is

subject to

$$\left. \begin{array}{l} Lg(x/\xi) = \delta(x - \xi) \text{ in } V \\ Bg(x/\xi) = 0 \text{ on } S \end{array} \right\} \quad (8.32b)$$

where we have the point source located at  $\xi = \xi_1, \xi_2, \dots, \xi_n$ .  $\delta(x - \xi)$  is the Dirac delta distribution for  $n$  independent variables. Let us see the form of this function for the three different coordinate systems.

**Rectangular cartesian coordinates.** In the cartesian coordinates for three dimensions

$$\delta(x - \xi) = \delta(x - x_0)\delta(y - y_0)\delta(z - z_0) \quad (8.33a)$$

This represents a unit source term, as the differential volume element is  $dx dy dz$  and

$$\int_V \delta(x - \xi) dx dy dz = 1$$

**Cylindrical coordinates.** The differential volume element here is  $r d\theta dr dz$ . If the Dirac-delta function is chosen as  $\delta(r - r_0)\delta(\theta - \theta_0)\delta(z - z_0)$ , then

$$\int_V \delta(r - r_0)\delta(\theta - \theta_0)\delta(z - z_0)r d\theta dr dz = r_0$$

Hence the Dirac delta function must be chosen as

$$\delta(x - \xi) = \delta(r - r_0)\delta(\theta - \theta_0)\delta(z - z_0)/r \quad (8.33b)$$

in the three-dimensional cylindrical coordinates. This is now a unit source term as it can be normalised to yield

$$\int_V \frac{\delta(r - r_0)\delta(\theta - \theta_0)\delta(z - z_0)}{r} r d\theta dr dz = 1$$

**Spherical coordinates.** Here, the volume element is  $r^2 \sin \theta dr d\theta d\phi$ . The corresponding unit source term is, therefore,

$$\delta(x - \xi) = \frac{\delta(r - r_0)\delta(\theta - \theta_0)\delta(\phi - \phi_0)}{r^2 \sin \theta} \quad (8.33c)$$

For a general curvilinear coordinate system  $\mu_1, \mu_2, \mu_3$ , the Dirac delta function is given by (see Kreyszig, 1982)

$$\frac{\delta(\mu_1 - \mu_1^0)\delta(\mu_2 - \mu_2^0)\delta(\mu_3 - \mu_3^0)}{|J|}$$

where  $|J|$  is the determinant of the Jacobian transformation.

$$\begin{bmatrix} \frac{\partial x_1}{\partial \mu_1} & \frac{\partial x_1}{\partial \mu_2} & \frac{\partial x_1}{\partial \mu_3} \\ \frac{\partial x_2}{\partial \mu_1} & \frac{\partial x_2}{\partial \mu_2} & \frac{\partial x_2}{\partial \mu_3} \\ \frac{\partial x_3}{\partial \mu_1} & \frac{\partial x_3}{\partial \mu_2} & \frac{\partial x_3}{\partial \mu_3} \end{bmatrix}$$

In the cylindrical coordinate system,  $\mu_1 = r$ ,  $\mu_2 = \theta$ ,  $\mu_3 = z$ .

$$x = x_1 = r \cos \theta$$

$$y = x_2 = r \sin \theta$$

$$z = x_3 = z$$

$$J = \begin{vmatrix} \cos \theta & -r \sin \theta & 0 \\ \sin \theta & r \cos \theta & 0 \\ 0 & 0 & 1 \end{vmatrix} = r$$

which yields (8.33b). Similarly, in the spherical coordinates,

$$\mu_1 = r, \quad \mu_2 = \theta, \quad \mu_3 = \phi$$

$$x_1 = x, \quad x_2 = y, \quad x_3 = z$$

$$x = r \sin \theta \cos \phi$$

$$y = r \sin \theta \sin \phi$$

$$z = r \cos \theta$$

In this case,

$$J = \begin{vmatrix} \sin \theta \cos \phi & r \cos \theta \cos \phi & -r \sin \theta \sin \phi \\ \sin \theta \sin \phi & r \cos \theta \sin \phi & r \sin \theta \cos \phi \\ \cos \theta & -r \sin \theta & 0 \end{vmatrix}$$

$$= r^2 \sin \theta$$

which gives rise to (8.33c). The form of the Dirac delta function in (8.29) can now be easily justified. The factor  $2\pi$  used there is to ensure normalisation of the same function. Consider for the sake of concreteness the general three-dimensional elliptic problem.

$$-\nabla^2 u(x) + k(x)u(x) = f(x) \quad \text{for } x \in V \quad (8.34a)$$

subject to

$$n \cdot \nabla u(x) + c(x)u(x) = h(x) \quad \text{on } S_R \quad (8.34b)$$

$$u(x) = p(x) \quad \text{on } S_D \quad (8.34c)$$

The boundary has been decomposed into two parts  $S_R$ ,  $S_D$  on which the Robin and Dirichlet boundary conditions hold respectively. We restrict  $k(x)$ ,  $c(x)$  to be non-negative. Here  $x$  denotes all spatial coordinates. This is really not a restriction as they are met in most problems due to physical considerations. Equation (8.34a) occurs in processes where diffusion is the main mode of transport (the inclusion of convective effects would necessitate incorporating first order derivatives).

The causal Green's function  $g(x/\xi)$  for this problem is

$$-\nabla^2 g(x/\xi) + k(x)g(x/\xi) = \delta(x - \xi) \quad \text{in } V \quad (8.35a)$$

subject to

$$n \cdot \nabla g(x/\xi) + c(x)g(x/\xi) = 0 \quad \text{on } S_R \quad (8.35b)$$

$$g(x/\xi) = 0 \quad \text{on } S_D \quad (8.35c)$$

To obtain the adjoint Green's function, we need to find the associated adjoint operator and the adjoint boundary conditions. We can define an adjoint operator for this three-dimensional

problem as we did for the finite dimensional vector space and the one-dimensional differential equation. Before going into the determination of the adjoint, we introduce a few identities from calculus, see Kreyszig (1982) and Stakgold (1968).

$$\nabla \cdot u \nabla v = \nabla u \cdot \nabla v + u \nabla^2 v \quad (8.36a)$$

$$\int_V \nabla \cdot u \, dV = \int_S n \cdot \nabla u \, ds \quad (8.36b)$$

The first identity is an extension of the product rule of differentiation for functions of more than one variable. The second identity is called Green's theorem and its proof can be found in any book of calculus. It is used to convert volume integrals to surface integrals. The inner-product in the space of **real functions**  $f(x), g(x)$  is defined as

$$\langle f(x), g(x) \rangle = \int_V f(x)g(x) \, dV \quad (8.36c)$$

This is an extension of the definition in Chapter 6, to real functions defined in a three-dimensional region. Our operator  $L = (-\nabla^2 + k(x))$ . We would like to determine  $L^*$ . Using the definition (8.36c), we get

$$\begin{aligned} \langle Lu, v \rangle &= \int_V v(-\nabla^2 u + k(x)u) \, dV \\ &= \int_V -v \nabla^2 u \, dV + \int_V k(x)uv \, dV \\ &= \int_S \left( u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) + \int_V (-\nabla^2 v + k(x)v)u \, dV \\ &= \int_{S_D} \left( u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) + \int_{S_R} \left( u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) + \langle u, L^*v \rangle \end{aligned}$$

The homogeneous version of (8.34b) and (8.34c) is

$$\frac{\partial u}{\partial n} + c(x)u = 0 \quad \text{on } S_R$$

$$u = 0 \quad \text{on } S_D$$

Remembering that we choose the homogeneous version of boundary conditions of original operators to determine adjoints, the above expression simplifies to

$$\int_{S_R} u \left( \frac{\partial v}{\partial n} + c(x)v \right) - \int_{S_D} v \frac{\partial u}{\partial n} + \langle u, L^*v \rangle$$

Clearly,

$$L^*v = -\nabla^2 v + k(x)v \quad \text{in } V$$

$$B^*v \text{ is } n \cdot \nabla v + c(x)v = 0 \quad \text{on } S_R$$

$$v = 0 \quad \text{on } S_D$$

Obviously,  $L = L^*$ ,  $B = B^*$ , and our system is self-adjoint (see Kaplan, 1962). The adjoint Green's function corresponding to the problem in (8.35) is defined as

$$-\nabla^2 g^*(x/x_1) + k(x)g^*(x/x_1) = \delta(x - x_1) \quad \text{in } V \quad (8.37\text{a})$$

$$, g^*(x/x_1) = 0 \quad \text{on } S_D \quad (8.37\text{b})$$

$$n \cdot \nabla g^*(x/x_1) + hg^*(x/x_1) = 0 \quad \text{on } S_R \quad (8.37\text{c})$$

Comparing this problem with (8.35), it follows that

$$g(x/x_0) = g^*(x/x_0)$$

Taking the inner-product of (8.35a) with  $g^*(x/x_1)$  and the equation (8.37a) with  $g(x/x_0)$  and subtracting we get

$$g(x_1/x_0) = g^*(x_0/x_1)$$

As our operator is self-adjoint, we have

$$g(x/x_0) = g(x_0/x)$$

We can relate  $u(x)$  to  $g(x/x_0)$  directly here as our operator is self-adjoint. Hence are not interested in determining the adjoint Green's function  $g^*(x/\xi)$  and directly relate  $u(x)$  to  $g(x/\xi)$ . Taking the inner-product of (8.34a) with  $g(x/\xi)$  and (8.35a) with  $u(x)$  and subtracting, we have

$$\int_V (u \nabla^2 g - g \nabla^2 u) = \int_V g(x/\xi) f(x) - u(\xi)$$

$$\int_S (un \cdot \nabla g - gn \cdot \nabla u) = \int_V g(x/\xi) f(x) - u(\xi)$$

$$\int_{S_R} (un \cdot \nabla g - gn \cdot \nabla u) + \int_{S_D} (un \cdot \nabla g - gn \cdot \nabla u) = \int_V g(x/\xi) f(x) - u(\xi)$$

From the boundary conditions on  $u$ , viz. (8.34b) and (8.34c)

$$u(\xi) = \int_V g(x/\xi) f(x) dV + \int_{S_R} g(x/\xi) h(x) dS_R - \int_{S_D} p(x) \frac{\partial g}{\partial n}(x/\xi) dS_D \quad (8.38)$$

The solution to the nonhomogeneous equation (8.34a) can be determined now in terms of the Green's function from (8.38). But how about determining the Green's function itself? There are, in general, two methods for determining it for problems in spatially bounded regions:

1. Full eigenfunction expansion
2. Partial eigenfunction expansion.

**Full eigenfunction expansion.** The full or complete eigenfunction expansion method is a generalisation of the finite Fourier transform (see Chapter 7), in which we seek an arbitrary function of a single variable  $f(x)$  in terms of the eigenfunctions of an associated ordinary differential equation. For a system with Dirichlet conditions, we get the Fourier sine series or the finite Fourier sine transform as the eigenfunctions are the sine functions. The extension of this idea to partial differential equations governing functions of more than one variable is the basis of the method of eigenfunction expansion, see Churchill (1963) and Weinberger (1965).

We illustrate the method for ordinary differential equations first. Consider the nonhomogeneous equation

$$\frac{d^2u}{dx^2} = f(x) \quad (8.39a)$$

subject to  $u(0) = u(1) = 0$ . This is a self-adjoint system. The corresponding eigenvalue problems defined by the adjoint operator (which is equal to the original operator) is

$$\frac{d^2\psi_n}{dx^2} + \lambda_n \psi_n = 0 \quad (8.39b)$$

subject to

$$\psi_n(0) = \psi_n(1) = 0$$

This equation has an infinity of eigenvalues

$$\lambda_n = n^2\pi^2, \quad n = 1, 2, 3, \dots$$

with the corresponding eigenfunctions

$$\psi_n(x) = A_n \sin(n\pi x)$$

The solution  $u(x)$  to (8.39a) can be sought in terms of  $\psi_i(x)$  as

$$u(x) = \sum_{i=1}^{\infty} c_i \psi_i(x) \quad (8.40a)$$

where

$$c_i = \frac{\langle u(x), \psi_i(x) \rangle}{\langle \psi_i(x), \psi_i(x) \rangle} \quad (8.40b)$$

Taking the inner-product of (8.39a) with  $\psi_i$  and of (8.39b) with  $u(x)$ , we have

$$\langle u(x), \psi_i(x) \rangle = - \frac{\langle f(x), \psi_i(x) \rangle}{\lambda_i}$$

Using (8.40a) and (8.40b), we obtain

$$u(x) = \sum_{i=1}^{\infty} - \frac{\langle f(x), \psi_i(x) \rangle}{\lambda_i \langle \psi_i(x), \psi_i(x) \rangle} \psi_i(x)$$

This is an extension of the approach described in Chapter 3, for solving nonhomogeneous linear algebraic equations of the form

$$Au = b$$

Equation (8.39a) is identical to this, where the operator  $A$  is  $d^2/dx^2$ , and the nonhomogeneity  $b$  is  $f(x)$ . The complete eigenfunction expansion can be viewed as an extension of the method of eigenvector expansion to solve nonhomogeneous partial differential equations.

This is easy to understand since (8.34b), the causal Green's function problem, is a nonhomogeneous equation. The causal Green's function equation for a typical elliptic problem is given by

$$\nabla^2 g(x/x_0) = -\delta(x - x_0) \quad \text{in } V \quad (8.41a)$$

subject to

$$g(x/x_0) = 0 \quad \text{on } S_D \quad (8.41b)$$

$$n \cdot \nabla g(x/x_0) + hg(x/x_0) = 0 \quad \text{on } S_R \quad (8.41c)$$

This problem is self-adjoint as we have seen and the associated eigenfunction problem is

$$\nabla^2 \psi_i = -\lambda_i \psi_i \quad (8.42a)$$

subject to

$$\psi_i = 0 \quad \text{on } S_D \quad (8.42b)$$

$$n \cdot \nabla \psi_i + h \psi_i = 0 \quad \text{on } S_R \quad (8.42c)$$

Adopting the same approach as in the Sturm-Louiville theory, we can prove that all the eigenvalues are real, and the eigenfunctions corresponding to distinct eigenvalues are orthogonal for the system (8.41). We seek  $g(x/x_0)$  as a linear combination of  $\psi_n(x)$  as

$$g(x/x_0) = \sum_{n=1}^{\infty} A_n(x_0) \psi_n(x) \quad (8.43a)$$

Extending the concept of finite Fourier transforms, we obtain

$$A_n(x_0) = \frac{\langle g(x/x_0), \psi_n(x) \rangle}{\langle \psi_n(x), \psi_n(x) \rangle} \quad (8.43b)$$

Taking the inner-product of (8.42a) with  $g(x/x_0)$  and (8.41a) with  $\psi_n(x)$  and subtracting yields

$$\langle g(x/x_0), \psi_n(x) \rangle = \frac{\psi_n(x_0)}{\lambda_n}$$

From (8.43a) and (8.43b), we obtain

$$g(x/x_0) = \sum_{n=1}^{\infty} \frac{\psi_n(x_0) \psi_n(x)}{\lambda_n \langle \psi_n(x), \psi_n(x) \rangle}$$

In a partial differential equation we have eigenfunctions in different directions. Hence, the eigenfunctions have more than one index. The summation in (8.43c) is over each eigenvalue, and therefore every possible combination of the index as we will see in the following examples. We conclude with the remark that this method is applicable to problems in spatially bounded domains where we have a countable infinity of eigenvalues.

**Example 8.5** Solve

$$\nabla^2 g(r, \theta/r_0, \theta_0) = -\frac{\delta(r - r_0) \delta(\theta - \theta_0)}{r} \quad (8.44a)$$

subject to

$$g(1, \theta/r_0, \theta_0) = 0 \quad (8.44b)$$

The corresponding eigenvalue problem is

$$\nabla^2 \psi(r, \theta) = -\lambda \psi(r, \theta) \quad (8.45a)$$

subject to

$$\psi(r = 1, \theta) = 0 \quad (8.45b)$$

The eigenvalue problem is solved using separation of variables. Since the eigenvalue problem is completely homogeneous, we have eigenvalues in both  $r$  and  $\theta$  directions. This yields in the  $\theta$ -direction the eigenvalues  $n^2$  with the eigenfunctions

$$\theta_n(\theta) = A_n \cos n\theta + B_n \sin n\theta, \quad n = 0, 1, 2, \dots, \infty$$

The solution in the  $r$ -direction is

$$R_{m,n}(r) = J_n(\lambda_{m,n} r)$$

where the  $\lambda_{m,n}$ 's are such that

$$J_n(\lambda_{m,n}) = 0$$

The eigenvalue problem associated with a particular  $m$  and  $n$  is given by

$$\nabla^2 \psi_{m,n}(r, \theta) = -\lambda_{m,n} \psi_{m,n}(r, \theta) \quad (8.46a)$$

subject to

$$\psi_{m,n}(1, \theta) = 0 \quad (8.46b)$$

Taking the inner-product of (8.44a) with  $\psi_{m,n}(r, \theta)$  and (8.46a) with  $g(r, \theta/r_0, \theta_0)$  and subtracting, we get

$$\langle \psi_{m,n}(r, \theta), g(r, \theta/r_0, \theta_0) \rangle = \frac{\psi_{m,n}(r_0, \theta_0)}{\lambda_{m,n}}$$

We seek the Green's function as a linear combination of the  $\psi_{m,n}$

$$\begin{aligned} g(r, \theta/r_0, \theta_0) &= \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} A_{m,n} \psi_{m,n} \\ &= \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} (B_{m,n} J_n(\lambda_{m,n} r) \cos n\theta + C_{m,n} J_n(\lambda_{m,n} r) \sin n\theta) \end{aligned}$$

where

$$\begin{aligned} B_{m,n} &= \frac{J_n(\lambda_{m,n} r_0) \cos n\theta_0}{\lambda_{m,n} \langle J_n(\lambda_{m,n} r) \cos n\theta, J_n(\lambda_{m,n} r) \cos n\theta \rangle} \\ C_{m,n} &= \frac{J_n(\lambda_{m,n} r_0) \sin n\theta_0}{\lambda_{m,n} \langle J_n(\lambda_{m,n} r) \sin n\theta, J_n(\lambda_{m,n} r) \sin n\theta \rangle} \end{aligned}$$

with

$$\langle \bar{u}(r, \theta), v(r, \theta) \rangle = \int_{-\pi}^{\pi} d\theta \int_0^1 dr u(r, \theta) v(r, \theta) r$$

since the orthogonality in the  $r$ -direction is with respect to weighting function  $r$ .

**Example 8.6** Find Green's function

$$\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = -\delta(x - x_0)\delta(y - y_0), \quad 0 < x, x_0 < 1 \quad (8.47)$$

$g = 0$  at  $x = 0, 1$  and  $y = 0, 1$ ,  $0 < y, y_0 < 1$ . The corresponding eigenvalue problem is

$$\frac{\partial^2 \phi_i}{\partial x^2} + \frac{\partial^2 \phi_i}{\partial y^2} = -\lambda_i \phi_i \quad (8.48)$$

$\phi_i = 0$  on the boundary. We seek the eigenfunction by separation of variables. Now

$$\phi = X(x)Y(y)$$

This yields

$$\frac{X_{xx}}{X} + \frac{Y_{yy}}{Y} = -\lambda \quad (8.49)$$

Both  $x$ - and  $y$ -directions are homogeneous in the boundary. Hence we need eigenvalues in both these directions. Once they are determined, we can get the eigenvalues  $\lambda$  of the problem in (8.49). In the  $x$ -direction we have  $X_{xx} + \alpha^2 X = 0$ . This yields as eigenvalues and eigenfunctions

$$\alpha_n^2 = n^2 \pi^2, \quad X_n = C_n \sin(n\pi x)$$

In the  $y$ -direction the eigenvalues are  $\beta^2 = m^2 \pi^2$ , with  $Y_m = E_m \sin(m\pi y)$ . So,

$$\phi_{n,m} = F_{n,m} \sin(n\pi x) \sin(m\pi y)$$

$$\lambda_{n,m} = n^2 \pi^2 + m^2 \pi^2$$

We determine  $F_{n,m}$  from the normalisation conditions, i.e.

$$\langle \phi_{n,m}, \phi_{n,m} \rangle = 1, \int_0^1 dx \int_0^1 \phi_{n,m}^2 dy = 1$$

This yields  $F_{n,m} = 2$ .

Proceeding as in the earlier example, we obtain

$$g(x, y/x_0, y_0) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \frac{4 \sin(n\pi x) \sin(m\pi y) \sin(n\pi x_0) \sin(m\pi y_0)}{n^2 \pi^2 + m^2 \pi^2}$$

**Partial eigenfunction expansion.** In Chapter 7, we saw how we can reduce a partial differential equation to an infinite set of ordinary differential equations by using finite Fourier transforms. This is possible as long as we are dealing with problems in spatially bounded domains. Here we are assured of a discrete, infinite number of eigenvalues in each of the spatial directions (see Weinberger, 1965).

The partial eigenfunction expansion is a generalisation of this technique where we use the eigenfunctions arising naturally in the problem to reduce the partial differential equation to an infinite system of ordinary differential equations. In this process, as we saw earlier, we introduce general transform pairs. The resulting ordinary differential equations are solved using first principles. This is a computationally efficient method as we expand the Green's function in an  $n$ -dimensional problem only in  $n - 1$  spatial directions and solve explicitly in the remaining direction using first principles. To solve problem (8.34), we consider the auxiliary homogeneous problem

$$\nabla^2 u = 0 \quad \text{in } V \quad (8.50a)$$

subject to

$$u = 0 \quad \text{on } S_D \quad (8.50b)$$

$$n \cdot \nabla u + hu = 0 \quad \text{on } S_R \quad (8.50c)$$

This problem can be solved by using separation of variables method. Here we would have eigenvalue problems in all spatial directions (since the boundary conditions are homogeneous in all directions). Let  $\psi_n(x)$  be the eigenfunction in the  $x_1$  direction. The Green's function  $g(x_1, x_2, \dots, x_n/x_{10}, x_{20}, \dots, x_{n0})$  can be expanded in terms of these eigenfunctions as

$$g(x_1, x_2, \dots, x_n/x_{10}, x_{20}, \dots, x_{n0}) = \sum_{i=1}^{\infty} g_i(x_2, \dots, x_n/x_{10}, \dots, x_{n0}) \psi_i(x_1) \quad (8.51a)$$

The dependence of  $g$  on  $x_1$  is now absorbed in the index  $i$ , and  $g_i$  is obtained from

$$g_i(x_2, \dots, x_n/x_{10}, \dots, x_{n0}) = \frac{\langle g(x_1, \dots, x_n/x_{10}, \dots, x_{n0}), \psi_i(x_1) \rangle}{\langle \psi_i(x_1), \psi_i(x_1) \rangle} \quad (8.51b),$$

The  $g_i$ 's here play the role of Fourier coefficients, and the inner-product in (8.51b) is an integral only in the  $x_1$  direction and not in all of  $V$  (This is therefore strictly not an inner-product).

Equations (8.51a) and (8.51b) formally constitute a transform pair. The infinite number of equations for  $g_i$  are generated, by taking the transform of the original equations. These may be partial differential equations. These are reduced to ordinary differential equations in  $x_n$  by repeatedly expanding in eigenfunctions along  $x_2, \dots, x_{n-1}$ . These equations can be solved by first principles and the Green's function is obtained by using (8.51a). For instance, consider the two-dimensional Laplacian in the cartesian coordinates as in (8.47). We have eigenfunctions in the  $x, y$  direction as

$$Y_m = D_m \sin(m\pi y) \quad m = 1, 2, 3\dots$$

$$X_n = C_n \sin(n\pi x) \quad n = 1, 2, 3\dots$$

Let us choose to expand  $g(x, y/x_0, y_0)$  in terms of the eigenfunctions in the  $x$ -direction (we can alternatively choose to expand in the  $y$ -direction). The transform pair which we use is

$$g(x, y/x_0, y_0) = \sum_{n=1}^{\infty} g_n(y/x_0, y_0) \sin(n\pi x) \quad (8.52a)$$

$$\left. \begin{aligned} g_n(y/x_0, y_0) &= \frac{\int_0^1 g(x, y/x_0, y_0) \sin(n\pi x) dx}{\int_0^1 \sin^2(n\pi x) dx} \\ &= 2 \int_0^1 g(x, y/x_0, y_0) \sin(n\pi x) dx \end{aligned} \right\} \quad (8.52b)$$

Every  $g_n(y/x_0, y_0)$  is governed by ordinary differential equations, which can be solved from first principles. Then  $g(x, y/x_0, y_0)$  can be obtained from (8.52a). We illustrate how to use the method to determine the Green's function in the following examples.

**Example 8.7** Solve Example 8.5 using partial eigenfunction expansions.

We can expand  $g(r, \theta/r_0, \theta_0)$  in terms of eigenfunctions in the  $r$ -direction or the  $\theta$ -direction. We have eigenvalue problems in both directions, as the boundary conditions are homogeneous. Let us expand in terms of eigenfunctions in the  $\theta$ -direction.

$$\theta_n(\theta) = A_n \sin n\theta + B_n \cos n\theta, n = 0, 1, 2, \dots, \infty$$

$$g(r, \theta/r_0, \theta_0) = \sum_{n=0}^{\infty} (g_n^1(r/r_0, \theta_0) \sin n\theta + g_n^2(r/r_0, \theta_0) \cos n\theta) \quad (8.53)$$

$g_n^1, g_n^2$  are determined by taking the transforms of (8.44a) in the  $\theta$ -direction (multiplying by  $\sin n\theta$ ,  $\cos n\theta$  and integrating from  $-\pi$  to  $\pi$ , after normalisation) to yield

$$\frac{1}{r} \frac{d}{dr} r \frac{dg_n^1}{dr}(r/r_0, \theta_0) - \frac{n^2}{r^2} g_n^1(r/r_0, \theta_0) = -\frac{\delta(r - r_0)}{\pi r} \sin n\theta_0$$

$$\frac{1}{r} \frac{d}{dr} r \frac{dg_n^2}{dr}(r/r_0, \theta_0) - \frac{n^2}{r^2} g_n^2(r/r_0, \theta_0) = -\frac{\delta(r - r_0)}{\pi r} \cos n\theta_0$$

The Green's functions determined by these ordinary differential equations are obtained as discussed earlier. Treating the case  $n = 0$  separately, we obtain

$$g_0^1 = 0$$

$$g_0^2 = \frac{1}{2\pi} \begin{cases} -\ln r_0, & r < r_0 \\ -\ln r, & r_0 < r \end{cases}$$

The factor  $2\pi$  arises because of normalisation in the angular direction. For  $n \geq 1$ ,

$$g_n^1(r/r_0, \theta_0) = \begin{cases} -\sin n\theta_0(r_0^n - r^{-n})r^n/(2n\pi), & r < r_0 \\ -\sin n\theta_0(r^n - r^{-n})r^n/(2n\pi), & r_0 < r \end{cases}$$

$$g_n^2(r/r_0, \theta_0) = \begin{cases} -\cos n\theta_0(r_0^n - r^{-n})r^n/(2n\pi), & r < r_0 \\ -\cos n\theta_0(r^n - r^{-n})r^n/(2n\pi), & r_0 < r \end{cases}$$

Substituting these in (8.53), we obtain  $g(r, \theta/r_0, \theta_0)$  as the sum of an infinite number of terms.

**Example 8.8** Solve

$$\left. \begin{array}{l} -g_{xx} - g_{yy} = \delta(x - x_0)\delta(y - y_0) \\ g = 0 \text{ at } x = 0, 1; y = 0, 1 \end{array} \right\} \quad (8.54)$$

subject to

The corresponding homogeneous problem is

$$\left. \begin{array}{l} u_{xx} + u_{yy} = 0 \\ u = 0 \text{ at } x = 0, 1; y = 0, 1 \end{array} \right\} \quad (8.55)$$

subject to

By separating the variables, we get

$$X''/X = -Y''/Y = -\lambda^2$$

We choose to get an eigenvalue problem in the  $x$ -direction and expand  $g(x, y/x_0, y_0)$  about these eigenfunctions. We now obtain

$$g(x, y/x_0, y_0) = \sum_{n=1}^{\infty} g_n(y/x_0, y_0) \sin(n\pi x) \quad (8.56)$$

Taking the inner-product of (8.54) with  $\sin(n\pi x)$ , we have

$$-\int_0^1 \sin(n\pi x) \frac{\partial^2 g}{\partial x^2} dx - \int_0^1 \sin(n\pi x) \frac{\partial^2 g}{\partial y^2} dy = \delta(y - y_0) \sin(n\pi x_0)$$

Integrating the first integral by parts twice and applying the boundary conditions on  $g$  and in the second integral, we are differentiating with respect to  $y$  so that we can take the second derivative with respect to  $y$  outside the integral sign (since we are integrating with respect to  $x$ ). Remembering

$$2 \int_0^1 g(x, y/x_0, y_0) \sin(n\pi x) dx = g_n(y/x_0, y_0)$$

we get

$$\frac{d^2 g_n}{dy^2} - n^2 \pi^2 g_n = -2 \sin(n\pi x_0) \delta(y - y_0) \quad (8.57)$$

The boundary conditions on  $g_n$  are

$$g_n(y = 0) = 0, g_n(y = 1) = 0$$

This is a Green's function problem in the  $x$ -direction. Now our operator is

$$\left( \frac{d^2}{dy^2} - n^2 \pi^2 \right)$$

We can solve this equation by the method discussed for ordinary differential equations and obtain

$$g_n(y/x_0, y_0) = \frac{2 \sin(n\pi x_0)}{n\pi \sin(hn\pi)} \begin{cases} \sin(hn\pi y) \sin(hn\pi(1 - y_0)), & 0 \leq y \leq y_0 \\ \sin(hn\pi y_0) \sin(hn\pi(1 - y)), & y_0 \leq y \leq 1 \end{cases} \quad (8.58a)$$

and

$$g(x, y/x_0, y_0) = \sum_{n=1}^{\infty} g_n(y/x_0, y_0) \sin(n\pi x) \quad (8.58b)$$

To summarise in the full eigenfunction expansion method, we expand the Green's function in terms of eigenfunctions in all directions. In partial eigenfunction expansion method, we expand the Green's function in terms of eigenfunctions in all but one direction. The resulting ordinary differential equation in the remaining direction is solved analytically to obtain the Green's function. This makes the latter a computationally superior technique.

### 8.2.2 Parabolic Equations

The typical parabolic problem in engineering systems is of the form

$$Lu = \frac{\partial u}{\partial t} - \nabla^2 u = f(x, t) \quad \text{in } V \quad (8.59a)$$

subject to

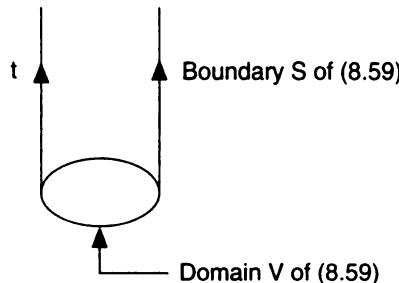
$$u(x, t = 0) = u_0(x) \quad \text{in } V \quad (8.59b)$$

$$u = u_D \quad \text{on } S_D \quad (8.59c)$$

$$n \cdot \nabla u + hu = hu_R \quad \text{on } S_R \quad (8.59d)$$

Such problems arise while modelling the transient behaviour of the systems. Here we are interested in studying the evolution of a system with time. For purposes of clarity, the time coordinate is given the distinct identity ' $t$ ', and all spatial coordinates are represented by  $x(x_1, x_2, x_3)$ . The region of interest to us can now be represented by a space-time cylinder. This is an abstract

concept. The time coordinate is depicted by the axis of the cylinder and all space coordinates which describe a three-dimensional volume are thought of as representing the cross-section of the cylinder. The region of interest to us is the interval  $0 < t < \tau$ , and  $x \in V \cup S$ . The base of the cylinder represents the plane  $t = 0$ , and the curved surface of the cylinder presents the boundary  $S$  of the volume  $V$  under consideration (Fig. 8.4).



**Fig. 8.4** Space-time cylinder of a general parabolic problem. Curved surface represents boundary of problem and cross-sectional area the domain of the problem.

We now discuss the methods of obtaining the Green's function for the parabolic differential operator in (8.59). Before embarking on that task we would like to obtain the relationship between the solution to (8.59) and the Green's function.

Our objective is to determine  $u$  in terms of a Green's function associated with the operator as earlier. We follow the same lines as we did for the elliptic equation. That is:

1. Determine the adjoint operator  $L^*$  and the associated boundary conditions  $B^*$ .
2. Relate the causal Green's function  $g(x, t/x_0, t_0)$  to the adjoint Green's function  $g^*(x, t/x_1, t_1)$ .
3. Relate the solution  $u$  to the adjoint Green's function.
4. Use steps 1 and 3 to determine  $u$  in terms of causal Green's function.
- Once again, it will not be necessary to obtain the adjoint Green's function and we obtain  $u$  directly in terms of the causal Green's function.
5. Determine the causal Green's function  $g(x, t/x_0, t_0)$ .

**Adjoint operator.** The parabolic operator in (8.59a) is defined on functions in the space-time cylinder. This is an abstract concept. Here time varies along the cylindrical axis as shown. The base of the cylinder corresponds to the coordinate  $t = 0$ . The cross-section of the cylinder represents all the spatial dimensions,  $x_1, x_2, x_3$ . Allowing the time coordinate to extend up to a finite value  $\tau$ , we define the inner-product.

$$\langle u(x, t), v(x, t) \rangle = \int_0^\tau dt \int_V u(x, t)v(x, t) dV \quad (8.60)$$

The adjoint operator is determined from (see Kaplan, 1962)

$$\langle v, Lu \rangle = \int_0^\tau dt \int_V v(x, t) \left( \frac{\partial u}{\partial t} (x, t) - \nabla^2 u(x, t) \right) dV$$

Changing the order of integration for the first term and using the identities (8.36a) and (8.36b), from vector calculus, we get

$$\begin{aligned} \langle v, Lu \rangle &= \int_0^\tau dt \int_V u(x, t) \left( -\frac{\partial v}{\partial t}(x, t) - \nabla^2 v(x, t) \right) dV + \int_V (u(x, \tau)v(x, \tau) - u(x, 0)v(x, 0)) \\ &\quad + \int_0^\tau dt \int_{S_D} (u \nabla v \cdot n - v \nabla u \cdot n) dS_D + \int_0^\tau dt \int_{S_R} (u \nabla v \cdot n - v \nabla u \cdot n) dS_R \end{aligned} \quad (8.61)$$

The adjoint operator is therefore defined as

$$L^*v = \left( -\frac{\partial}{\partial t} - \nabla^2 \right) v \quad (8.62a)$$

Setting the bilinear concomitant to zero yields

$$v(x, t = \tau) = 0 \quad (8.62b)$$

$$v(x, t) = 0 \quad \text{on } S_D \quad (8.62c)$$

$$n \cdot \nabla v(x, t) + hv(x, t) = 0 \quad \text{on } S_R \quad (8.62d)$$

This leads us to the following definitions of the causal Green's function:

$$\left( -\frac{\partial}{\partial t} - \nabla^2 \right) g(x, t/x_0, t_0) = -\delta(x - x_0)\delta(t - t_0) \quad (8.63a)$$

subject to

$$g(x, 0/x_0, t_0) = 0 \quad \text{in } V \quad (8.63b)$$

$$g(x, t/x_0, t_0) = 0 \quad \text{on } S_D \quad (8.63c)$$

$$n \cdot \nabla g(x, t/x_0, t_0) + hg(x, t/x_0, t_0) = 0 \quad \text{on } S_R \quad (8.63d)$$

The only nonhomogeneity in (8.63) is the unit source function in the equation. In particular, the initial conditions and the boundary conditions are homogeneous. As the unit source function acts only at  $t = t_0$ , the Green's function is zero for  $t < t_0$ , i.e. till the system feels the effect of the source term. Thus

$$g(x, t/x_0, t_0) = 0 \quad \text{for } t < t_0 \quad (8.63e)$$

One can therefore replace (8.63b) with (8.63e). The adjoint Green's function  $g^*(x, t/x_1, t_1)$  is defined as

$$\left( -\frac{\partial}{\partial t} - \nabla^2 \right) g^*(x, t/x_1, t_1) = -\delta(x - x_1)\delta(t - t_1) \quad (8.64a)$$

subject to

$$g^*(x, 0/x_1, t_1) = 0 \quad \text{in } V \quad (8.64b)$$

$$g^*(x, t/x_1, t_1) = 0 \quad \text{on } S \quad (8.64c)$$

$$n \cdot \nabla g^*(x, t/x_1, t_1) + hg^*(x, t/x_1, t_1) = 0 \quad \text{on } S \quad (8.64d)$$

The problem-defining  $g$  describes the evolution of  $g$  which is forward in time, i.e. with increasing time. The problem defining  $g^*$  describes its backward evolution in time. Arguing like we did earlier for  $g$ , we obtain,

$$g^*(x, t/x_1, t_1) = 0 \quad \text{for } t > t_1. \quad (8.64e)$$

The following identity follows from (8.61).

$$\begin{aligned} & \int_0^\tau dt \int_V dV \left[ v(x, t) \left( \frac{\partial u}{\partial t} - \nabla^2 u \right) - u(x, t) \left( -\frac{\partial v}{\partial t} - \nabla^2 v \right) \right] \\ &= \int_V [u(x, t)v(x, t)] \Big|_{t=0}^{t=\tau} + \int_0^\tau dt \int_S (u \nabla v \cdot n - v \nabla u \cdot n) dS \end{aligned} \quad (8.65)$$

To relate  $g(x, t/x_0, t_0)$  and  $g(x, t/x_1, t_1)$ , we set

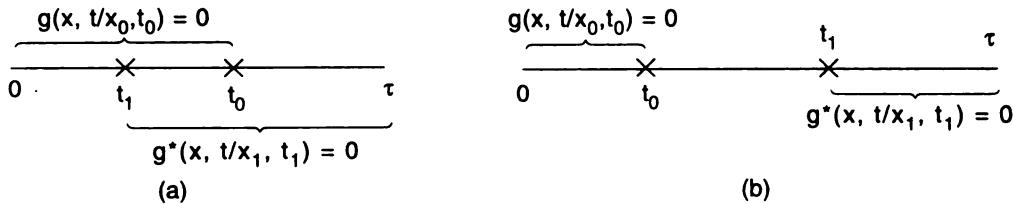
$$u(x, t) = g(x, t/x_0, t_0) \quad (8.66a)$$

$$v(x, t) = g^*(x, t/x_1, t_1) \quad (8.66b)$$

Two distinct possibilities arise:

1.  $0 < t_1 < t_0 < \tau$ . For this case the distribution of  $g, g^*$  is shown in Fig. 8.5(a). Substituting (8.66a,b) in (8.65), we obtain for this case the trivial identity  $0 = 0$ .

2.  $0 < t_0 < t_1 < \tau$ . The variation of  $g, g^*$  is depicted in Fig. 8.5(b). The integral from 0 to  $\tau$  in (8.65) is now reduced to the interval from  $t_0$  to  $t_1$ .



**Fig. 8.5** Distribution of causal and adjoint Green's functions: (a)  $t_1 < t_0$ ; (b)  $t_1 > t_0$ .

Using (8.66), we obtain

$$\begin{aligned} & \int_{t_0}^{t_1} dt \int_V [g^*(x, t/x_1, t_1) (-\delta(x - x_0)\delta(t - t_0)) - g(x, t/x_0, t_0) (-\delta(x - x_1)\delta(t - t_1))] dV \\ &= \int_V [g(x, \tau/x_0, t_0)g^*(x, \tau/x_1, t_1) - g(x, 0/x_0, t_0)g^*(x, 0/x_1, t_1)] dV \\ &+ \int_{t_0}^{t_1} \int_{S_D} g(x, t/x_0, t_0) n \cdot \nabla g^*(x, t/x_1, t_1) dS + \int_{t_0}^{t_1} \int_{S_R} g(x, t/x_0, t_0) n \cdot \nabla g^*(x, t/x_1, t_1) dS \\ &- \int_{t_0}^{t_1} \int_{S_D} g^*(x, t/x_0, t_0) n \cdot \nabla g(x, t/x_0, t_0) dS - \int_{t_0}^{t_1} \int_{S_R} g^*(x, t/x_0, t_0) n \cdot \nabla g(x, t/x_0, t_0) dS \end{aligned}$$

Using the conditions (8.63(b-e)) and (8.64(b-e)), we have

$$g(x_1, t_1/x_0, t_0) = g^*(x_0, t_0/x_1, t_1) \quad (8.67)$$

The relation between  $u(x, t)$  and  $g^*(x, t/x_1, t_1)$  is obtained from the identity (8.51) by setting

$v = g^*(x, t/x_1, t_1)$ . Remembering that  $g^*(x, t/x_1, t_1) = 0$  for  $t > t_1$ , the integral is nonzero only for  $t \in (0, t_1)$ . The identity (8.61) now is

$$\begin{aligned} & \int_0^{t_1} dt \int_V \left[ g^*(x, t/x_1, t_1) \left( \frac{\partial u}{\partial t}(x, t) - \nabla^2 u(x, t) \right) - u(x, t) \left( -\frac{\partial g^*}{\partial t}(x, t/x_1, t_1) - \nabla^2 g^*(x, t/x_1, t_1) \right) \right] dV \\ &= \int_V [u(x, \tau)g^*(x, \tau/x_1, t_1) - u(x, 0)g^*(x, 0/x_1, t_1)] dV \\ &+ \int_0^{t_1} dt \int_{S_D} [u(x, t)n \cdot \nabla g^*(x, t/x_1, t_1) - g^*(x, t/x_1, t_1)n \cdot \nabla u(x, t)] \\ &+ \int_0^{t_1} dt \int_{S_R} [u(x, t)n \cdot \nabla g^*(x, t/x_1, t_1) - g^*(x, t/x_1, t_1)n \cdot \nabla u(x, t)] \end{aligned}$$

Using the boundary conditions and the initial conditions (8.59b)–(8.59d), (8.64b)–(8.64e), we get

$$\begin{aligned} u(x_1, t_1) &= - \int_0^{t_1} dt \int_V g^*(x, t/x_1, t_1)f(x, t) - \int_V u_0(x)g^*(x, 0/x_1, t_1)dV \\ &+ \int_0^{t_1} dt \int_{S_D} u_D(x, t)n \cdot \nabla g^*(x, t/x_1, t_1) - \int_0^{t_1} dt \int_{S_R} g^*(x, t/x_1, t_1)hu_R(x, t) \end{aligned}$$

Substituting from (8.67), we get

$$\begin{aligned} u(x_1, t_1) &= - \int_0^{t_1} dt \int_V g(x_1, t_1/x, t)f(x, t) - \int_V u_0(x)g(x_1, t_1/x, 0) \\ &+ \int_0^{t_1} dt \int_{S_D} u_D n \cdot \nabla g(x_1, t_1/x, t) - \int_0^{t_1} dt \int_{S_R} g(x_1, t_1/x, t)hu_R(x, t) \end{aligned}$$

To recover  $u(x, t)$ , we interchange  $x$  with  $x_1$  and  $t$  with  $t_1$ , and obtain

$$\begin{aligned} u(x, t) &= - \int_0^t dt_1 \int_{V_1} g(x, t/x_1, t_1)f(x_1, t_1)dV_1 - \int_{V_1} u_0 g(x, t/x_1, 0)dV_1 \\ &+ \int_0^t dt_1 \int_{S_{D1}} u_D(x_1, t_1) n_1 \cdot \nabla g(x, t/x_1, t_1) - \int_0^t dt_1 \int_{S_{R1}} g(x, t/x_1, t_1)hu_R(x_1, t_1) \quad (8.68) \end{aligned}$$

The subscript 1 on  $V$ ,  $S_D$ ,  $S_R$  is used to denote that the integrations are with respect to  $x_1$ ,  $t_1$  and are not w.r.t.  $x$ ,  $t$ .

**Determination of  $g(x, t/x_0, t_0)$ .** So far we have discussed methods for computing Green's functions for elliptic problems. Parabolic problems can be dealt with in a similar way. There are two ways of solving for the Green's function in the parabolic case:

1. Eigenfunction expansion

2. Laplace transform.

**Eigenfunction expansion.** When solving first order linear ordinary differential equations, we expanded the solutions in terms of the eigenvectors (Chapter 4). Here we assumed the coefficients were time dependent and we determined their dependence on time. In this approach for computing the Green's function we use the same idea and seek the Green's function as a linear combination of the eigenfunctions with the associated operator. Consider

$$-\nabla^2 g(x, t/x_0, t_0) + \frac{\partial g}{\partial t} = \delta(x - x_0)\delta(t - t_0) \quad \text{in } V \quad (8.69a)$$

subject to

$$g = 0 \quad \text{in } V, t < t_0, x \in V \quad (8.69b)$$

$$g = 0 \quad \text{on } x \in S_D \quad (8.69c)$$

$$n \cdot \nabla g + k(x)g = 0, \quad x \in S_R \quad (8.69d)$$

The eigenfunction problem for this equation is

$$\nabla^2 \phi_i + \lambda_i \phi_i = 0 \quad \text{in } V \quad (8.70a)$$

subject to

$$\phi_i = 0 \quad x \in S_D \quad (8.70b)$$

$$n \cdot \nabla \phi_i + k(x)\phi_i = 0, \quad x \in S_R \quad (8.70c)$$

We assume that  $\phi_i$  form a complete set of orthonormal eigenfunctions (remember that the elliptic system in (8.70) is self-adjoint). We seek

$$g(x, t/x_0, t_0) = \sum_{i=1}^{\infty} a_i(t) \phi_i(x) \quad (8.71a)$$

$$a_i(t) = \int_V g(x, t/x_0, t_0) \phi_i(x) dV \quad (8.71b)$$

The pair of equations (8.71) again constitute a transform pair. We multiply (8.69a) by  $\phi_i(x)$  and (8.70a) by  $g(x, t/x_0, t_0)$  and integrate over the volume  $V$  and subtract. Since we have homogeneous boundary conditions for  $g(x, t/x_0, t_0)$  and  $\phi_i(x)$ , the bilinear concomitant vanishes. We can take the time derivative outside the integral as we are integrating over space only, and obtain the evolution equation for  $a_i(t)$  as

$$\frac{da_i}{dt} + \lambda_i a_i = \phi_i(x_0) \delta(t - t_0) \quad (8.72)$$

$$a_i = 0 \quad \text{for } t < t_0$$

The homogeneous version of (8.72) has a solution

$$a_i(t) = A e^{-\lambda_i t}$$

We thus write

$$a_i(t) = \begin{cases} A e^{-\lambda_i t} & \text{for } t < t_0 \\ B e^{-\lambda_i t} & \text{for } t > t_0 \end{cases} \quad (8.73)$$

From (8.69b)

$$a_i(t) = 0 \quad \text{for } t < t_0 \text{ or } A = 0 \quad (8.74)$$

The jump condition on  $a_i(t)$  comes by integrating (8.72) over a time interval  $(t_0 - \varepsilon, t_0 + \varepsilon)$ . This yields

$$a_i(t_0+) - a_i(t_0-) = \phi_i(x_0)$$

Using (8.73) and (8.74), we obtain

$$Be^{-\lambda_i t_0} = \phi_i(x_0)$$

Thus

$$a_i(t) = H(t - t_0)e^{-\lambda_i(t - t_0)} \phi_i(x_0)$$

where

$$\begin{aligned} H(t - t_0) &= 0 \quad \text{for } t < t_0 \\ &= 1 \quad \text{for } t > t_0 \end{aligned}$$

and is called the Heaviside function. Once  $a_i(t)$  is known, we have

$$g(x, t) = \sum_{i=1}^{\infty} H(t - t_0)e^{-\lambda_i(t - t_0)} \phi_i(x)\phi_i(x_0) \quad (8.75)$$

**Example 8.9** Solve

$$-\partial^2 g / \partial x^2 + \partial g / \partial t = \delta(x - x_0)\delta(t - t_0) \quad (8.76a)$$

subject to

$$g(t < t_0) = 0 \quad (8.76b)$$

$$g(x = 0) = g(x = 1) = 0 \quad (8.76c)$$

The corresponding eigenvalue problem is given by

$$\left. \begin{aligned} \frac{d^2 \phi_i}{dx^2} + \lambda_i \phi_i &= 0 \\ \phi_i(0) &= \phi_i(1) = 0 \end{aligned} \right\} \quad (8.77)$$

Its section is,

$$\phi_i(x) = A_i \sin(i\pi x), \quad \lambda_i = i^2\pi^2 \quad (8.78)$$

where  $i = 1, 2, \dots, \infty$ . To determine an orthonormal set of eigenfunctions, we define

$$\langle \phi_i, \phi_j \rangle = 1$$

This yields  $A_i = \sqrt{2}$ . Consider now

$$-\partial^2 g / \partial x^2 + \partial g / \partial t = \delta(x - x_0)\delta(t - t_0)$$

Let

$$g(x, t/x_0, t_0) = \sum_{i=1}^{\infty} a_i(t)\phi_i(x)$$

Multiplying (8.77) by  $g(x, t/x_0, t_0)$  and (8.76a) by  $\phi_i$  and integrating by parts and adding, we get

$$(d/dt)a_i + \lambda_i a_i = A_i \sin(i\pi x_0)\delta(t - t_0) \quad (8.79)$$

Solving this we obtain

$$a_i(t) = H(t - t_0) e^{-(t - t_0)\lambda_i} A_i \sin(i\pi x_0)$$

Thus

$$g(x, t/x_0, t_0) = \sum_{i=1}^{\infty} H(t - t_0) e^{-(t - t_0)^2 \pi^2} 2 \sin(i\pi x) \sin(i\pi x_0) \quad (8.80)$$

This method is similar to the method for solving a system of linear first order ordinary differential equation as seen in Chapter 4. It can also be thought of as a partial eigenfunction expansion method, since we solve analytically for the Green's function in time.

**Laplace transforms.** The Laplace transform method converts the parabolic problem to an elliptic problem. In this method we use this transform to integrate out the time dependency. This reduces the parabolic partial differential equation to a problem dependent only on the space direction. We can solve this problem by the methods discussed for elliptic problems, full eigenfunction expansion or partial eigenfunction expansion.

**Example 8.10** Solve

$$-\frac{\partial^2 g}{\partial x^2} + \frac{\partial g}{\partial t} = \delta(x - x_0)\delta(t - t_0)$$

Taking the Laplace transform of this equation we have

$$\begin{aligned} - \int_0^{\infty} \frac{\partial^2 g}{\partial x^2} e^{-st} dt + \int_0^{\infty} \frac{\partial g}{\partial t} e^{-st} dt &= \int_0^{\infty} e^{-st} \delta(x - x_0)\delta(t - t_0) dt \\ - \frac{d^2 \hat{g}}{dx^2} + s\hat{g} &= e^{-st_0} \delta(x - x_0) \end{aligned}$$

with  $\hat{g}(0, s) = 0$ ,  $\hat{g}(1, s) = 0$ . Here we have used  $g(x, 0/x_0, t_0) = 0$ . This ordinary differential equation in  $x$  can be solved using first principles to obtain

$$\begin{aligned} \hat{g}(x, s) &= \frac{e^{-st_0}}{\sqrt{s}} \frac{\sin h(\sqrt(s)(1-x_0)) \sin h(\sqrt(s) \cdot x)}{\sin h \sqrt(s)}, \quad x < x_0 \\ &= \frac{e^{-st_0}}{\sqrt{s}} \frac{\sin h(\sqrt(s)x_0) \sin h(\sqrt(s)(1-x))}{\sin h \sqrt(s)}, \quad x > x_0 \end{aligned}$$

The Green's function  $g(x, t)$  is obtained by taking the inverse Laplace transform.

### 8.3 UNBOUNDED DOMAINS

The eigenfunction expansions are used to obtain solutions to problems in spatially bounded domains. These can also be viewed as a finite Fourier transform as discussed in Chapter 7. The Green's function for spatially unbounded domains can be determined using Fourier transform techniques. We conclude this chapter with an example illustrating this.

**Example 8.11** Find Green's function in the quarter plane

$$-\frac{\partial^2 g}{\partial x^2} - \frac{\partial^2 g}{\partial y^2} = \delta(x - x_0)\delta(y - y_0), \quad x > 0, y > 0$$

subject to

$$g = 0 \quad \text{at } x = 0, x = \infty$$

$$g = 0 \quad \text{at } y = 0, y = \infty$$

Since this is a problem in semi-infinite intervals in  $x$ ,  $y$ -directions, we take the Fourier sine transform in the  $x$ ,  $y$ -directions. Defining

$$\hat{g}(\alpha, y) = \int_0^\infty g(x, y/x_0, y_0) \sin(\alpha x) dx$$

where  $\alpha$  is the transform variable in the  $x$ -direction. Taking the Fourier sine transform in the  $x$ -direction, we obtain the ordinary differential equation

$$\alpha^2 \hat{g} - \frac{d^2 \hat{g}}{dy^2} = \delta(y - y_0) \sin \alpha x_0$$

Defining

$$\hat{g}(\alpha, \beta) = \int_0^\infty \hat{g}(\alpha, y) \sin(\beta y) dy$$

and taking the Fourier sine transform in the  $y$ -direction, we have

$$\hat{g} = \frac{\sin \alpha x_0 \sin \beta y_0}{\alpha^2 + \beta^2}$$

The Green's function can be obtained using the inverse transform (see Chapter 7)

$$g(x, y/x_0, y_0) = \frac{2}{\pi} \cdot \frac{2}{\pi} \int_0^\infty d\alpha \int_0^\infty \frac{\sin \alpha x_0 \sin \alpha x \sin \beta x_0 \sin \beta x}{\alpha^2 + \beta^2} d\beta$$

## PROBLEMS

$$1. \quad Lu = \frac{d^2 u}{dx^2} = x$$

subject to

$$u'(0) = 2u(1) + 3, \quad u(0) = 1$$

(a) Find the adjoint operator and boundary condition  $L^*$ ,  $B^*$ .

(b) Find the causal Green's function  $g(x/x_0)$  and the adjoint Green's function  $g^*(x/x_0)$ . Verify  $g(x_0/x) = g^*(x/x_0)$ .

(c) Using the Green's functions  $g(x/x_0)$ ,  $g^*(x/x_0)$  determine  $u(x)$ .

2. Consider

$$\frac{d^2 g}{dx^2} = -\delta(x - x_0),$$

subject to

$$g'(0/x_0) = g(1/x_0), \quad g(0/x_0) = 0$$

Find  $g(x/x_0)$ . Analyse the result obtained.

**3.**  $d^2u/dx^2 + d^2u/dy^2 = 0 \quad \text{in } 0 < x < 1, \quad 0 < y < 1$

subject to

$$u(y=1) = x(1-x)$$

$$u = 0 \quad \text{at } x = 0, 1 \text{ and } y = 0$$

Solve using Green's functions. Normally this problem is solved using separation of variables. The nonhomogeneous boundary condition at  $y = 1$ , ensures that the bilinear concomitant will not vanish and we can get  $u$  in terms of Green's function.

**4.** Using eigenfunction expansions, solve

$$\frac{\partial g}{\partial t} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial g}{\partial r} \right) + \frac{\delta(r - r_0)\delta(t - t_0)}{r}$$

subject to

$$g(r=1) = 0, \quad g(t < t_0) = 0$$

**5.**  $\frac{\partial g}{\partial t} = \frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} - \delta(x - x_0)\delta(t - t_0)(y - y_0) \quad \text{in } 0 < x < 1, 0 < y < 1$

subject to

$$g = 0 \quad \text{at } x = 0, 1; y = 0, 1$$

$$g = 0 \quad \text{for } t < t_0$$

**6.** Using Green's function, solve

$$u''(x) + u'(x) = x$$

subject to

(a)  $u(0) = 2, \quad u(1) = 0$ ; (b)  $u(0) = 2, \quad u'(1) = 0$

**7.** Find  $g(x/x_0)$  from

$$\frac{d^2 g}{dx^2}(x/x_0) = -\delta(x - x_0)$$

subject to

$$g(0/x_0) = 0, \quad g'(1/x_0) + B_i g(1/x_0) = 0$$

**8.** Solve

$$\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} = \delta(x - x_0)\delta(y - y_0)$$

Find  $g(x, y/x_0, y_0)$  using

(a) the partial eigenfunction expansion,

(b) the total eigenfunction expansion

subject to

$$g(x=0) = g(x=1) = g(y=1) = 0, \quad g'(y=0) = 0$$

Explain why the two answers are not identical?

**9.** Solve

$$\nabla^2 g(x, y/x_0, y_0) = -\delta(x - x_0)\delta(y - y_0)$$

subject to

$$g(x=0) = g(x=1) = g(y=0) = 0$$

$$\frac{\partial g}{\partial y}(y=1) + g(y=1) = 0$$

- 10.** Solve the Green's function problem in the cylindrical coordinate system, subject to  
 $g(r = 1) = g(z = 0) = g(z = a) = 0$   
use (a) the partial eigenfunction expansion,  
(b) the total eigenfunction expansion.

- 11.** Using Green's function, solve

$$u''(x) = 8x^5$$

subject to

$$u(x = 0) = 2 \quad u'(x = 1) = 4$$

- 12.** Solve  $u''(x) = x^3$ ,

subject to

$$u(0) = 0, \quad u'(0) = u'(1)$$

Find the adjoint operator and boundary conditions. Find the causal Green's function and solve for  $u(x)$ .

- 13.** Solve

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{\partial^2 u}{\partial z^2} = 0$$

subject to

$$u(r = 1) = z(1 - z), \quad u(z = 0) = 0, \quad u(z = 1) = 0$$

using (a) separation of variables, and (b) Green's functions.

- 14.** Solve

$$g''(x/x_0) = \delta(x - x_0)$$

subject to

$$g'(0) = 0, \quad g'(1) + g(1) = 0$$

## REFERENCES

- Brebbia, C.A. and Dominguez, J., Boundary Elements and Introductory Course, Computational Mechanics Publishers, Southampton (1989).
- Churchill, R.V., Fourier Series and Boundary Value Problems, McGraw-Hill, New York (1963).
- Kaplan, W., Operational Methods for Linear Systems, Addison-Wesley, Reading, Mass. (1962).
- Kreyszig, E., Advanced Engineering Mathematics, Wiley, New York (1982).
- Stakgold, I., Boundary-value Problems of Mathematical Physics, Macmillan, New York (1968).
- , Green's Functions and Boundary-value Problems, Wiley, New York (1979).
- Weinberger, H.F., A First Course in Partial Differential Equations: With complex variables and transform methods, Wiley, New York (1965).

# 9

## Uniqueness Conditions for Linear and Nonlinear Systems

---

---

The emphasis in this text till now has been on methods of solving equations which arise in modelling different systems. In this endeavour we have restricted ourselves to linear systems. We are interested now in determining whether these systems can admit multiple solutions. If the equation can be proven to have a unique solution, the solutions constructed in the earlier chapters will be the only permissible solutions and no other solution will exist. We would like to obtain the conditions under which an equation can possess multiple solutions.

In this chapter we present the basic principles and arguments which enable us to establish uniqueness of solutions for linear systems. Some of these ideas can be generalised to obtain conditions for nonlinear systems as well. Once again our emphasis is on showing how these principles are sufficiently general. This permits their ready extension from an algebraic system to a differential system. In this chapter we introduce four basic methods which allow us to establish uniqueness of solutions to systems.

**(i) Maximum principles.** We study their applications to linear elliptic and parabolic equations. We will also discuss how they can be used to get uniqueness conditions for some nonlinear systems.

**(ii) Energy methods.** We discuss the applications of this method to linear elliptic and parabolic equations.

**(iii) Fredholm alternative.** We show how the solvability conditions for linear nonhomogeneous equations presented in Chapter 3 in terms of adjoint operators for matrices can be extended and generalised to differential operators.

**(iv) Monotone iteration methods.** This method allows us to establish the existence of solutions to nonlinear equations. We show how this method can be used to construct solutions to algebraic and differential equations and obtain uniqueness conditions.

Our aim is not to present a detailed exposition of the different methods and the associated intricacies, but to explain the basic principles and show how these methods are sufficiently general and can deal with different classes of problems. This will help develop an overall philosophy and enable the student to develop a wider perspective of the applications of mathematics to different problems. It will also sharpen his ability to analyse the results obtained during the course of his research.

## 9.1 MAXIMUM PRINCIPLES

The maximum principles can be used to obtain uniqueness conditions for elliptic and parabolic equations. They were developed primarily for linear systems. They can however be used for obtaining some uniqueness results on some nonlinear elliptic systems as well. We will see this and understand what the equivalent criterion is for a nonlinear algebraic system.

Consider the function  $f(x)$  in  $[a, b]$  such that

$$f''(x) > 0 \quad (9.1a)$$

This clearly implies  $f(x)$  cannot have a maximum in  $(a, b)$  as here  $f''(x)$  would be negative. The maximum of  $f(x)$  is therefore necessarily attained at either of the two end-points,  $x = a$  or  $x = b$ . Hence we can assert

$$f(x) \leq \max (f(a), f(b)) \quad (9.1b)$$

The inequality in (9.1a) is a strict inequality. We would like to extend the above argument to the case when

$$f''(x) \geq 0 \quad (9.2)$$

This is because we will be ultimately interested in obtaining the results for  $f''(x) = 0$ . Define

$$g(x) = f(x) + \varepsilon x^2 \quad (9.3)$$

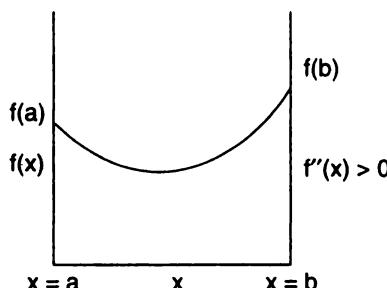
where  $\varepsilon > 0$ . Clearly,  $g(x)$  satisfies (9.1a), i.e.  $g''(x) > 0$ . It follows from (9.1a) and (9.1b) that

$$g(x) \leq \max (g(a), g(b))$$

By definition, from (9.3)

$$f(x) \leq g(x) \leq \max (f(a) + \varepsilon a^2, f(b) + \varepsilon b^2)$$

In the limit  $\varepsilon \rightarrow 0$ , we recover the result in (9.1b), when  $f(x)$  satisfies (9.2). Once again we have proven that the maximum of  $f(x)$  does not occur in  $(a, b)$  but at the end-points  $x = a$  or  $x = b$ . The



**Fig. 9.1** Schematic diagram depicting  $f(x)$  in  $(a, b)$  such that  $f''(x) > 0$ . Maximum of  $f$  occurs at  $x = a$  or  $x = b$ .

results discussed here can be best understood by the graphical depiction in Fig. 9.1. We next see the implication of this result to elliptic systems.

### 9.1.1 Second Order Ordinary Differential Equations

Consider the two point boundary value problem

$$\frac{d^2u}{dx^2} = f(x) \quad \text{in } 0 < x < 1 \quad (9.4a)$$

subject to the Dirichlet conditions

$$u(0) = \alpha \quad (9.4b)$$

$$u(1) = \beta \quad (9.4c)$$

Let  $f(x) > 0$ . This implies that  $u$  does not possess any local maxima in  $(0, 1)$  and that its maxima lies on the boundary, i.e. at  $x = 0$  or  $x = 1$ .

$$u(x) < \max(\alpha, \beta) \quad (9.5)$$

$u(x)$  cannot violate (9.5), as then it would have a local maxima where  $u''(x) < 0$  violating the supposition that  $f(x) > 0$ . We extend this argument for the case  $f(x) \geq 0$ . Define

$$v(x) = u(x) + \varepsilon x^2 \quad (9.6)$$

Clearly,

$$v''(x) > 0$$

and it follows that

$$v(x) < \max(\alpha, \beta + \varepsilon)$$

Once again by definition (9.6),  $u(x) \leq v(x)$ , and in the limit  $\varepsilon \rightarrow 0$ , we obtain (9.5). The validity of (9.5) when  $f(x) \geq 0$  allows us to establish the uniqueness of the solution to (9.4) without actually solving the system. This is important, especially in the context of partial-differential equations on complex geometries where it may not be possible to obtain closed-form analytical solutions.

Assume the contrary, i.e. (9.4) has two distinct solutions  $u_1(x), u_2(x)$ . Define

$$v(x) = u_1(x) - u_2(x) \quad (9.7)$$

Clearly,  $v(x)$  is the solution to the completely homogeneous system

$$\frac{d^2 v}{dx^2} = 0, \quad \text{in } 0 < x < 1 \quad (9.8a)$$

subject to

$$v(0) = 0 \quad (9.8b)$$

$$v(1) = 0 \quad (9.8c)$$

If we can prove that (9.8) has only the trivial solution  $v = 0$ , then it follows from (9.7) that (9.4) has a unique solution. We use maximum principles to prove that the completely homogeneous system (9.8) has only the trivial solution. From the maximum principles it follows that the maximum of  $v(x)$  is on the boundary, and hence

$$v(x) \leq 0 \quad (9.9a)$$

Defining  $\omega(x) = -v(x)$  we see that  $\omega(x)$  also satisfies (9.8). Once again from the maximum principle it follows that  $\omega(x) \leq 0$ , or

$$v(x) \geq 0 \quad (9.9b)$$

The only way the two conflicting statements in (9.9a) and (9.9b) can be satisfied is if  $v(x) = 0$ . This implies  $u_1(x) = u_2(x)$ , and that (9.4) has a unique solution.

This approach enables us to extend the results to more general problems, i.e. elliptic partial differential equations with arbitrary boundaries, without actually solving the problem. We will discuss this next.

### 9.1.2 Second Order Elliptic Partial Differential Equations

We next see how we can generalise these results and extend them to the classic elliptic operator,

i.e. the Laplacian in an arbitrary domain. For the sake of simplicity and concreteness, we assume Dirichlet boundary conditions at all points on the boundary. Consider

$$\nabla^2 u = F(x, y, z) \quad \text{in } V \quad (9.10a)$$

subject to

$$u = f(x, y, z) \quad \text{on } S \quad (9.10b)$$

Let  $F > 0$  and suppose  $u$  attains a maximum at a point  $(x_0, y_0, z_0)$  in  $V$ . At this point

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial y} = \frac{\partial u}{\partial z} = 0$$

$$\frac{\partial^2 u}{\partial x^2} < 0, \quad \frac{\partial^2 u}{\partial y^2} < 0, \quad \frac{\partial^2 u}{\partial z^2} < 0$$

These conditions imply  $\nabla^2 u < 0$  and violate the assumption  $F > 0$ . Consequently,  $u$  cannot have a maximum in  $V$ ; its maximum occurs on  $S$ , the surface bounding  $V$ . Hence

$$u(x, y, z) \leq \max f(x, y, z), \quad \text{for } x, y, z \in S \quad (9.11)$$

Once again we extend this result for the case where  $F$  is non-negative. Define the auxiliary function

$$v(x, y, z) = u(x, y, z) + \varepsilon(x^2 + y^2 + z^2) \quad (9.12a)$$

For  $\varepsilon > 0$ ,

$$\nabla^2 v = F + 6\varepsilon > 0 \quad \text{in } V \quad (9.12b)$$

Let  $R$  be the minimum radius of a three-dimensional sphere which encloses the three-dimensional domain  $V$  (containing the origin (see Weinberger, 1965). Hence it follows from (9.10) and (9.11) that

$$v(x, y, z) < \max (f(x, y, z) + \varepsilon R^2), \quad \text{for } x, y, z \in S$$

Using the definition of  $v$ , from (9.12a) we have

$$u(x) \leq v(x) \leq \max (f(x, y, z) + \varepsilon R^2), \quad \text{for } x, y, z \in S$$

Taking the limit  $\varepsilon \rightarrow 0$ , we again recover (9.11).

We can now establish the uniqueness of the solution to (9.10) as we did earlier. Assume the contrary, i.e.  $u_1, u_2$  are two distinct solutions to the system (9.10). Consider the equation for

$$v(x, y, z) = u_1(x, y, z) - u_2(x, y, z)$$

Clearly,  $v$  satisfies the completely homogeneous system

$$\nabla^2 v = 0 \quad \text{in } V \quad (9.13a)$$

subject to

$$v = 0 \quad \text{on } S \quad (9.13b)$$

From the maximum principles it follows that the maximum of  $v$  occurs on  $S$  and so  $v(x, y, z) \leq 0$ . Once again,  $\omega(x, y, z) = -v(x, y, z)$  satisfies (9.13), and so we conclude  $\omega(x, y, z) \leq 0$  or  $v(x, y, z) \geq 0$ . Arguing as earlier we prove that  $v = 0$  is the only solution to (9.13) and so the nonhomogeneous equation (9.10) has a unique solution. As already seen, (9.10) can be solved by applying the principle of linearity and superposition and using the method of separation of variables and/or Green's functions. We have established now that the solution thus obtained as projections on an eigenbasis is the only solution and no other solution exists.

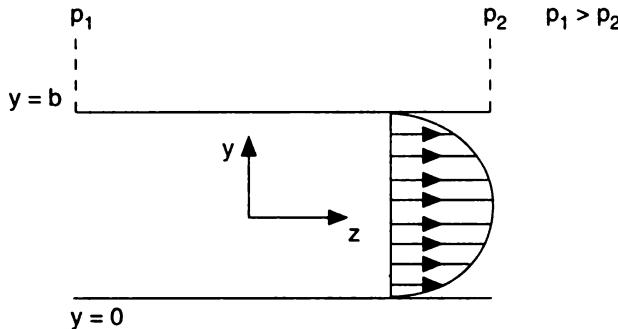
The maximum principles can be applied to elliptic problems where the boundary conditions are not necessarily Dirichlet everywhere on the boundary. Uniqueness is assured for elliptic problems when we have Dirichlet conditions on a part of the boundary, Neumann conditions on some parts and Robin conditions on the remaining parts. It is not the purpose of this book to deal with these cases individually. The interested reader must refer to Protter and Weinberger (1967) for a detailed discussion and application of the maximum principle to these cases. In the next section we will be applying “energy methods” to some of these boundary conditions.

We conclude this section by remarking that the elliptic problem, (9.10), when subject to Neumann conditions everywhere on the boundary, does not have a unique solution. The completely homogeneous equation (9.13a) subject to homogeneous Neumann conditions admits a constant as a nontrivial solution. So the corresponding nonhomogeneous equation (9.10a) has a nonunique solution if at all it has any.

**Example 9.1** Consider the laminar flow between two horizontal plates. The governing equation is given by

$$\frac{\partial^2 v_z}{\partial y^2} = \frac{\rho}{\mu} \frac{\partial p}{\partial z}$$

Determine the direction of fluid flow when the pressure decreases with  $z$  as shown in Fig. 9.2.



**Fig. 9.2** Fluid flow between two stationary parallel plates, with  $\partial p / \partial z < 0$ .

$$\frac{\partial p}{\partial z} < 0, \text{ as pressure decreases with increasing } z.$$

$$\frac{\partial^2 v_z}{\partial y^2} < 0$$

From the no-slip boundary conditions  $v_z(y = 0)$ ,  $v_z(y = b) = 0$ . These are Dirichlet at both end points. From maximum principles it follows that  $v_z$  cannot have a minimum in  $(0, b)$  and hence  $v_z > 0$ . The fluid flows in the direction of increasing  $z$  or from left to right. We have established mathematically in a formal way what we expected intuitively from Physics.

### 9.1.3 Parabolic Systems

The maximum principles can also be applied to parabolic equations (see Protter and Weinberger, 1967). Consider the one-dimensional heat conduction equation

$$Lu = \frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} = f(x, t) \quad \text{in } 0 < x < 1, t > 0 \quad (9.14a)$$

$$u(t = 0) = u_0(x) \quad (9.14b)$$

$$u(x = 0) = \alpha(t) \quad (9.14c)$$

$$u(x = 1) = \beta(t) \quad (9.14d)$$

This represents the equation governing the temperature ( $u$ ) profile in a rod of length 1 with a source of heat,  $f(x, t)$ . The temperature at the two ends ( $x = 0, x = 1$ ) are specified, see (9.14c) and (9.14d), as is the initial temperature (9.14b). The principle of causality tells us that the temperature distribution at any fixed time  $\tau$  is unaffected by changes which occur for  $t > \tau$ . It is therefore sufficient to consider the rectangular region  $0 < x < 1, 0 < t < \tau$ , as shown in Fig. 9.3.  $u(x, t)$  is specified on the three boundaries  $S_1, S_2, S_3$ , where

$$S_1 \text{ is } (x = 0, 0 < t \leq \tau) \quad (9.15a)$$

$$S_2 \text{ is } (0 < x < 1, t = 0) \quad (9.15b)$$

$$S_3 \text{ is } (x = 1, 0 < t \leq \tau) \quad (9.15c)$$

**Theorem 9.1** Let  $Lu \geq 0$  in the rectangular region ABCD (Fig. 9.3), where  $L$  is the linear operator in (9.14a). Then the maximum of  $u$  on the closed rectangular region  $E$  in Fig. 9.3 must occur on one of the three sides  $S_1, S_2$  or  $S_3$ .

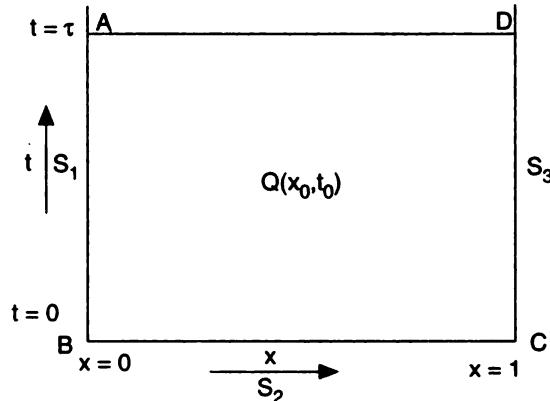


Fig. 9.3 Domain of problem for parabolic systems.

**Proof.** Let  $M$  be the maximum value of  $u$  on the three sides  $S_1, S_2, S_3$ . Assume that there is a point  $Q(x_0, t_0)$  in  $E$ , where  $u$  is  $M_1 > M$ . Define the auxiliary function

$$\omega(x) = \frac{M_1 - M}{2} (x - x_0)^2$$

and denote

$$v(x, t) = u(x, t) + \omega(x)$$

Since  $u \leq M$  on  $S_1, S_2, S_3$ , we have

$$v(x, t) = u(x, t) + \omega(x) \leq M + \frac{M_1 - M}{2} < M_1 \quad \text{on } S_1, S_2, S_3$$

Also,  $v(x_0, t_0) = M_1$ , by definition.

Assume the maximum of  $v$  occurs along the line  $t = \tau$  or inside the rectangular region. Since  $Lv > 0$ ,  $v$  cannot have an interior maximum. At a maximum along  $t = \tau$ , we would have  $\partial^2 v / \partial x^2 < 0$ , implying  $\partial v / \partial t < 0$ . Thus  $v$  must have been larger at an earlier instant of time so that the maximum in  $E$  cannot be on the surface  $t = \tau$ . So we have a contradiction because of the assumption  $u(x_0, t_0) > M$ . Hence,  $u(x_0, t_0) < M$ , the maximum value on the surface  $S_1, S_2, S_3$ . This result in turn can be used to establish the uniqueness of the solution to (9.14). Let  $u_1, u_2$  be two solutions to the system (9.14). Define

$$v(x, t) = u_1(x, t) - u_2(x, t)$$

This satisfies the completely homogeneous system

$$Lv = 0 \quad (9.16a)$$

subject to

$$v(t = 0) = 0 \quad (9.16b)$$

$$v(x = 0) = 0 \quad (9.16c)$$

$$v(x = 1) = 0 \quad (9.16d)$$

From the maximum principles it follows that  $v(x, t) \leq 0$  in  $E$ . Similarly,  $\omega(x, t) = -v(x, t)$  also satisfies the system (9.16), and so we have  $\omega(x, t) \leq 0$  or  $v(x, t) \geq 0$  in  $E$ . Hence the solution to (9.16) is only the trivial solution  $v(x, t) = 0$ .

These results can be easily extended to arbitrary three-dimensional spatial domains  $V(x, y, z)$ , where

$$Lu = \nabla^2 u - \frac{\partial u}{\partial t} > 0$$

The region of interest now is the four-dimensional space-time cylinder represented abstractly in Fig. 8.4, and not the two-dimensional region of Fig. 9.3. This is denoted as  $Vx(0, \infty)$ . The principle of causality again allows us to restrict ourselves to the finite cylinder  $Vx(0, \tau)$ . The function  $u$  in this cylinder is determined by the values of  $u$  in  $V$  at  $t = 0$ , and the values of  $u$  on the cylindrical wall  $Sx(0, \tau)$ .

The maximum principle states that the maximum of  $u$  occurs on the surface  $t = 0$ , or along the curved surface of the cylinder. If  $Lu > 0$ , then the maximum cannot occur at an interior point of  $E$ , since here  $\partial u / \partial t = 0$ ,  $\nabla^2 u < 0$ , and this would violate  $Lu > 0$ . The maximum cannot be attained at  $t = \tau$ , as here  $\nabla^2 u < 0$  and  $Lu > 0$  implies  $\partial u / \partial t < 0$ . This in turn means that the maximum of  $u$  occurred at an earlier instant of time. Hence once again the maximum of  $u$  occurs on the bottom surface or the curved surface of the space-time cylinder  $Vx(0, \tau)$ .

The maximum principle for the parabolic problem can be easily extended to the case when  $Lu \geq 0$  and used to obtain uniqueness conditions for the general three-dimensional problem just as in the case of the one-dimensional problem. We refer the interested reader to Protter and Weinberger (1967) for a detailed discussion on the maximum principles for parabolic systems.

#### 9.1.4 Physical Basis of Maximum Principles

The maximum principles which we have explained so far have a physical basis. Consider the steady state temperature distribution in a slab with a constant uniform heat sink  $q$  (where  $q > 0$ ). The temperature is governed by

$$\nabla^2 T = q \quad \text{in } V \text{ with } T = T_0 \text{ on } S$$

Since we have a uniform heat sink,  $q > 0$  throughout  $V$ . As a result, heat is removed at every point in  $V$  and so the temperature in  $V$  cannot exceed  $T_0$ , the value on the boundary. If  $q$  were to vanish at some points in  $V$ , then again we would have no source term present in  $V$ , and so the temperature again cannot exceed  $T$ . Of course, this result is not valid if  $q$  is negative in some region in  $V$  because this would correspond to a source of heat somewhere in  $V$ . The maximum principle is a formal mathematical statement of this feature. A similar physical basis can be established for the maximum principles applied to parabolic systems. We leave this as an exercise to the reader.

### 9.1.5 Elliptic Nonlinear Systems

The maximum principle has allowed us to establish the uniqueness of the solutions to the linear elliptic and parabolic problems under most conditions. The only situation where we have nonuniqueness is in the elliptic problem when the boundary conditions are Neumann everywhere on  $S$ . We would also like to extend and apply maximum principles to obtain uniqueness conditions to some nonlinear systems (see Protter, 1967). Consider the nonlinear one-dimensional elliptic problem

$$\frac{d^2u}{dx^2} + f(u) = 0 \quad \text{in } 0 < x < 1 \quad (9.17a)$$

subject to

$$u(0) = \gamma \quad (9.17b)$$

$$u(1) = \beta \quad (9.17c)$$

Such equations describe the steady state behaviour of systems. Let us assume that the equation possesses at least one solution (since the equation is nonlinear it may have no solution). We would like to determine under what conditions the solution is unique, i.e. we have exactly one solution. Assume  $u_1(x), u_2(x)$  are two solutions to the system (9.17). Consider

$$v(x) = u_1(x) - u_2(x)$$

This satisfies

$$\frac{d^2v}{dx^2} + f'(\omega)v = 0 \quad (9.18a)$$

subject to

$$v(0) = 0, \quad v(1) = 0 \quad (9.18b)$$

where  $f'(\omega(x))$  is the derivative  $\partial f / \partial u$  evaluated at  $u(x) = \omega(x)$ , with  $\omega(x) = u_1(x) + \alpha(x)(u_2(x) - u_1(x))$ , where  $0 \leq \alpha(x) \leq 1$ . This is an extension to the composite function  $f(u(x))$  of the mean-value theorem in calculus, which states that for a function  $f(x)$  in  $(a, b)$

$$f(x) = f(x_0) + f'(c)(x - x_0)$$

for some point  $c$  satisfying  $x \leq c \leq x_0$  and when  $x, x_0$  both lie in  $[a, b]$ . In particular, (9.18) is not a linearisation of (9.17a).  $v$  satisfies a linear homogeneous equation and is subject to homogeneous boundary conditions. Assume now that  $f'(\omega) < 0$ . We now prove that under this restriction, (9.18) admits only the trivial solution  $v = 0$ .

Let  $v(x)$  be nonzero. It can be either negative for some or all  $x$  or positive for some or all  $x$ . Let  $v(x)$  be negative in some interval of  $(0, 1)$ . Then clearly, since  $v(x)$  vanishes at the end points, it attains a minimum in this interval. Here  $v''(x) > 0$ , and so  $v'' + f'(\omega)v > 0$ . Hence (9.18a) is not satisfied in this interval at the minimum point. Consequently,  $v(x)$  cannot be negative in any part

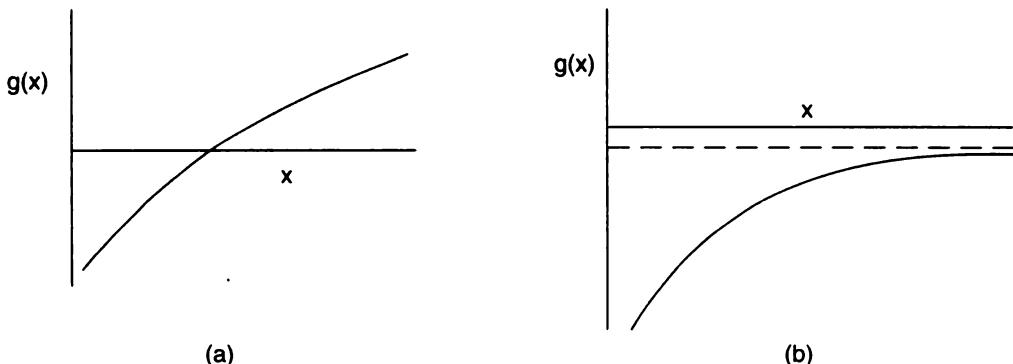
of  $(0, 1)$ . Let  $v(x)$  be positive in some interval of  $(0, 1)$ . Then again,  $v(x)$  must have a maximum somewhere in this interval as  $v$  vanishes at the two end-points. Here,  $v''(x) < 0$  and  $v''(x) + f''(\omega)v < 0$ , and again  $v(x)$  cannot satisfy (9.18a) at this maximum point. So  $v(x)$  cannot be positive in any sub-interval of  $(0, 1)$ . Hence  $v(x)$  has to identically vanish in  $(0, 1)$ . The nonlinear elliptic equation has a unique solution (if at all it has one) when  $f'(u) < 0$ . An application of this method to a chemical engineering reaction system can be found in Pushpavanam and Narayanan (1988a).

It must be emphasised that in establishing this result we have assumed that (9.17) has a solution. In particular, we take the existence of a solution for granted. We make special mention of this here since nonlinear equations in general may have no solutions. Let us now see if there is any analogous uniqueness criterion for nonlinear algebraic equations.

Consider the equation

$$g(x) = 0 \quad (9.19)$$

In Fig. 9.4 we show graphically two situations for  $g(x)$  which are such that the functions are monotonically increasing. In the first case the function has a zero and in the second case it has no



**Fig. 9.4** Monotonic function  $g(x)$  such that: (a) a solution exists, (b) no solution exists.

solution. Thus monotonicity by itself does not guarantee the uniqueness of a solution. Uniqueness is assured if a solution exists and if the function is monotonic. Let us assume once again that (9.19) has at least one solution. If  $g(x)$  is a monotonic function, then clearly (9.19) has only a unique solution. Once again if we take the existence of the solution for granted as we did for (9.17), we have uniqueness when

$$g'(x) > 0 \quad \text{or} \quad g'(x) < 0$$

This is in contrast to (9.17) which had a unique solution only for  $f'(u) < 0$ . This apparent inconsistency or paradox can be explained if we realise that the two nonlinearities  $f(u)$  and  $g(u)$  in the equations play different roles in (9.17) and (9.19). This can best be seen if we approximated the second derivative in (9.17a) using a central-difference scheme. Assuming only one grid point at the centre, we have

$$4(u(0) + u(l) - 2u^*) + f(u^*) = 0 \quad (9.20a)$$

Using the boundary conditions (9.17b)–(9.17c) on  $u$ , we obtain

$$4(\gamma + \beta - 2u^*) + f(u^*) = 0 \quad (9.20b)$$

or

$$h(u^*) = 0 \quad (9.20c)$$

Here,  $u^* = u(1/2)$ , and the grid spacing  $1/2$  yields the factor 4. This nonlinear algebraic equation in  $u^*$  is the true (actually approximate) representation of the elliptic equation and can be viewed as  $h(u^*) = 0$ . This has a unique solution (if at all it has a solution) when  $h$  is monotonic. We are guaranteed the monotonicity of  $h(u^*)$  only when  $f(u^*)$  is monotonically decreasing, i.e.  $f'(u^*) < 0$ . If  $f'(u^*) > 0$ ,  $h(u^*)$  is not necessarily monotonic, we are not assured of uniqueness. The function  $g(x)$  in (9.19) is therefore equivalent to  $h(u)$  in (9.20c), and not  $f(u)$  in (9.17a).

For both the algebraic equation and the boundary-value problem, these are sufficient conditions for uniqueness. In both cases, we have tacitly assumed that a solution exists.

**Example 9.2** The temperature profile in a catalytic pellet sustaining a zeroth-order reaction is governed by (see Aris, 1975)

$$\frac{d^2T}{dx^2} + \delta e^{-E/RT} = 0$$

subject to

$$T = T_0 \text{ at } x = 0, \quad (9.21a)$$

Assume that a solution exists.

For an exothermic reaction  $\delta > 0$ , and for an endothermic reaction,  $\delta < 0$ . The nonlinearity arises from the Arrhenius temperature dependency of the reaction rate. Identifying  $f(T) = \delta e^{-E/RT}$ , we have

$$f'(T) = \frac{\delta E}{RT^2} e^{-E/RT}$$

For an endothermic reaction  $\delta < 0$ ,  $f'(T) < 0$ , and so we have a unique solution. For an exothermic reaction we cannot prove uniqueness using this method. This does not mean that the exothermic reaction does not have a unique solution, as  $f'(T) < 0$  is only a sufficient condition for uniqueness. The temperature in a CSTR sustaining a zeroth-order reaction is

$$0 = T_0 - T + \delta e^{-E/RT} \quad (9.21b)$$

Here  $T_0$  represents the temperature of the inlet stream.

For  $\delta < 0$  (i.e. an endothermic reaction), we are once again guaranteed uniqueness since the function is monotonic, and changes sign in  $(-\infty, T_0)$ . For an exothermic reaction we again cannot prove uniqueness. A comparison of (9.21a), (9.21b), and (9.20b) will give some insight into how the same uniqueness conditions hold for nonlinear algebraic and boundary-value problems arising in modelling physically similar systems. Pushpavanam and Narayanan (1988), have applied this method to establish uniqueness conditions for the  $m$ th-order reaction in a catalyst pellet with a temperature-dependent transport coefficient.

## 9.2 ENERGY METHODS

We have so far seen how the maximum principles can be used to obtain the uniqueness conditions to linear nonhomogeneous problems. The same results can be obtained by using energy methods. We briefly describe the applications of these methods to elliptic and parabolic systems and refer the interested reader to Weinberger (1965) for more details.

### 9.2.1 Elliptic Problems

Consider the nonhomogeneous equation (9.10) for  $u$  and the associated homogeneous equation

(9.13) for  $v$ . Taking the inner-product of (9.13a) with  $v$  (or simply multiplying the equation by  $v$  and integrating over the volume  $V$ ), we have

$$\int_V v \nabla^2 v = 0 \quad (9.22a)$$

Using the identities (8.36a), (8.36b), it follows that

$$\int_S v \nabla v \cdot n \, ds = \int_V |\nabla v|^2 \, dv \quad (9.22b)$$

When we have Dirichlet boundary conditions everywhere on the surface,  $v = 0$  on  $S$ . Substituting this in (9.22b), we obtain

$$\int_V |\nabla v|^2 \, dv = 0 \quad (9.23)$$

Since the integrand, i.e.  $|\nabla v|^2$ , can never be negative, it has to be identically zero throughout  $V$  to satisfy (9.23). This implies

$$v = \text{constant}$$

However, since  $v$  equals zero on the surface, it has to vanish identically, as we are interested only in continuous solutions.

Consider the problem (9.10a) with Robin conditions on the part  $S_R$  of the boundary and Dirichlet conditions on the part  $S_D$  of the boundary

$$u = u_D \quad \text{on } S_D$$

$$n \cdot \nabla u + hu_R = hu_R \quad \text{on } S_R$$

Applying the resulting homogeneous boundary conditions for  $v$  (the difference of two solutions), we now have

$$\int_V |\nabla v|^2 \, dv + h \int_{S_R} v^2 \, dS_R = 0 \quad (9.24)$$

where  $h > 0$ .

This is the sum of two non-negative quantities. Consequently, (9.24) can hold only if

$$v = 0 \quad \text{on } S_R \quad (9.25a)$$

$$\nabla v = 0 \quad \text{in } V \quad (9.25b)$$

Equation (9.25b) implies  $v = \text{constant}$ . It follows again from (9.25a) (since we are looking only for continuous solutions) that  $v$  must identically vanish in  $V \cup S$ . We have established that the elliptic system has a unique solution when a part of the boundary has Dirichlet conditions and a part has Robin boundary conditions.

## 9.2.2 Parabolic Equations

Consider the completely nonhomogeneous parabolic problem

$$\frac{\partial u}{\partial t} - \nabla^2 u = F \quad \text{in } V \quad (9.26a)$$

subject to

$$u(t=0) = u_0 \quad \text{in } V \quad (9.26b)$$

$$u = u_D \quad \text{on } S \quad (9.26c)$$

Here we follow the same approach as we did for elliptic equations. Let there be two solutions  $u_1$ ,

$u_2$  to this equation. Define

$$v = u_1 - u_2$$

Clearly,  $v$  satisfies the completely homogeneous system

$$\frac{\partial v}{\partial t} - \nabla^2 v = 0 \quad \text{in } V \quad (9.27a)$$

subject to

$$v(t = 0) = 0 \quad \text{in } V \quad (9.27b)$$

$$v = 0 \quad \text{on } S \quad (9.27c)$$

Multiplying (9.27a) by  $v$  and integrating over the volume  $V$ , we obtain

$$\frac{d}{dt} \int_V \frac{v^2}{2} dV = \int_S v \nabla v \cdot n - \int_V |\nabla v|^2 dV \quad (9.28a)$$

(This is not the same as taking an inner-product with  $v$ . Why?)

Using (9.27c), we reduce this expression to

$$\frac{d}{dt} \int_V \frac{v^2}{2} dV = - \int_V |\nabla v|^2 dV \quad (9.28b)$$

The term on the right is non-positive. Consequently, the integral on the left decreases with time. The integral is zero at  $t = 0$  from (9.27b). So for  $t > 0$ , (9.28b) implies the integral is non-positive. But the integral, being that of a squared quantity, has to be non-negative. So  $v$  has to identically vanish for all  $x, t$ . This establishes a unique solution for (9.26). This result is valid for all combinations of boundary conditions, i.e. Neumann, Robin, Dirichlet, as also the completely Neumann problem.

The term on the left of (9.28) is representative of the change in ‘kinetic energy’ of the system. Hence this method is called an energy method (see Weinberger, 1965).

### 9.3 FREDHOLM ALTERNATIVE

In Chapter 3 we established solvability conditions for nonhomogeneous linear algebraic systems of the form

$$Ax = b \quad (9.29a)$$

The nature of the solutions (their existence, uniqueness or multiplicity) of (9.29a) is related to the solutions of the two homogeneous problems

$$Au = 0 \quad (9.29b)$$

$$A^*v = 0 \quad (9.29c)$$

We will extend this solvability condition to partial differential equations now. Once again this is possible only because of the operator notion introduced in this book. Otherwise, we would be forced to determine the solvability condition of (9.29a) in terms of the rank of the matrix  $A$  and the rank of the augmented matrix  $(A, b)$  as done in classical books on linear algebra. The definition of rank is not applicable to differential operators and so it is necessary to cast the problem in operator notation (see Noble and Daniel (1977), Stakgold (1979)). We restrict ourselves to elliptic problems where the operator is  $L$ . This is only fair since (9.29) is the solution to a steady state problem, as is the elliptic problem. We can associate with the completely nonhomogeneous problem.

$$Lu = F(x, y, z) \quad \text{in } V \quad (9.30a)$$

$$Bu = u_D \quad \text{on } S \quad (9.30b)$$

Consider the two problems:

1. The homogeneous problem

$$Lu_H = 0 \quad \text{in } V \quad (9.31a)$$

subject to

$$Bu_H = 0 \quad \text{on } S \quad (9.31b)$$

2. The homogeneous adjoint problem

$$L^*v = 0 \quad \text{in } V \quad (9.32a)$$

subject to

$$B^*v = 0 \quad \text{on } S \quad (9.32b)$$

where  $L^*$ ,  $B^*$  represent the adjoint operator, and the boundary operator associated with  $L$ ,  $B$ . The Fredholm alternative states that: (a) If the system (9.31) has only the trivial solution, so does the system (9.32), and the system (9.30) then has a unique solution, and (b) if the system (9.31) admits a nontrivial solution then so does the system (9.32). The nonhomogeneous system (9.30) has a solution if and only if

$$\langle v, F \rangle = J(u, v) \quad (9.33)$$

for all  $v$  which satisfy (9.32).

It is easy to prove why the condition is necessary. Taking the inner-product of (9.30a) with  $v$  and (9.32a) with  $u$  and subtracting, we obtain (9.33). The bilinear concomitant in (9.33) does not vanish since  $u$  does not satisfy the homogeneous conditions (9.30b). We refer the reader to Stakgold (1979) for a proof of the sufficiency of (9.33) as a solvability condition.

The most frequently encountered form of  $L$  is the Laplacian  $\nabla^2$ . This operator subject to the most frequently (classically) encountered set of boundary conditions is self-adjoint, as seen in Chapter 8. Consequently, it has a complete set of eigenfunctions. The solution  $v$  in (9.33) satisfies the original and adjoint problem.

The Fredholm alternative for the matrix operator can be interpreted as constraining (see Chapter 3) the nonhomogeneity such that it satisfies the linear dependencies in the columns or rows of  $A$  if any. We next see an example which explains the physical significance of the solvability condition for the differential operator.

### Example 9.3 Determine the conditions under which

$$\nabla^2 T = f(x) \quad \text{in } V \quad (9.34a)$$

subject to

$$n \cdot \nabla T = q \quad \text{on } S \quad (9.34b)$$

will have a solution. Discuss the physical significance of the result.

The homogeneous equation is

$$\nabla^2 u = 0 \quad \text{in } V \quad (9.35a)$$

$$n \cdot \nabla u = 0 \quad \text{on } S \quad (9.35b)$$

The homogeneous adjoint equation is also given by the system (9.35). This admits a nonzero solution, i.e.  $u = c_1$ , a constant. Taking the inner-product of both sides of (9.34a) with  $c_1$  and of (9.35a) with  $T$  and subtracting, we obtain

$$c_1 \int_S \nabla T \cdot n \, dS = c_1 \int_V f(x) \, dV$$

or

$$\int_S q \, dS = \int_V f(x) \, dV \quad (9.35c)$$

Equations (9.34) have a solution if this condition is satisfied. This states that the total amount of heat supplied to the body across the surface (LHS) must equal the total rate at which it is consumed inside the volume (RHS) for the system to possess a steady state. Therefore, the solvability condition (9.33) is only a mathematical statement of what we expect from the physics of the problem.

A frequently encountered application of this solvability condition is while solving nonlinear differential equations using the method of perturbation. Here we solve for the dependent variable in terms of a series expansion where each term of the series is given by the solution to a linear equation. The original nonlinear equation is broken up into a sequence of linear problems and the solvability condition (9.33) plays an important role in determining the solution here (see Nayfeh 1973, 1981). The Fredholm alternative can also be used in solving optimal control problems (San and Stephanopoulos, 1984).

## 9.4 MONOTONE-ITERATION METHODS

In deriving the uniqueness conditions for the nonlinear elliptic problem (9.17a), we have assumed that the equation had a solution. This was necessary as nonlinear equations may have no solutions. In this section we discuss a method which will prove the existence of solutions to nonlinear equations, when certain conditions are satisfied. We will then obtain conditions for the uniqueness of the solutions to these equations. To show the versatility of the technique, we discuss both algebraic and differential equations. The maximum principle plays a vital role in the latter (refer Section 9.4.2).

### 9.4.1 Algebraic Systems

The nonlinear equation in one variable can be written as

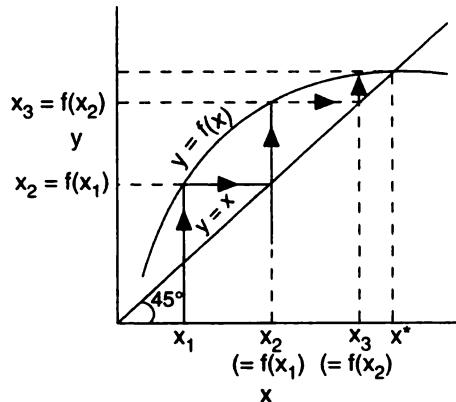
$$x = f(x) \quad (9.36a)$$

The solution to this equation has to be found using a numerical procedure. One such method is the method of successive substitutions. Here we generate a sequence of iterates starting from a given  $x_0$ , using the relation

$$x_{n+1} = f(x_n) \quad (9.36b)$$

The function  $f$  is called the map. The sequence  $\{x_i\}$  generated by (9.36b), i.e.  $(x_1, x_2, x_3, \dots, x_n)$ , may converge to an  $x^*$ . This  $x^*$  is called the fixed point of the map (9.36b) and is a solution of (9.36a). The monotone iteration method enables us to construct the solution to the nonlinear equation (9.36a), using (9.36b) and also establish its existence. The graphical basis of this method is shown in Fig. 9.5. The fixed point is the intersection of the curve  $f(x)$  with the bisectrix ( $45^\circ$  line).

To eliminate any confusion existing in the mind of the reader, we reiterate the notation which will be used while discussing maps and the Newton-Raphson method. The subscript/superscript  $n$  refers not to the  $n$ th coordinate of a vector as in Chapters 2 and 3, but the  $n$ th iterate of the map. Most of our discussion will centre around one-dimensional maps. The change in notation is necessitated for consistency with the literature.



**Fig. 9.5** Approach to fixed point  $x^*$  of a one-dimensional map starting from  $x_1$ , geometric representation.

**Definition 9.1** An **upper solution** of (9.36a)  $x^0$  is such that  $x^0 \geq f(x^0)$ , and a **lower solution** of (9.36a)  $x_0$  is such that  $x_0 \leq f(x_0)$ .

Consider the function  $f(x)$  shown in Fig. 9.6. Clearly,  $f(a) < a, f(b) < b$ . So  $a, b$  are candidates for upper solutions  $x$ . Also,  $f(c) > c, f(d) > d$ . So  $c, d$  are candidates for lower solutions  $x_0$ .

Assume now that  $x_0 \leq x^0$  and  $f(x)$  is an increasing function in  $(x_0, x^0)$ . We will now discuss the behaviour of the two sequences  $\{x_n\}$  and  $\{x^n\}$  obtained by using the map (9.36b), starting with the initial guess of  $x_0$  and  $x^0$  respectively.

$$x^1 = f(x^0) \leq x^0 \quad (9.37a)$$

$$x_1 = f(x_0) \geq x_0 \quad (9.37b)$$

Moreover,  $x^0 \geq x_0$ , and since  $f(x)$  is an increasing function in  $(x_0, x^0)$ ,

$$f(x^0) \geq f(x_0)$$

or

$$x^1 \geq x_1 \quad (9.37c)$$

So we now have the ordered sequence

$$x_0 \leq x_1 \leq \dots \leq x^1 \leq x^0$$

Similarly,

$$x^2 = f(x^1), \quad x_2 = f(x_1)$$

From the inequalities  $x^1 \geq x_1, x^1 \leq x^0, x_0 \leq x_1$  and as  $f(x)$  is an increasing function in  $(x_0, x^0)$ , we have

$$x^2 \geq x_2, \quad x^2 \leq x^1, \quad x_1 \leq x_2$$

Using mathematical induction, we now prove that

$$x^{n+1} \geq x_{n+1}, \quad x_{n+1} \geq x_n, \quad x^{n+1} \leq x^n \quad (9.38a)$$

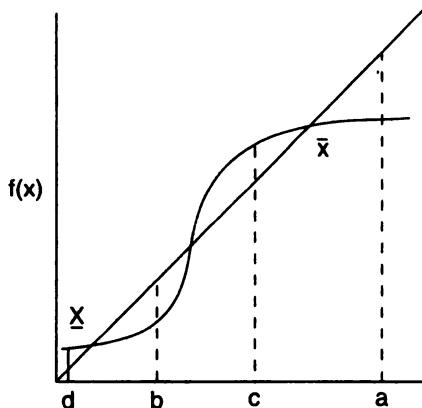
given that

$$x^n \geq x_n, \quad x_n \geq x_{n-1}, \quad x^n \leq x^{n-1} \quad (9.38b)$$

Remembering that  $f(x)$  is an increasing function in  $(x_0, x^0)$  and that  $x_n, x_{n-1}, x^n, x^{n-1} \in (x_0, x^0)$ , and by applying the map ' $f$ ' on both sides of the inequalities in (9.38b), we get the inequalities in (9.38a). The elements  $x_n, x^n$  can be arranged as

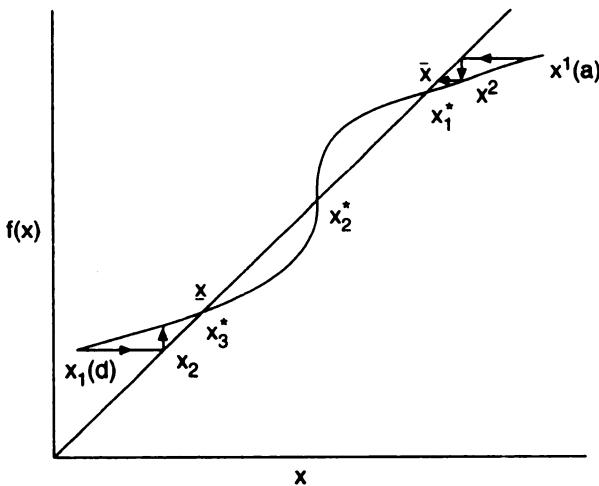
$$x_0 \leq x_1 \leq x_2 \dots \leq x_n \dots \leq x^n \leq \dots \leq x^2 \leq x^1 \leq x^0$$

The sequence  $\{x_i\}$  is a monotonically nondecreasing sequence bounded from above and  $\{x^i\}$  is a monotonic nonincreasing sequence bounded from below. Therefore, both sequences converge. The limit of the sequence  $\{x_i\}$  is called the *minimal solution* and is denoted by  $\underline{x}$  and that of  $\{x^i\}$  is called the *maximal solution* and is denoted by  $\bar{x}$ . The minimal and maximal solutions are fixed points of (9.36a), see Fig. 9.6. They are called so, as all other fixed points in  $(x_0, x^0)$  lie in  $(\underline{x}, \bar{x})$ .



**Fig. 9.6** Maximal  $\bar{x}$  and minimal  $\underline{x}$  solutions of map, generated by upper solution  $a$ , and lower solution  $d$ . Maximal and minimal solutions are fixed points.

Sequence  $x^i$  starting from  $a$  converges to the fixed point  $x_1^*$ . The sequence  $x_i$  starting from  $d$  converges to the fixed point  $x_3^*$  (Fig. 9.7). Thus in the interval  $(d, a)$ , where  $d$  is the lower solution and  $a$  is the upper solution, the maximal solution is  $x_1^*$ , and the minimal solution is  $x_3^*$ . The other fixed point  $x_2^*$  lies in between these solutions. Stakgold (1979) contains a detailed exposition of the monotone iteration method, the maximal and minimal solutions, for algebraic as well as elliptic systems.



**Fig. 9.7** Maximal and minimal solutions of a map.

### 9.4.2 Elliptic Systems

We show how this method can be extended to an elliptic problem. This extension is rendered feasible only by an application of the maximum principles. A similar technique is also valid for parabolic problems, but we will not be concerned with this here. Consider a nonlinear boundary-value problem of the form

$$\left. \begin{array}{l} \nabla^2 u + f(u) = 0 \quad \text{in } V \\ \text{subject to} \\ u = 0 \quad \text{on } S \end{array} \right\} \quad (9.39)$$

The nonlinearity is assumed to be confined to the term  $f(u)$ . The boundary conditions are assumed to be Dirichlet. A simple transformation can always render the boundary condition homogeneous (if the boundary condition is linear).

We assume  $f(u)$  to be an increasing function of  $u$ . An upper solution  $u^0$  is defined so that

$$\left. \begin{array}{l} \nabla^2 u^0 + f(u^0) \leq 0 \quad \text{in } V \\ \text{satisfying} \\ u^0 \geq 0 \quad \text{on } S \end{array} \right\} \quad (9.40a)$$

A lower solution  $u_0$  is defined as

$$\left. \begin{array}{l} \nabla^2 u_0 + f(u_0) \geq 0 \quad \text{in } V \\ \text{satisfying} \\ u_0 \leq 0 \quad \text{on } S \end{array} \right\} \quad (9.40b)$$

Also, let  $u_0 \leq u^0$ . Consider the two sequences generated by

$$\left. \begin{array}{l} \nabla^2 u^n + f(u^{n-1}) = 0 \quad \text{in } V \\ \text{subject to} \\ u^n = 0 \quad \text{on } S \end{array} \right\} \quad (9.41a)$$

$$\left. \begin{array}{l} \nabla^2 u_n + f(u_{n-1}) = 0 \quad \text{in } V \\ \text{subject to} \\ u_n = 0 \quad \text{on } S \end{array} \right\} \quad (9.41b)$$

Two sequences are generated by the linear problems (9.41a) and (9.41b).

By definition

$$\left. \begin{array}{l} \nabla^2 u^1 + f(u^0) = 0 \quad \text{in } V \\ \text{subject to} \\ u^1 = 0 \quad \text{on } S \end{array} \right.$$

$$\left. \begin{array}{l} \nabla^2 u^0 + f(u^0) \leq 0 \quad \text{in } V \\ \text{subject to} \\ u^0 \geq 0 \quad \text{on } S \end{array} \right.$$

Subtracting,

$$\left. \begin{array}{l} \nabla^2(u^1 - u^0) \geq 0 \quad \text{in } V \\ \text{subject to} \\ u^1 - u^0 \leq 0 \quad \text{on } S \end{array} \right.$$

From maximum principles, the maximum of  $u^1 - u^0$  occurs on the boundary  $S$ . Since it is nonpositive on  $S$ , we conclude that it is non-positive in  $V$ , and we have

$$u^1 \leq u^0 \quad \text{in } V \quad (9.42a)$$

Similarly, we can obtain

$$\nabla^2(u_0 - u_1) \geq 0 \quad \text{in } V$$

subject to

$$u_0 - u_1 \leq 0 \quad \text{on } S$$

and conclude from maximum principles that

$$u_0 \leq u_1 \quad \text{in } V \quad (9.42b)$$

We would like to see how  $u_1$  and  $u^1$  are ordered. From the definition of  $u^1$  and  $u_1$ , we have

$$\nabla^2(u_1 - u^1) = f(u^0) - f(u_0) \quad \text{in } V$$

subject to

$$u_1 - u^1 = 0 \quad \text{on } S$$

Since  $f(u)$  is an increasing function of  $u$ , we have for  $u^0 \geq u_0$

$$\nabla^2(u_1 - u^1) \geq 0 \quad \text{in } V$$

subject to

$$u_1 - u^1 = 0 \quad \text{on } S$$

So that again from maximum principles we have

$$u_1 \leq u^1$$

We again prove by mathematical induction the properties of the sequences  $\{u^i\}$  and  $\{u_i\}$  generated by (9.41). Assume that

$$u^n < u^{n-1}, \quad u_{n-1} < u_n, \quad u_n < u^n$$

We have proven this for  $n = 1$  and will now extend this to any arbitrary  $n$ .

$$\left. \begin{array}{l} \nabla^2 u^{n+1} + f(u^n) = 0 \quad \text{in } V \\ u^{n+1} = 0 \quad \text{on } S \end{array} \right\} \quad (9.43a)$$

and

$$\left. \begin{array}{l} \nabla^2 u_{n+1} + f(u_n) = 0 \quad \text{in } V \\ u_{n+1} = 0 \quad \text{on } S \end{array} \right\} \quad (9.43b)$$

also

$$\left. \begin{array}{l} \nabla^2 u^n + f(u^{n-1}) = 0 \quad \text{in } V \\ u^n = 0 \quad \text{on } S \end{array} \right\} \quad (9.43c)$$

and

$$\left. \begin{array}{l} \nabla^2 u_n + f(u_{n-1}) = 0 \quad \text{in } V \\ u_n = 0 \quad \text{on } S \end{array} \right\} \quad (9.43d)$$

Remembering that  $f(u)$  is an increasing function of  $u$ , we obtain by subtracting the relevant equations

$$\nabla^2(u^n - u^{n+1}) \leq 0 \quad \text{in } V$$

subject to

$$u^n - u^{n+1} = 0 \quad \text{on } S$$

and

$$\nabla^2(u_{n+1} - u_n) \leq 0 \quad \text{in } V$$

subject to

$$u_{n+1} - u_n = 0 \quad \text{on } S$$

and

$$\nabla^2(u_{n+1} - u^{n+1}) \geq 0 \quad \text{in } V \text{ with } u_{n+1} - u^{n+1} = 0 \quad \text{on } S.$$

So it follows from maximum principles that

$$u^n \geq u^{n+1}, \quad u_{n+1} \geq u_n, \quad u_{n+1} \leq u^{n+1}$$

Thus we have the following ordering of the sequences  $\{u_n\}$  and  $\{u^n\}$

$$u_0 \leq u_1 \leq u_2 \dots \leq u_n \leq \underline{u} \dots \leq \bar{u} \leq u^n \dots \leq u^1 \leq u^0$$

Once again the sequence  $\{u_n\}$  is a monotonically non decreasing sequence bounded from above and  $\{u^n\}$  is a monotonically non increasing sequence bounded from below. So they converge to the minimal and maximal solutions  $\underline{u}$ ,  $\bar{u}$  respectively of (9.39). Pushpavanam and Narayanan (1988b) have shown how this method can be extended to conjugate fluid-solid systems governed by integro-differential equations.

So far we have seen how the existence of the solution to nonlinear equations can be established using similar techniques for algebraic and differential equations. We will now see how we can derive uniqueness conditions for these two classes of systems.

#### 9.4.3 Uniqueness Conditions

We have seen how the existence of a solution can be established using the monotone iteration methods for nonlinear systems. All the solutions to the nonlinear system in  $[u_0, u^0]$  lie in the interval  $[\underline{u}, \bar{u}]$ . The upper and lower solutions  $u^0, u_0$  are usually determined by some physical reasoning. Consequently, all feasible solutions that exist lie in  $[\underline{u}, \bar{u}]$ . Uniqueness conditions can be obtained if we can determine conditions under which

$$\underline{u} = \bar{u} \tag{9.44}$$

**Elliptic systems.** We consider the differential system first, as the uniqueness criterion can be established easily here. By definition,

$$\begin{aligned} & \left. \begin{aligned} \nabla^2 \underline{u} + f(\underline{u}) &= 0 && \text{in } V \\ \underline{u} &= 0 && \text{on } S \end{aligned} \right\} \\ \text{subject to} \end{aligned} \tag{9.45a}$$

$$\begin{aligned} & \left. \begin{aligned} \nabla^2 \bar{u} + f(\bar{u}) &= 0 && \text{in } V \\ \bar{u} &= 0 && \text{on } S \end{aligned} \right\} \\ \text{subject to} \end{aligned} \tag{9.45b}$$

Taking the inner-product of the  $\underline{u}$  equation with  $\underline{u}$  and  $\bar{u}$  equation with  $\bar{u}$ , subtracting, and using the boundary conditions, we get

$$\int_V \bar{u} \underline{u} \left[ \frac{f(\underline{u})}{\underline{u}} - \frac{f(\bar{u})}{\bar{u}} \right] dV = 0 \tag{9.46}$$

Equation (9.46) is always satisfied by  $\underline{u}, \bar{u}$ . Now assume  $u_0 > 0$ . Then

$$\underline{u} \geq 0 \quad \text{in } V$$

$$\bar{u} \geq 0 \quad \text{in } V$$

Equation (9.46) can hold only if the term in the square brackets changes sign in  $V$ . If we further impose that  $f(u)/u$  is monotonic in  $u$ , then (9.46) will be satisfied only when  $\underline{u} = \bar{u}$ . The nonlinear system will have a unique solution when  $(f(u)/u)$  is monotonic in  $u$  since all solutions must lie between  $(\underline{u}, \bar{u})$  in  $u_0, u^0$ . The sufficient condition which assures uniqueness to the nonlinear system (9.39) is

$$(f(u)/u)' > 0 \text{ in } V \text{ or } (f(u)/u)' < 0 \text{ in } V \quad (9.47)$$

The student interested in applications of this method is referred to Aris (1975).

**Example 9.4** Apply the uniqueness criterion (9.47) to the zeroth order exothermic reaction in a catalytic pellet. Assuming Dirichlet conditions, the temperature is governed by an equation of the form

$$\left. \begin{array}{l} \nabla^2 T + \delta e^{-\gamma T} = 0 \text{ in } V \\ \text{subject to} \\ T = T_{in} \text{ on } S \end{array} \right\} \quad (9.48)$$

where  $\delta$  represents a dimensionless heat of reaction,  $\gamma$  the dimensionless activation energy, and  $T$  the dimensionless temperature. The transformation  $u = T - T_{in}$  renders the boundary condition homogeneous. This is important to ensure that the bilinear concomitant vanishes (equals zero) in (9.46).

$$\left. \begin{array}{l} \nabla^2 u + \delta e^{-1/u+T_{in}} = 0 \text{ in } V \\ \text{subject to} \\ u = 0 \text{ on } S \end{array} \right\} \quad (9.49)$$

An upper solution  $u^0$  is obtained as

$$\left. \begin{array}{l} \nabla^2 u^0 + \delta = 0 \text{ in } V \\ \text{subject to} \\ u^0 = 0 \text{ on } S \end{array} \right\} \quad (9.50)$$

and a lower solution  $u_0$  as  $u_0 = 0$ , i.e.

$$\left. \begin{array}{l} \nabla^2 u_0 = 0 \text{ in } V \\ \text{subject to} \\ u_0 = 0 \text{ on } S \end{array} \right\} \quad (9.51)$$

Since the reaction is exothermic, it follows that the heat generation in the pellet serves to raise the temperature above  $T_{in}$ , i.e.  $T > T_{in}$  in (9.48) or  $u > 0$ . The heat generation rate per unit volume  $\delta e^{-\gamma T}$  is bounded above by  $\delta$  (since  $e^{-x} < 1$  for  $x > 0$ ). So we expect the actual temperature  $T$  to be lower than the temperature that would be obtained when the heat source term is replaced by  $\delta$  throughout  $V$ . These ideas based on physical reasoning are used in obtaining the upper and lower solutions of (9.49). It can be verified that these solutions satisfy (9.40). Also,  $f(u) = \delta \exp[-1/(u + T_{in})]$  is an increasing function in  $u$  and we can construct the sequences  $\{u_n\}$  and  $\{u^n\}$  which converge to  $\underline{u}$  and  $\bar{u}$ . The equation has a unique solution in  $(u_0, u^0)$  if  $\exp[-1/(u + T_{in})]/u$  is monotonic in  $u$ , i.e. when

$$-\frac{\exp[-1/(u + T_{in})]}{(u + T_{in})^2 u^2} (u^2 + (2T_{in} - 1)u + T_{in}^2) < 0 \quad (9.52)$$

This is assured when the discriminant of the quadratic in brackets is negative, i.e. for  $T_{in} > 1/4$ . This quadratic is positive as long as  $T_{in} > 1/4$ . So we have a unique solution to the equation for all  $\delta$  when  $T_{in} > 1/4$ . So unlike the maximum principles which did not yield any uniqueness

criterion for the exothermic reaction (see Example 9.2), the monotone-iteration method does. Aris (1975) discusses the application of this method to analyse the first order reaction in a non-isothermal catalytic pellet. Pushpavanam and Narayanan (1988b) discuss the application of the monotone-iteration method to the case of a conjugate-fluid solid reaction system. They also have derived uniqueness conditions for this system.

**Algebraic systems.** The solution to the algebraic equation (9.36a) can be obtained by constructing the sequence

$$x_{n+1} = f(x_n) \quad (9.53)$$

starting from a suitable guess  $x_0$ . The function  $f(x)$  is said to be a one-dimensional map as it maps points from the real-line  $R$  to other points on the real-line  $R$ . The fixed points of the map  $f$  are the solutions of (9.36a), see Stakgold (1979).

**Definition 9.2** A map is said to be a **contraction** if

$$d(f(x), f(y)) \leq kd(x, y), \quad 0 \leq k < 1 \quad (9.54)$$

Here  $k$  is a scalar and  $d(x, y)$  represents a suitable metric. Equation (9.54) implies  $f$  maps two points  $x, y$  from its domain to two points  $f(x), f(y)$ , such that the distance between  $f(x)$  and  $f(y)$  is not greater than the distance between  $x$  and  $y$ , i.e.  $f$  contracts distances between points.

**Definition 9.3** A sequence is said to be a **Cauchy** sequence if for every  $m, p$  such that  $m, p > N$  we have  $d(x_m, x_p) < \varepsilon$ , where  $\varepsilon > 0$ .

**Definition 9.4** A metric space, as we have seen in Chapter 2, is a space (not necessarily a linear space) which has a metric defined on it. The notion of the metric automatically generates the concept of convergence. A metric space is said to be **complete** if every Cauchy sequence in that space converges to an element in it. For a Cauchy sequence, if

$$\lim_{n \rightarrow \infty} (x_n, x^*) < \varepsilon \quad \text{for } n > N$$

and  $x^*$  belongs to the metric space we have a complete space. Since the function  $f(x)$  in (9.54) is defined on the real line  $R$ , we are dealing with real numbers (and not just rational numbers) and we usually are working with complete spaces.

**Theorem 9.2** A contraction map  $f:T \rightarrow T$  defined on a complete metric space  $T$  has a unique fixed point.

*Proof.* Let  $f:T \rightarrow T$  be a contraction map which satisfies (9.54). If we can prove that starting from an initial guess  $x_0$ , a Cauchy sequence is generated, the sequence will converge to the fixed point by definition as our space is complete. Using (9.54) and (9.55), we have

$$\begin{aligned} d(x_2, x_1) &= d(f(x_1), f(x_0)) \leq kd(x_1, x_0) \\ d(x_3, x_2) &= d(f(x_2), f(x_1)) \leq kd(x_2, x_1) \leq k^2 d(x_1, x_0) \end{aligned}$$

Similarly, for two consecutive points

$$d(x_m, x_{m-1}) \leq k^{m-1} d(x_1, x_0) \quad (9.55)$$

To establish if the sequence is Cauchy, we have to obtain the metric between two arbitrary (not necessarily consecutive) points  $x_m, x_p$ . Assume for the sake of concreteness  $m > p$ . From the triangle inequality

$$\begin{aligned} d(x_m, x_p) &\leq d(x_m, x_{m-1}) + d(x_{m-1}, x_{m-2}) + \dots + d(x_{p+1}, x_p) \\ &\leq (k^{m-1} + k^{m-2} + \dots + k^p) d(x_1, x_0) \\ &\leq k^p (1 + k^1 + k^2 + \dots + k^{m-p-1}) d(x_1, x_0) \\ &\leq \frac{k^p}{1-k} d(x_1, x_0) \end{aligned} \quad (9.56)$$

We have replaced the finite geometric series with the infinite geometric series in the last step above. For a fixed  $x_0$ ,  $d(x_1, x_0)$  is fixed, the metric  $d(x_m, x_p)$  can be made smaller than  $\epsilon$ , by choosing  $N$  sufficiently large (where  $p, m > N$ ) from (9.56) as  $k < 1$ . This assures us that the sequence is a Cauchy sequence and since we are working in a complete space, it converges to an element  $x^*$  in it. This element is the fixed point of the map. To prove that the fixed point is unique, assume the contrary, i.e. it is not unique. Let  $x^{**}$  be another distinct fixed point of  $f(x)$ . Then we have

$$x^{**} = f(x^{**})$$

Also

$$d(x^*, x^{**}) = d(f(x^*), f(x^{**})) \leq kd(x^*, x^{**}) \text{ or } (1-k)d(x^*, x^{**}) \leq 0$$

For  $k < 1$ , this implies  $d(x^*, x^{**}) \leq 0$ . Since the metric between two distinct points cannot be negative, it can at best be zero, thereby implying that the two points are identical. This proves  $x^* = x^{**}$ , and the fixed point is unique. The application of this technique to a reactor problem can be found in Ramkrishna and Amundson (1985).

**Example 9.5** The temperature in a CSTR sustaining an exothermic zeroth order reaction is given by

$$T = T_{in} + \delta e^{-\gamma T} \quad (9.57)$$

where the parameters  $T, \gamma, \delta, T_{in}$  have the same qualitative significance as explained earlier. This equation is written as

$$T = f(T)$$

The function  $f$  maps points from  $(T_{in}, T_{in} + \delta)$  to points in  $(T_{in}, T_{in} + \delta)$ .  $T_{in}$  can be seen now as the lower solution and  $T_{in} + \delta$  as the upper solution. This again follows from the physical arguments used in Example 9.4. Just as the real line is a complete metric space, so is the segment  $(T_{in}, T_{in} + \delta)$ . Using the  $d_1$  metric, we have

$$\begin{aligned} d(f(x), f(y)) &= |f(x) - f(y)| = |\delta e^{-\gamma/x} - \delta e^{-\gamma/y}| \\ &= \left| \frac{\gamma \delta e^{-\gamma/c}}{c^2} \right| |x - y| \end{aligned}$$

where  $c$  is some point in  $(x, y)$ . This follows from the mean-value theorem of calculus. The map  $f$  is a contraction if

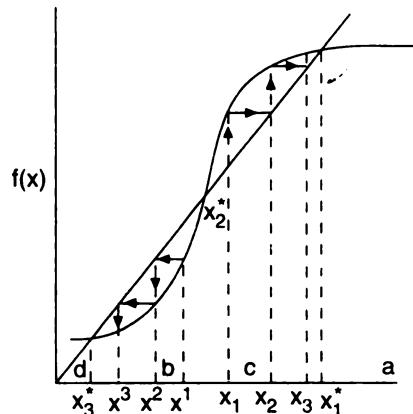
$$\left| \frac{\gamma \delta e^{-\gamma/c}}{c^2} \right| < 1 \text{ for all } c \in (T_{in}, T_{in} + \delta) \quad (9.58)$$

The maximum value of the left-hand side occurs for  $c = \gamma/2$ . Substituting this in (9.58), we have

$$\delta < \frac{\gamma}{4} e^2 \quad (9.59)$$

This condition tells us that we have a unique solution for an exothermic reaction whose rate of heat generation is sufficiently low as to satisfy (9.59). Violation of (9.59) does not assure us of multiple solutions, as the condition (9.59) is only a sufficient condition and not a necessary condition.

For the choice of ' $a$ ' as upper solution, Fig. (9.8), and ' $c$ ' as the lower solution, the maximal and minimal solutions coincide at  $x_1^*$ . The solution is unique in the interval  $(c, a)$ . The map is a contraction here. In the interval  $(d, a)$ , where  $d < x_3^*$ , the maximal and minimal solutions do not coincide. They are  $x_1^*$ ,  $x_3^*$  respectively. Here the map is not a contraction. This is easy to see since the slope of the function  $f(x)$  exceeds unity in the interval  $(d, a)$ , Fig. 9.8.



**Fig. 9.8** Uniqueness of the fixed point of  $f(x)$ , maximal and minimal solutions.

An important point to note here is that the fixed point  $x_2^*$  can never be obtained by the method discussed here. Points starting close to it and to its right converge to  $x_1^*$ , while points to its left converge to  $x_3^*$ . The implications of this will be discussed in detail in Chapters 11 and 12.

Before concluding this chapter and section, it will be worthwhile to summarise all the results found so far:

1. The methods of solving linear algebraic systems and linear partial differential equations are analogous. In the former, a vector is sought as a linear combination of the naturally occurring eigenvector basis set. In the latter, we use the eigenfunctions of the operator, and seek the solution in terms of the projections on these eigenfunctions. This constitutes the method of separation of variables.
2. The eigenvalues and eigenvectors of a self-adjoint matrix operator satisfy similar theorems as a self-adjoint differential operator. We have also seen that all the mathematical results have a physical basis.
3. The operator notion used here enables us to establish Rayleigh's quotient and Fredholm alternative for linear algebraic and linear partial differential equations.
4. The equivalent of the matrix inverse is the Green's function of the differential operator.

The methods of determining the Green's function for a wide variety of systems were discussed in Chapter 8.

5. The uniqueness proofs for nonlinear algebraic systems and partial differential equations are analogous. The two methods, one based on the maximum principle and the other on the monotone iteration methods, are applicable to finite and infinite dimensional systems.

To summarise, it is reasonable to expect that the mathematical techniques for finite dimensional problems can be extended to infinite dimensional systems. So by mathematically complicating a problem, i.e. changing an algebraic system to a partial differential system, we are not introducing any new mathematical techniques. The basis of these techniques is the same as has been shown so far.

Hence the real challenge in the modelling of process to study system behaviour lies not in mathematical complexity of a problem, but in the physical interactions considered in the system model. If the "same" physical processes are considered in a well-stirred system (algebraic system), and in a spatially distributed system (partial differential system), we expect the same qualitative behaviour. The mathematical techniques for investigating this behaviour are identical (at least as far as what we have seen).

The modelling expert is thus interested in incorporating the different interactions in as simple a model as possible without complicating it mathematically. This forms the basis of and motivation for Chapters 11 and 12, where we restrict ourselves *only* to finite dimensional nonlinear dynamical systems. Here we are not interested in the actual methods of solution (as these have to be numerical). Our aim is to determine the different kinds of qualitative behaviour that can arise by nonlinear interactions of system behaviour. It is hoped that the student will be able to use the insight given to him in these two sections to extend the same mathematical theory, i.e. "bifurcation theory," in the next section to infinite dimensional systems comfortably.

## PROBLEMS

1.  $u''(x) = \sin(\pi x) \sin(2\pi x)$

subject to

$$u(0) = 0, \quad u'(0) = u'(1)$$

Does this equation have a unique solution? Use: (a) maximum principles and (b) energy methods.

2. The steady temperature for a zeroth-order reaction in a CSTR is given by

$$x = x_{in} + b e^{-1/x}, \quad b > 0, \quad x_{in} > 0$$

(a) For  $x \in (x_{in}, \infty)$  plot the LHS, and RHS.

(b) Get an estimate using this on the bounds for  $x$ .

(c) Construct a monotone iteration sequence and prove the existence of the solution.

(d) Argue with clear reasons, and find conditions for unique solutions, graphically.

(e) Compare the results obtained with the contraction mapping proof.

3. Consider two slabs  $-a < x < a, -b < x < b$  where  $a > b$ . Both have a constant rate of heat generation ' $q$ ' and have their surfaces maintained at  $0^\circ\text{C}$ . For which slab will the center-line

temperature ( $x = 0$ ) be more? Discuss without solving the problem. Give mathematical and physical reasons.

4. (a) Consider an isothermal first order reaction in a catalyst pellet. Model the pellet as a rectangular slab. Prove that the concentration in the slab cannot be negative when the mass transfer coefficient at slab surface is high.

(b) Consider a rectangular slab sustaining a zeroth-order reaction. Neglect the heat transfer resistance on slab surface. Establish that (i) for an exothermic reaction the temperature in pellet exceeds the ambient value, and (ii) for an endothermic reaction the temperature is lower than the ambient value.

## REFERENCES

- Aris, R., *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Clarendon Press, Oxford (1975).
- Nayfeh, A.H., *Introduction to Perturbation Techniques*, Wiley, New York (1981).
- \_\_\_\_\_, *Perturbation Methods*, Wiley, New York (1973).
- Noble, B. and Daniel, J.W., *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, New Jersey (1977).
- Protter, M.H. and Weinberger, H.F., *Maximum Principles in Differential Equations*, Prentice-Hall, Englewood Cliffs, New Jersey (1967).
- Pushpavanam, S. and Narayanan, R., Comparison theorems for ignition and extinction in conjugate fluid solid systems, *IMA Journal of Applied Mathematics*, **40**, 37 (1988).
- \_\_\_\_\_, Uniqueness conditions for steady-solutions of  $m$ th order reactions—Non-isothermal pellets with variable transport coefficients, *Chemical Engineering Science*, **43**, 394 (1988).
- Ramkrishna, D. and Amundson, N.R., *Linear Operator Methods in Chemical Engineering: With applications to transport and chemical reaction systems*, Prentice-Hall, Englewood Cliffs, New Jersey (1985).
- San, K.Y. and Stephanopoulos, G., A note on the optimality criterion for maximum biomass production in a fed-batch fermentor, *Biotechnology and Bioengineering*, **XXVI**, 1261 (1984).
- Stakgold, I., *Green's Function and Boundary-value Problems*, Wiley, New York (1979).
- Weinberger, H.F., *A First Course in Partial Differential Equations with Complex Variables and Transform Methods*, Wiley, New York (1965).

# 10

## Steady State Characteristics of Nonlinear Dynamical Systems

---

### INTRODUCTION

So far, we have been primarily concerned with analytical methods for solving linear equations. The approach and the treatment we have emphasised until now pertained to the underlying universal nature of the techniques used in constructing analytical solutions to different classes of linear equations. The motivation for using this approach is to help the student understand and appreciate how the methods for solving partial differential equations can be viewed as a simple extension of solving systems of linear algebraic equations and ordinary differential equations. This enables him to develop a philosophy and get a better understanding of the basis of numerical methods which arise while solving complex systems of linear or nonlinear equations. It prevents the student from blindly using a computational technique or software (e.g. finite differences) as a black-box algorithm or tool.

The restriction to linear systems is a severe constraint since in reality most systems are nonlinear. Typically, in reaction systems, nonlinearities arise due to the Arrhenius temperature dependency of reaction rates, in autocatalytic reactions and other rate expressions (see Scott, 1991). In heat transfer they arise when thermal conductivity depends on temperature, or when heat transfer occurs due to radiation. In fluid mechanics the friction factor depends on flow conditions, i.e. Reynolds number (see Drazin and Reid, 1981) and in separation process the nonidealities in phase equilibrium relations are examples of ways in which nonlinearities are generated (see Chapter 1).

The actual methods for predicting system behaviour by solving the modelling nonlinear equations are numerical. The various numerical techniques for solving the equations are also universal in the sense that they are applicable to algebraic equations as well as ordinary differential and partial differential equations. Each system, however, is usually treated by specialised techniques which exploit the problem structure to ensure an efficient solution methodology. The emphasis on the various numerical techniques involved in solving these equations and their advantages and disadvantages will be kept at a minimum in this text. Besides, we are not interested in the methods of solution of nonlinear equations unlike in Chapters 3 and 4 dealing with linear equations. Here we use the modelling equations to predict the system behaviour. We would like to understand the different qualitative features such systems can exhibit and discuss methods of quantifying their behaviour. In particular, we are interested to see how a change in an intrinsic parameter of a system can affect its response or behaviour qualitatively as well as quantitatively.

By system behaviour we mean in particular the open loop response of the system which involves studying the evolution of the dynamic system when subject to changes in parameter

values. In chemical industry most processes are operated in a continuous mode. Such systems are therefore usually at steady state. If a steady state is not feasible or is not attainable, we would like to study the nature of the resulting behaviour of the system. We would also like to see why a particular state may not be feasible. Bifurcation theory helps us answer these questions in detail. This theory is sufficiently general and can be used to analyse different dynamical systems, irrespective of whether the governing equations are ordinary differential equations or partial differential equations or discrete maps. This theory is discussed because it is in keeping with the general approach adopted in the first two sections of the book, where we discussed a universal method of solution for different kinds of linear equations. Moreover, the emphasis on bifurcation theory helps us maintain the treatment at an analytical level and present important concepts in a formal manner. Many results from the first two sections can be directly used while studying nonlinear systems using this theory.

The last two decades have seen considerable amount of development in analysing nonlinear dynamical systems using bifurcation theory. Although this theory can be and has been applied to partial differential equations (Iooss and Joseph, 1980), we will restrict ourselves to treating two-dimensional maps and three-dimensional dynamical systems (system of three-coupled nonlinear ordinary differential equations) as the most complex systems. The motivation for this is that we expect the system behaviour to be dictated by the physical interactions present in the system and not by its mathematical complexity. Consequently, we keep our model at the simplest level from the mathematical point of view. The different kinds of terminal system states (states as time tends to infinity) we will study are—steady states, periodic states, quasi-periodic states and chaotic states. These states arise due to the nonlinearities present in the system. Bifurcation theory enables us to determine the conditions under which these states arise and the mechanism by which they are born. This study has gained importance because the dynamic or time dependent operation of many a system in chemical engineering can lead to a better performance than its steady operation (Lee et al., Hagedorn (1981)). We can extend this theory to isolate regions in parameter space where the system behaviour has a specific characteristic. This results in an optimum performance of the system (Pushpavanam, 1992). It also helps us understand how the initial conditions and the initial transients can possibly result in a pathological behaviour of the system.

The study of nonlinear dynamics has found applications in weather forecasting, ecological modelling, electrical engineering, to name a few disciplines (Vidyasagar (1979), May (1973), and Lorenz (1963)). In chemical engineering it has helped in the analysis of ignition-extinction phenomena in reaction systems and has proved to be a valuable tool to the control engineer to avoid pathological behaviour of a plant, see Poore (1973), and Scott (1991). The area of nonlinear dynamics has a very close bearing on nonlinear control theory (Fleming, 1988).

## 10.1 DYNAMIC SYSTEMS

A dynamical system is one that evolves with time. Nonlinear dynamical systems can be of two kinds: (a) discrete maps, and (b) continuous dynamical systems.

### 10.1.1 Discrete Maps

The evolution of the variables characterising a system at discrete points in time is described here. For simplicity and concreteness, consider the system to be an isothermal CSTR sustaining a single irreversible reaction. The progress of the reaction can be monitored by measuring a single variable, i.e. the concentration of the reactant or the concentration of the product.

For the sake of convenience, let us assume that one samples the system at equal intervals of time. Suppose now that the concentration of the reactant at any instant of time depends only on its value at the previous instant of time. Then, mathematically,

$$C(t + \Delta t, p) = f(C(t), p) \quad (10.1)$$

This represents an abstract discretised evolution equation. Here  $p$  represents the set of parameters which characterise the system, i.e. residence time, temperature of operation, etc. The function  $f$  is an unknown function. Its analytical form is known only when reaction kinetics are known and can be obtained in principle by integrating the governing system of differential equations. Alternatively,  $f$  can be obtained graphically by plotting the time series data measured experimentally. From this graphical depiction it is possible to follow the evolution of a system from a given initial state, provided all the system parameters are maintained at the same values. The steady state or the time invariant state is obtained as the value of  $C^*$  such that

$$C^* = f(C^*, p) \quad (10.2)$$

Graphically,  $C^*$  is obtained as the intersection of the  $45^\circ$  line (the bisectrix) with the map  $f$  for fixed  $p$ .  $C^*$  is called the fixed point of the map  $f$  (see Chapter 9). The function  $f$  maps the set of real numbers  $R$  to  $R$  itself (if we are dealing with mole-fractions  $R$  is the interval  $[0, 1]$ ).

Equation (10.1) is a recursive relation. The map  $f$  can be used to get  $C(t + \Delta t)$  once  $C(t)$  is known. Using  $C(t + \Delta t)$ , the map  $f$  can be applied again to obtain  $C(t + 2\Delta t)$ . The subscript  $n$  is often used to indicate the time instant  $t$  (see Iooss, 1979).

Suppressing the dependence on the parameter since we are interested in the system evolution for a fixed  $p$ , we can recast (10.1) as

$$C_{n+1} = f(C_n) \quad (10.3)$$

Incrementing the subscript  $n$  by unity in (10.3) corresponds to progressing forward in time by  $\Delta t$  in (10.1). The map  $f$  in (10.3) generates a sequence of real numbers  $\{C_i\}$  from an initial value  $C_0$ . The sequence represents successive states of the system. If it converges to a value, then it can be interpreted as the system attaining a steady state. The system behaviour is therefore described by the properties of the sequence  $\{C_i\}$ . The approach to a steady state is qualitatively similar to the method of successive substitutions seen in Chapter 9.

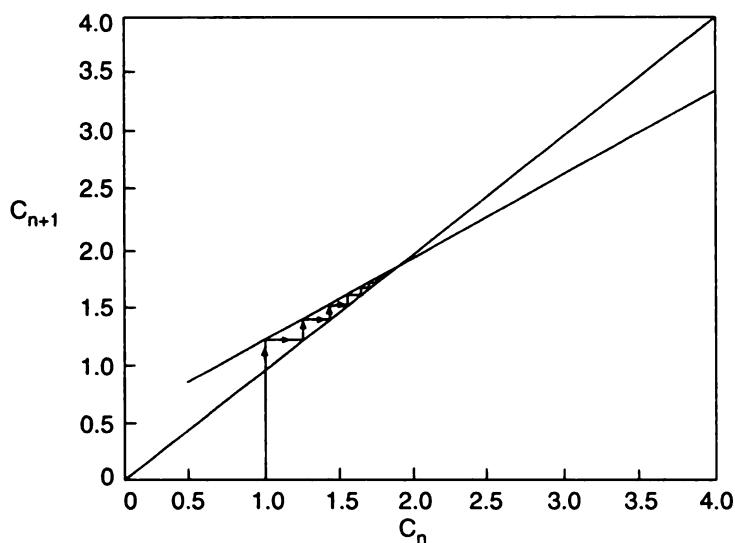
**Example 10.1** Consider the first order isothermal reaction in a CSTR. Table 10.1 contains the experimental data of two runs. The initial concentrations of these runs are .5 and 4 g mol/cc, respectively. From the data of the two runs, obtain the data which would be generated if the initial concentration were to be 1.0 gm mol/cc.

The data given in Table 10.1 is sampled at a time interval of .1 hrs. Plotting  $C(t + .1)$  vs.  $C(t)$  for each run generates the map shown in Fig. 10.1. The points are joined to give a smooth curve. The intersection of the map with the bisectrix is the steady state of the CSTR. The trajectory sampled at every 6 minutes starting from  $C_0 = 1.0$  can be constructed geometrically as shown and is reproduced in Table 10.1 in the third column. The implicit assumption is that the system is one-dimensional, i.e. no other variable determines the evolution of concentration.

A system is usually described by many variables. A distillation column, for example, is characterised by the composition of the various streams and their flow rates. An experimentalist would ideally measure these variables at discrete points in time to study his system. For concreteness let us assume that a distillation column is characterised by  $k$  variables. The system state is determined uniquely by specifying all these  $k$ -variables at an instant of time. These include the

**Table 10.1** Evolution of Concentration with Time for the First Order Isothermal Reaction in CSTR (Example 10.1)

$t$ (hrs)	$C$ (g mol/l)	$C$ (g mol/l)	$C$ (g mol/l)
0	4.0000	0.5000	1.0000
0.1000	3.6375	0.7719	1.1813
0.2000	3.3406	0.9945	1.3297
0.3000	3.0976	1.1768	1.4512
0.4000	2.8987	1.3260	1.5507
0.5000	2.7358	1.4482	1.6321
0.6000	2.6024	1.5482	1.6988
0.7000	2.4932	1.6301	1.7534
0.8000	2.4038	1.6972	1.7981
0.9000	2.3306	1.7521	1.8347
1.0000	2.2707	1.7970	1.8647
1.1000	2.2216	1.8338	1.8892
1.2000	2.1814	1.8639	1.9093
1.3000	2.1485	1.8886	1.9257
1.4000	2.1216	1.9088	1.9392
1.5000	2.0996	1.9253	1.9502
1.6000	2.0815	1.9389	1.9592
1.7000	2.0667	1.9499	1.9666
1.8000	2.0546	1.9590	1.9727
1.9000	2.0447	1.9664	1.9776
2.0000	2.0366	1.9725	1.9817
2.1000	2.0300	1.9775	1.9850
2.2000	2.0246	1.9816	1.9877
2.3000	2.0201	1.9849	1.9899



**Fig. 10.1** Map generated by the data of Example 10.1. Vertical and horizontal lines depict evolution from an initial point.

temperatures of each stream and its composition in each stage. The system state at each instant of time is assumed to be dependent only on the state at the previous instant as before. This can be represented as

$$\left. \begin{array}{l} x^1(t + \Delta t) = f_1(x^1(t), x^2(t), \dots, x^k(t)) \\ x^2(t + \Delta t) = f_2(x^1(t), x^2(t), \dots, x^k(t)) \\ \vdots \\ x^k(t + \Delta t) = f_k(x^1(t), x^2(t), \dots, x^k(t)) \end{array} \right\} \quad (10.4)$$

The  $k$ -variables  $x^1, x^2, \dots, x^k$  determine the system state uniquely. The set of equations (10.4) is a  $k$ -dimensional map and is written compactly in vectorial form as

$$x(t + \Delta t) = F(x(t)) \quad (10.5a)$$

where  $x$  now represents the  $k$ -tuple  $x^1, x^2, \dots, x^k$ . The map  $F$  transforms elements in  $\mathbb{R}^k$  to other elements in  $\mathbb{R}^k$ . The matrix  $A$  which we saw in Chapter 3 is a linear operator or a linear map operating in  $\mathbb{R}^k$ .  $F$  in (10.5) is a nonlinear map and it may not satisfy the axioms of a linear operator (see Chapter 3). Rewriting the time variable  $t$  as a subscript  $n$ , we obtain the  $k$ -dimensional map

$$x_{n+1} = F(x_n) \quad (10.5b)$$

We use  $F$  to denote a  $k$ -dimensional map ( $k > 1$ ) and  $f$  to denote a one-dimensional map. It must be emphasised that the subscript  $n$  does not represent the  $n$ th coordinate of the vector but the vector  $x$  at the discretised instant of time denoted by  $n$ . Similarly,  $x^i$  does not denote the  $i$ th vector, but the  $i$ th coordinate of the vector  $x$ . We use this nomenclature which is different from that used in Chapter 2 to be consistent with the notation in the literature.

The evolution of a  $k$ -dimensional system is described by the successive elements of the sequence, where each element now is a  $k$ -dimensional vector. If the sequence converges to  $x^*$ , we say that the system evolves to a steady state. By analogy to the one-dimensional map, the fixed point of the  $k$ -dimensional map satisfies

$$x^* = F(x^*, p). \quad (10.6)$$

A number of industries these days use digital computers in process industries. The emerging trend of using computers in processing information has accentuated the importance of maps which use discretised information of a system. Examples of systems modelled by maps can be found in looss (1979).

### 10.1.2 Dynamical Systems

In this class of nonlinear equations the variables evolve continuously with time. Most chemical engineering systems belong to this class. Here ordinary differential equations or other evolution equations model the system. The independent variable here is only time  $t$  when the systems are taken to be spatially homogeneous, i.e. well stirred like the CSTR. The elimination of the dependence on the spatial coordinates helps us focus our attention on the temporal behaviour of the system investigated. The system in which the variables depend on spatial coordinates also belongs to this class as long as they change smoothly or continuously with time. The equations describing the evolution now are partial differential equations (parabolic). Examples of systems modelled by ordinary differential equations can be found in May (1973) and Scott (1991).

The space independent, dynamical system is modelled by evolution equations of the form

$$\left. \begin{array}{l} \dot{x}_1 = f_1(x_1, x_2, \dots, x_n, p) \\ \dot{x}_2 = f_2(x_1, x_2, \dots, x_n, p) \\ \vdots \\ \dot{x}_n = f_n(x_1, x_2, \dots, x_n, p) \end{array} \right\} \quad (10.7)$$

The dot over  $x$  represents the derivative with respect to time. The above system is said to be an autonomous system as the variable  $t$  does not occur explicitly in any of the  $f_i$ 's. The system can be written more compactly in vectorial form as

$$\dot{x} = F(x, p) \quad (10.8)$$

where  $x$  is an  $n$ -dimensional vector. As a specific example, consider a nonisothermal CSTR sustaining an irreversible first order exothermic reaction



The evolution of the dimensionless conversion  $C$  and temperature  $T$  of the reactor are given by

$$\dot{C} = -C + Da(1 - C) e^T \quad (10.9a)$$

$$\dot{T} = -(1 + \beta)T + BDa(1 - C) e^T \quad (10.9b)$$

These equations have been derived in Poore (1973).

We identify  $x_1$  with  $C$  and  $x_2$  with  $T$  (alternatively, we could use  $x_1$  for  $T$  and  $x_2$  for  $C$ ). This is a two-dimensional autonomous dynamical system. The dimensionless parameters  $B$ ,  $Da$  and  $\beta$  constitute the parameter vector  $p$ .  $B$  represents the dimensionless heat of reaction,  $Da$  is the Damkohler number and  $\beta$  the dimensionless heat transfer coefficient. Each  $x_i$  varies continuously with time. The system traces out a trajectory. To study the dynamic behaviour of a continuous system, we investigate the trajectory traced by the system as it emanates from the initial point. This is analogous to the situation in a map where the dynamic behaviour is obtained by starting from an initial point and studying the sequence generated. The steady state or the time invariant state of a system is obtained by setting  $\dot{x} = 0$ . This yields the algebraic equation

$$F(x, p) = 0 \quad (10.10)$$

This is similar to the fixed point equation of the map (10.6), which is also an algebraic equation.

In the study of the dynamic behaviour of systems, we are primarily interested in their long time behaviour. The asymptotic behaviour of the systems as  $t \rightarrow \infty$  may not always be time invariant. This behaviour depends on the parameter set  $p$ . The system may be at a time-dependent state for some parameters. It is the origin and characteristic of this dynamic behaviour that bifurcation theory helps us understand. In particular, the initial transient behaviour is not the dynamic state which is of interest to us. The initial transients are relevant in start-up studies. The different terminal (i.e. as  $t \rightarrow \infty$ ) dynamic states can be described qualitatively and quantitatively. The emphasis in this text will be on one-dimensional and two-dimensional maps and 1-, 2-, 3-dimensional autonomous dynamical systems. We will also see how the theories of the study of maps and dynamical systems are interconnected. The "dimension" of a system here refers to the number of variables describing it. For example, the map (10.4) is  $k$ -dimensional and the system (10.7) is

$n$ -dimensional. May (1973) provides a variety of excellent examples in ecology modelled by simple equations which show diverse dynamic features.

The change in system behaviour occurs normally when we vary operating conditions. This manifests itself as a change of parameters in the governing equations. A systematic analysis of a system can be done by varying one of the parameters in the system, keeping all other parameters constant. Bifurcation theory essentially studies how the system behaviour changes as we vary a parameter. In this chapter we restrict ourselves to investigating only the steady-state behaviour. In Chapter 11 we study the occurrence of a limit cycle or a periodic state. In Chapter 12 we discuss the origin of more complex dynamic behaviour such as quasi-periodic solutions and chaotic solutions (Lorenz, 1963). Examples of systems exhibiting instabilities and modelled by partial differential equations can be found in Drazin and Reid (1981), Chandrasekhar (1950), and Kuramoto (1984).

The remaining part of this chapter is devoted to: (a) the numerical evaluation of the steady state for a fixed set of parameters, and (b) to homotopy continuation methods which allow us to study the effect of varying a parameter on the steady state.

The numerical method we describe here is the Newton-Raphson method (see Gupta (1995) and Hildebrandt (1956)). It is sufficiently general and can be extended to determine periodic solutions as well (as we will see in Chapter 11 (refer Roberts and Shipman, 1972)). Continuation methods offer an efficient way to track the variation of a solution as a parameter changes. Besides it is a general method and can be used to determine variations in the dynamic states of the system. It can also be used for distributed parameter systems governed by partial differential equations. Further, this method can be employed to obtain good initial guesses for the Newton-Raphson method (see Kubicek and Marek, 1983).

The efficient implementation of the methods requires access to a computer. In this section our emphasis will be primarily on the philosophy and the basis of the numerical methods. This facilitates their generalisation and extension to studying complex dynamic behaviour and enables one to get a comprehensive picture of system behaviour and to analyse it. It also illustrates how a student can intuitively analyse system behaviour on the basis of mathematical results obtained by him.

## 10.2 STEADY STATES (NUMERICAL EVALUATION)

The steady state of a dynamical system or the fixed point of a map is given by the roots of (10.6) or (10.10). A popular numerical method used in solving the system of nonlinear algebraic equations is the Newton-Raphson method. The method is discussed first in the context of a single algebraic equation and is later generalised to a system of  $n$ -equations.

The steady state of a one-dimensional system is the root  $x$  of an equation of the form

$$f(x, p) = 0 \quad (10.11)$$

where  $p$  is a fixed set of parameters. The fixed point equation (10.6) can also be rearranged to be in the above form and does not have to be treated separately. Equation (10.11) being nonlinear has to be solved iteratively. Since we are interested in obtaining  $x$  for a fixed  $p$ , we suppress the dependency on  $p$  in what follows. Let  $x_1$  be a guess for the root of the above equation. Expanding  $f(x)$  in a Taylor series about  $x_1$  and retaining only the linear term, we obtain

$$f(x) = f(x_1) + f'(x_1)(x - x_1) + \dots \quad (10.12)$$

Starting with the guess  $x_1$  in the above equation, we like to obtain  $x$  from (10.12) such that  $f(x) = 0$ . We use (10.12) to generate the next iterate  $x_2$  such that  $f(x) = 0$ . Substituting  $x_2 = x$ , this yields.

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \quad (10.13)$$

Clearly, this  $x_2$  may not satisfy (10.11) because of the approximation made in (10.12) in retaining only the linear term. Should  $x_1$  be close to the actual root, then  $x_2$  will be closer to the root than  $x_1$ . We can use this  $x_2$  to similarly generate the next iterate  $x_3$ . The successive iterates are obtained by using the recursive relation

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \quad (10.14)$$

If the generated sequence  $\{x_n\}$  is convergent, the sequence converges to the root of (10.11), see Problem 10.3. The notation used here is the same as that for maps. Note that the subscript  $n$  denotes the  $n$ th iterate, and not the  $n$ th coordinate of the vector.

In an  $n$ -dimensional system, we have to solve a system of  $n$ -nonlinear algebraic equations. This is written compactly as

$$F(x, p) = 0 \quad (10.15)$$

We start with a guess vector  $x_1$ .

Proceeding in a similar to the one-dimensional case, we expand each of the  $f_i$ 's in a Taylor series around  $x_1$  and retain only the first order or linear terms. This yields, for the first equation in the system (10.15), the relation

$$f_i(x_1^1, x_1^2, \dots, x_1^n) = f_i(x_1^1, x_1^2, \dots, x_1^n) + \sum_{i=1}^n \frac{\partial f_i}{\partial x^i}(x_1^1, x_1^2, \dots, x_1^n)(x^i - x_1^i) \quad (10.16)$$

Here,  $x_i^n$  represents the  $n$ th coordinate of the vector at the  $i$ th iterate. We have  $n$  such equations for each of the  $f_i$ 's. Starting from the guess vector  $x_1$ , we generate the vector  $x_2$  at the next iterate such that it is a better approximation to the root, i.e. we impose  $F(x_2, p) = 0$ .

In the iterative sequence here, the successive iterates are obtained by the recursive relation

$$x_{m+1} = x_m - J^{-1}(x_m)F(x_m) \quad (10.17)$$

The matrix  $J$  is the Jacobian matrix whose elements are given by

$$J_{mn} = \frac{\partial f_m}{\partial x^n} \quad (10.18)$$

The convergence of the sequence generated by the Newton-Raphson method depends on the initial guess value. The choice of a good initial guess  $x_0$  is very important for the success of this scheme. This choice can be difficult for systems of large dimensions. When an initial guess is sufficiently close to the root, the sequence generated by Newton-Raphson is likely to converge to it. Under other conditions the iterates may diverge away from the root or oscillate about it without convergence. For a detailed presentation of these features of the Newton-Raphson method, we refer the interested student to Gupta (1995). Before concluding this section, we would like to emphasise that this method is extremely versatile. It can be easily extended to obtain solutions of two-point boundary value problems, calculation of periodic orbits of nonlinear dynamical systems, etc. We will discuss the latter in detail in Chapter 11.

**Example 10.2** Determine the steady states of the CSTR described by (10.9), for  $B = 7$ ,  $\beta = 1$ ,  $Da = .1$ .

The steady states of the system are determined as the roots of the coupled algebraic system

$$f_1(C, T) = -C + Da(1 - C) e^T = 0 \quad (10.19a)$$

$$f_2(C, T) = -(1 + \beta)T + BDa(1 - C) e^T = 0 \quad (10.19b)$$

Multiplying the first equation by  $(-B)$  and adding it to the second equation, we have

$$T = \frac{BC}{1 + \beta}$$

This linear relationship can be used to determine the conversion  $C$  from the single scalar equation

$$0 = -C + Da(1 - C) e^{BC/(1+\beta)} \quad (10.19c)$$

The Newton-Raphson method described is now implemented on: (a) the scalar equation (10.19c) and (b) the vectorial system (10.19a) and (10.19b).

The scalar equation (10.19c) is written as  $f(C)$ . The dependence on various parameters is suppressed as we are interested in a solution for a fixed set of parameters. Clearly,

$$f'(C) = -1 - Da e^{BC/(1+\beta)} + \frac{DaB(1 - C)}{1 + \beta} e^{BC/(1+\beta)}$$

The conversion  $C$  must lie in  $[0, 1]$ . We use  $C_0 = .9$  as an initial guess. The recursive relation

$$C_{n+1} = C_n - \frac{f(C_n)}{f'(C_n)}$$

is used to generate the sequence  $\{C_i\}$ . The sequence generated is (.9, .6351, .2344, .1366, .1406). This is reproduced here to enable the interested reader to verify his results. The sequence converges to the value .1406. This is the steady state conversion of the system. The temperature at this state is obtained as .492.

The  $d_1$  metric between two successive elements is  $d_1(C_n, C_{n+1}) = |C_n - C_{n+1}|$ . The convergence of the sequence should be tested using the criterion for the relative error

$$\frac{d_1(C_n, C_{n+1})}{\|C_n\|} < \varepsilon$$

and not the absolute error  $d_1(C_n, C_{n+1}) < \varepsilon$ .

The solution determined using  $\varepsilon = 10^{-6}$  is accurate up to  $10^{-4}\%$ . It must be emphasised that the use of the absolute error criterion, i.e.

$$d_1(C_n, C_{n+1}) < \varepsilon,$$

can result in incorrect answers or in no solutions in some problems. If the elements of the sequence are themselves small, i.e. infinitesimal, the absolute convergence criterion will be easily satisfied. On the other hand, when the elements of the sequence are large, the absolute convergence criterion will never be satisfied and is a very stringent condition. These problems do not arise while using the relative error as a criterion, which fixes the percentage error of the system.

We now discuss the implementation of the Newton-Raphson method on the coupled system. The solution is sought as a two-dimensional vector  $[C, T]'$ . Since  $C \in [0, 1]$ , it is clear that  $T \in [0, B/(1 + \beta)]$ . The vector  $[.9, .9]'$  is chosen as a candidate for an initial guess. The sequence of vectors is now generated by the recursive relation

$$\begin{bmatrix} C \\ T \end{bmatrix}_{n+1} = \begin{bmatrix} C \\ T \end{bmatrix}_n - J_n^{-1} \begin{bmatrix} f_1(C, T) \\ f_2(C, T) \end{bmatrix}_n$$

Here the matrix  $J$  is

$$\begin{bmatrix} -1 - Dae^T & Da(1 - C)e^T \\ -Bdae^T & -(1 + \beta) + BDa(1 - C)e^T \end{bmatrix}$$

and the subscript  $n$  denotes the  $n$ th iterate. The sequence generated is [.193, .675]', [.1393, .4877]', [.1406, .492]'. Since (10.19a)–(10.19c) are nonlinear, they can possess multiple roots. In Chapter 11, we will prove that for the chosen parameters the system has a unique solution. This is the solution we have found.

### 10.3 CONTINUATION METHODS

Nonlinear equations can possess multiple roots. For example, a cubic can have up to three real roots. When the nonlinear equations are transcendental in nature, they can have an infinite number of roots. For a fixed set of parameter values, the existence of multiple roots to an algebraic system of equations modelling the steady states of a system implies that the system can be in one of several possible steady states. The total number of roots or steady states depends on the parameter values. For some parameters the system can have only one root while for others it may have multiple roots. Quite often, it is important to determine all the roots for each set of parameters. A method of doing this is by studying the variation of a solution as we change a parameter. These parameters can be those which occur intrinsically in the system. Homotopy continuation methods are useful in determining the dependence of a state on a parameter. The method is very efficient since it provides a good initial guess for the solution for each parameter in this method. The models of systems can have many parameters. The system is studied methodically by fixing all but one of these parameters and varying the last parameter. The effect of this last parameter can be then determined for a fixed set of other parameters. If we had varied more than one parameter simultaneously, we would not have been able to isolate the cause which induced the change in system behaviour. The technique of determining steady states as we change a parameter is called *homotopy continuation method*. This technique has been widely used in simulating the behaviour of distillation columns and reaction systems (Lin et al., 1987).

#### 10.3.1 Continuation along an Intrinsic Parameter $p$

Homotopy methods are discussed now for the one-dimensional equation, where the solution  $x$  depends on one parameter  $p$ . The generalisation to a system of  $n$ -equations is straightforward and only the algebra is involved. We leave this as an exercise to the reader. Let the equation whose steady state we are interested in be represented by

$$f(x, p) = 0 \quad (10.20a)$$

In case more than one parameter occurs in (10.20a), we fix all of them except  $p$ . We now determine how the solution  $x$  or state changes with the scalar parameter  $p$ .

We assume that we know  $x_0$  to be a solution to (10.19) for  $p = p_0$ . We also assume that small changes in  $p$  result only in small changes in  $x$ . This implies  $dx/dp \neq \infty$ . In a small neighbourhood

of  $p_0$ , i.e. for  $p_0 + dp$ , the solution  $x$  to the above equation is not going to be very different from  $x_0$ . The solution would have changed by a small amount, say  $dx$ . Hence  $x_0$  would be a good initial guess for  $x$ , when  $p = p_0 + dp$ . This is a zeroth order guess for  $p = p_0 + dp$ . The homotopy continuation method uses an “improved” (actually a first order) estimate for the initial guess to obtain the solution for  $p_0 + dp$ . The derivative  $dx/dp$  is used to obtain a better initial guess for the solution at  $p_0 + dp$ . Differentiating (10.20a) with respect to  $p$  and remembering that  $x$  depends upon  $p$ , we have

$$\frac{\partial f}{\partial x} \cdot \frac{dx}{dp} + \frac{\partial f}{\partial p} = 0 \quad (10.20b)$$

This implies that

$$\frac{dx}{dp} = -f_p/f_x \quad (10.20c)$$

where  $f_p = \partial f / \partial p$ . The derivative  $dx/dp$  can be determined at  $(x_0, p_0)$  once the analytical form of  $f$  is known. This tells us how the solution  $x$  varies with  $p$  at this point. A better approximation for the root  $x$  at  $p_0 + dp$  is obtained by using this information which yields

$$x \Big|_{p_0 + dp} = x_0 + \frac{dx}{dp} \Big|_{x=x_0, p=p_0} dp \quad (10.21)$$

Using (10.21) to yield an initial guess results in a fewer number of iterations in the Newton-Raphson to converge on the root for  $p_0 + dp$  than if we were to use  $x_0$  for the guess. This can be advantageous for higher dimensional problems. Since we are using the information from the derivative, this can be viewed as a first order method.

To summarise, the homotopy continuation method is used to obtain the variation of the solution with respect to a parameter. The algorithm consists of the following steps:

1. Determine the solution  $x_0$  for an initial  $p = p_0$ . Obtain  $f_x, f_p, dx/dp$  at  $x_0, p_0$ .
2. Use (10.21) to obtain an initial guess for  $x$  at  $p_0 + dp$ .
3. Use Newton-Raphson on (10.19) for a fixed  $p = p_0 + dp$ , with this initial guess to converge on the root  $x$  within the desired tolerance.
4. The value of  $x$  at this  $p$  is now initialised to  $x_0, p_0$ , and we repeat steps 1–3. This generates the solution branch depicting the dependence of  $x$  on  $p$ . Alternatively, we can integrate (10.20c) as an initial-value problem and determine this branch. More details can be found in Lin, et al. (1987).

The homotopy method described is based on the following two assumptions:

1. It needs a prior knowledge of a solution  $x_0$  at a parameter value  $p_0$ .
2. It assumes that small changes in the parameter  $p$  yield small changes in  $x$ , i.e.  $dx/dp \neq \infty$ .

For a nonlinear problem, it may not be easy to always obtain an  $x_0$  for a  $p_0$ . Besides, small changes in  $p$  can yield drastic differences in  $x$ . In these problems  $x$  varies discontinuously across critical values in  $p$ . The continuation method discussed so far cannot be used directly when these assumptions are violated. We now see how these limitations can be overcome by modifying the above algorithm.

### 10.3.2 Obtaining a Good First Initial Guess ( $x_0$ for $p_0$ )

The parameter  $p$  in (10.20a) is an intrinsic parameter in the problem, such as the reflux ratio in a distillation column, or the coolant temperature in a heat exchanger, or the residence time in a reactor. This parameter  $p$  (whose effect on the system is of interest) is chosen such that the experimentalist or the engineer has a control over it. This will enable him to vary it experimentally to meet and study the system behaviour as a function of this parameter. The solution branch, i.e. the dependence of  $x$  on  $p$ , can be generated once we know the solution  $x_0$  at a particular  $p_0$ . This allows us to obtain good initial estimates for the other  $p$  values. It may not be always possible to obtain  $x_0$  for  $p_0$  easily. We now discuss how we can get this solution by extending the concept behind continuation along an intrinsic parameter, by introducing an artificial parameter. To obtain the solution  $x_0$  for a fixed  $p_0$ , we seek the root of

$$f(x, p_0) = 0 \quad (10.22a)$$

Now, since the parameter  $p$  is fixed at  $p_0$ , we suppress the explicit dependence on it and rewrite (10.22a) as

$$f(x) = 0 \quad (10.22b)$$

Let  $t$  be an artificial parameter such that  $0 < t < 1$ . We generate an auxiliary function  $h(x, t)$

$$h(x, t) = tf(x) + (1 - t)g(x) \quad (10.23)$$

The auxiliary equation is constructed such that  $g(x)$ , is an arbitrarily chosen function whose root is known. Now

$$h(x, 0) = g(x), h(x, 1) = f(x)$$

From the construction of (10.23), i.e. the choice of  $g(x)$ , we know the root of  $h(x, 0)$ . To obtain the root of  $f(x) = 0$ , we consider continuing along the parameter  $t$  in steps of  $\Delta t$  and obtain the root of  $h$  as  $t$  varies. The artificial parameter  $t$  in  $h$  now plays the role of the intrinsic parameter  $p$  we mentioned earlier. The root of  $h(x, t)$  for  $t = 1$  is identical with the root of  $f(x)$ . This root is the root  $x$  of  $f(x, p_0) = 0$ . The function  $g(x)$  chosen in (10.23) can even be a linear function. Differentiating (10.23) with respect to  $t$ , we have

$$f(x) + t \frac{df}{dx} \cdot \frac{dx}{dt} + (1 - t) \frac{dg}{dx} \frac{dx}{dt} - g(x) = 0$$

or

$$\frac{dx}{dt} = \frac{g(x) - f(x)}{t \frac{df}{dx} + (1 - t) \frac{dg}{dx}} \quad (10.24)$$

The right-hand side of (10.24) is a known function of  $x, t$  as  $g(x), f(x)$  are known. This equation can be integrated by using a numerical procedure like Runge-Kutta, subject to the initial condition

$$t = 0, x = x_0 \text{ (the root of } g(x))$$

Integrating up to  $t = 1$ , we get the root of  $f(x)$ .

Alternatively, we can finite-difference (10.24) to obtain an initial guess for the root of  $h(x, t + \Delta t)$  as

$$x(t + \Delta t) = x(t) + \left. \frac{dx}{dt} \right|_{x=x(t)} \Delta t \quad (10.25)$$

Starting with this initial guess for the solution at  $t + \Delta t$  we use Newton-Raphson to iterate on the solution of (10.23). We keep incrementing  $t$  by  $\Delta t$ , and repeat this procedure till we reach  $t = 1$ . This procedure is analogous to the original continuation method (i.e. with respect to an intrinsic parameter) we discussed for  $f(x, p)$ , with  $f(x, p)$  now being replaced by  $h(x, t)$ .

We conclude that the construction of the solution  $x_0$  for a fixed  $p_0$  can be easily accomplished by constructing the auxiliary function  $h(x, t)$ .

**Example 10.3** Obtain the solution to the equation

$$3x = e^x \quad (10.26a)$$

The roots of this system can be obtained graphically as shown in Fig. 10.2. To compute them algebraically we need a good initial guess for the Newton-Raphson algorithm. ♦

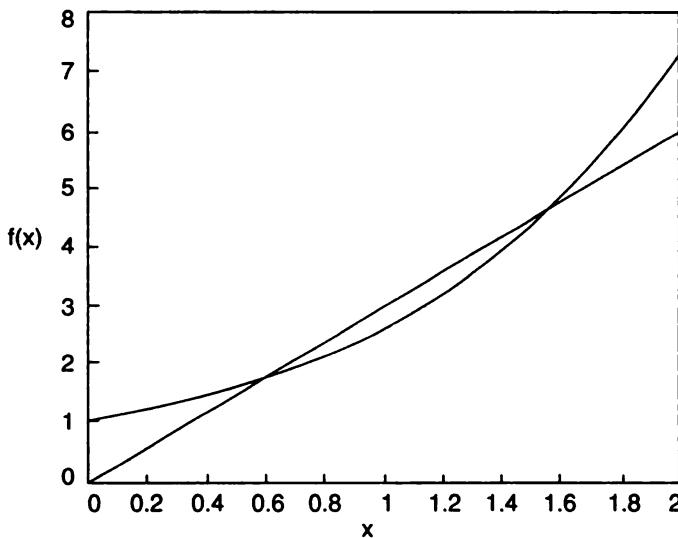


Fig. 10.2 Graphical representation of (10.26a).

Define  $f(x) = 3x - e^x$ ,  $g(x) = x - 0.1$ . The auxiliary function  $h(x, t)$  is given by

$$h(x, t) = t(3x - e^x) + (1 - t)(x - 0.1) \quad (10.26b)$$

So  $x = 0.1$  is a root of  $h(x, 0) = 0$ . The root of  $f(x)$  that we seek is the root of  $h(x, 1) = 0$ . Continuing along the parameter  $t$ , we have

$$\frac{dx}{dt} = \frac{(x - 0.1) - (3x - e^x)}{t(3 - e^x) + (1 - t)} \quad (10.26c)$$

Using  $\Delta t = 0.1$ , we increment  $t$  from  $t = 0$ . The initial guesses for each value of  $t$  are generated from (10.26c) as

$$x^{\text{guess}}(t + \Delta t) = x(t) + \left. \frac{dx}{dt} \right|_{x=x(t)} (\Delta t)$$

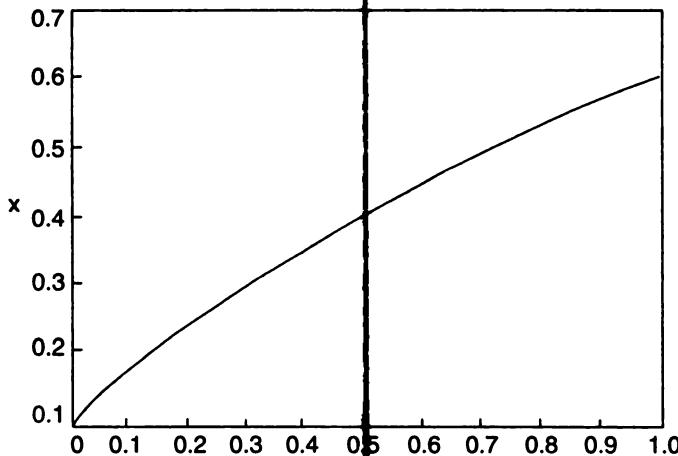
Using this guess estimate, we iterate on (10.26a) for a fixed  $t$  using the Newton-Raphson method.

For  $t = 0.1$ ,  $x^{\text{guess}} = .1739$  and  $x(0.1) = .1742$ . Using this converged value for  $x(0.1)$  to evaluate the derivative at  $t = 0.1$  in, we obtain  $x^{\text{guess}}(0.2) = .238$ . The converged  $x(0.2) = .2385$ . This process is repeated till we get to  $t = 1$ . The guess values and the converged values are given in Table 10.2. The solution obtained to this system is not unique as shown in Fig. 10.2. This can be seen graphically since the two curves intersect at two points.

**Table 10.2** Guess and Converged Values for  $f_x$  of Example 10.3

$t$	$x^{\text{guess}}$	$x^{\text{conv}}$
.1	0.1739	0.1742
.2	0.2380	0.2385
.3	0.2953	0.2958
.4	0.3475	0.3481
.5	0.3961	0.3967
.6	0.4421	0.4429
.7	0.4865	0.4873
.8	0.5300	0.5309
.9	0.5734	0.5745
1	0.6176	0.6191

The second solution can be obtained by choosing a different trial function  $g(x)$ . The method presented here is an elegant method to construct solutions to equations, when a good initial guess is not available. The dependence of  $x$  on  $t$ , obtained from (10.26c), is depicted in Fig. 10.3. The interested reader is referred to (Seader, 1990) for details.



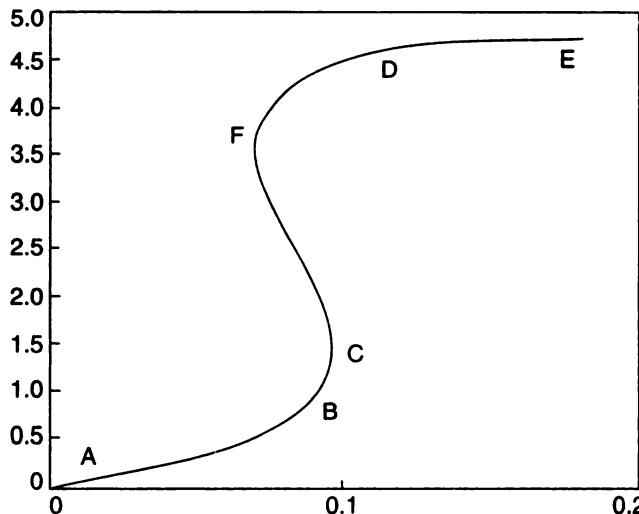
**Fig. 10.3** Dependence of solution  $x$  on  $t$  of  $h(x, t) = 0$  (see 10.26b).

### 10.3.3 Continuation across Parametrically Sensitive Regions $\left(\text{the case of } \frac{dx}{dp} = \infty\right)$

Nonlinear problems frequently display a sharp sensitivity to parameters. Here small changes in parameter values yield drastic changes in the solution as well as the system behaviour. For example,

in a reaction system, as we vary a parameter across an ignition (extinction) point, the system shows a sharp increase (decrease) in temperature. The variation of the temperature  $x$  is discontinuous at these points and  $dx/dp$  is not defined here. The continuation method described earlier cannot be used across these points because of this limitation.

A schematic picture of the dependence of  $x$  on  $p$  showing this behaviour is provided in Fig. 10.4. Starting from the point  $A$ , we can use the continuation method discussed in Section 10.3.1, to trace the solution branch AB. The method fails as we approach the point  $C$ , the ignition point. To the right of  $C$  the solution is given by the branch DE. Here a point on the lower branch



**Fig. 10.4** Schematic diagram showing sensitivity to parameter across critical points.

is a poor initial guess for the upper branch. This will be a big limitation for a higher dimensional problem. Besides, the existence of the middle branch would be undetected if we continue along  $p$  because then we would only trace the upper branch DE to the right of  $C$  and the lower branch AC to its left. All possible solutions (with respect to the parameter  $p$ ) cannot be detected using the continuation method involving the intrinsic parameter.

This problem can be overcome by using the arc-length continuation method which enables us to trace the entire solution branch systematically. In this method, we continue along the arc generated by the solution branch. Let  $x_0$  be a solution of the equation for  $p = p_0$ . This point is indicated as  $A$  in Fig. 10.4 and can be obtained as discussed earlier. The parameter  $s$  represents the arc-length along the solution curve measured from  $A$ . As we move along the curve,  $x$ ,  $p$  and  $s$  vary. We treat  $x$  and  $p$  as dependent variables and  $s$  as the independent variable. Now  $s$  is the parameter along which we continue. Equation (10.20a) can be written as

$$f(x(s), p(s)) = 0 \quad (10.27a)$$

Differentiating this equation with respect to  $s$ , we get

$$\frac{\partial f}{\partial x} \frac{dx}{ds} + \frac{\partial f}{\partial p} \frac{dp}{ds} = 0$$

or

$$f_x \frac{dx}{ds} + f_p \frac{dp}{ds} = 0 \quad (10.27b)$$

This is one equation that relates  $dx/ds$  to  $dp/ds$ . To trace the solution branch, we have to know independently how  $x, p$  vary with  $s$ . For this we need another relationship between  $dx/ds$  and  $dp/ds$ . Here calculus comes to our rescue. It provides the normalisation equation

$$\left(\frac{dx}{ds}\right)^2 + \left(\frac{dp}{ds}\right)^2 = 1 \quad (10.28)$$

Using (10.27b), (10.28), we obtain

$$\frac{dp}{ds} = \pm \frac{f_x}{\sqrt{(f_x^2 + f_p^2)}} \quad (10.29a)$$

$$\frac{dx}{ds} = \mp \frac{f_p}{\sqrt{(f_x^2 + f_p^2)}} \quad (10.29b)$$

We treat the system (10.29) as an initial value problem in  $s$ , and solve for  $x(s), p(s)$  using the initial condition at  $s = 0, p = p_0$  and  $x = x_0$ , at  $s = 0$ .

Alternatively, we can use a finite difference or a Eulerian approximation to obtain an initial guess for  $x$  and  $p$  at  $s + \Delta s$ . We would then iterate on 'x' for fixed  $p$  (as predicted by (10.29a)) using the Newton-Raphson method on (10.26). The choice of the sign in (10.29) indicates the direction in which we are traversing the curve. A positive (negative) value for  $dp/ds$  indicates that  $p$  increases (decreases) with  $s$ , i.e. we are moving to the right (left) of  $A$ . Once  $dp/ds$  is determined,  $dx/ds$  is uniquely obtained from (10.27b).

At the turning point  $C$  (Fig. 10.4),  $dp/ds$  vanishes. As we approach  $C$  from the left,  $dp/ds$  changes sign from positive to negative,  $dx/ds$  also vanishes at  $C$ . Equations (10.29) can be solved for  $x, p$  across  $C$  without any numerical difficulties. To trace the middle branch we choose the sign for  $dp/ds$  in (10.29a), so that it is negative, then  $dx/ds$  is again uniquely determined as earlier. A second turning point is encountered at  $F$  where the  $dp/ds$  changes from negative to positive. The branch  $FE$  is tracked by choosing an appropriate sign for  $dp/ds$ . We continue along the top branch by increasing  $s$ , till the parameter  $p$  exceeds the value at the ignition point  $C$ . The corresponding value of  $x$  is obtained by the simultaneous solution of (10.29a), (10.29b), and (10.27a). Using the arc-length continuation method, we can systematically obtain the solution on the upper branch to the right of  $C$ , by tracing the branch  $ABCDFE$ . It is a slightly laborious method since to get to point  $E$  we have to trace the branch  $CFDE$ . However, this technique is systematic and guarantees us that all branches will be traced. This is very useful in the context of higher (say,  $n$ ) dimensional systems where it may be difficult to get a good initial guess in  $\mathbb{R}^n$  which will converge on  $E$  although we may know the solution at  $C$ .

The salient features of the homotopy continuation method have been described, with the objective of explaining the basis of the method. The philosophy behind the different variations and modifications of the basic methods has been explained. This will permit the extension of the method to more complex systems of algebraic equations, two-point boundary-value problems and tracing limit cycles. It can also be used in the numerical solution of partial differential equations where the equation is converted to a finite number of algebraic equations.

As is perhaps the case of any numerical method, there are many subtleties involved in the application of homotopy methods. A detailed exposition of these facets can be found in Kubicek and Mareck (1983). Some important aspects are:

1. The step-size and the initial guess must be chosen carefully across turning points (i.e.  $C$ ,  $F$  in Fig. 10.4) while using arc-length continuation methods.
2. The method can be used to track solutions on a branch, when we start from a point on the branch. It is not possible to generate a completely isolated branch unless we have a point on it.
3. When different branches of solutions emanate from a point on the basic solution curve, all of them can be tracked by a careful selection of step-size and initial guesses. This enables one to switch branches and continue along any branch.

**Example 10.4** The cubic autocatalysis with catalyst decay has been studied by Scott (1991) in detail. This system consists of the reactions



The dimensionless equations governing the evaluation of  $A$ ,  $B$  in a CSTR are

$$\frac{da}{d\tau} = \frac{1 - a}{\tau_{\text{res}}} - ab^2$$

$$\frac{db}{d\tau} = \frac{b_0 - b}{\tau_{\text{res}}} + ab^2 - k_d b$$

The steady states are obtained from

$$\frac{1 - a_{\text{ss}}}{\tau_{\text{res}}} - a_{\text{ss}} b_{\text{ss}}^2 = 0$$

$$\frac{b_0 - b_{\text{ss}}}{\tau_{\text{res}}} + a_{\text{ss}} b_{\text{ss}}^2 - k_d b_{\text{ss}} = 0$$

Adding these two equations and rearranging, we get

$$b_{\text{ss}} = \frac{1 + b_0 - a_{\text{ss}}}{1 + k_d \tau_{\text{res}}}$$

where  $a_{\text{ss}}$  is obtained from the cubic

$$(1 + k_d \tau_{\text{res}})^2 (1 - a_{\text{ss}}) - a_{\text{ss}} (1 + b_0 - a_{\text{ss}})^2 \tau_{\text{res}} = 0 \quad (10.30)$$

The three parameters occurring here are  $k_d$ ,  $b_0$ ,  $\tau_{\text{res}}$ . For a fixed  $k_d$  and  $b_0$ , we study the dependence of the steady state on  $\tau_{\text{res}}$ . The residence time  $\tau_{\text{res}}$  is a preferable choice as it can be easily varied and controlled by an experimentalist to verify the theoretical predictions. For  $\tau_{\text{res}} = 0$ ,  $a_{\text{ss}} = 1$  and  $b_{\text{ss}} = b_0$ . This corresponds to the state of no reaction as the flow rate is very large and the reactants hardly spend any time in the reactor. Starting from  $\tau_{\text{res}} = 0$ ,  $a_{\text{ss}} = 1$ , we construct the steady state branch for  $k_d = .05$ ,  $b_0 = .25$  by continuing along the intrinsic parameter  $\tau_{\text{res}}$ .

We view (10.30) as

$$f(a_{ss}, \tau_{res}) = 0$$

Differentiating with respect to  $\tau_{res}$ , we get

$$\frac{\partial f}{\partial \tau_{res}} + \frac{\partial f}{\partial a_{ss}} \frac{da_{ss}}{d\tau_{res}} = 0 \quad (10.31a)$$

where

$$\frac{\partial f}{\partial \tau_{res}} = (1 - a_{ss}) 2(1 + k_d \tau_{res}) k_d - a_{ss}(1 + b_0 - a_{ss})^2 \quad (10.31b)$$

$$\frac{\partial f}{\partial a_{ss}} = -(1 + k_d \tau_{res})^2 - (1 + b_0 - a_{ss})^2 \tau_{res} + 2a_{ss}(1 + b_0 - a_{ss})\tau_{res} \quad (10.31c)$$

The algorithm for the generation of the branch is

1. Fix  $\Delta\tau_{res}$  the increment in  $\tau_{res}$  at which we need to determine the solution. Set  $a_{ss}^0 = 1$ ,  $\tau_{res}^0 = 0$ .

2. Obtain

$$\left. \frac{da_{ss}}{d\tau_{res}} \right|_{a_{ss} = a_{ss}^0, \tau_{res} = \tau_{res}^0} = \tau_{res}^0$$

from (10.31a)–(10.31c).

3. Set  $\tau_{res}^1 = \tau_{res}^0 + \Delta\tau_{res}$

The guess for  $a_{ss}$ , the root of  $\tau_{res} = \tau_{res}^1$ , is

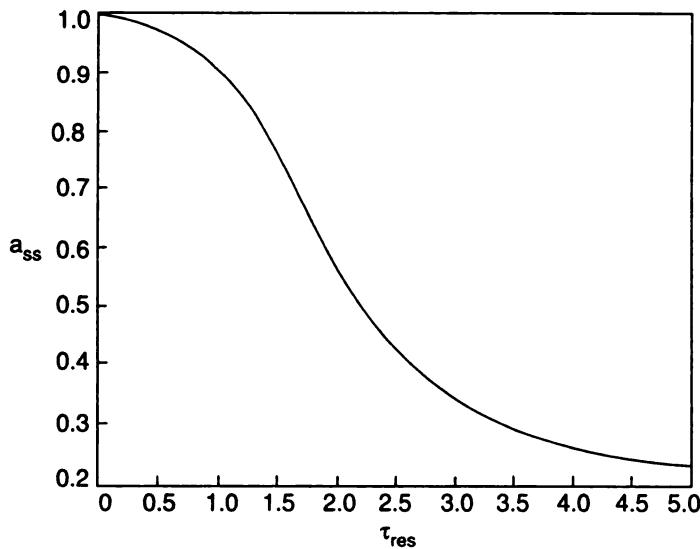
$$a_{ss}^{1 \text{ guess}} = a_{ss}^0 + \left. \frac{da_{ss}}{d\tau_{res}} \right|_{a_{ss}^0, \tau_{res}^0} \Delta\tau_{res}$$

4. Use Newton-Raphson and iterate on (10.30) to obtain  $a_{ss}^1$  for  $\tau_{res} = \tau_{res}^1$ . Store this value.

5. Update  $a_{ss}^0 = a_{ss}^1$  and  $\tau_{res}^0 = \tau_{res}^1$  and repeat steps 2–5. The steady state dependence of  $a_{ss}$  on  $\tau_{res}$  for  $\Delta\tau_{res} = 0.1$  is shown in Fig. 10.5.

A second possible choice of the bifurcation parameter is  $b_0$ . This represents the dimensionless inlet concentration of  $B$  to the CSTR. It can also be varied easily by the experimentalist, keeping the other two parameters  $k_d$ ,  $\tau_{res}$  constant. This feature will allow the engineer to experimentally validate the theoretical results.

The algorithm described here is ineffective in tracing the steady state branch of the curve for  $k_d = .02$ ,  $b_0 = .1$ . There is a sharp jump in the steady state value of  $a_{ss}$ ,  $b_{ss}$  at a critical  $\tau_{res}$ . Here  $da_{ss}/d\tau_{res} = \infty$ . This problem can be overcome by using the arc-length continuation method. We explain the algorithm and the basis of this method in the context of the first order exothermic reaction in a CSTR.



**Fig. 10.5** Dependence of  $a_{\text{ss}}$  on  $\tau_{\text{res}}$  of autocatalytic system.

**Example 10.5** The evolution of the dimensionless conversion (and temperature  $T$  of a first order exothermic irreversible reaction in a CSTR, is given by

$$\frac{dC}{d\tau} = -C + Da(1 - C)e^T$$

$$\frac{dT}{d\tau} = -(1 + \beta)T + BDa(1 - C)e^T$$

The steady states of the system  $C_{\text{ss}}$ ,  $T_{\text{ss}}$  are obtained from

$$C_{\text{ss}} = \frac{(1 + \beta)T_{\text{ss}}}{B}$$

where  $T_{\text{ss}}$  is obtained from

$$(1 + Da e^{T_{\text{ss}}})(1 + \beta)T_{\text{ss}} - BDa e^{T_{\text{ss}}} = 0 \quad (10.32)$$

This equation again has three parameters  $B$ ,  $\beta$ ,  $Da$ . The Damkohler number,  $Da$ , is a good choice for the bifurcation parameter as it contains the volumetric flow rate term  $q$ . This can be varied easily by an experimentalist to study the system behaviour. We use the arc-length continuation method to trace  $T_{\text{ss}}$  vs.  $Da$  branch for  $B = 10$ ,  $\beta = 1$ . Now  $T_{\text{ss}} = 0$  for  $Da = 0$ . This corresponds to a point on the solution curve, and we do not have to construct the auxiliary function  $h(x, t)$ . This is taken as the initial point on the branch, i.e. the arc-length parameter  $s = 0$  here. Viewing  $T_{\text{ss}}$ ,  $Da$  to be dependent on the arc-length parameter  $s$ , we differentiate (10.32) with respect to  $s$ . This yields

$$\frac{\partial f}{\partial T_{\text{ss}}} \frac{dT_{\text{ss}}}{ds} + \frac{\partial f}{\partial Da} \frac{dDa}{ds} = 0 \quad (10.33a)$$

$$\begin{aligned}\frac{\partial f}{\partial T_{ss}} &= (1 + \beta)(1 + Da \exp(T_{ss})) + T_{ss}(1 + \beta)Da \exp(T_{ss}) - BDa \exp(T_{ss}) \\ &= f_{T_{ss}}\end{aligned}\quad (10.33b)$$

$$\frac{\partial f}{\partial Da} = T_{ss}(1 + \beta) \exp(T_{ss}) - B \exp(T_{ss}) = f_{Da} \quad (10.33c)$$

Another relation between  $dT_{ss}/ds$  and  $dDa/ds$  comes from calculus, i.e.

$$\left(\frac{dT_{ss}}{ds}\right)^2 + \left(\frac{dDa}{ds}\right)^2 = 1 \quad (10.33d)$$

From (10.33a) and (10.33d)

$$\frac{dDa}{ds} = \pm \frac{f_{T_{ss}}}{\sqrt{(f_{T_{ss}}^2 + f_{Da}^2)}} \quad (10.34a)$$

$$\frac{dT_{ss}}{ds} = \mp \frac{f_{Da}}{\sqrt{(f_{T_{ss}}^2 + f_{Da}^2)}} \quad (10.34b)$$

The arc-length continuation algorithm now consists of the following steps:

1. Choose  $Da = 0.0$ ,  $T_{ss} = 0.0$  for  $s = 0$  as the initial condition.
2. Integrate (10.34) numerically using an explicit Euler method with a step-size  $\Delta s$  in  $s$ . This yields a value for  $Da$  and  $T_{ss}$ . Fix the value of  $Da$  as determined by (10.34a) and iterate for  $T_{ss}$  using a Newton-Raphson method on (10.32) and obtain the corresponding  $T_{ss}$ .
3. This  $Da$ ,  $T_{ss}$  is a point on the solution branch for  $s = \Delta s$ .

We now repeat steps 2-3 until  $dDa/ds = 0$ . To continue along this branch across this turning point, we ensure that  $dDa/ds < 0$  and repeat steps 2-3.

This algorithm helps us trace the entire solution branch (refer Fig. 10.4). The ignition point occurs at  $Da = .0959$ . Different initial guesses are given near this point to ensure that the solution converges on the intermediate branch CF, and not on AC.

It may also be necessary to use a finer spacing in  $\Delta s$  near these critical points across which the system behaviour is sensitive. We continue along this branch till we reach the extinction point. Here  $dDa/ds = 0$  again. Along the branch FE, this derivative is positive. This branch can also be traced using the method described.

The choice of the initial guess plays a crucial role in obtaining the solution for a system of nonlinear equations. The homotopy method described provides us with a valuable tool to obtain good initial guesses in tracing the solution branch. An equally effective method of generating a good initial guess is by using the physical insight from the processes occurring in a system. This is based on using a common sense approach, which as we have already seen, is the basis of the homotopy continuation methods.

The temperature of nonadiabatic CSTR sustaining an exothermic reaction will exceed the feed temperature. This provides the engineer with a lower bound on the initial guess. The adiabatic temperature rise provides him with an upper bound on the temperature. Similarly, in simulating a

distillation column, he can use the information about relative volatilities to decide as to which component will have a higher mole fraction as we move up the column. The astute engineer must use this kind of physical insight in not only making the initial guess but also in analysing the results obtained.

Nonlinear equations can, in general, have multiple solutions for a fixed set of parameters. The reactor, for example, has three steady states for a range of  $Da$ . These represent the different possible states of the system. The co-existing states can be all steady states, or some of the states can be dynamic, as we will see in Chapter 11. Two obvious questions arise in this situation:

1. Are all the admissible states attainable by the system? What determines if a particular state is attainable or not?
2. If more than one state is attainable, what determines which of these states is the terminal state of a system?

The answer to these questions can be obtained only by considering the dynamic or transient behaviour of the equations, and not just the steady state considerations. This gives rise to the notion of stability. We analyse this in detail in Chapters 11 and 12 and see how the loss of stability of a steady state can yield a dynamic state.

## PROBLEMS

1. Enzyme catalysed and fermentation processes are usually governed by Monod kinetics. They are also carried out isothermally. Consider the reaction



in a CSTR, where the rate expression is,

$$-r_A = \frac{\mu S}{K_f + S}$$

(a) Plot the bifurcation diagram depicting dependence of  $S$  on residence time for a fixed  $\mu$ ,  $K_1$ . Assume the feed concentration to be  $S_f$ .

(b) Repeat part (a) assuming the rate expression to be,

$$-r_A = \frac{\mu S}{K_1 + K_2 S + S^2}$$

2. Consider the enzyme catalysed reaction to be of the form



Assume the reaction to be first order in  $X$ , and the feed to contain pure  $S$  at concentration  $S_f$ .

- (a) Obtain the steady state bifurcation diagram for Monod dependency of rate on  $S$ .
- (b) Redo part (a) for Haldane dependency of rate on  $S$ .

3. Use the Newton-Raphson method to obtain solutions to  $x^2 + 2 = 0$ .
4. Rework Example 10.3 such that you can obtain the other solution.
5. Rework Example 10.4 using the arc-length continuation method.

**REFERENCES**

- Chandrasekhar, S., *Hydrodynamic and Hydromagnetic Instabilities*, Dover, New York (1950).
- Drazin, P.G. and Reid, W.H., *Hydrodynamic Stability*, Cambridge University Press, Cambridge (1981).
- Fleming, W.H., Report of the Panel on Future Directions in Control Theory: A mathematical perspective, Society for Industrial and Applied Mathematics (1988).
- Gupta, S.K., *Numerical Methods for Engineers*, Wiley Eastern, New Delhi (1995).
- Hagedorn, Peter, *Non-linear Oscillations*, translated by Stadler, W., Clarendon Press, Oxford (1981).
- Hildebrand, F.B., *An Introduction to Numerical Analysis*, McGraw-Hill, New York (1956).
- Iooss, G., *Bifurcations of Maps and Applications*, North-Holland, Amsterdam (1979).
- Iooss, G. and Joseph, D.D., *Elementary Stability and Bifurcation Theory*, Springer-Verlag, New York (1980).
- Kubicek, M. and Marek, M., *Computational Methods in Bifurcation Theory and Dissipative Structures*, Springer-Verlag, Berlin (1983).
- Lee, C.K., Yeung, S.Y.S. and Bailey, J.E., Experimental studies of consecutive-competitive reactions in steady state and forced periodic CSTRs, *The Canadian Journal of Chemical Engineering*, **58**, 212 (1980).
- Lin, W.J., Seader, J.D. and Wayburn, T.L., Computing Multiple Solutions to Systems of Interlinked Separator Columns, *AIChE*, **33**, 886 (1987).
- Lorenz, E.N., Deterministic non-periodic flow, *J. Atmos. Sci.*, **20**, 130 (1963).
- May, R.M., *Stability and Complexity in Model Ecosystems*, Princeton University Press, Princeton (1973).
- Poore, A., A Model Equation Arising from Chemical Reactor Theory, *Archives of Rational Mechanics and Analysis*, **52**, 358 (1973).
- Pushpavanam, S., The D-partition Method: An application to the first order reaction in a CSTR, *Chemical Engineering Science*, **49**, 502 (1992).
- Roberts, S.M. and Shipman, J.S., *Two-point Boundary-value Problems: Shooting methods*, American Elsevier Publishing Company, New York (1972).
- Scott, S.K., *Chemical Chaos*, Clarendon Press, Oxford (1991).
- Vidyasagar, M., *Non-linear Systems Analysis*, Prentice-Hall, Englewood Cliffs, New Jersey (1979).

# 11

## Linear Stability and Limit Cycles

---

---

In Chapter 10 we have seen the steady state behaviour of nonlinear systems. Under some conditions these systems admit multiple steady states. Theoretically, the system can be in any one of these steady states. We would like to determine on what basis the system decides which steady state to wind up in. Are all the possible steady states attainable by the system? If so, why? If not, why not? Is it possible for the system to be at a terminal state which is not a steady state?

These questions can be addressed only by looking at the dynamic behaviour of systems. The concept of the stability of a state is based on the dynamic behaviour of the system near this state.

### 11.1 LINEAR STABILITY OF DYNAMICAL SYSTEMS

#### 11.1.1 One Dimensional System

By a one-dimensional system we mean a system with one dependent variable. The governing equation is a single nonlinear ordinary differential equation. We illustrate the concept of stability of a steady state for a hypothetical one-dimensional system.

$$\dot{x} = -(x - 1)(x - 2)(x - 3) = f(x) \quad (11.1)$$

This system has three steady states:  $x = 1$ ,  $x = 2$ , and  $x = 3$ . Figure 11.1(a) is a plot of  $f(x)$  vs.  $x$ . The graph intersects the  $x$ -axis at the three roots of  $f(x)$ , i.e  $x = 1, 2, 3$ . At these three points,  $\dot{x} = 0$ , and so if the system originates at these points, it remains there. We would like to determine how the system is going to behave or evolve when it is away from these points. This is reflected in the variation of  $x$  with respect to time. The real line is divided into four segments:  $(-\infty, 1)$ ,  $(1, 2)$ ,  $(2, 3)$  and  $(3, \infty)$ , by these three points. The behaviour of generic points in each of these four segments is as follows:

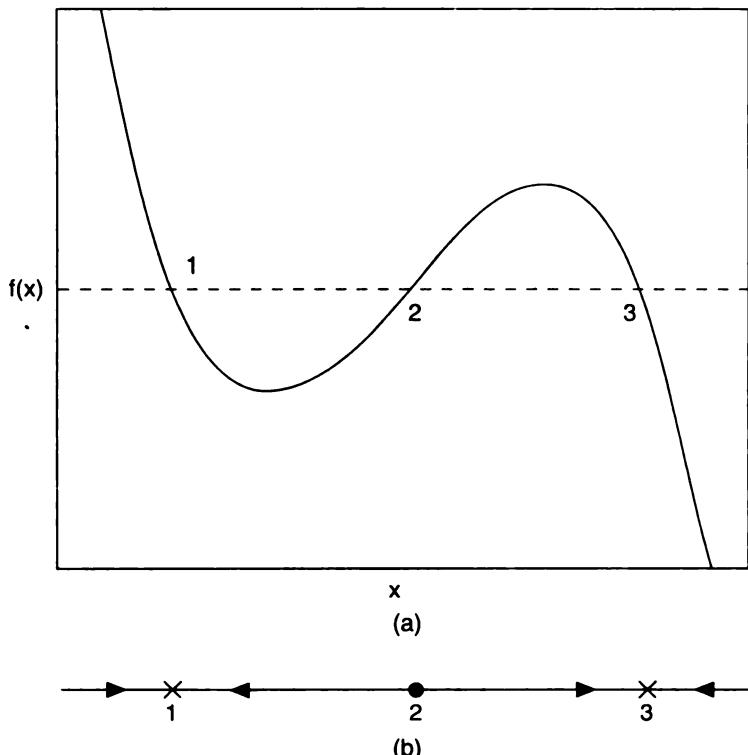
$(-\infty, 1)$ : In this segment,  $f(x)$  is positive; consequently,  $\dot{x} > 0$  and  $x$  increases with time monotonically till it reaches the value 1.

$(1, 2)$ : Here  $f(x)$  is negative and  $x$  decreases with time till it reaches  $x = 1$ .

$(2, 3)$ : In this interval  $f(x) > 0$ , and if the system starts evolving from any point here, it winds up at  $x = 3$ .

$(3, \infty)$ : We determine, in a similar fashion, that points on this interval will tend to  $x = 3$  as  $t$  increases to infinity.

This situation is succinctly represented by the real line, see Fig. 11.1(b). The arrows indicate the direction in which the system will evolve if left to itself. The steady state at 2, though a feasible



**Fig. 11.1** (a) Dependence of  $f(x)$  of (11.1) on  $x$ ; (b) evolution or behaviour of points on different intervals of the real line.

steady state, cannot be attained by the system. The state at  $x = 2$  is unstable as a deviation however small from it such that  $x > 2$ , gets amplified with time and the system tends to 3. Similarly, a negative deviation from  $x = 2$  such that  $x < 2$  drives the system away from 2 to 1. Points very close to the system are repelled away by it, i.e. they diverge away from it. Such a state is said to be unstable. In a realistic system there are always small perturbations which are present. These cannot be eliminated. Theoretically, the system therefore can never attain a steady state but will always hover around a small neighbourhood around it. The nonlinearities present in the system amplify the deviations near the steady state  $x = 2$ , and hence the system is driven away from the steady state at 2 when it is left to itself. In this text, when we say a state is unstable, we mean small deviations present in the neighbourhood of the state get amplified and the system is driven away from the state by the nonlinear interactions present. The state at 3 is a stable steady state. Consider a deviation of the system such that  $x$  becomes greater than 3. Now since  $f(x) < 0$ ,  $x$  decreases to 3, the deviation decays, and the state is stable. If the system were to be given a small perturbation such that the state  $x$  satisfies  $2 < x < 3$ , then  $x$  increases back to 3 as here  $f(x) > 0$ .

Similarly, the system state at 1 is stable. Points in the neighbourhood of the states  $x = 1$ ,  $x = 3$  are attracted to them as shown in the phase-line (see Fig. 11.1b). Any deviation sufficiently small from these steady states decays, and the system relaxes back to these steady states. These states are therefore called *attractors* as they possess the properties of attracting the neighbouring points. We conclude, in general, that stable states are attractors and unstable states are repellers, see Bhatia (1967) and Minorsky (1969).

A system state is **stable**, when infinitesimal disturbances imposed on the system in terms of state or dependent variables decay and the system converges or attains the original state. Since we are discussing the behaviour of the system in a small vicinity of its state, this stability is called a **local stability**. It is also called **linear stability** since in the vicinity of this state the evolution of variables is well represented by the linearised equations, see Hagedorn (1981) and Minorsky (1969).

We will discuss the concept of stability in this text only in terms of local stability. A state is said to be **globally stable** if it is stable to all perturbations and not only to infinitesimal ones. It is difficult to prove global stability for most problems. We illustrate this idea with reference to (11.1). The states at  $x = 1, x = 3$  are locally stable but globally unstable. There exist sufficiently big disturbances which can drive the system away from each of these states. Consider the system operating at  $x = 3$ . If a sufficiently big disturbance were to render  $x < 2$ , then the system would eventually tend to 1. The system at  $x = 3$  is unstable to big disturbances and is globally unstable but locally stable. Similarly, the state  $x = 1$  is globally unstable. Here disturbances  $x$  such that  $x > 2$ , terminate at  $x = 3$ .

Two facts emerge from the above discussion:

1. The state to which a given system tends to is determined solely by the initial conditions.
2. In particular, not all steady states (and, in general, terminal states) which are admissible will be attained by the system. Moreover, a system can attain one of many possible states for a fixed set of operating conditions. Philosophically, while solving the steady state equation

$$f(x) = 0$$

we lose some information about the system. This information is contained in the initial condition of the system.

For a one-dimensional continuous dynamical system, the only possible terminal states are steady states. Moreover, the approach to the steady state is monotonic with respect to time. This kind of a steady state is formally called a *node*. A node can be stable or unstable. The state at  $x = 2$  is an unstable node and the states at  $x = 1, 3$  are stable nodes. We will formally define the concept of node while discussing two-dimensional systems.

The discussion so far is based on the geometrical approach and uses the form of  $f(x)$ . It helps in understanding the concept of stability physically. This concept is introduced in terms of the evolution of disturbances. It thus has a physical and a geometrical basis. We will now see how to determine the stability of a state mathematically. This will allow us to generalise the ideas presented so far to higher dimensional systems. Here graphical representations and the geometric approach is not possible and is not elegant. In these cases it is difficult to understand system behaviour using the geometric approach of the one-dimensional system. The mathematical theory on which stability is determined is called **linear stability analysis** (see Haken (1983) and Nicolis (1986)). This has an algebraic basis and is valid for dynamical systems of any order of complexity. What we are trying to do now is similar to the extension of the geometric concepts of metrics, norms and inner-products to higher dimensional spaces using the algebraic representation in Chapter 2.

The theory of linear stability is sufficiently general and can be applied to systems of any degree of complexity. We first consider its application to a single autonomous equation

$$\dot{x} = f(x) \quad (11.2a)$$

The steady state of the above system satisfies

$$0 = f(x_{ss}) \quad (11.2b)$$

We would now like to determine if the given steady state is stable. Let  $\hat{x}$  represent the deviation of the system state  $x$  from the steady state  $x_{ss}$ , i.e.

$$\hat{x} = x - x_{ss} \quad (11.2c)$$

By substituting (11.2c) in (11.2a), the evolution of  $\hat{x}$  can be obtained as

$$\dot{\hat{x}} = f(\hat{x} + x_{ss})$$

For sufficiently small disturbances  $\hat{x}$ , we can linearise  $f$  around  $x_{ss}$ , i.e. expand  $f(x)$  in a Taylor series around  $x_{ss}$  and retain only the first order terms. This yields

$$\dot{\hat{x}} = \frac{df}{dx} \Big|_{x=x_{ss}} \hat{x} \quad (11.3)$$

The derivative in (11.3) is a constant since it is evaluated at the known steady state ( $x_{ss}$ ) whose stability we are interested in. This is a linear equation with a constant coefficient, and its solution is

$$\hat{x}(t) = \hat{x}(0) \exp [f'(x_{ss})t] \quad (11.4)$$

Here,  $\hat{x}(0)$  represents the deviation from the steady state at  $t = 0$ . The state is stable if  $\hat{x}(t) \rightarrow 0$  as  $t \rightarrow \infty$ . This will ensure  $x(t) \rightarrow x_{ss}$ . This is assured if  $f'(x_{ss}) < 0$ . It must be remembered that the linearisation restricts the validity of (11.3) to small  $\hat{x}(0)$ , and so this is only a local condition.

The condition for the stability of a steady state here is

$$f'(x_{ss}) < 0$$

and a condition for instability is

$$f'(x_{ss}) > 0$$

A state that satisfies the stability condition may be unstable globally. The state  $x_{ss} = 3$  in (11.1) has  $f'(3) < 0$ . It is locally stable but globally unstable. The state at  $x = 2$  is locally unstable as  $f'(2) > 0$ . It is unstable to small disturbances and consequently to all disturbances. So it is globally unstable.

To summarise, a locally unstable system will be globally unstable, but a locally stable system may be globally stable or unstable. Hence the conditions determining stability properties are the necessary conditions for stability and sufficient conditions for instability.

### 11.1.2 *N*-Dimensional Systems ( $N \geq 2$ )

Consider the general two-dimensional autonomous system represented by

$$\dot{x}_1 = f_1(x_1, x_2), \quad \dot{x}_2 = f_2(x_1, x_2) \quad (11.5)$$

We have suppressed the dependence of the parameter vector  $p$  in the above equation, as we are interested in determining the steady state and its stability for a fixed set of parameters  $p$ . Since the system is autonomous, it can admit a steady state. The steady state is characterised by the 2-tuple  $(x_{1ss}, x_{2ss})$ . Denoting the deviations of the system from the steady state by  $\hat{x}$ , we have

$$\hat{x}_1 = x_1 - x_{1ss}, \quad \hat{x}_2 = x_2 - x_{2ss}$$

To determine the evolution of the deviations, we substitute this in (11.5) to obtain

$$\left. \begin{aligned} \dot{\hat{x}}_1 &= f_1(\hat{x}_1 + x_{1ss}, \hat{x}_2 + x_{2ss}) \\ \dot{\hat{x}}_2 &= f_2(\hat{x}_1 + x_{1ss}, \hat{x}_2 + x_{2ss}) \end{aligned} \right\} \quad (11.6)$$

For small deviations in (11.6) about the steady state, we linearise the functions around the steady state, to obtain

$$\left. \begin{aligned} \dot{\hat{x}}_1 &= f_1(x_{1ss}, x_{2ss}) + \frac{\partial f_1}{\partial x_1} \hat{x}_1 + \frac{\partial f_1}{\partial x_2} \hat{x}_2 \\ \dot{\hat{x}}_2 &= f_2(x_{1ss}, x_{2ss}) + \frac{\partial f_2}{\partial x_1} \hat{x}_1 + \frac{\partial f_2}{\partial x_2} \hat{x}_2 \end{aligned} \right\} \quad (11.7a)$$

The higher order terms can be neglected as long as  $\hat{x}_1, \hat{x}_2$  are small (infinitesimal). The partial derivatives are all evaluated at the steady state of interest. This set of linear equations with constant coefficients can be cast in vectorial form as

$$\frac{d\hat{x}}{dt} = J\hat{x} \quad (11.7b)$$

Here  $J$ , the Jacobian matrix, is given by

$$J = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}$$

and  $\hat{x} = [\hat{x}_1, \hat{x}_2]'$ . The elements of the Jacobian are evaluated at the steady state. Equation (11.7) governs the evolution of  $\hat{x}$  as long as it is in an infinitesimal neighbourhood of the steady state. It follows from the results in Chapter 4 that the solution to the above equation is of the form

$$\hat{x}(t) = c_1 u^1 e^{\lambda_1 t} + c_2 u^2 e^{\lambda_2 t} \quad (11.8)$$

Here  $\lambda_1, \lambda_2$  represent the eigenvalues of  $J$  and  $u^1, u^2$  are its eigenvectors. The constants  $c_1, c_2$  are evaluated from the initial conditions as discussed earlier in Chapter 4. This is determined by the perturbation at  $t = 0$ . The eigenvalues can, in general, be complex. The long time behaviour of the deviation  $\hat{x}(t)$  is then governed by the real part of the eigenvalues. If

$$\operatorname{Re}(\lambda_i) > 0$$

for even one eigenvalue, then the perturbation will get amplified, rendering the system unstable.

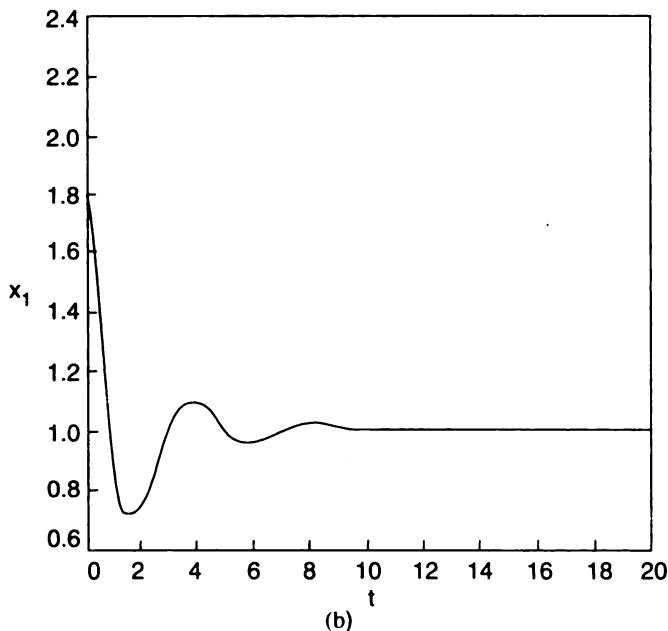
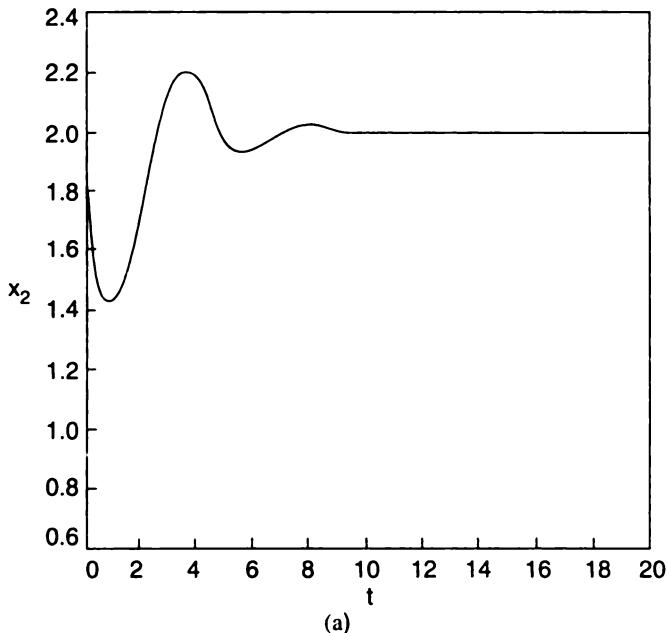
In representing  $\hat{x}(t)$  by (11.8), we are decomposing  $\hat{x}(0)$ , the initial perturbation, in the direction of the eigenvectors. If all the eigenvalues have negative real parts, the components of the perturbation decrease in magnitude with time in all directions, and the system winds up on the steady state. If even one of the eigenvalues has positive real parts, then the perturbation component in the direction of the corresponding eigenvector increases with time. After some time, the perturbation becomes sufficiently large and cannot be governed by the linear equations any more. The nonlinear effects come into picture and the new system state is found by solving the nonlinear equations (11.5). The stability of a steady state for a general system is therefore related to the eigenvalues of the Jacobian matrix. For an arbitrary  $n$ -dimensional system, a necessary condition for stability of a steady state is

$$\operatorname{Re}(\lambda_i) < 0 \text{ for all } i = 1, \dots, n \quad (11.9a)$$

A sufficient condition for instability, on the other hand, is

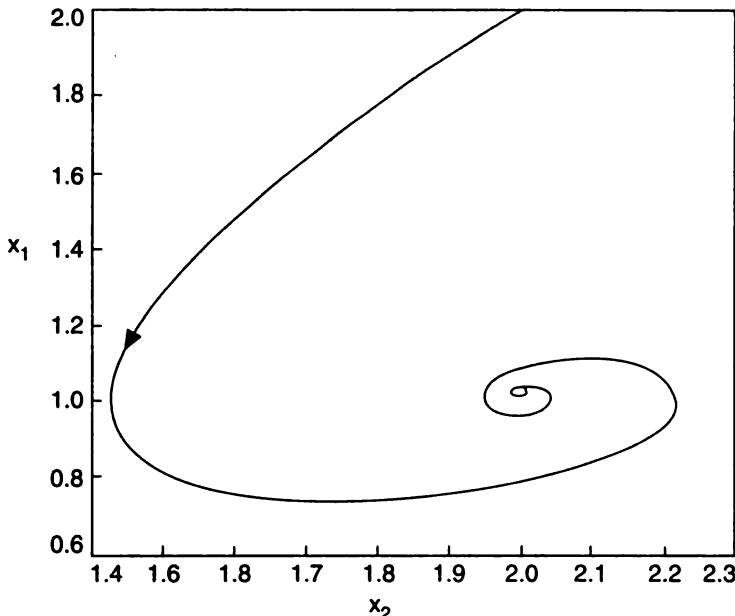
$$\operatorname{Re}(\lambda_i) > 0 \text{ for at least one } i \quad (11.9b)$$

In a two-dimensional system, the system behaviour can be understood by studying both the variables  $x_1$ ,  $x_2$ . A possible way to represent this is to plot  $x_1$  and  $x_2$  as a function of time as shown in Fig. 11.2. In this representation we require one plot for each variable; so for an  $n$ -dimensional system we need  $n$  such figures.



**Fig. 11.2** Possible evolution of state variables to a stable steady state in a two-dimensional system:  
 (a) evolution of  $x_2(t)$ ; (b) evolution of  $x_1(t)$ .

A compact way to represent system behaviour involves representing the trajectory in the  $x_1 - x_2$  plane. Such a plane is called a **phase plane** (Fig. 11.3). The system state at any instant of time is characterised by a point in this plane. As the system evolves with time, it goes through different states and we trace out different points in this plane. The locus of these points represents



**Fig. 11.3** Phase plane representation of system trajectory corresponding to Fig. 11.2 ( $t$  is a parameter along the trajectory; arrow indicates direction of evolution).

the trajectory of the system, see Haken (1983) and Verhulst (1980). The trajectory is a directed line segment. The direction is represented by an arrow along the curve and time is a parameter along this trajectory in, the phase plane. The phase plane can be mathematically viewed as a parametric representation of the trajectory of a system where time is the parameter. A unique trajectory passes through every point in phase plane and the system behaviour is uniquely determined by the initial condition, see Minorsky (1969).

The steady state of a two-dimensional system is classified as being a node or a focus or a saddle by the behaviour of the trajectories in its vicinity. This classification can be best understood geometrically in terms of phase plane behaviour, and mathematically in terms of the eigenvalues. These definitions are valid only for a two-dimensional system, and they can not directly be extended to higher dimensional systems, see Minorsky (1969).

(i) **Unstable node.** This state is characterised by two real eigenvalues  $\lambda_1, \lambda_2$  both positive, i.e.  $\lambda_1 > 0, \lambda_2 > 0$ . The behaviour of the trajectories in the phase plane is shown in Fig. 11.4(a). Here the trajectory diverges from the steady state monotonically in all directions. The two eigendirections corresponding to these two eigenvalues are depicted as  $u^1, u^2$  in the figure. Two trajectories are tangential to these eigendirections at the steady state as shown in the figure.

(ii) **Stable node.** At this steady state, the two real eigenvalues are both negative, i.e.  $\lambda_1 < 0,$

$\lambda_2 < 0$ . The trajectories converge to the steady state from all directions in phase plane. The eigenvectors corresponding to these eigenvalues are marked as  $u^1$ ,  $u^2$  in Fig. 11.4(b). The eigenvectors are tangential to two unique trajectories at the steady state. The approach to the steady state from nearby points is monotonic for this case.

**(iii) Saddle.** At this steady state, we have two real eigenvalues: one is positive, and the other is negative, i.e.  $\lambda_1 > 0$ ,  $\lambda_2 < 0$ . By definition a saddle is unstable. Trajectories in phase-plane diverge away from a saddle in all directions except one Fig. 11.4(c). This direction along which trajectories converges to the saddle, is tangential to the eigenvector  $u^2$  (corresponding to the stable eigenvalue) near the steady state. A disturbance along this direction (if we could control it exactly) would decay to the saddle. All other disturbances grow away from it.

So far we have discussed the case where both the eigenvalues are real. Another possibility which occurs in two-dimensional systems is when the eigenvalues form a complex-conjugate pair. The eigenvectors now have complex elements and cannot be represented in the real phase-plane. The steady state is characterised as described now.

**(iv) Stable focus.** Here the real part of the complex eigenvalues are negative. The trajectories in the vicinity of this focus spiral in towards it. There can be no representation of the two eigendirections in the phase-plane for this case as the eigenvector has complex elements. The approach of each variable to the steady value is oscillatory as shown in Fig. 11.4(d).

**(v) Unstable focus.** Here the two eigenvalues are complex and they have positive real parts. The trajectories emanating from the steady state now spiral outward from it. Each variable diverges in an oscillatory manner from the steady state as shown in Fig. 11.4(e). This behaviour results from the complex part of the eigenvalues.

These concepts are useful to visualise trajectories in a phase-plane. They do not have any direct relevance and cannot be easily extended to higher dimensional systems. The representation of trajectories is now in phase-space. In an  $n$ -dimensional system, for example, we could have some real eigenvalues and some complex-conjugate pair of eigenvalues. The different combinations of eigenvalues possible are enormous and we cannot classify system states elegantly.

**Example 11.1** Consider the system of coupled ordinary differential equations

$$\dot{x} = x - xy, \quad \dot{y} = -y + xy - y^2$$

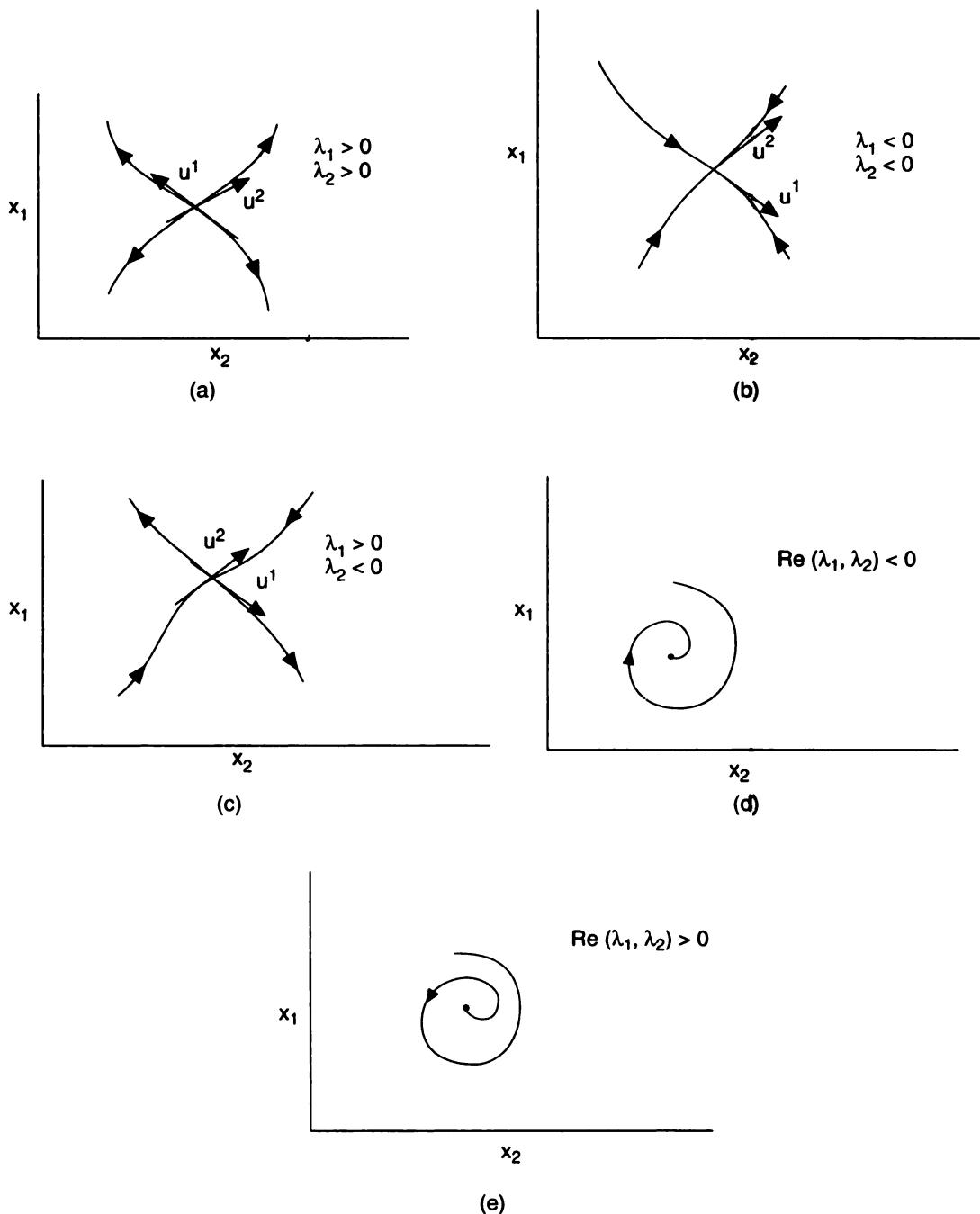
Such systems occur in the modelling of autocatalytic systems in a CSTR.

The system admits two steady states: (i)  $(0, 0)$ , (ii)  $(2, 1)$ . Let us determine the nature of these states. The Jacobian matrix which determines the system stability is given by

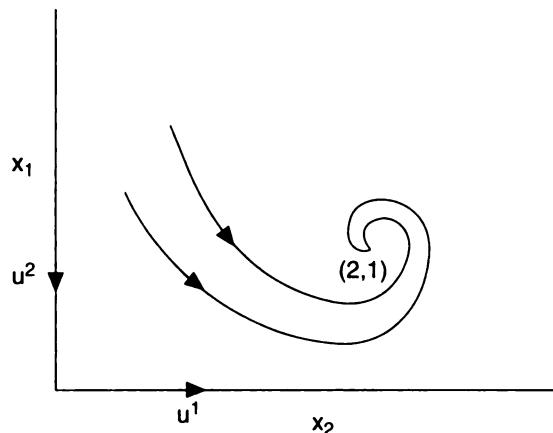
$$\begin{bmatrix} 1-y & -x \\ y & -1+x-2y \end{bmatrix}$$

At the trivial state  $(0, 0)$ , the matrix  $J$  has eigenvalues  $\lambda_1 = +1$ ,  $\lambda_2 = -1$ , and so it is a saddle. The corresponding eigendirections are  $u^1 = (1, 0)'$  and  $u^2 = (0, 1)'$ . The eigenvalues at the nontrivial state  $(2, 1)$  are  $(-1 + \sqrt{7}i)/2$  and  $(-1 - \sqrt{7}i)/2$ . Clearly, this is a stable focus. The trajectories in phase-space are shown schematically in Fig. 11.5.

The linear stability analysis is valid near the steady state of interest. It indicates only the local features of the system. The evolution of trajectories from each point in the system can be represented



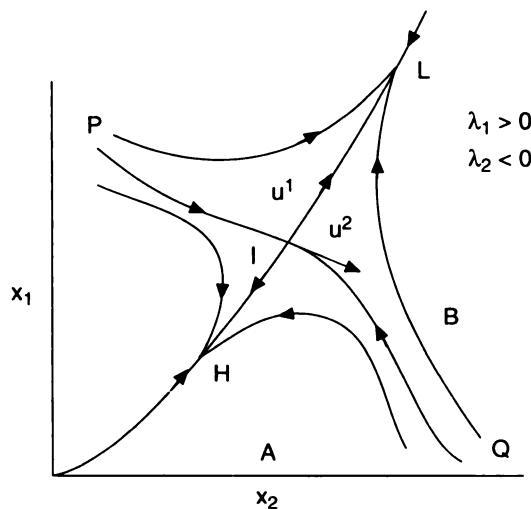
**Fig. 11.4** Schematic phase-plane portraits illustrating the different steady states of two-dimensional systems: (a) unstable node; (b) stable node; (c) saddle; (d) stable focus; (e) unstable focus.



**Fig. 11.5** Phase-plane portrait of system of Example 11.1. The origin is a saddle and the point (2, 1) is a stable focus.

in the phase-plane (or space). It therefore contains the global features of the system, which includes the nonlinear interaction of the variables. The trajectories in a phase-plane have the same properties as the lines of force of a magnetic field or streamlines in fluid flow. Analogous to these disciplines, no two trajectories can intersect at a point in phase-space.

Consider a phase-plane where we have many stable steady states. For concreteness let us consider Example 10.5. The dependence of the system on the parameter  $Da$  is shown in Fig. 10.6. A schematic phase-plane portrait of this system for a  $Da$ , where multiple steady states exist, is drawn in Fig. 11.6. Points in the neighbourhood of each stable steady state are attracted to it. The collection of points in the phase-plane, which are attracted to a state, are said to lie in the ‘basin of attraction’ of that state. In Fig. 11.6, the basin of attraction of the stable steady state  $H$  are all points lying to the left of the line  $PQ$  (Scott, 1991). If a trajectory were to start at a point in this



**Fig. 11.6** Basin of attraction when co-existing stable states are present  $PQ$  is a separatrix and is the boundary of the basin of attraction.

region  $A$ , i.e. if the system were to be initially at a point in  $A$ , it would eventually evolve to the steady state  $H$ . The points to the right of the line  $PQ$  in region  $B$  will be attracted to the other steady state  $L$ . From this it becomes clear that the initial condition determines the terminal state of a system when many stable states co-exist for a higher dimensional system. The boundary of the basin, of attraction (in this case  $PQ$ ), therefore, presents a critical surface across which the trajectories are very sensitive. A small change in the initial condition across  $PQ$  drastically changes the system trajectory. The line  $PQ$  passes through the intermediate steady state  $I$ , which is a saddle. The trajectories along  $PQ$  evolve as indicated by arrows. Near steady state  $I$ ,  $PQ$  is tangential to the eigenvector corresponding to the negative real eigenvalue. Since the points along  $PQ$  are attracted to steady state  $I$ , this locus is called the stable manifold of steady state  $I$ . Similarly, the unstable manifold represents the trajectory tangential to eigenvector of the positive real eigenvalue of the Jacobian at steady state  $I$ . This trajectory evolves to the other stable states like all other points in phase-plane. We will see some schematic phase portraits in two-dimensional systems in Example 11.5. The bifurcation diagram depicts the dependence of a state variable on a parameter. Hence, if a system has many possible states, these are depicted in such a diagram. The phase-plane portrait, on the other hand, represents the behaviour of the system for a fixed set of parameters. It is therefore a constant parameter section of the bifurcation diagram, and contains information about how the state variables interact to determine the trajectories of the system. When solving for the steady states of a system using the algebraic equation (11.2b), we lose the information about the initial condition of the system (in fact, it is not used). The evaluation of the dynamic system (11.2a), on the other hand, is tracked subject to a given initial condition and is uniquely determined.

The phase-plane portraits must be consistent. This implies that a unique trajectory passes through every point in the phase-plane. Every trajectory must end on one of the terminal stable states or diverge to infinity. As we will establish in Section 11.2.2, no two trajectories can intersect at a point. Hence the trajectories in phase-plane depicting system evolution are similar to streamlines in a fluid flow field.

The stability of a given steady state is determined by the eigenvalues of the Jacobian matrix evaluated at the steady state. The steady state is stable (locally) if all eigenvalues have negative real parts. Although it is quite easy to compute the eigenvalues explicitly and determine if a state is stable or unstable, it is desirable from a practical and an operational point of view to obtain stability conditions in terms of intrinsic parameters occurring in a system. The stability condition (11.9) is in terms of the eigenvalues. We will now see how to obtain this in terms of the intrinsic operational parameters  $p$  of the system. The intrinsic parameters in the system influence the steady state of the system. This, in turn, affects the elements of the Jacobian matrix and its eigenvalues. It is therefore clear that the stability of a steady state can change as we change the system parameters. The engineer can then choose the parameters  $p$  which occur in the original nonlinear dynamical system so that he can operate it at a stable steady state.

The eigenvalues of a matrix are determined in terms of the characteristic polynomial of the form

$$s^n + a_1 s^{n-1} + a_2 s^{n-2} + \dots + a_n = 0 \quad (11.10)$$

The coefficients of the polynomial  $a_i$  contain the parameters  $p$  and the variable  $s$  describing the system state. The parameter set  $p$  can affect the eigenvalues: (i) directly or explicitly since  $a_i$ 's contain  $p$ , or (ii) indirectly by influencing the steady state which depends on  $p$ . The following theorem enables us to obtain a necessary criterion for the stability of a system in terms of the coefficients  $a_i$  (see Porter, 1967).

**Theorem 11.1** A necessary condition for all roots of (11.10) to lie in the left-half plane, i.e. to have negative real parts, is that the coefficient of (11.10) should not have any sign changes and no coefficient should vanish.

Consider the characteristic equation, with real coefficients  $a_i$

$$F(s) = a_0 s^n + a_1 s^{n-1} + a_2 s^{n-2} + \dots + a^n = 0 \quad (11.11a)$$

Since the coefficients are real, the complex roots occur in pairs as complex conjugates. Let this equation have  $p$  real roots  $\alpha_1, \alpha_2, \dots, \alpha_p$  and  $q$  pairs of complex conjugate roots  $\beta_i \pm iv_1, \beta_2 \pm iv_2, \dots, \beta_q \pm iv_q$ . This implies that the above polynomial can be factorised as

$$F(s) = a_0(s - \alpha_1)(s - \alpha_2)\dots(s - \alpha_p)(s^2 - 2\beta_1 s + \beta_1^2 + v_1^2)\dots(s^2 - 2\beta_q s + \beta_q^2 + v_q^2) \quad (11.11b)$$

If the systems were stable, then clearly,

$$\beta_i, \alpha_i \leq 0 \quad \text{for all } i$$

Hence all coefficients will be positive in the factors of  $F(s)$ . Consequently, on multiplication of these factors, all the coefficients  $a_i, i = 1, \dots, n$ , will have the same sign as  $a_0$ . Hence if a system with a real characteristic polynomial is stable, then all the coefficients in the polynomial will be nonzero and of the same sign.

**Remarks.** The condition that all coefficients of a polynomial be of the same sign for stability is a necessary condition (see Porter, 1967). If this condition is violated, then the system is definitely unstable and if the condition is satisfied, the system may or may not be stable.

If there are no sign changes in  $a$ , the state could be stable or unstable. The presence of a sign change, however, guarantees that the state is unstable. This theorem is frequently used to establish instability rather than stability. The equation

$$s^4 - 3s^3 + 2s^2 + s + 1 = 0$$

is the characteristic equation of an unstable system.

$$s^3 + 2s^2 + s + 1 = 0$$

Here, the system may or may not be stable.

For a two-dimensional system, the eigenvalues are obtained from the quadratic equation

$$s^2 + a_1 s + a_2 = 0 \quad (11.12a)$$

If  $a_1 > 0$  and  $a_2 > 0$ , the above theorem states that we may have a stable steady state. For this case, i.e. the two-dimensional system and the quadratic characteristic equation, these conditions are both necessary and sufficient. Violation of the conditions ensures us of instability and satisfying these guarantees stability. The cubic equation

$$s^3 + a_1 s^2 + a_2 s + a_3 = 0 \quad (11.12b)$$

determines the eigenvalues of a three-dimensional system. In order for a state to be stable,  $a_1, a_2, a_3$  have to be all positive. As the following example shows, this is not a sufficient condition for stability. The cubic characteristic equation

$$s^3 + s^2 + 4s + 30 = 0$$

has no sign changes in the coefficients. It satisfies the necessary condition for stability. In spite of

this it is unstable. Here the three roots are  $-3, 1 - 3i, 1 + 3i$ . The complex-conjugate pair of eigenvalues have a positive real part rendering the system unstable.

To obtain sufficient conditions, we need the following condition as well:

$$a_1 a_2 - a_3 > 0$$

It is easy to establish that this additional condition for stability arises from the Routh-Hurwitz criterion encountered in process control.

The complete set of conditions

$$a_1 > 0, \quad a_2 > 0, \quad a_3 > 0 \quad \text{and} \quad a_1 a_2 - a_3 > 0 \quad (11.12c)$$

form the necessary and sufficient set of conditions required for stability of a steady state of a three-dimensional system. Similarly, for other higher order polynomials we get other additional conditions (see Porter, 1967).

The stability conditions in terms of the eigenvalues has now been translated in terms of the coefficients of the characteristic polynomial. We have mapped the stability boundary (the imaginary axis) from the eigenvalue plane to the coefficient ( $a_i$ ) plane. This can be used to get the stability conditions in terms of system parameters occurring in the coefficients. It must be remembered that changing a system parameter may change the steady state. This will in turn affect the different coefficients implicitly and thereby the stability of the steady state.

It appears from what we have discussed so far that the steady state is the only plausible state to which an autonomous system will tend to as  $t \rightarrow \infty$ . The phase portrait near an unstable focus suggests the possibility of the existence of a periodic state. The outward spiralling trajectories can possibly terminate on a closed curve. The variables  $x_1, x_2$  would then repeat themselves periodically. Such a state is called a limit cycle. The time period of the motion would be the time the trajectory takes to return from a point on this curve and back to itself. The long time (limiting) behaviour of the system here is cyclic with respect to time. The variables oscillate with time at a fixed frequency in this state. An important point to note here is that the entire system is autonomous. There is no periodic input or variation in any parameter. The oscillatory behaviour of the system is a result of the inherent nonlinearities in the system and is not externally induced.

The bifurcation theory helps us study the origin of limit cycles in a system and analyse it. It is again a general theory and we will now describe it in the context of finite-dimensional systems. The theory helps us study how a system state gets destabilised as we vary a parameter across critical points. It helps us determine these critical points and discusses qualitative and quantitative characteristics system behaviour across these points.

## 11.2 BIFURCATION THEORY

The word ‘bifurcation’ means “breaking up into two”. In bifurcation theory we study how a solution branch breaks up into two, as we vary a parameter, giving rise to another solution. The steady state of a system is determined by a large number of parameters which occur in the model. The dependence of a steady state on a parameter is generated by varying it and keeping all others constant. This enables us to study how a steady state solution changes. Homotopy continuation methods can be used to trace a steady state branch, as discussed in Chapter 9. The stability of the solution for a fixed set of parameters is determined using linear stability analysis at each point. As discussed earlier, the eigenvalues of the solution depend on the parameter value. The loss of stability of a solution occurs across critical points called bifurcation points (see Haken, 1983). Here the original or basic solution becomes unstable and a new solution is born. Bifurcation

theory deals with the location of these points and determination of the nature of the solutions emanating from these points. The number of solutions of a system changes across a bifurcation point.

### 11.2.1 Dynamic Systems

A dynamical system is characterised by the evolution of all the dependent variables continuously with time. The governing equations are a system of first order equations (initial-value problems). A stable steady state of an  $n$ -dimensional dynamical system is characterised by all its  $n$  eigenvalues lying in the left-half plane. As we vary a parameter of the system, the steady state itself may change, and these eigenvalues move around in the left-half plane. A steady state can become unstable in two basic ways: (i) a real eigenvalue crossing the imaginary axis; (ii) a complex conjugate pair of eigenvalues crossing the imaginary axis. It is clearly conceivable that more than one eigenvalue may cross the imaginary axis simultaneously. These could be all real, or all complex, or a combination of both. We will not consider these as they are degenerate cases and rarely occur. The two basic mechanisms by which an eigenvalue crosses the imaginary axis give rise to the following two bifurcations from a steady state, see Haken (1983) and Kuramoto (1984).

**(i) Saddle-node bifurcation.** This is also called a steady state bifurcation. At this bifurcation point, a real eigenvalue crosses the imaginary axis from the left to the right, and all other  $(n - 1)$  eigenvalues remain in the left-half plane. At this point, the basic steady state solution becomes unstable and a new steady state is born. This evolves as another steady state branch from the basic solution branch.

**(ii) Hopf bifurcation.** Here, as we vary the parameter across a critical point, a complex-conjugate pair of eigenvalues crosses the imaginary axis and all other  $n - 2$  eigenvalues remain in the left-half plane. At this point the steady state solution becomes unstable and a periodic state or a limit cycle is born as we cross the critical point. A detailed discussion of this bifurcation can be found in Marsden and McCracken (1976).

The linear-stability analysis coupled with the dependence of the eigenvalues on the parameter is used to locate the bifurcation points. Across these points new solutions are born, i.e. a number of solutions emerge. The nature of the emerging solutions, their stability, the directions in which they exist, etc. can only be determined by a higher order analysis. We will restrict ourselves in this text to only the qualitative aspects of these facets without going through the quantitative analysis. The interested reader can find this analysis detailed in Kuramoto (1984).

Let  $p$  be the bifurcation parameter (the parameter we vary) and  $p_c$  be the critical value of  $p$  at which the steady state becomes unstable. For the sake of concreteness, let all eigenvalues be in the left-half plane for  $p < p_c$ , and let the system be rendered unstable as we increase beyond  $p_c$ . Let us arrange the eigenvalues such that their real parts are in descending order, i.e.

$$\operatorname{Re}(\lambda_1) \geq \operatorname{Re}(\lambda_2) \dots$$

A perturbation close to the steady state is governed by the linearised equations of the form (11.7b). It will evolve according to

$$\hat{x}(t) = c_1 u^1 e^{\lambda_1 t} + c_2 u^2 e^{\lambda_2 t} + \dots \quad (11.13)$$

For  $p < p_c$ , all the eigenvalues are in the left-half plane and so the perturbations decay to the trivial state. We now discuss the behaviour for  $p > p_c$  for the above types of bifurcation.

(i) *Saddle node bifurcation.* Consider now the case where for  $p > p_c$   $\text{Re}(\lambda_1) > 0$ , and all other eigenvalues are in the left-half plane. Although other modes try to dampen the perturbation,  $\lambda_1$  predominates after some time and the perturbation increases monotonically such that it could possibly be attracted to another steady state (which would be stable). In the two-dimensional case, such a bifurcation corresponds to a stable node (with both eigenvalues negative, real) being transformed to a saddle (with one real positive eigenvalue and the other real negative eigenvalue). Hence the name “saddle-node bifurcation”. The steady state changes its characteristic from a stable node to a saddle at this bifurcation point.

(ii) *Hopf bifurcation.* Here again we assume that the steady state is stable for  $p < p_c$  and that a complex conjugate pair of eigenvalues with positive real part exist for  $p > p_c$ . The perturbations are again governed by equations of the form (11.13). Here,  $\lambda_1, \lambda_2$  have positive real parts. This dominates the evolution of the perturbation after some time. The imaginary part ensures us that the perturbation has a periodic dependence on time. The new solution emanating from this point is *usually* a periodic solution. Close to the bifurcation point the linear equations are valid. Should the limit cycle emanating be stable, these linear equations govern the evolution of the system to the small amplitude periodic state close to  $p_c$  (see Marsden and McCracken (1976) for more details).

### 11.2.2 Case Study: First Order Nonisothermal Reaction in a CSTR

So far, we have discussed the different features of nonlinear dynamical systems, as part of a mathematical theory. We now discuss the new features which this theory reveals. These cannot be predicted or explained otherwise. This brings out the importance of nonlinear dynamics. For the sake of clarity and simplicity, we discuss a specific problem, i.e. the first order reaction in a CSTR. This was analysed by Poore (1973).

The equations governing the dynamic evolution are now reproduced for convenience:

$$\dot{C} = -C + Da(1 - C) e^T \quad (11.14a)$$

$$\dot{T} = -(1 + \beta)T + BDa(1 - C) e^T \quad (11.14b)$$

The temperature and conversion at steady state are related by

$$T = BC/(1 + \beta) \quad (11.14c)$$

The stability features of this system are investigated in most text books on chemical reaction engineering by studying the system at steady state. Eliminating  $C$  between the steady state equations

$$0 = -C + Da(1 - C) e^T$$

$$0 = -(1 + \beta)T + BDa(1 - C) e^T$$

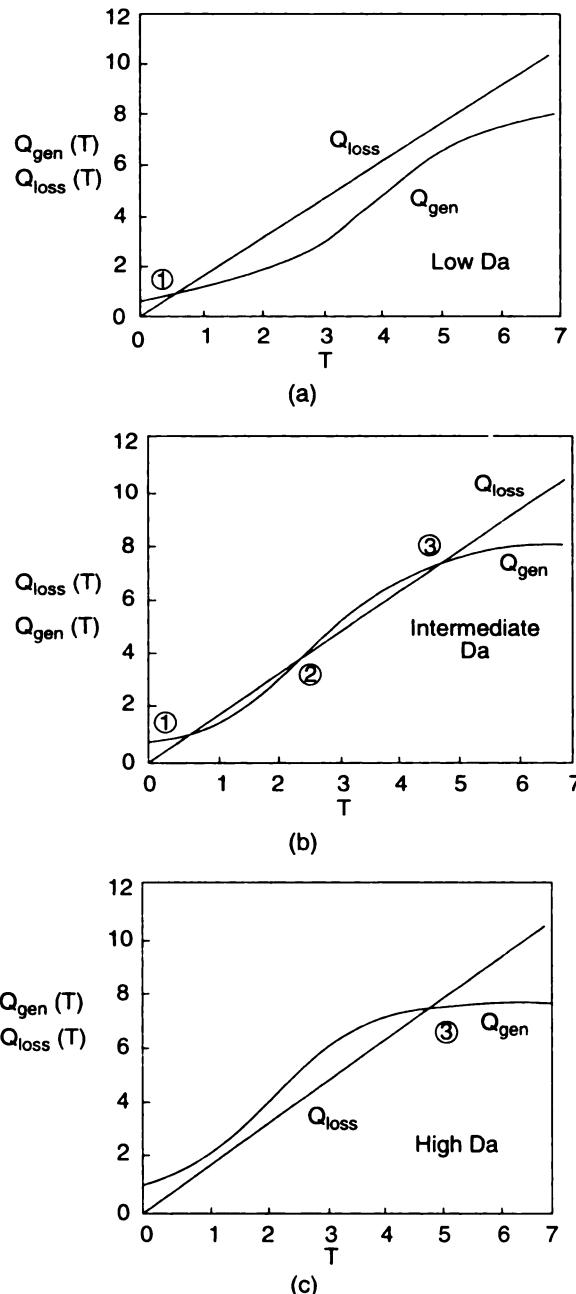
we have

$$(1 + \beta)T = \frac{BDa e^T}{1 + Da e^T}$$

The linear term on the left is identified as the rate of heat loss from the system  $Q_{\text{loss}}$ . This accounts for the heat loss due to the actual flow of fluid, as well as the heat transferred to the coolant. The nonlinear term, on the RHS of the last equation signifies the net rate of heat generation due to the exothermicity of the reaction,  $Q_{\text{gen}}$ .

The number of solutions is determined by the number of intersections of these two curves. For a fixed  $B, \beta$  we depict the effect of  $Da$  on  $Q_{\text{loss}}$  and  $Q_{\text{gen}}$ . For very high, and low values of  $Da$ ,

the curves intersect once on the upper and lower temperature branches respectively and the system possesses a unique steady state. For intermediate  $Da$ , the curves can intersect thrice as shown in Fig. 11.7(a)–11.7(c), and the system exhibits multiple steady states.



**Fig. 11.7** Heat loss, heat generation curves: (a and c) unique steady state for all  $Da$ ; (b) multiple steady states for some  $Da$ .

The stability of this system is studied in the geometric approach by considering a perturbation which results in an increase in  $T$ . When the system is at the state denoted by 1 and subject to this perturbation, we observe that  $Q_{\text{gen}}$  is lower than  $Q_{\text{loss}}$ . Since the heat loss is higher than the heat generated here,  $T$  decreases and the system reverts to the original state at 1. Similarly, for a perturbation which results in a decrease in temperature  $T$  from the steady state value,  $Q_{\text{gen}}$  exceeds  $Q_{\text{loss}}$ . This again causes the system to revert to the steady state, thereby confirming its stability.

It can be argued similarly that the state at 2 is unstable and that at 3 is stable. This analysis predicts the intermediate states to be unstable and the upper and lower states to be stable. This is generalised to yield the “slope condition” for stability as

$$\frac{dQ_{\text{gen}}(T)}{dT} < \frac{dQ_{\text{loss}}(T)}{dT}$$

An implicit assumption made in the analysis so far is that  $C, T$  are always related by the steady state equations (since we eliminated  $C$  using these). This implies that  $C, T$  at every instant of time during the dynamic evolution of perturbation satisfies (11.14c). In general, of course, an arbitrary disturbance in a dependent variable (such as  $T$ ) induces a change in the other variables (such as  $C$ ). The evolution of the two variables is dictated by the governing ordinary differential equations (and not by the steady state equation).

The intermediate steady state is unstable to the class of perturbations which satisfies the steady state equation. It is therefore unstable to all classes of perturbations. Similarly, the upper and lower steady states are stable only to the class of perturbations which satisfy (11.14c) at every instant of time. These could be unstable to the general class of disturbances when the evolution is determined by the coupled system. The upper and lower steady state branches can therefore get destabilised. The coupling between the two variables can possibly yield a dynamic instability. This results in the system exhibiting a dynamic behaviour. In fact, such a dynamic instability can occur even when the system has a unique steady state. This dynamic instability can only be predicted by looking at linear-stability since it considers the interaction between the different variables to determine the evolution of the trajectory.

The instability of the intermediate steady state is usually called a static instability, since it can be obtained using only the steady state relationships. Here the system usually evolves from one steady state to another. The instability of the upper and lower steady states, on the other hand, is a dynamic instability. Here we consider the interaction between the different variables; a steady state is destabilised and the resulting state is usually a dynamic state.

For the two-dimensional system, (11.7b) the stability condition is obtained using linear stability analysis as discussed in Section 11.1.2:

$$\text{tr } J < 0, \det J > 0.$$

as the governing characteristic equation is

$$s^2 - \text{tr } Js + \det J = 0$$

The stability boundaries are thus given by: (a)  $\det J = 0$ , and (b)  $\text{tr } J = 0$ . The former corresponds to the static instability criterion and the latter to the dynamic stability condition. We will now show how the former is equivalent to the condition obtained from the steady state analysis.

Consider the steady state equation in  $C$ , obtained by eliminating  $T$ . This can be written as

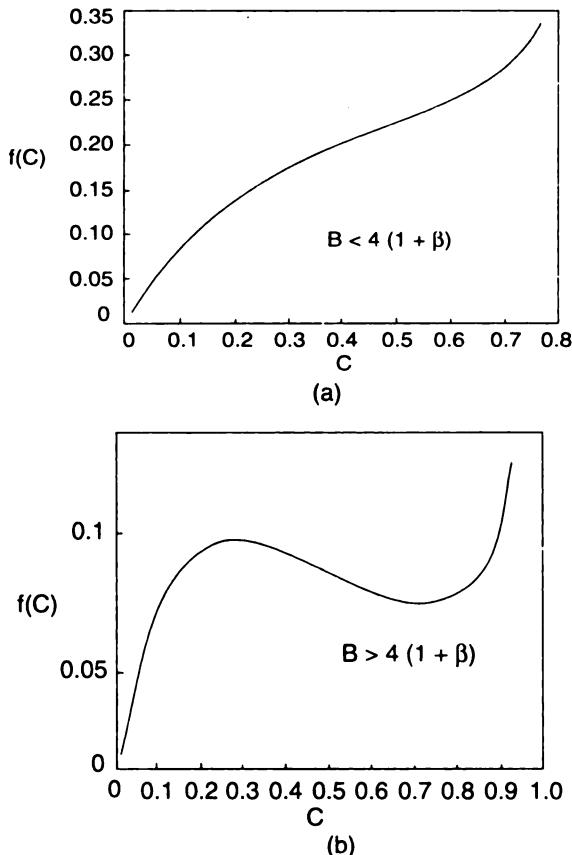
$$Da = \frac{Ce^{BC/(1+\beta)}}{1-C} = f(C)$$

This equation will have multiple solutions for a range of  $Da$  when  $f(C)$  is not monotonic. For a monotonic  $f(C)$ , we have a unique solution for all  $Da$ . The intermediate steady state exists only when we have three solutions. This is obtained when  $f'(C)$  changes sign (see Fig. 11.8). Setting

$$f'(C) = 0$$

we get

$$BC^2 - BC + (1 + \beta) = 0$$



**Fig. 11.8** Variation of  $f(C)$  with  $C$ : (a) monotonic; (b) nonmonotonic.

This equation has two roots in  $(0, 1)$  for  $B > 4(1 + \beta)$ . Consequently, we have a unique steady state for  $B < 4(1 + \beta)$ , and the criterion for static instability is  $B > 4(1 + \beta)$ .

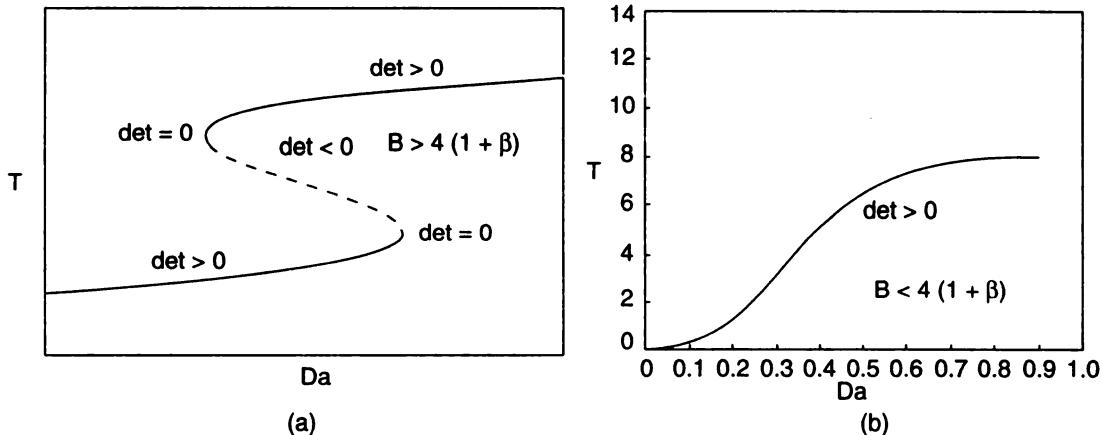
The Jacobian matrix  $J$  at the steady state is given by

$$\begin{bmatrix} -1/(1-C) & C \\ BC/(1-C) & -(1+\beta) + BC \end{bmatrix}$$

So,

$$\det J = BC^2 - BC + (1 + \beta)$$

The determinant condition is thus identical with the static stability criterion. The variation of  $\det J$  along the bifurcation diagrams is illustrated schematically in Fig. 11.9. The intermediate steady state is unstable since the “det condition” is violated. The upper and/or lower branches of the steady state curve of Fig. 11.9a, and the steady state branch of Fig. 11.9b can be rendered unstable only by the trace condition.



**Fig. 11.9** Variation of determinant of Jacobian matrix along bifurcation diagrams depicting: (a) multiple; (b) unique steady states.

The Hopf bifurcation helps us understand the birth of a limit cycle. When a system is in a limit cycle, every state variable characterising the system shows a periodic variation. The period of the variation is a constant for fixed values of parameters. Near the bifurcation point the frequency of oscillation is determined approximately by the imaginary part  $\omega$  of the complex eigenvalue ( $2\pi/\omega$ ). In the phase-plane such a behaviour appears like a closed curve. The two characteristics of such a behaviour are the amplitude of the oscillations and its time period. Both these characteristics vary as we vary the parameter beyond the bifurcation point  $p_c$ .

The limit cycle emerging at a Hopf bifurcation point may be stable or unstable, depending on the nature of the bifurcation. The stable limit cycle has the property of attracting points in the phase plane in its neighbourhood similar to a stable steady state. Hence, a direct numerical integration of the equations will yield the periodic solution as long as the initial condition lies in its basin of attraction. This limit cycle is also an attractor and depicts the long-time behaviour of the system, when the initial condition is chosen in its basin of attraction. An unstable limit cycle, on the other hand, is a repeller and it cannot be obtained by direct integration. We discuss a numerical technique called the shooting method to explain how it is obtained (see Appendix at the end of the chapter). The stability of a limit cycle is determined using the Floquet theory. This theory is based on systems of linear equations with periodic coefficients and is discussed in the Appendix to Chapter 12.

In a bifurcation diagram, we represent the dependence of the solution or a system state on a parameter. It is a two-dimensional plot. The steady state value of a variable versus the bifurcation parameter is usually plotted. On this curve a steady state is depicted as a point and the maximum or the minimum of the limit cycle is plotted when it exists at a particular parameter value. The usual sign convention is to represent stable steady state branches by solid lines, unstable steady state branches by dashed lines, the maxima of stable limit cycles by filled circles and those of unstable

limit cycles by empty circles. The phase-plane representation depicts all co-existing states, stable as well as unstable, for a fixed parameter value. It represents a cross-section of the bifurcation diagram at a given parameter value.

The discussion so far concerns dissipative systems. In such systems energy gets dissipated. This concept can best be illustrated by considering the phase-plane portrait. Consider the set of initial conditions denoted by a closed region  $A_0$ . For a conservative system, the trajectories emanating from this set at any time  $t$  will also span the same area  $A_0$ . This is true for all time instants. In a dissipative system, on the other hand, as time  $t$  progresses, the area  $A_0$  keeps contracting and the system would terminate on an attractor. Besides this, for most physical systems the state variables cannot become unbounded and must remain in a confined region in phase-space. In this text we restrict our attention to dissipative systems only, since they are representative of realistic engineering systems.

Consider the two-dimensional dynamical system

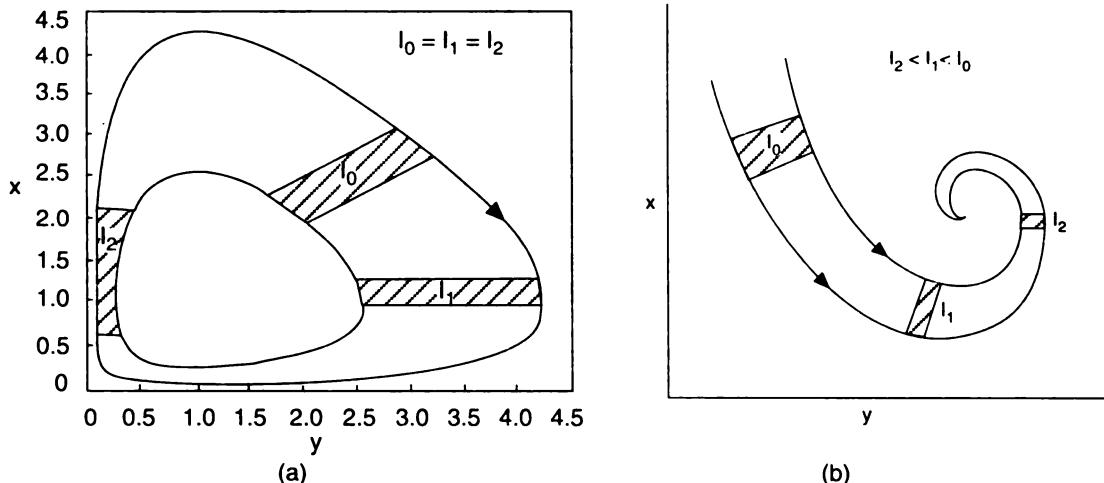
$$\dot{x} = y, \quad \dot{y} = -k^2x$$

This system of linear equations models the motion of a frictionless pendulum. The evolution of the system from a point on this phase-plane is given by the trajectory emanating from this point (see May, 1976). Consider the set of all initial conditions in the region  $I_0$  (Fig. 11.10a). At a later instant of time  $t_1$ , the system states starting from initial conditions in  $I_0$  spans the region  $I_1$ . Similarly,  $I_2$  represents the states of the system starting from points in  $I_0$  at  $t_2$ . For the system shown, the areas  $I_0, I_1, I_2$  are all equal. This is true no matter which portion of the phase-plane we choose as  $I_0$ . The above system is therefore conservative.

Consider now the nonlinear system

$$\dot{x} = x - xy, \quad \dot{y} = -y + xy - y^2$$

The presence of the ' $y^2$ ' term renders the system dissipative. Now the set of initial conditions  $I_0$  evolve as depicted in Fig. 11.10(b). The system states starting from  $I_0$  at two later instants of time  $t_1 < t_2$  are shown in Fig. 11.10. The areas  $I_1, I_2$ , satisfy  $I_2 < I_1 < I_0$ . Eventually, all the trajectories terminate on the steady state as shown in Fig. 11.10(b). This is clearly a state of zero area. Areas in phase-space thus get contracted. This is a characteristic of a dissipative system.

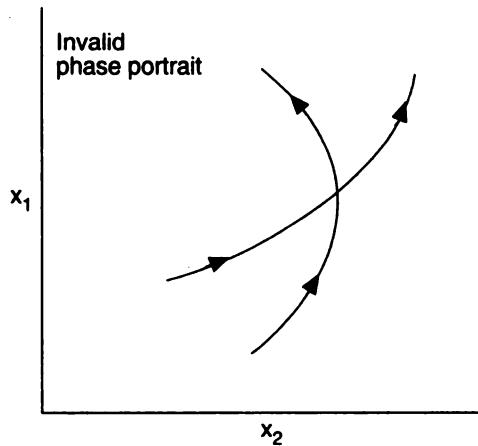


**Fig. 11.10** Phase-plane of: (a) conservative systems; (b) dissipative systems.

A one-dimensional dynamical system can exhibit only steady state behaviour. In particular, it cannot show oscillations or limit cycles as seen earlier. The two-dimensional system, on the other hand, can exhibit multiple steady states as well as limit cycles. Limit cycles may co-exist with each other and steady states, as we will see in the examples that follow. It is not possible to have any other kind of dynamic behaviour in such a system. This is precluded by the fact that through each point a unique trajectory passes. The slope of a trajectory at a point in phase-plane is uniquely determined by

$$\frac{dx_1}{dx_2} = \frac{f_1(x_1, x_2)}{f_2(x_1, x_2)} \quad (11.15)$$

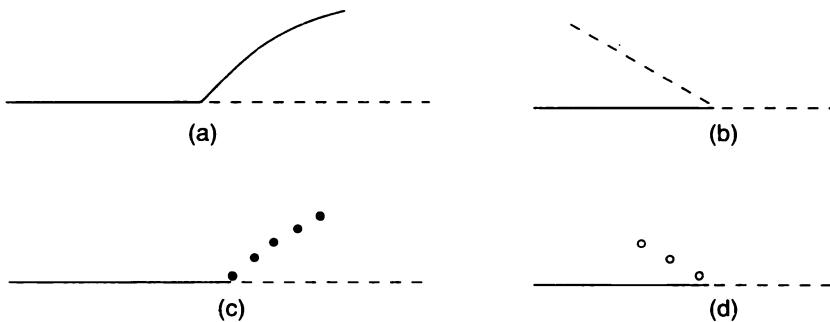
This basic criterion must be satisfied by any system at all points in the phase-plane. If at any point two trajectories were to intersect as shown in Fig. 11.11, then the system would not know in which direction to evolve once it reaches the point. The intersection of the trajectories implies that here the system has two slopes, which is precluded by the fact that the  $f_i$ 's in (11.15) are single-valued functions of  $x_i$ 's. This implies that for a fixed  $x_1, x_2$  the  $f_i$ 's are uniquely determined. The trajectories in phase-plane as mentioned earlier, are similar to the streamlines in fluid flow field. The streamlines depict the actual path traversed by fluid particles in a steady-flow field and no two streamlines can intersect. An idealised sink in fluid flows is analogous to a stable node, while an idealised source is analogous to an unstable node (see Gupta and Gupta, 1984).



**Fig. 11.11** Nonintersection of trajectories in a phase-plane.

Each of the two bifurcations defined above can be either: (a) **sub-critical**, or (b) **supercritical**. This is determined by the nature of the bifurcating solution. If the new solution branch (whether it is steady state or a limit cycle) emerging from the basic solution is stable, we call it a supercritical bifurcation; else it is subcritical (Fig. 11.12). A detailed discussion on determining the nature of the bifurcation can be found in Kuramoto (1984).

The more complex dynamic behaviour like sub-harmonic solutions, quasi-periodic attractors etc. are exhibited only by higher dimensional systems (see Scott, 1991). A minimum of three dimensions are required to see such behaviour. The origin of such behaviour through secondary bifurcations will be the focus of Chapter 12.



**Fig. 11.12** Typical bifurcation diagram indicating sub-critical and super-critical bifurcations:  
 (a) supercritical saddle-node bifurcation; (b) subcritical saddle-node bifurcation;  
 (c) supercritical Hopf bifurcation; (d) subcritical Hopf bifurcation.

A clear understanding of the different concepts that we have introduced can be obtained from the following examples.

**Example 11.2** The system of equations

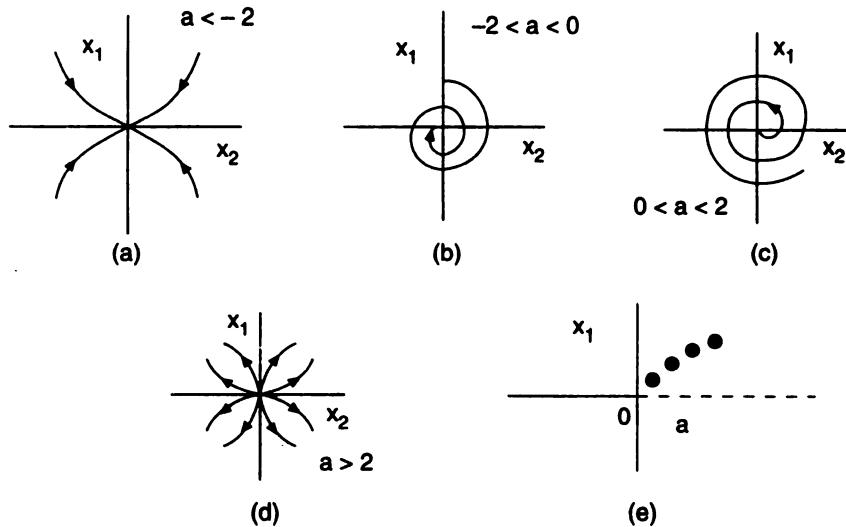
$$\dot{x}_1 = x_2, \quad \dot{x}_2 = a(1 - x_1^2)x_2 - x_1$$

is called the van der Pol oscillator. It arises in the study of vibrations in mechanical engineering. The trivial solution  $x_1 = x_2 = 0$  is the nonoscillatory state, i.e. a steady state. This is valid for all values of the real parameter  $a$ . The stability of this state is determined by the eigenvalues which satisfy the characteristic equation

$$\lambda^2 - a\lambda + 1 = 0$$

Thus from (11.11a), the trivial state is stable for  $a < 0$ . At  $a = 0$ , we have a Hopf bifurcation. This renders the trivial state unstable for  $a > 0$ . In  $-\infty < a < -2$ , the steady state is a stable node, and for  $-2 < a < 0$ , it is a stable focus. For  $0 < a < 2$ , it is an unstable focus; for  $2 < a < \infty$ , it becomes an unstable node. The typical phase-portraits corresponding to each of these four regions are depicted in Fig. 11.13(a)–(d). For  $a < 0$ , all the points in the phase-plane are attracted to the origin. The basin of attraction is therefore the entire phase-plane. For  $a > 0$ , all points are attracted to the limit cycle which is shown as the closed curve. A trajectory starting from a point within this curve or outside this curve is attracted to the limit cycle. Here again the basin of attraction is the entire phase-plane. The concept of the basin of attraction plays a significant role only when we have co-existing stable states. In this example, it is not significant as we have only one stable state (either a steady state or a limit cycle) for all  $a$ . In Fig. 11.13(e), we depict the dependence of the solution  $x_1$  on  $a$ .

**Example 11.3** The Lorenz equations have a special place in the theory of nonlinear equations. They occur in the modelling of the natural convection problem or the Rayleigh Benard problem (see Lorenz, 1963). While working with these equations Lorenz discovered the phenomena of chaos. The chaotic behaviour exhibited by deterministic systems and its features are studied in detail in Chapter 12. Here we restrict ourselves to the preliminary bifurcations and see how a limit



**Fig. 11.13** Steady state characteristics of van der Pol oscillator for different  $a$  values: (a)  $a < -2$ , stable node; (b)  $-2 < a < 0$ , stable focus; (c)  $0 < a < 2$ , unstable focus; (d)  $a > 2$ , unstable node; (e) bifurcation diagram.

cycle is born for this system. The three-dimensional Lorenz system is

$$\frac{dx_1}{dt} = a(x_2 - x_3)$$

$$\frac{dx_2}{dt} = -x_1x_3 + rx_1 - x_2$$

$$\frac{dx_3}{dt} = x_1x_2 - bx_3$$

One of the salient features of this system is that its behaviour can be studied analytically till the first Hopf-bifurcation. The equations have three parameters  $a$ ,  $r$ ,  $b$ . The bifurcation parameter (the one we vary) is chosen to be  $r$ , and we fix the other two parameters at  $a = 10$ ,  $b = 8/3$ . The system has three steady states:

- (a) The trivial state  $x_1 = x_2 = x_3 = 0$
- (b)  $x_1 = +\sqrt{r-1}$ ,  $x_2 = \sqrt{r-1}$ ,  $x_3 = r-1$
- (c)  $x_1 = -\sqrt{r-1}$ ,  $x_2 = -\sqrt{r-1}$ ,  $x_3 = r-1$

The stability of the steady state is determined by the eigenvalues of the Jacobian matrix

$$J = \begin{bmatrix} 0 & a & -a \\ r - x_3^{ss} & -1 & -x_1^{ss} \\ x_2^{ss} & x_1^{ss} & -b \end{bmatrix}$$

For the trivial steady state  $A$ , the stability is governed by the eigenvalues which are the roots of the characteristic equation

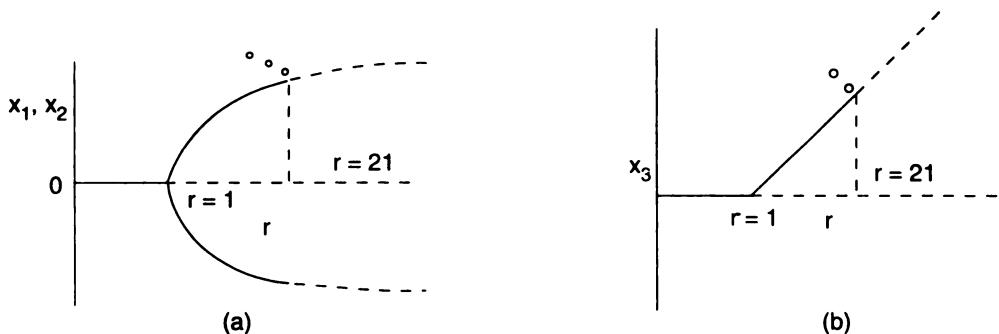
$$(-b - \lambda)(\lambda^2 + (a + 1)\lambda + a(1 - r)) = 0 \quad (11.16)$$

One of the eigenvalues is always  $-b$ , and is therefore negative for  $b > 0$ . For  $r < 1$  and  $a > 0$ , the quadratic factor contributes two eigenvalues in the left-half plane (as there are no sign changes now). At  $r = 1$ , a real eigenvalue crosses the imaginary axis and the trivial steady state becomes unstable for  $r > 1$ . The two other steady states  $B, C$  are nonexistent for  $r < 1$ , as  $x_1, x_2$  are complex. These states therefore do not co-exist with the trivial state for  $r < 1$ . Their stability for  $r > 1$  is determined by the roots of the characteristic equation

$$\lambda^3 + (a + b + 1)\lambda^2 + \lambda(ab + b + r - 1) + 2ab = 0$$

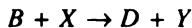
This equation is obtained by substituting the expressions for  $x_i^{ss}$  which indicate the state  $B$ . For  $r > 1$ , all coefficients are positive. Being a cubic equation we are now not assured that all the eigenvalues are in the left-half plane. The additional condition that must be satisfied for the stability comes from (11.13). This implies that this state is stable for  $1 < r < 21$ .

At  $r = 21$ , these steady states become unstable due to a sub-critical Hopf bifurcation. Here a complex-conjugate pair of eigenvalues crosses the imaginary axis. The bifurcation diagram for this system is depicted in Fig. 11.14. For  $0 < r < 1$ , the only admissible state is the trivial state. For  $r > 1$ , the two nontrivial states are stable and they co-exist with the unstable trivial state.



**Fig. 11.14** Steady state bifurcation diagrams of Lorenz system dependence on  $r$  of: (a)  $x_1, x_2$ ; (b)  $x_3$ .

**Example 11.4** The system or network of reactions



was studied by a group in Brussels and called it the *Brusselator* (see Scott, 1991). The reactions occur under well-stirred conditions in a batch reactor. The batch reactor concentrations of  $A, B, D, E$  are maintained constant in the reactor. The reaction rates are therefore independent of their concentrations. The concentrations of  $X, Y$  are modelled by ( $x_1 = x, x_2 = y$ ).

$$\frac{dx_1}{dt} = A - Bx_1 + x_1^2 x_2 - x_1$$

$$\frac{dx_2}{dt} = Bx_1 - x_1^2 x_2$$

The nonlinearity here arises due to the third reaction which is auto-catalytic. The steady state of the system is given by

$$x_1^{ss} = A, \quad x_2^{ss} = B/A$$

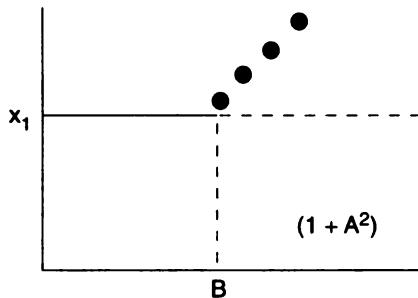
The stability of this unique state is governed by the Jacobian matrix

$$J = \begin{pmatrix} B - 1 & A^2 \\ -B & -A^2 \end{pmatrix}$$

The eigenvalues of the matrix are governed by the characteristic equation

$$\lambda^2 + (A^2 + 1 - B)\lambda + A^2 = 0 \quad (11.17)$$

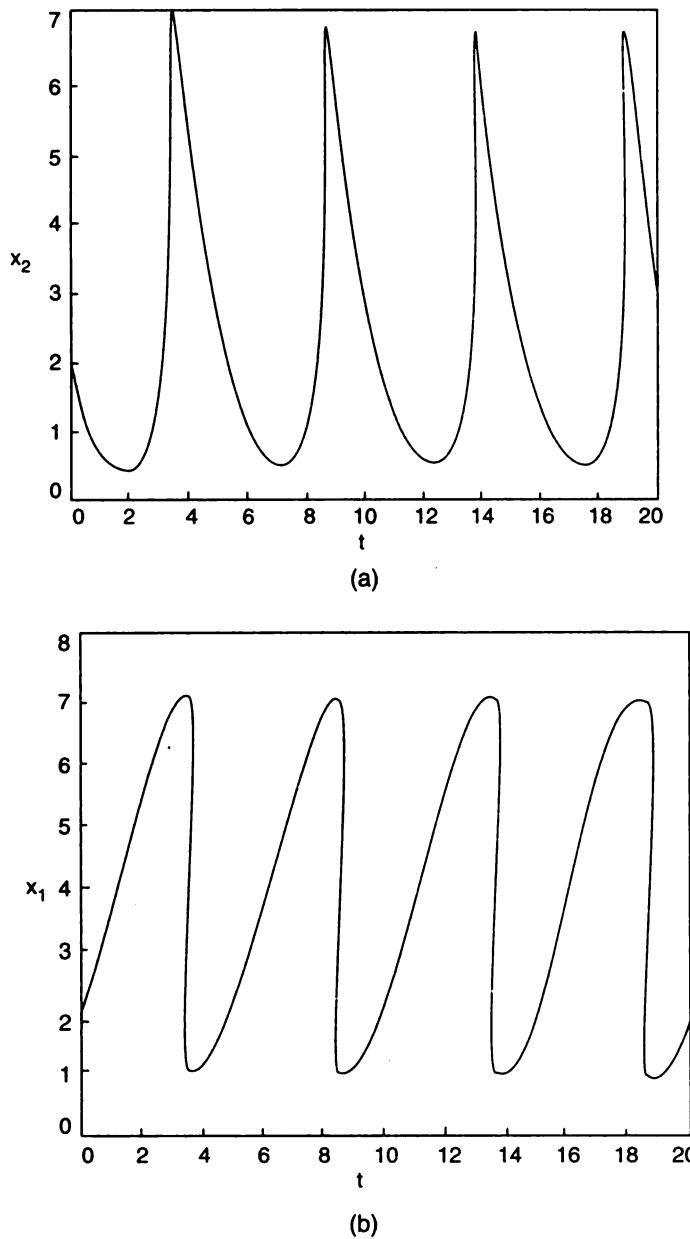
The state is stable for  $B < 1 + A^2$ . It is unstable when this condition is violated. The Hopf bifurcation at  $B = 1 + A^2$  is a super-critical bifurcation and the limit cycle emerging here is stable. For a fixed  $A$  the bifurcation diagram is shown in Fig. 11.15.



**Fig. 11.15** Bifurcation diagram of Brusselator. Dependence on  $B$  of  $x_1$  and stability.

The steady state is a stable node for  $B < (1 - A)$  and a stable focus for  $(1 - A)^2 < B < 1 + A^2$ . For  $B > 1 + A^2$ , the steady state is an unstable focus. The typical variation with time in the unstable region is depicted in Fig. 11.16.

In the examples seen so far we were able to determine the steady state explicitly. This has enabled us to obtain the stability directly in terms of the eigenvalues or the characteristic equation. These examples help in understanding the applications of the various concepts developed so far. In most nonlinear problems, however, this cannot be expected. In Chapter 10 we saw the examples of the cubic autocatalysis reaction and the first order nonisothermal reaction in a CSTR. In these two examples, the steady states have to be obtained numerically for fixed values of the parameters. The steady state dependence on the bifurcation parameter can be obtained from the homotopy continuation methods. The stability of the different branches or segments is determined by the eigenvalues of the Jacobian matrix at each state and are computed numerically. More examples of such systems can be found in Murray (1977).



**Fig. 11.16** Dynamic behaviour of Brusselator: (a)  $x_2$ ; (b)  $x_1$ ,  $B > 1 + A^2$

**Example 11.5** The steady state bifurcation diagram and stability features of the Salmikov model (see Scott, 1991) will be discussed. These are governed by the equations

$$\dot{x}_1 = \mu - kx_1 \exp [x_2/(1 + \varepsilon x_2)]$$

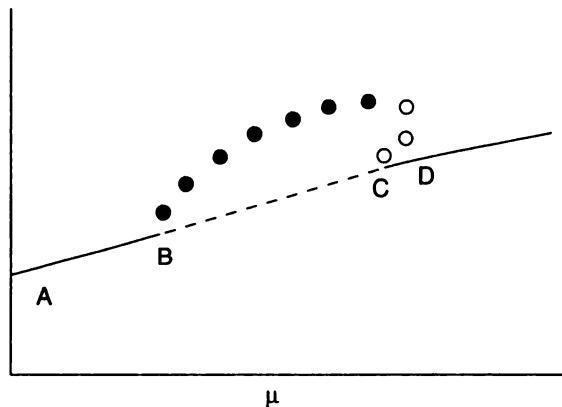
$$\dot{x}_2 = x_1 \exp [x_2/(1 + \varepsilon x_2)] - x_2$$

The stability of the steady states is determined by the eigenvalues of the Jacobian matrix

$$A = \begin{bmatrix} -ka & -kx_1 a / (1 + \varepsilon x_2)^2 \\ a & x_1 a / (1 + \varepsilon x_2)^2 - 1 \end{bmatrix}$$

where  $a = e^{x_2 / (1 + \varepsilon x_2)}$ .

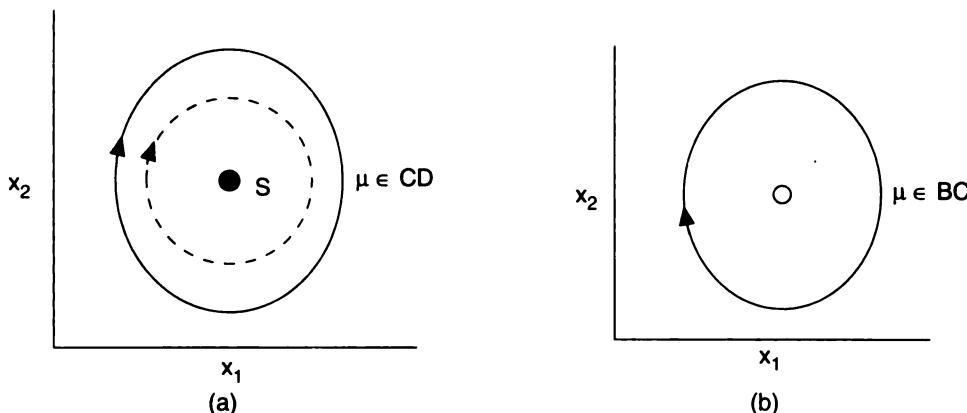
The steady state branch (Fig. 11.17) can be computed numerically by using the continuation method described in Chapter 10. The stability of the steady state at a particular value of the



**Fig. 11.17** Bifurcation diagram of the Salmikov model system showing super-critical Hopf bifurcation at  $B$  and sub-critical Hopf bifurcation at  $C$  ( $k = .001$ ,  $\varepsilon = .21$ ).

bifurcation parameter  $\mu$  is determined by the eigenvalues of the Jacobian matrix  $A$ . These eigenvalues depend on it through  $x_1, x_2$ . For the parameter range in  $(A, B)$ , we have a unique stable steady state and the system converges to this steady state for all initial conditions. In the interval  $(B, C)$ , the only steady state of the system is rendered unstable and is surrounded by a stable limit cycle. This is born out of a supercritical Hopf bifurcation at  $B$ , where a small amplitude periodic state is born. The amplitude of this periodic state increases as we go past  $B$ . The steady state branch regains its stability at  $C$  where a complex-conjugate pair of eigenvalues crosses the imaginary axis from the right-half plane to the left-half plane and the steady state regains its stability. In the interval  $(C, D)$ , we have a stable steady state co-existing with an unstable limit cycle born out of a subcritical bifurcation at  $C$  and with the stable limit cycle born from the super-critical bifurcation at  $B$ . The phase-plane diagram is shown for a typical  $\mu$  in  $(C, D)$  in Fig. 11.18(a). Here all points within the unstable limit cycle shown by the dashed curve terminate on the steady state  $S$ , and all points outside the dashed curve terminate on the stable limit cycle shown as a solid curve. Points inside (outside) the unstable limit cycle are in the basin of attraction of the steady state (stable limit cycle). The boundary of the basin of attraction is the unstable limit cycle and the trajectories are very sensitive across this locus. At the point  $C$  the unstable limit cycle collides with the stable steady state as we vary the parameter to the left, and this renders the steady state unstable. A schematic phase-plane portrait in  $(B, C)$  is shown in Fig. 11.18(b).

At the point  $D$  the two limit cycles—stable and unstable—collide and disappear, and so, for values of  $\mu$  to the right of  $D$ , we have a unique steady state. At  $D$  we have a saddle-node bifurcation



**Fig. 11.18** Phase-plane portraits at different values of bifurcation parameter for  $k = .001$ ,  $\varepsilon = .21$ :  
 (a)  $\mu \in CD$ ; (b)  $\mu \in BC$ .

of a limit cycle (*not a steady state* which is what we have seen so far). We will discuss this in detail in Chapter 12 when we deal with the stability of a limit cycle.

If the parameter  $\mu$  were to be in the interval  $(C, D)$  and if it were to be suddenly changed to slightly less than  $C$  and in  $(B, C)$  the system will show large amplitude oscillations. The sudden appearance of these oscillations can have a detrimental effect on system performance. For example, in a mechanical system it can cause fatigue and failure. These large amplitude oscillations are called hard oscillations. The small amplitude oscillations which arise when we gradually increase the parameter beyond  $B$  in contrast are called soft-mode oscillations. It must be remembered that these oscillations arise due to certain feedback mechanisms generated by the nonlinear interactions in the system and are self-sustained, i.e. never die out.

Many different bifurcation diagrams are possible for a fixed system (as we vary the parameters). These have to be obtained numerically. We refer the interested reader to Scott (1991) for an excellent review of the subject and a discussion on different chemical systems.

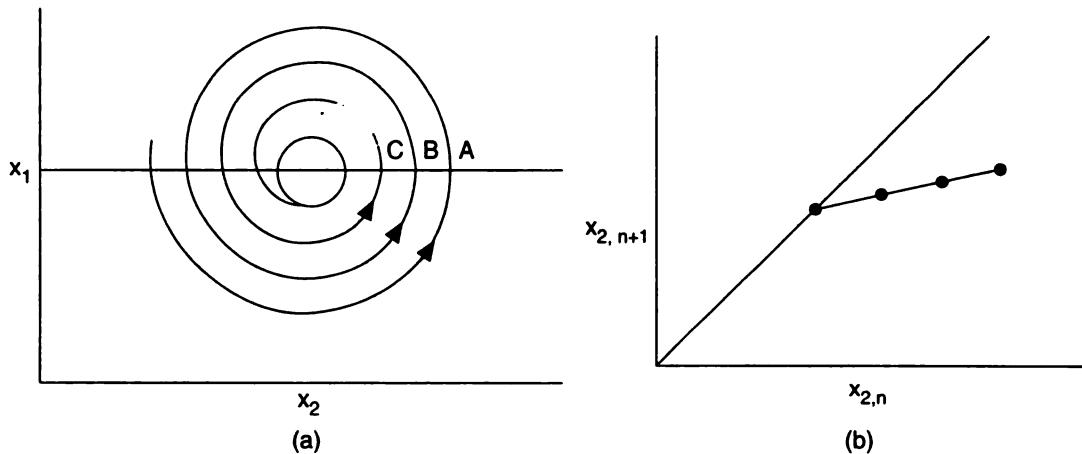
### 11.2.3 Poincare Maps

A dynamic evolution system can also be represented as a map. This idea is easily understood by the concept of the Poincare map or Poincare section. In the phase-plane of a two-dimensional system, a section is chosen and the intersections of a trajectory with it are determined. In Fig. 11.19, the section chosen is the line  $x_1 = \text{constant}$ . The points at which the trajectory intersects the section are of two kinds: (i) where  $\dot{x}_1 > 0$ , and (ii) where  $\dot{x}_1 < 0$ . Let us look at only those points where  $x_1 = \text{constant}$  and  $\dot{x}_1 > 0$ . The direction of intersection of the trajectory is the same at these points, i.e. upwards. The  $x_2$  coordinate of the first point of intersection A uniquely determines the  $x_2$ -coordinate of the second point of intersection B, and so on (Fig. 11.19). This is represented as

$$x_2(B) = f(x_2(A)), \quad x_2(C) = f(x_2(B))$$

Starting from any one point on the section, we generate a sequence of points. From a given point we can also uniquely determine the previous point of intersection as well as the subsequent point of intersection of the trajectory with the section. This implies that the one-dimensional map  $f(x)$  is invertible. The successive points of intersection are obtained from the one-dimensional map

$$x_{n+1} = f(x_n) \tag{11.18}$$



**Fig. 11.19** (a) Construction of Poincare map from trajectory of a system; (b) depiction of Poincare map.

The map can be obtained graphically by plotting the successive  $x_2$  coordinates of the sequence from a trajectory. Such a map is called a Poincare map. By choosing different sections we generate different maps  $f$ . The intersection of the map with the bisectrix or the  $45^\circ$ -line is the fixed point of the map. A trajectory emanating from this point is mapped back onto it by the map. Such a fixed point would determine the intersection of a limit cycle with the section. A stable limit cycle would generate a stable fixed point and an unstable one an unstable fixed point. The unidirectional motion of the trajectory ensures us that each point of intersection is obtained uniquely from the previous point. In other words, the map is invertible. A two-dimensional dynamical system is therefore equivalent to a one-dimensional invertible map.

The concept of the Poincare maps is useful in studying higher dimensional systems. A three-dimensional dynamical system is represented by a two-dimensional invertible map. Since it is difficult to visualise phase-plane portraits in higher dimensional systems, these lower dimensional maps help in visualising the trajectories effectively. We will discuss the role of maps in more detail in Chapter 12, where we will be studying the other kinds of complex dynamic behaviour due to secondary bifurcations. For the present we restrict ourselves to the linear stability and bifurcation characteristics of fixed points of maps.

### 11.3 MAPS

The Poincare map shows the connection between maps and dynamical systems. Besides, these maps occupy a place of importance by themselves as they occur in modelling many systems. We now discuss the dynamic behaviour of these systems. The basic state of the map is the fixed point. This is equivalent to the steady state of a dynamical system. Proceeding along the lines of the dynamical systems analysis, we address the following two issues:

1. Obtaining the fixed point, and determining its stability characteristics.
2. Determining the nature of the solution emanating from the fixed point at a bifurcation point when it is rendered unstable. (This is obtained from a linear stability analysis.)

In this chapter we restrict our discussion to the primary bifurcations. These are bifurcations from the basic state, i.e. the fixed point of the map. In Chapter 12, we will discuss the role of secondary bifurcations or further bifurcations which occur from the bifurcated solutions and yield more complex behaviour.

### 11.3.1 Linear Stability of Maps

We follow the same treatment for maps as we did for continuous dynamical systems. We first discuss the mathematical criterion which helps determine the stability of a fixed point of a map. We then discuss how a fixed point of a map can get destabilised when we vary an intrinsic parameter of the system.

Consider the  $k$ -dimensional map

$$x_{n+1} = F(x_n) \quad (11.19)$$

Here  $x$  is a  $k$ -dimensional vector.

The fixed point of a map can be obtained using the Newton-Raphson method as discussed in Chapter 10. Let  $x^*$  be a fixed point. To determine its stability we consider the evolution of a nearby point  $x_1$  by the  $k$ -dimensional map  $F$ .

$$x_2 = F(x_1) \quad (11.20a)$$

$$x^* = F(x^*) \quad (11.20b)$$

Subtracting, we have

$$x_2 - x^* = F(x_1) - F(x^*)$$

For  $x_1$  close to  $x^*$ , we expand  $F$  in a Taylor series around  $x^*$ . Retaining only the linear terms, we obtain

$$\hat{x}_2 = F'(x^*)\hat{x}_1 \quad (11.21)$$

where  $\hat{x}_i = x_i - x^*$ . Taking the norms on both sides of (11.21), we have

$$\|\hat{x}_2\| = \|F'(x^*)\| \|\hat{x}_1\|$$

Clearly, if  $\|F'(x^*)\| \leq 1$ , the iterates generated by (11.20a) will converge to  $x^*$ . Since the norm of a matrix is related to its eigenvalues (Noble and Daniel, 1977), we have the stability condition

$$|\lambda_i| < 1$$

for all  $i \in 1, n$ .

For the case of the one-dimensional map, this reduces to  $|f'(x^*)| < 1$  at the fixed point. A fixed point of a map is therefore stable if all the eigenvalues of the Jacobian at the fixed point lie within the unit circle (see Iooss (1979) and Nicolis (1986)).

The stability condition obtained is again a local stability condition. If even one eigenvalue exceeds unity, we are assured of an unstable fixed point and if all eigenvalues are within the unit circle, points "close by" are attracted towards it. This condition obtained is again a necessary condition for stability, and its violation is a sufficient one for instability. At this stage it is beneficial to recall the discussion on maps in Chapter 9. We saw that iterates close to the intermediate fixed point  $x_2^*$  diverge from it (Fig. 9.8). At this fixed point  $|f'(x_2^*)| > 1$ , and is unstable.

### 11.3.2 Bifurcations in Maps

When we studied dynamical systems earlier, we investigated the stability of the steady state in

terms of its eigenvalues. We saw how the steady state and its stability depended upon parameter values and discussed the evaluation of bifurcation points. We follow the same lines here in the study of maps. We discuss the dependence of a fixed point and its stability characteristics upon a parameter. This gives insight into the different kinds of bifurcation a map can exhibit when a parameter is varied causing the eigenvalues to move across the unit circle (see Iooss, 1979).

A map can have many parameters. To determine the stability characteristics of the map, all the parameters except one are held constant. The parameter which is varied and whose effect on the system we seek is called the bifurcation parameter. A systematic approach to studying the behaviour of the map is to obtain a fixed point and study its variation as the parameter is varied using continuation methods. The fixed point can be obtained numerically using the Newton-Raphson method discussed in Chapter 10. The stability of the fixed point is determined by the eigenvalues. The loss of stability of a fixed point can occur in three possible ways.

**(i) Saddle-node bifurcation.** Here as we vary the bifurcation parameter, a real eigenvalue leaves the unit circle through +1. At this point another fixed point branches out of the main branch of solutions.

**(ii) Period-doubling bifurcation.** This arises when a real eigenvalue leaves the unit circle through -1. At this point a period-two solution branches out of the main branch. This solution is not a fixed point but a state which oscillates between two points  $x_1^*, x_2^*$ . The map, maps the first point  $x_1^*$  into the second  $x_2^*$  and the second point back into the first, i.e.

$$x_2^* = F(x_1^*) \quad (11.22a)$$

$$x_1^* = F(x_2^*) \quad (11.22b)$$

The long-time behaviour of the sequence generated *if the period-two solution is stable* is  $\dots x_1^*, x_2^*, x_1^*, x_2^* \dots$ . The two states of the period-two solution are fixed points of the composite map  $F^2(x)$ . This follows from (11.22a) and (11.22b). Hence

$$x_1^* = F(F(x_1^*)) \quad (11.22c)$$

$$x_2^* = F(F(x_2^*)) \quad (11.22d)$$

**(iii) Torus bifurcations.** A third way in which the eigenvalues can cross the imaginary axis is as a complex-conjugate pair. This yields a quasi-periodic solution, which is characterised by the motion on a torus. We will discuss more about this bifurcation in Chapter 12.

For the present we will concentrate on only the first two ways in which a fixed point can become unstable. These methods are analogous to the two mechanisms by which a steady state can become unstable through primary bifurcations in a dynamical system, i.e. saddle-node bifurcation and the Hopf bifurcation. We now show intuitively how a period-two state can arise across a period doubling bifurcation point. We approximate the nonlinear map (11.19) by the linear map (11.21) near a fixed point. The linearisation is valid for points close to the fixed point. Using the matrix  $L$  to represent the Jacobian matrix  $F'(x^*)$  in (11.21), we have

$$\tilde{x}_{n+1} = L(x^*)\tilde{x}_n \quad (11.23)$$

$\tilde{x}_n$  (the perturbation from the fixed point  $x^*$ ) can be resolved (see Chapter 4) in terms of the basis vectors chosen as the eigenvectors ( $u^i$ ) of  $L$ , as

$$\tilde{x}_n = \sum_{i=1}^k c_i u^i$$

Then we have

$$\begin{aligned}\tilde{x}_{n+1} &= L \sum_{i=1}^k c_i u^i = \sum_{i=1}^k c_i L u^i = \sum_{i=1}^k c_i \lambda_i u^i \\ \tilde{x}_{n+2} &= L \tilde{x}_{n+1} = \sum_{i=1}^k c_i \lambda_i^2 u^i\end{aligned}$$

This yields

$$\tilde{x}_{n+m} = \sum_{i=1}^k c_i \lambda_i^m u^i$$

as  $m \rightarrow \infty$   $\|\tilde{x}_{n+m}\|$  is determined solely by the magnitude of the eigenvalues  $\lambda_i$ . If all eigenvalues are within the unit circle, then clearly,  $\|\tilde{x}_{n+m}\| \rightarrow 0$ . When the critical eigenvalue leaves the unit circle from +1, the perturbation grows in the direction of the eigenvector corresponding to it. It is likely to converge on another fixed point. This is similar to the saddle-node bifurcation or steady state bifurcation of dynamical systems. There is one direction in which perturbations grow and in all other directions they decay. (For one-dimensional maps, the eigenvalue multiplies the perturbation and hence must be less than unity for stability, while for a dynamical system the eigenvalue occurs in the exponential term and must therefore lie in the left-half plane.)

The second possibility corresponds to the critical eigenvalue  $\lambda_i = -1$ . Here for  $m$  even (odd), the perturbation grows in the direction of (or opposite of) the eigenvector (if  $c_i > 0$ ). The perturbation now oscillates in the direction of the eigenvector and the solution obtained is one that repeats itself alternately. This gives rise to the period-doubling solution or an oscillatory solution. This is analogous to the situation where we have a pair of eigenvalues on the imaginary axis for a dynamical system. The perturbations now grow periodically.

The nature of the eigenvalue gives some insight into the growth of the new solution as in dynamical systems. It predicts the point where the new solution is born. However, this new solution can co-exist along with the original or basic solution branch and may be stable or unstable. This decides whether the bifurcating solution is stable or unstable, i.e. whether it is super-critical or sub-critical. These aspects can be determined only by analysing the higher order terms (beyond the linear term).

The period-two solutions can be determined numerically as fixed points of the composite map ( $F \circ F(x)$ ). This can again be done using a Newton-Raphson technique. The fixed point of  $F(x)$  is also a fixed point of  $F^2(x)$ . We discuss the torus bifurcation in Chapter 12 along with secondary bifurcations.

### **Example 11.6** The one-dimensional logistic map

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

has been studied extensively in the literature. It occurs in the modelling of ecosystems, where  $x_n$  represents the population of a species. The maximum of  $f(x)$  is  $a/4$ . We restrict  $x$  to lie between 0 and 1. For this we impose  $0 < a < 4$ . This assures us that the maximum value of  $f(x)$  is less than 1 and the map  $f$  maps the interval  $(0, 1)$  to  $(0, 1)$ . The fixed points  $x^*$  of the map correspond to the roots

$$x^* = f(x^*) \text{ i.e., } x^* = ax^*(1 - x^*),$$

This yields the two fixed points  $x_1^* = 0$ ,  $x_2^* = 1 - 1/a$ . The trivial-fixed point  $x_1^*$  is independent

of the parameter  $a$ . This is a fixed point for all  $a$  in  $(0, 4)$ . However, it may not be a stable fixed point for all  $a$ . Its stability is determined from

$$\left| f'(x_1^*) \right| < 1$$

This yields  $|a| < 1$ , or in our case  $0 < a < 1$  (see Nicolis, 1986). Hence the iterates in the neighbourhood of  $x_1^* = 0$  will converge to  $x_1^*$  as  $n \rightarrow \infty$ . In this particular case, all points in  $0 < x < 1$  converge to  $x_1^*$  for  $0 < a < 1$ . The trivial state is the only attractor for this range of  $a$ . The point  $a = 1$  is a bifurcation point. Here  $x_1^*$  becomes unstable as an eigenvalue (in fact, the only eigenvalue) leaves the unit circle through  $+1$ . At this bifurcation point, the other fixed point  $x_2^*$  emerges, and the trivial fixed point is rendered unstable. The fixed point  $x_2^*$  is stable for  $a > 1$ . Clearly, it does not exist for  $a < 1$ . The fixed point  $x_2^*$  is stable for

$$\left| f'(x_2^*) \right| < 1 \text{ or } |2 - a| < 1, \text{ i.e. } 1 < a < 3$$

In the interval  $1 < a < 3$ , the two fixed points co-exist. Here,  $x_1^*$  is unstable and  $x_2^*$  is stable. We call the bifurcation at  $a = 1$ , primary bifurcation as it is the first one occurring from the trivial fixed point, the basic state. At  $a = 3$ , the second fixed point, viz.  $x_2^*$ , is rendered unstable as the eigenvalue leaves the unit circle through  $-1$  as  $a$  exceeds 3. Here a period-two solution is born. This corresponds to a dynamic state. We will discuss this state and the properties of the map for  $a > 3$  in detail in Chapter 12. The bifurcations at  $a = 1$  (from the trivial state) and at  $a = 3$  from  $x_2^*$  are both super-critical. These bifurcations render the respective base solutions unstable. The new solutions emanating from these points are stable. So for  $a < 1$ , the trivial fixed point  $x_1^*$  is the only feasible fixed point. For  $1 < a < 3$ , the stable  $x_2^*$  co-exists with the unstable  $x_1^*$ . We depict the bifurcation diagram of this map in Fig. 11.20(d). Figures 11.20(b), 11.20(c) show how iterates starting in  $(0, 1)$  get attracted to the trivial state 0 for  $a < 1$  and to the fixed point  $1 - 1/a$  for  $1 < a < 3$ .

For the one-dimensional map there is only one direction along which perturbations can either grow or decay. Hence the perturbations cannot grow in some direction and decay in another. No fixed point can be a saddle for a one-dimensional map. Therefore, strictly speaking, the bifurcation at  $a = 1$  cannot be called a saddle-node bifurcation as it does not convert a node to a saddle. The bifurcation features of this system are discussed in detail in May (1976).

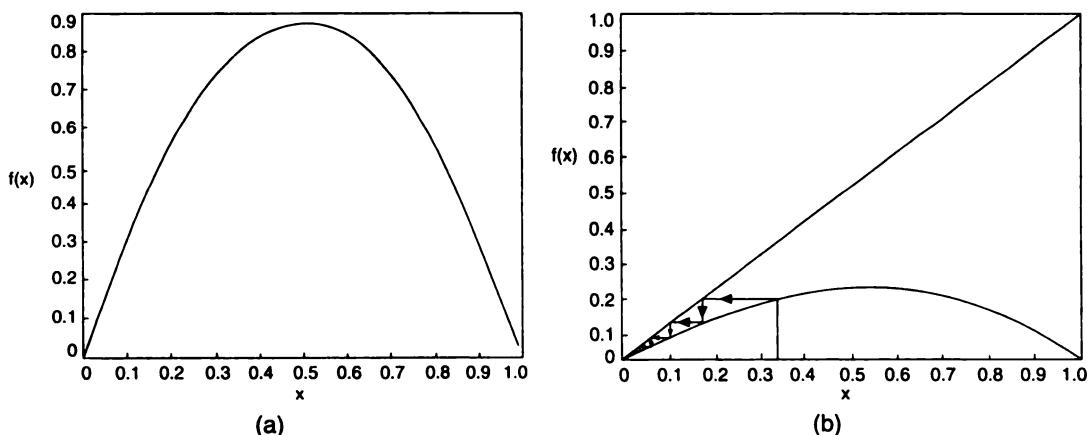
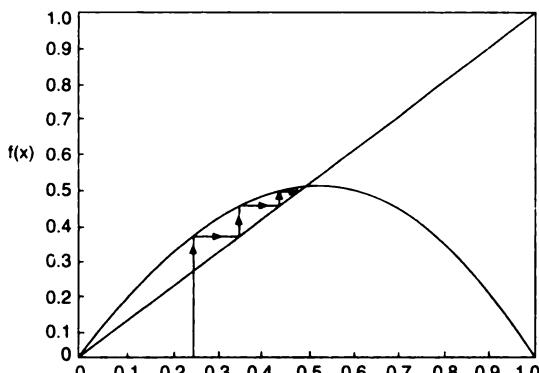
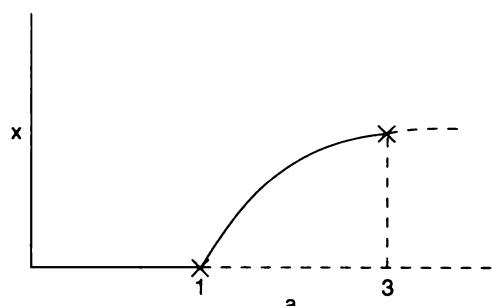


Fig. 11.20 Cont.



(c)



(d)

**Fig. 11.20** Features of logistic map: (a)  $f(x)$  versus  $x$  for  $a < 4$ ; (b) evolution of iterates for  $0 < a < 1$ ; (c) evolution of iterates for  $1 < a < 3$ ; and (d) bifurcation diagram of  $x$  vs.  $a$ .

## Appendix

### Numerical Computation of Unstable Limit Cycle

The unstable limit cycle cannot be obtained by a direct numerical integration using standard routines like the Runge-Kutta as it is a repeller. It has to be obtained using an iterative Newton-Raphson scheme. We explain this technique, called the *shooting method*, to demonstrate the versatility of the Newton-Raphson method in obtaining periodic solutions. This method is an extension of the “shooting method” used in solving two-point boundary-value problems. This extension is rendered possible since we pose the problem of determining the limit cycle as a boundary-value problem.

A limit cycle in any system is a closed curve in a phase plane. Consider a two-dimensional system for the sake of simplicity:

$$\dot{x}_1 = f_1(x_1, x_2) \quad (11.24a)$$

$$\dot{x}_2 = f_2(x_1, x_2) \quad (11.24b)$$

An unstable limit cycle appears as depicted in Fig. 11.21. Take a section of this limit cycle as shown by the line  $x_1 = x_{10}$  (a constant). The constant value must be chosen suitably so that the plane intersects the limit cycle (Scott, 1991). This Poincare section of the limit cycle is represented by the point  $A$ . A point such as  $B$  chosen close to  $A$  evolves, as shown by the trajectory in Fig. 11.21. It intersects the Poincare section, with  $\dot{x}_1 > 0$  at  $C, D$ , etc. The point  $A$  is the only point in the phase plane which is mapped by the evolution equations, onto itself after a suitable period  $T$ . The limit cycle is characterised uniquely by the  $x_2$ -coordinate of  $A$  (the  $x_1$ -coordinate being fixed at  $x_{10}$ ) and the time period  $T$ . The determination of these two quantities uniquely defines the limit cycle of the system.

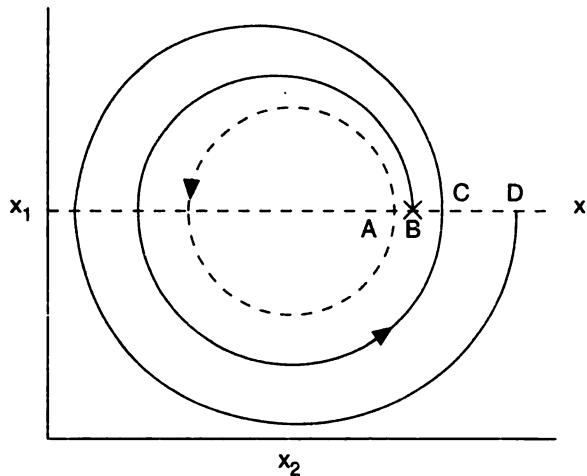


Fig. 11.21 Evolution from close to an unstable limit cycle.

The dynamical system of equations (11.24) can be perceived as a map. For instance, the point  $B$  is mapped by these equations onto point  $C$  after some time. To obtain  $(x_{20}, T)$ , we start with a guess value for  $[x_{20}, T]^t$ . We integrate the equations for a time  $T$  with the initial conditions  $x_{10}$ ,  $x_{20}^{\text{guess}}$ , and check if  $x_1$ ,  $x_2$  obtained after  $T$  equal  $x_{10}$  and  $x_{20}^{\text{guess}}$ . This yields two equations for  $x_{20}$  and  $T$  and we iterate on them using the Newton-Raphson technique to converge on the solution. These equations can be succinctly represented as

$$x_1(x_{10}, x_{20}, T) - x_{10} = 0 \quad (11.25a)$$

$$x_2(x_{10}, x_{20}, T) - x_{20} = 0 \quad (11.25b)$$

Equations (11.25) state that  $x_1$ ,  $x_2$  of the system depend on the initial condition as well as the time period of integration. The unknowns here are  $x_{20}$ ,  $T$ . The values of  $x_1$ ,  $x_2$  obtained by integrating (11.24) clearly depend upon  $x_{10}$ ,  $x_{20}$ ,  $T$ . The Newton-Raphson method can be applied on this as

$$\begin{bmatrix} x_{20} \\ T \end{bmatrix}_{n+1} = \begin{bmatrix} x_{20} \\ T \end{bmatrix}_n - \begin{bmatrix} \frac{\partial x_1}{\partial x_{20}} & \frac{\partial x_1}{\partial T} \\ \frac{\partial x_2}{\partial x_{20}} - 1 & \frac{\partial x_2}{\partial T} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - x_{10} \\ x_2 - x_{20} \end{bmatrix}_n \quad (11.26)$$

The elements of the Jacobian matrix are evaluated by differentiating (11.24) with respect to  $x_{10}$ ,  $x_{20}$ . Then

$$\frac{d}{dt} \left( \frac{\partial x_1}{\partial x_{10}} \right) = \frac{\partial f_1}{\partial x_1} \frac{\partial x_1}{\partial x_{10}} + \frac{\partial f_1}{\partial x_2} \frac{\partial x_2}{\partial x_{10}} \quad (11.27a)$$

$$\frac{d}{dt} \left( \frac{\partial x_1}{\partial x_{20}} \right) = \frac{\partial f_1}{\partial x_1} \frac{\partial x_1}{\partial x_{20}} + \frac{\partial f_1}{\partial x_2} \frac{\partial x_2}{\partial x_{20}} \quad (11.27b)$$

$$\frac{d}{dt} \left( \frac{\partial x_2}{\partial x_{10}} \right) = \frac{\partial f_2}{\partial x_1} \frac{\partial x_1}{\partial x_{10}} + \frac{\partial f_2}{\partial x_2} \frac{\partial x_2}{\partial x_{10}} \quad (11.27c)$$

$$\frac{d}{dt} \left( \frac{\partial x_2}{\partial x_{20}} \right) = \frac{\partial f_2}{\partial x_1} \frac{\partial x_1}{\partial x_{20}} + \frac{\partial f_2}{\partial x_2} \frac{\partial x_2}{\partial x_{20}} \quad (11.27d)$$

$$\frac{dx_1}{dT} = f_1(x_1(T), x_2(T)) \quad (11.27e)$$

$$\frac{dx_2}{dT} = f_2(x_1(T), x_2(T)) \quad (11.27f)$$

The partial derivative in (11.27e) and (11.27f) is the same as the total derivative since we are evaluating these derivatives for a fixed set of initial conditions. The four equations (11.27a)–(11.27d) are linear but with time varying coefficients. The coefficients depend on  $x_1, x_2$  which vary with time as given by (11.24). Hence all six equations—(11.24a) and (11.24b), and (11.27a)–(11.27d)—have to be solved simultaneously, until we obtain convergence. A suitable choice of  $x_{10}$  is necessary to assure the convergence of the scheme. More details of the shooting method can be found in Roberts and Shipman (1972).

## PROBLEMS

- 1.** Find the stability of all the steady states of

$$\dot{x}_1 = 3x_1 - x_1^2 - x_1x_2$$

$$\dot{x}_2 = x_2x_1 - x_2$$

Classify the different steady states as being node or focus and draw a composite phase plane picture.

- 2.** For a steady state, the characteristic equation is given by

$$(i) \quad s^2 + (a - b)s + (ab - 1) = 0$$

$$(ii) \quad s^2 - (a^2 - 4a + 4)s + (a - 6) = 0$$

$$(iii) \quad s^2 - (a^2 - 5a + 4)s + (a^2 - 5a + 6) = 0$$

$$(iv) \quad s^2 + (a + 2)s + (a - 1) = 0$$

$$(v) \quad s^3 + s^2 + 2s + 1 = 0$$

$$(vi) \quad s^2 + (a + 1)s + (b + 1) = 0$$

Determine the parameter space-stable regions and the loci of different bifurcations ( $a, b$  are real).

- 3.** Determine the steady state stability and draw the bifurcation diagrams ( $x, y, z$  vs. ' $a$ ') for

$$(i) \quad \dot{x} = 10(y - x), \quad \dot{y} = -xz + ax - y, \quad \dot{z} = 3xy - 8z$$

$$(ii) \quad \dot{x} = -(y + z), \quad \dot{y} = x + .2y, \quad \dot{z} = .2 + xz - az$$

- 4.** Determine the conditions on the parameters  $a, b$ , where  $a, b > 0$ , for which the following equation has a unique solution:

$$x = a + b e^{-1/x}$$

5.  $\frac{dx}{dt} = -x^3 + 3x^2 - 2x$

Find all possible steady states of the system. Find their stability. What are the initial conditions which will get attracted to each of these states?

6. For what values of the parameters  $a, b$  will the systems described by the characteristic equations be stable?

(i)  $s^2 + (2 + a)s + (a - 3) = 0$

(ii)  $s^2 - (a + 2)s + (a - 3) = 0$

(iii)  $s^3 + s^2 + (a + 2)s + b = 0$

7. When does the equation

$s^3 + as^2 + bs + c = 0$  have purely imaginary roots?

8.  $\frac{dx}{dt} = 3(x - y)$

$$\frac{dy}{dt} = -xz + rx - y$$

$$\frac{dz}{dt} = xy - z$$

Find the steady states and their stability. Draw the bifurcation diagram of  $x$  vs.  $r$ ,  $z$  vs.  $r$ .

9. Determine all possible steady states and their stability. Then

(i)  $\frac{dx_1}{dt} = x_1 - 2x_1x_2$

$$\frac{dx_2}{dt} = -x_2 + x_1x_2$$

(ii)  $\frac{dx_1}{dt} = x_1^2 - x_1x_2 - x_1$

$$\frac{dx_2}{dt} = x_2^2 + x_1x_2 - 2x_2$$

(iii)  $\frac{d^2y_1}{dt^2} = a(1 - y_1^2)\dot{y}_1 - y_1 \quad \text{for } a = 1$

(iv)  $\dot{x} = -y - z, \quad \dot{y} = x + ay, \quad \dot{z} = b + z(x - c), \quad \text{for } a = 1, b = 1, c = 1$

10. Determine the steady state bifurcation diagrams of

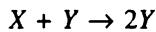
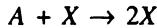
(i)  $\dot{x} = x^3 + x^2 + ax$

(ii)  $\dot{x} = x^2 + ax + 2$

11. For the first order reaction in a CSTR, prove that a necessary condition for limit cycles to be exhibited is that  $\beta > 0$ , or the reactor must be nonadiabatic.

**12.** Consider the quadratic and the cubic characteristic equations respectively. Obtain the necessary and sufficient conditions for stability using the Routh-Hurwitz criterion. Compare these with the criterion presented in the text.

**13.** Consider the following reactions occurring in a batch reactor.



We assume  $A$  is present in excess so that its concentration can be treated as a constant. The equations modelling the batch reactor exhibit temporal oscillations. Verify this by numerical integrations.

**14.** Determine the stability of the different branches of the steady state in Problem 1 (Chapter 10) for both kinetic expressions. Does the system have Hopf-bifurcation points? Can we observe sustained oscillations? Can we obtain oscillations if the kinetics are given by more complex expressions?

**15.** Determine the nature of the steady state along different branches of Problem 2 (Chapter 10). Prove that this system cannot exhibit Hopf bifurcations when the feed is independent of  $X$  for Monod kinetics? Can this system exhibit Hopf bifurcation for Haldane kinetics? If so when can this occur?

**16.** Rework Problems 14 and 15 when the feed contains  $X$  at a concentration of  $X_f$ .

## REFERENCES

- Bhatia, N.P., *Dynamical Systems, Stability Theory and Applications*, Springer-Verlag, Berlin (1967).
- Coddington, E.A. and Levinson, N., *Theory of Ordinary Differential Equations*, McGraw-Hill, New York (1955).
- Gupta, V. and Gupta, S.K., *Fluid Mechanics and Its Applications*, Wiley Eastern, New Delhi (1984).
- Hagedorn, Peter, *Non-linear oscillations*, translated by Stadler W., Clarendon Press, Oxford (1981).
- Haken, H., *Advanced Synergetics: Instability hierarchies of self-organising systems and devices*, Springer-Verlag, Berlin (1983).
- Ince, E.L., *Ordinary Differential Equations*, Dover, New York (1956).
- Ion, G., *Bifurcations of Maps and Applications*, North Holland, Amsterdam (1979).
- Kubicek, M. and Marek, M., *Computational Methods in Bifurcation Theory and Dissipative Structures*, Springer-Verlag, Berlin (1983).
- Kuramoto, Y., *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, Berlin (1984).
- Lorenz, E.N., *Deterministic Non-periodic Flow*, *J. Atmos. Sci.*, **20**, 130 (1963).

- Marsden, J.E. and McCracken, M., The Hopf-bifurcation and Its Applications, Springer-Verlag, New York (1976).
- May, R.M., Simple mathematical models with very complicated dynamics, *Nature*, **261**, 459 (1976).
- Minorsky, N., Non-linear Oscillations, D. Van-Nostrand, New York (1969).
- Murray, J.D., Lectures on Non-linear Differential Equation Models in Biology, Clarendon Press, Oxford (1977).
- Nicolis J.S., Dynamics of Hierarchical Systems, Springer-Verlag, Berlin (1986).
- Noble, B. and Daniel, J.W., Applied Linear Algebra, Prentice-Hall, Englewood Cliffs, New Jersey (1977).
- Poore, A.B., A model problem arising in chemical reactor theory, *Archives of Rational Mechanics and Analysis*, **52**, 358 (1973).
- Porter, B., Stability Criteria for Linear Dynamical Systems, Oliver & Boyd, Edinburgh (1967).
- Roberts, S.M. and Shipman, J.S., Two-point Boundary-value Problems: Shooting methods, American Elsevier Publishing Co., New York (1972).
- Scott, S.K., Chemical Chaos, Clarendon Press, Oxford (1991).
- Verhulst, F., Non-linear Differential Equations and Dynamical Systems, Springer-Verlag, New York (1980).

# 12

## Secondary Bifurcations and Chaos

---

---

In Chapter 11 we saw how the nonlinear interactions in a system can generate spontaneous sustained oscillatory behaviour. This behaviour occurs across critical points of the system. These are generically called bifurcation points and more specifically called Hopf-bifurcation points. In this chapter we will see how more complex time dependent behaviour can arise in nonlinear systems. As discussed in Chapter 11, two-dimensional systems can exhibit time periodic solutions as the most complex behaviour. Other kinds of dynamic states like quasi-periodic solutions and chaotic solutions can be exhibited only by higher dimensional dynamic systems. The transitions giving rise to the new behaviour again occur across bifurcation points. At these points the existing solutions get destabilised. These solutions may not be steady state solutions. Since the bifurcations may occur from limit cycles, we call them secondary bifurcations, as opposed to primary bifurcations which occur from the steady state solution branch. We will be focussing on the different transitions in the context of dynamical systems as well as maps. Before we have a detailed mathematical analysis, let us see some physical situations where such complex dynamic behaviour arises. Gleick (1988) provides an elementary introduction to the theory of chaotic behaviour of systems.

The phenomenon of turbulence is perhaps one of the least understood areas in fluid mechanics. The origin or transition to turbulence and its characteristics are not well understood (Landau and Lifschitz, 1959). Qualitatively, turbulence is characterised by the presence of a wide range of time scales and spatial scales. In this chapter we are concerned primarily with the evolution of time scales. Turbulence is normally considered to be synonymous with randomness or noise. Noise is a high frequency low amplitude signal which corrupts the original signal. Its frequency is not fixed. The phenomena of chaotic behaviour (which we discuss in this chapter) deals with purely deterministic systems. The behaviour is not induced by noise or any random input but by the nonlinear interactions in the system. This is characterised by a low frequency component in the signal. The different variables characterising the system are time dependent. The time series of a “chaotic” signal contains a wide distribution of time scales. This distribution has a low frequency component of oscillations as opposed to noise induced randomness which has a high frequency component. In chemical engineering, many reaction systems have been investigated and are known to exhibit chaotic behaviour. In such systems the inputs to the reactor are constant, yet the outputs from the reactor, viz. conversion, temperature, etc., vary “chaotically”, i.e. they have no fixed frequency. (We will formally define “chaos” later.) This can be analysed by using the Fast Fourier Transform (FFT) which generates the power spectra of the signal.

In this chapter the complex dynamic behaviour that we analyse includes: (a) **large period** or

**sub-harmonic solutions, (b) quasi-periodic solutions, and (c) chaotic solutions** (see Schuster, 1984). We shall primarily focus on two issues:

1. The different ways in which a chaotic or a complex periodic signal can arise.
2. Methods of quantifying and characterising a signal.

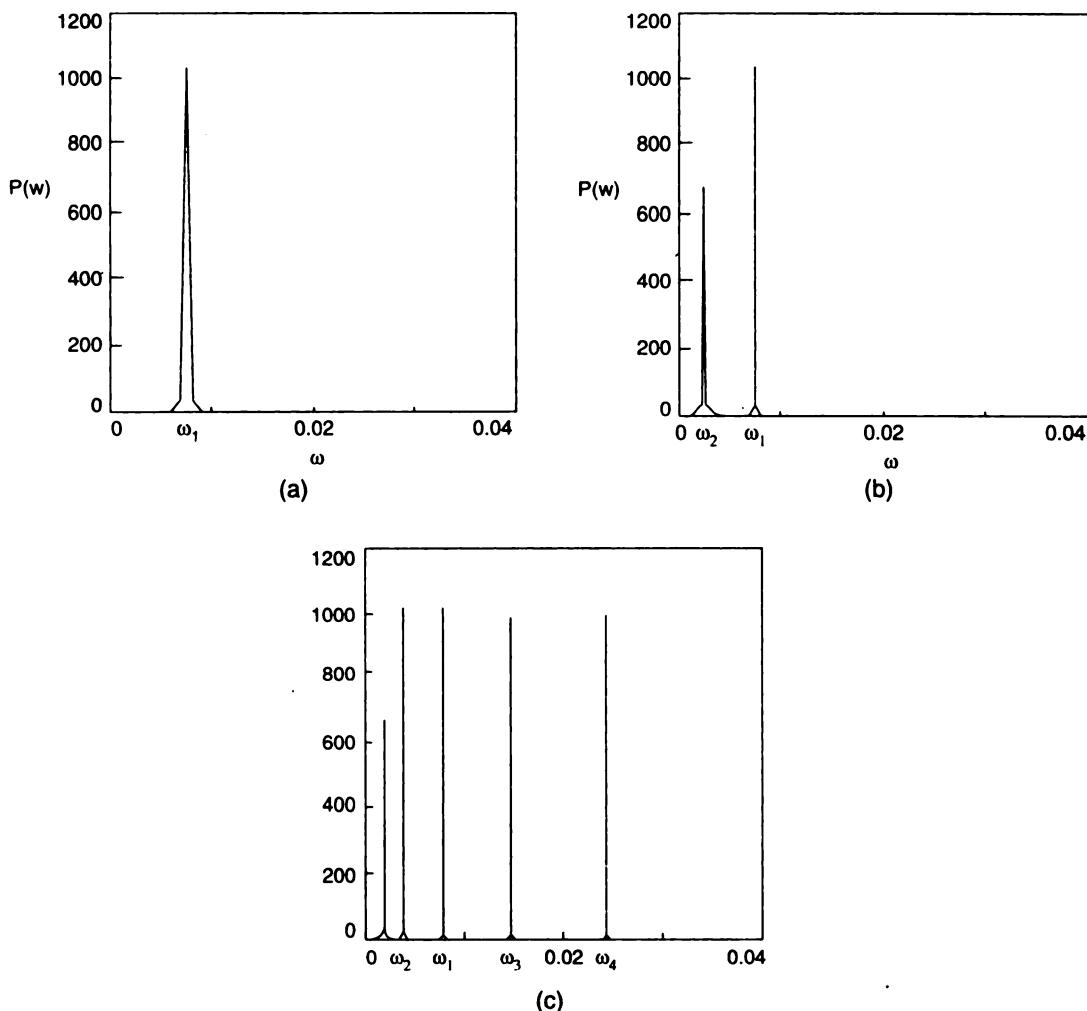
This subject of chaos has received a lot of attention in the past 30 years or so. Many chemical systems have been shown to exhibit the features described here. The theory presented here is sufficiently general and is applicable to dynamical systems in a wide variety of disciplines like sociology, ecology, and physical sciences. We will first discuss the different scenarios or mechanisms which can give rise to chaotic behaviour.

## 12.1 LANDAU-HOPF SCENARIO

The Landau-Hopf scenario was conjectured by Landau and is now widely recognised as being incorrect. It is not found in any real physical system. It is a mathematical hypothesis or a conjecture with no physical basis. This scenario is described here primarily because it was the first attempt made to explain hypothetically the origin of a wide spectrum of time scales in a system. The scenario was conceived to model the onset of turbulence.

Consider a system at a steady state. As we vary (i.e. increase) the bifurcation parameter  $p$ , assume a Hopf bifurcation at  $p_1$  destabilises the steady state. If the bifurcation is super-critical, we have a stable limit cycle for  $p > p_1$ . This is a periodic signal with a fixed frequency, say  $\omega = \omega_1$ . This explains the origin of a time scale in the signal. The power spectra  $P(\omega)$  of such a signal (of infinite length) show a peak at a fixed frequency  $\omega_1$  (Fig. 12.1a). The conjecture in the Landau-Hopf scenario is that there exist many more bifurcation points (in fact, an infinity of them)— $p_2, p_3, \dots, p_n$ . As the bifurcation parameter  $p$  is increased further beyond  $p_1$ , a new frequency or time-period is generated as we cross each of these critical points. The basic limit cycle after the first Hopf bifurcation gets destabilised at  $p_2$ , and for  $p_2 < p < p_3$ , we have a new solution composed of two frequencies. This new frequency is generated by a second bifurcation at  $p_2$ . The power spectra of such a signal shows peaks at two discrete frequencies, not necessarily related to each other (see Fig. 12.1(b)). This second bifurcation at  $p_2$  gives rise to a second time scale in the system. At  $p_3$  we have another (or a third) bifurcation generating a third frequency or time scale. This process continues *ad infinitum* till we obtain a signal which contains countably infinite number of frequencies. These frequencies are independent of each other, and here the FFT shows an infinite number of peaks (see Fig. 12.1c).

As already explained this scenario is incorrect. Here, even after an infinite number of bifurcations we cannot obtain a power spectrum with a continuous band, but only one with a number of discrete peaks close to each other. The continuous band in a power spectrum is the real characteristic of a chaotic signal. So this hypothesis can never really explain chaos. Besides, there is no known mechanism for the generation of the secondary frequencies at  $p_2, p_3$ , etc. from bifurcations from an unstable steady state or the existing limit cycle. We describe next the period-doubling bifurcation scenario which is frequently encountered in many dynamical systems and maps.



**Fig. 12.1** Power spectra obtained from an FFT of periodic signal (of infinite length) with: (a) one frequency; (b) two frequencies; (c) a large number of frequencies.

## 12.2 PERIOD-DOUBLING CASCADES

The period-doubling cascade is one of the most extensively studied routes to chaotic behaviour. It received considerable attention on the theoretical front and has been quantified accurately. The existence of this route has also been confirmed experimentally in many systems. We discuss this route to chaos for maps as well as dynamical systems.

### 12.2.1 Map

The one-dimensional logistic map

$$x_{n+1} = ax_n(1 - x_n) \quad (12.1)$$

looks very innocuous, with only a quadratic nonlinearity. We restrict  $0 < a < 4$  and  $0 < x_n < 1$  as

explained in Chapter 11. This simple looking dynamical system exhibits a period-doubling cascade and chaotic behaviour (see Feigenbaum, 1978). In Chapter 11, we saw that the logistic map has a period-doubling bifurcation at  $a = 3$ . For  $a > 3$ , the two fixed points  $x = 0$  and  $x = 1 - 1/a$  are both unstable. The obvious question that arises here is: How does the sequence  $\{x_n\}$  generated by the map behave for large  $n$ , when we start with an arbitrary and general  $x_0 \in [0, 1]$ ? The sequence will remain bounded by virtue of the restriction on  $a$ . It also cannot converge to either fixed point. For  $a = 3.1$ , and  $x_0 \in [0, 1]$  and  $x_0$  not equal to either fixed point, the map generates a sequence  $\{x_n\}$  such that for large  $n$  the sequence alternates between the two elements  $b, c$ , where  $b = 0.7645665$  and  $c = 0.558014$ . Here

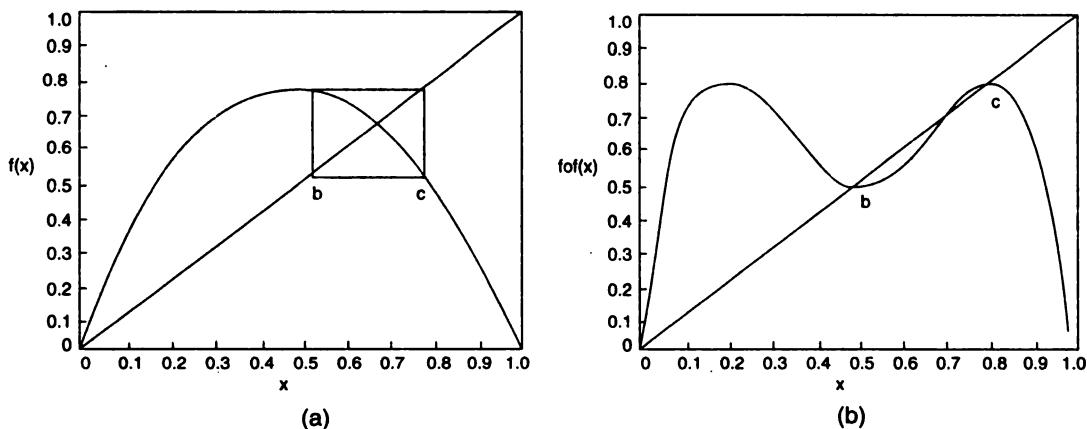
$$b = f(c) \quad (12.2a)$$

$$c = f(b) \quad (12.2b)$$

$$c = f(f(c)) \quad (12.2c)$$

$$b = f(f(b)) \quad (12.2d)$$

This terminal state is called a period-two state as here every second iterate repeats itself. The point 'c' is mapped on to point 'b' by the map  $f$  and the point 'b' is mapped back on to point 'c' as illustrated in Fig. 12.2. So the sequence  $\{x_n\}$  generated oscillates between  $c$  and  $b$  for large  $n$ . Since the parameter  $n$  plays the role of time, this solution can be visualised as a periodic solution. This period-two solution (or  $2p$ -solution) is a stable state and an attractor. All points  $x_0$  are attracted to



**Fig. 12.2** (a) Evolution to the period two states of logistic map,  $c = f(b)$ ,  $b = f(c)$ ; (b) composite map  $f(f(x))$ , which has fixed points at  $b, c$ .

this state for a fixed  $a$ . This qualitative behaviour prevails for  $a \in (3, 3.449)$ . In this interval the terminal state of the sequence is such that the elements always oscillate between these two points. The value of these points ( $b, c$ ) depends on the actual value of the parameter  $a$  (see Nicolis, 1986). The map  $f(f(x))$  denoted by  $f^2(x)$  has four fixed points. Two of these are the same as the fixed points of  $f(x)$  and are unstable. The remaining two fixed points of  $f^2(x)$  are the points  $b, c$  which generate the period-two state. At  $a = a_1$  (3.449), this period-two state is rendered unstable again by a period-doubling bifurcation. This can be easily verified now as

$$\frac{d}{dx} (f(f(x))) = -1 \quad (12.3)$$

for  $a = a_1$  at  $x = b$ , and  $x = c$ .

The composite map  $f(f(x))$  is destabilised by a super-critical period-doubling bifurcation at this bifurcation point (see Nicolis, 1986). This gives rise to a stable period-four state (double the period-two state) for  $a > a_1$ . We term this bifurcation a secondary bifurcation as the new period-four state is born out of the period-two state, and not the fixed point  $x = 1 - 1/a$ , which we consider as our basic branch. The period-four state is an attractor for  $a \in a_1$  (3.449),  $a_2$  (3.544).

In this interval the iterates starting from a point in  $0 < x < 1$  generate a sequence such that now every fourth iterate repeats itself for large  $n$ . For  $a = 3.51$ , the terminal state is characterised by the four numbers

$$d = 0.37722, \quad e = 0.825079, \quad g = 0.506713, \quad h = 0.877342$$

Each of these four numbers satisfies

$$x = f^4(x) \quad (12.4)$$

Each of the period-two states,  $x = b$ ,  $x = c$  gets destabilised by a period-doubling bifurcation at  $a = a_1$  and gives rise to the period-four state. Thus 'b' gives rise to 'd', 'g' and 'c' give rise to 'e' and 'h'. The order in which the system visits these four states can only be determined by using (9.1) for a fixed  $a$ . As  $a$  is increased beyond  $a_2$  (3.544), the period-four solution is rendered unstable by a further period-doubling bifurcation. This can be easily verified as

$$f^4(x) = -1 \text{ at } x = d, e, g, h \text{ for } a = a_2$$

This bifurcation is also super-critical and gives rise to a stable period-eight solution which gets destabilised at  $a = 3.564$ . As  $a$  is varied further, we generate more period-doubling bifurcations, in fact, an infinite cascade of bifurcations. The successive bifurcation points  $a_i$  occur with increasing rapidity as we go along the cascade. The points  $a_i$  converge and accumulate at a critical point  $a^*$  (approximately 3.57), where we have a  $2^\infty$  solution or an infinite period solution. For  $a > a^*$ , the terminal state of the map is such that the state repeats itself only after an infinite number of iterations i.e. the trajectory is aperiodic.

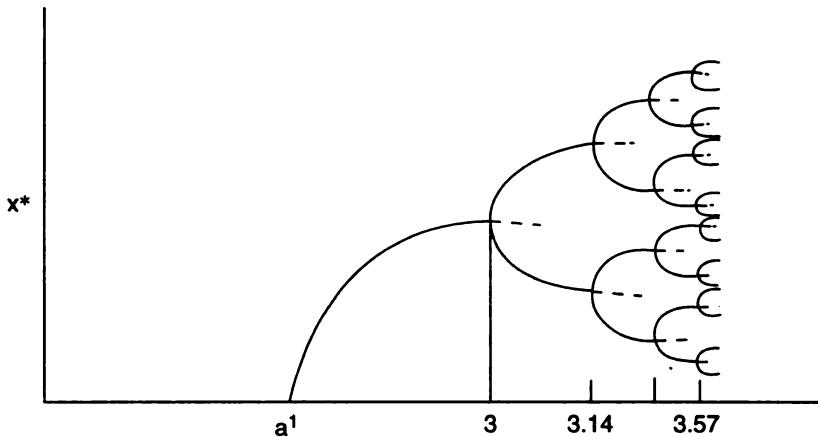
Each period-doubling bifurcation in the cascade for the above system is super-critical. This implies that at every bifurcation point the new solution (doubled period solution) emanating is stable and coexists with the original solution which is destabilised. This is in contrast to a sub-critical bifurcation where at each bifurcation point the new solution emerging is unstable, and surrounds the stable basic branch which is getting destabilised. The cascade of bifurcations and the transitions can be succinctly represented as:

$$\begin{aligned} (\text{fixed point}) &\rightarrow \text{period-1} \rightarrow \text{period-2 state} \rightarrow \text{period-2}^2 \rightarrow \\ &\text{period-2}^3 \rightarrow \text{period-2}^n \rightarrow \text{period-2}^\infty \text{ (chaos or aperiodic)} \end{aligned}$$

The bifurcation diagram of the logistic map is shown in Fig. 12.3. As we go along the bifurcation cascade, the interval over which a particular stable solution exists decreases. In particular, if the  $2^k-p$  solution exists in  $(a_k, a_{k+1})$  and the  $2^{k-1}-p$  solution exists in  $(a_{k-1}, a_k)$ , then

$$|a_{k+1} - a_k| < |a_k - a_{k-1}|$$

Each periodic solution exists over a parameter range or an interval of  $a$ , however small. Hence the periodic behaviour can be observed in an experimental system provided the parameter can be controlled and maintained in that interval. This property is called *structural stability*. This is in contrast with the bifurcation that occurs at a fixed bifurcation point. Since it is not possible to fix the parameter with infinite precision, the bifurcation point can never be observed. One can only



**Fig. 12.3** Bifurcation diagram of the logistic map, showing the period-doubling cascade.

infer about the bifurcation by determining system behaviour across the bifurcation point. Solutions having a period  $2^n$  are also called *sub-harmonic solutions*.

Feigenbaum (1978) analysed the bifurcation behaviour of the period-doubling cascade of the logistic map quantitatively. He established that the bifurcation points  $a_i$  converged geometrically to the accumulation point  $a^*$ . More precisely, the bifurcation points  $a_n$  satisfy

$$(a_n - a^*) \propto \delta^{-n} \text{ as } n \rightarrow \infty$$

This leads to

$$\lim_{n \rightarrow \infty} \frac{a_n - a_{n-1}}{a_{n-1} - a_{n-2}} = \delta^{-1} \quad (12.5)$$

Feigenbaum (1978) determined the constant  $\delta$  as  $4.6992\dots$  for the logistic map. He proved that this was a universal constant in the sense that it characterises the period-doubling cascade of any map which has a single quadratic maximum. He also showed that there was a second universal constant. We refer the interested reader to his original paper for the physical significance of this constant and for more details.

The behaviour of the sequence  $\{x_n\}$  for  $a > a^*$  is said to be chaotic. The attracting set (or state), viz. the set of points to which the sequence  $\{x_n\}$  tends to for large  $n$ , is called a chaotic attractor. Unlike the sub-harmonic state, it is not composed of a discrete finite set of points. The chaotic attractor is a bounded set. We are assured of this since for  $a \in (0, 4)$ , the maximum of  $f(x) < 1$ . This assures us that  $f(x)$  maps the interval  $(0, 1)$  into itself. The sequence generated here does not repeat itself and hence it is not a periodic state. This behaviour is not representative for all  $a \in (a^*, 4)$ . In this interval the behaviour of the map is even more interesting. There are regions of periodic behaviour interspersed with regions of chaotic behaviour. Periodic solutions with periods which are not multiples of 2 arise. Each of these undergoes period-doubling cascades. A detailed discussion of the behaviour of the system in this region is beyond the scope of this text, which aims at only giving the basics of bifurcation theory and secondary bifurcations.

An important characteristic of the chaotic state is its **sensitivity to initial conditions** (see Lorenz (1963) and Scott (1991)). Consider two points  $x_0^1, x_0^2$  which are sufficiently close by, i.e.

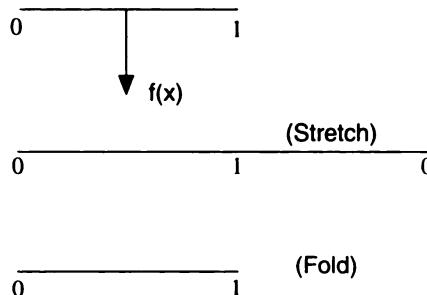
$$|x_0^1 - x_0^2| < \epsilon$$

The map  $f(x)$  in (12.1a) can be used to generate two sequences  $\{x_n^1\}$  and  $\{x_n^2\}$  from these two initial points. For any  $a$ , such that  $a < a^*$ , both the sequences converge to the same sub-harmonic solution since now this is the only possible attractor. The terminal state is the  $2^p$ -state and the  $2^p$ -points are visited in the same cyclic order.

For  $a > a^*$ , the terminal state or the attractor consists of an infinite set of points. Here the two sequences  $\{x_n^1\}$  and  $\{x_n^2\}$  are attracted to the same chaotic attractor. The attractor now is not composed of a finite set of points but an infinite bounded set. The above two sequences are such that for large  $n$  the elements in the sequence hover around the attractor. The attractor is the set of points in the neighbourhood of which every trajectory starting from some  $x_0$  will converge. It is not the set of points on which every trajectory will lie exactly. In fact, the two trajectories will be attracted to different sets of points, but these points will be close to the attractor after a large  $n$ . This is called the **shadowing property** of the trajectory.

For sufficiently small  $n$ ,  $|x_n^1 - x_n^2|$  stays small but as  $n$  increases,  $|x_n^1 - x_n^2|$  for a fixed  $n$  becomes large. The first few elements of the two sequences remain close to each other and the distance between the elements increases for large  $n$ . This property, in which small changes in the initial condition give rise to large changes in the trajectory (i.e. the dependence of  $x$  on  $n$ ), is called **sensitivity** to initial conditions. This is a feature of a deterministic chaotic system. This sensitivity is exhibited by the attractor for every initial condition in its basin of attraction.

This chaotic attractor and the sensitivity to initial condition arise because the map (12.1) has the property of stretch and fold. The function  $f(x)$  maps the points from  $(0, 1/2)$  to the interval  $(0, 1)$  and again the points  $(1/2, 1)$  to the interval  $(1, 0)$ , for  $a = 4$ . This is depicted in Fig. 12.4. The action of the map  $f$  is to increase the length of the interval on which it acts and folds this interval back



**Fig. 12.4** Stretch and fold property of the logistic map.

onto  $(0, 1)$ . This folding property ensures that the sequence generated is always bounded. The stretching property of  $f$  increases the distance between two nearby points. There are points in the attractor where there is stretch and other points where there is fold. A system is chaotic when in some “average” sense stretching predominates over the folding. The sensitivity to initial conditions of these chaotic systems is different from that discussed in Chapter 11, with reference to Fig. 11.6. There the sensitivity was only across a critical surface in phase-space and the system gets attracted to two different attractors. In chaotic systems the sensitivity is shown in a region of phase-space and points in the region are attracted to the same attractor.

### 12.2.2 Dynamical Systems

In Chapter 11, we saw how a steady state of a dynamical system loses stability due to a Hopf bifurcation and gives rise to a periodic state or a limit cycle. The nature of the bifurcation, whether

it is sub-critical or super-critical, determines the stability of the limit cycle emerging from the bifurcation point. We saw in Chapter 11, that the stability of a steady state is determined by the eigenvalues of the Jacobian matrix evaluated at the steady state. Let us see how the stability of a limit cycle is determined once it is found. This is obtained by studying linear differential equations with periodic coefficients. This is in contrast to the stability of a steady state where the linearised system of equations has constant coefficients (since we linearised around a steady state). The Floquet theory formally deals with linear equations with periodic coefficient (since here we linearise around a period solution). (We refer the interested reader to the Appendix for a detailed discussion of Floquet theory.) The stability of a limit cycle as mentioned there is obtained from the eigenvalues of the monodromy matrix  $A$ . If all the eigenvalues  $\lambda_i$  of  $A$  called Floquet multipliers are such that

$$|\lambda_i| < 1 \quad (12.6)$$

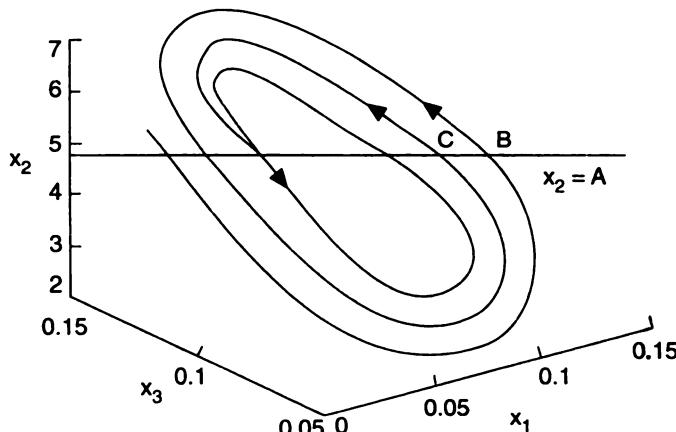
and only one eigenvalue (critical Floquet multiplier) equals +1, the limit cycle is stable. If even one of the eigenvalues is outside the unit circle, the limit cycle is unstable. In contrast to the stability of a steady state where the imaginary axis acts as the stability boundary in the eigenvalue plane, here the stability boundary is the unit circle. The stability condition is similar to that of maps. This is only to be expected since we have seen that maps allow us an equivalent way to represent limit cycles.

The stability condition for a limit cycle is identical with that of a fixed point of a map. The  $\lambda_i$ 's in (12.6) are the eigenvalues of the monodromy matrix. The mathematical basis for this condition comes from the Floquet theory. The analogy between the stability conditions of a fixed point of a map and that of a limit cycle of a dynamical system can be easily seen using Poincare maps (see Scott, 1991).

Consider a three-dimensional dynamical system. Let  $x_1, x_2, x_3$  be the three state variables. Let the dynamical system be given by

$$\left. \begin{array}{l} \dot{x}_1 = f_1(x_1, x_2, x_3) \\ \dot{x}_2 = f_2(x_1, x_2, x_3) \\ \dot{x}_3 = f_3(x_1, x_2, x_3) \end{array} \right\} \quad (12.7)$$

A limit cycle in the three-dimensional phase-space is a closed curve as shown in Fig. 12.5.



**Fig. 12.5** A limit cycle in three-dimensional phase-space.

Consider the Poincare section defined as  $x_2 = A$  and  $\dot{x}_2 > 0$ . If the constant  $A$  is chosen suitably, the limit cycle will intersect the section and the  $x_1$ ,  $x_3$  coordinates of the point will define the Poincare map. To determine the stability of the limit cycle, we study the evolution of the trajectory from a point nearby. The trajectory intersects the Poincare section at a series of points and converges to the limit cycle if it is stable or diverges from the limit cycle if it is unstable. The stability of the limit cycle is now transformed to that of determining the stability of the fixed point of a two-dimensional invertible map. It is a two-dimensional map since we generate a sequence of points  $(x_{1,n}, x_{3,n})$  by taking the Poincare section of a three-dimensional system. The map is invertible because a unique trajectory passes through every point in phase-space. So the trajectory emanating from point  $B$  on the Poincare section uniquely determines the point  $C$ , the next point of intersection, of the trajectory with the Poincare section. Once we are at point  $C$ , we could theoretically go back in time (this will be numerically very difficult) and uniquely obtain point  $B$ , because the trajectory passing through  $C$  is again unique. This implies that the dynamical system can be viewed as being equivalent to an invertible map.

$$x_{1,n+1} = g_1(x_{1,n}, x_{3,n})$$

$$x_{3,n+1} = g_2(x_{1,n}, x_{3,n})$$

Thus the stability of the limit cycle becomes equivalent to the stability of the fixed point of the two-dimensional map. The stability condition of the limit cycle of a dynamical system it follows is similar to that of the fixed point of a map. In general, an  $n$ -dimensional dynamical system can be viewed as analogous to an  $n - 1$  dimensional invertible map. A limit cycle which is stable can be destabilised when the Floquet multipliers cross the unit circle. This can occur in three possible ways and result in the following three kinds of bifurcations:

**(i) Saddle-node bifurcation.** Here a real eigenvalue other than the critical Floquet multiplier crosses the unit circle through  $+1$ . At this bifurcation point an unstable limit cycle with the same period as the original limit cycle branches out. This bifurcation is similar to the saddle-node bifurcation of a steady state where a real eigen-value crosses the imaginary axis. Here a new steady state branch evolves at the bifurcation point. This bifurcation is also called a limit point since usually it is sub-critical. The periodic solution now turns back at this point and no-limit cycle exists on one side of the point as shown in Fig. 12.6.

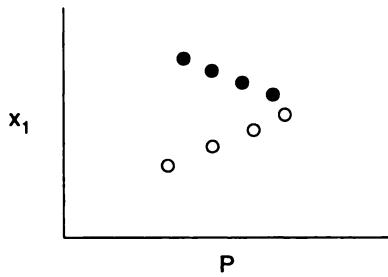


Fig. 12.6 Saddle-node bifurcation of a limit cycle.

**(ii) A period-doubling bifurcation.** Here a real eigenvalue crosses the unit circle through  $-1$  and all other Floquet multipliers are within the unit circle. At this bifurcation point, a limit cycle whose period is twice that of the original or base limit cycle is born. The base limit cycle gets destabilised at this point. The doubled period solution that emerges may be stable or unstable depending on whether the bifurcation is super-critical or sub-critical.

**(iii) Torus bifurcation.** This third possibility arises when a complex-conjugate pair of eigenvalues can cross the unit circle. At this bifurcation point the base limit cycle gets destabilised and a quasi-periodic solution is born. This solution can be represented as the motion on a toroidal surface and hence the name “torus bifurcation”. This bifurcation gives rise to a new frequency and is sometimes called a secondary Hopf bifurcation. We will study this in detail shortly.

Period doubling cascades can give rise to chaos in dynamical systems. Let us understand possibly how a period-doubling cascade can occur for a three-dimensional system. Let  $x^*(t)$  represent a time-periodic solution of period  $T$  of this system. This limit cycle, let us say, is stable. So it has Floquet multipliers which lie within the unit circle in addition to the CFM. As we vary a parameter  $p$  in the system, the period of the solution changes and we reach a period-doubling bifurcation point at  $p = p_1$ , where the period is  $T_1$ . Here one of the Floquet multipliers leaves the unit circle through  $-1$ . At this point a new periodic solution branches out. The period of this solution is  $2T_1$ , and this branch is called the period-two state (Fig. 12.7). Assuming the bifurcation to be super-critical, we follow this stable solution with period  $2T_1$  for  $P > P_1$ . The period of the solution changes continuously till we reach the point  $P_2$  where the period is  $T_2$  (different from  $2T_1$ ). Here again the stable-limit

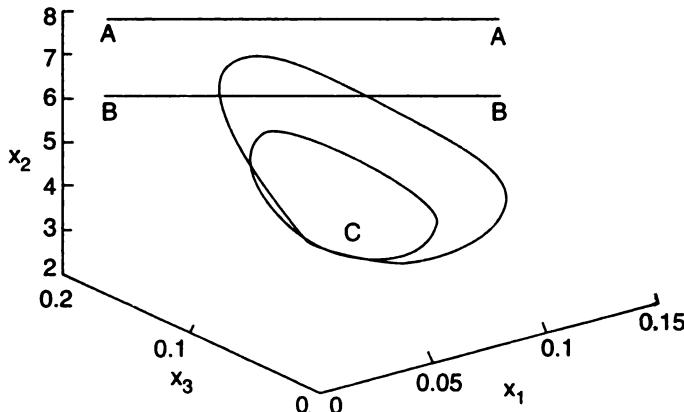
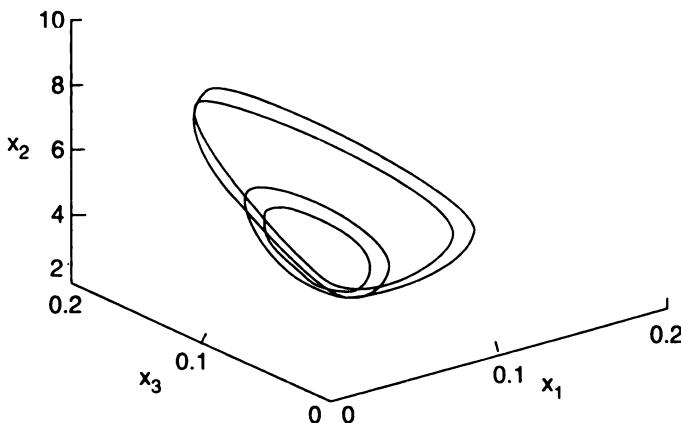


Fig. 12.7 A period-two limit cycle.

cycle exhibits a period-doubling bifurcation and a new solution with period  $2T_2$  (not  $4T_1$ ) is born. The period-two solution is now destabilised and the period-four state is born. This period-four state is stable if this period-doubling bifurcation is also super-critical (Fig. 12.8). The bifurcation from the period-two state to the period-four state is determined by the Floquet multipliers of the period-two solution, and not by the Floquet multipliers of the period-one solution at  $p = p_2$  (as bifurcation theory is a local theory). Once the period-one solution is rendered unstable, we lose our interest in it and continue along the period-two solution. There are instances when the unstable solutions can cause further bifurcations due to some global phenomena. This is beyond the scope of the present text which aims only at giving an introduction to the local theory.

Varying the parameter further can give rise to the period-doubling bifurcation cascade as we saw for the logistic map. We see a succession of period- $2^n$  states. These bifurcations occur with increasing rapidity again and the bifurcation points accumulate at  $p^*$ , beyond which we have a  $2^\infty$  solution, i.e. an aperiodic or a chaotic solution. The solution or the trajectory does not repeat itself no matter how long we wait. A trajectory starting from a point in phase space is attracted to a confined region in it. This region is called the *chaotic attractor*.



**Fig. 12.8** A period-four limit cycle.

For a periodic solution, the period  $T$  varies continuously as we vary the bifurcation parameter  $p$ . When we take the Poincare section of the limit cycle, the number of points at which the trajectory intersects the Poincare section remains constant. What varies is the time interval between successive piercings (the time period) and the location of the piercing, i.e. the fixed point. This is again consistent with the behaviour of maps as we have seen for the logistic map.

The Poincare map of a basic limit cycle (Fig. 12.9a) is a point and that of a period-two state consists of two points (Fig. 12.9b). The Poincare map is determined by the chosen section. An improper choice of the section for a period-two state may yield only one intersection or no intersection. Proper care must be exercised to determine the Poincare map. It must be emphasised that the trajectory does not intersect itself at the point  $C$  (Fig. 12.7). It appears to do so since the figure is a two-dimensional projection of a trajectory in three-dimensional phase-space.



**Fig. 12.9** Schematic Poincare maps of basic limit cycle and period-two limit cycle.

**Example 12.1** Chemburkar et al. (1987) investigated the consecutive reaction system, where the first reaction is exothermic and the second endothermic. When the activation energies of the two reactions are equal, this system is modelled by

$$\begin{aligned}\dot{x}_1 &= 1 - x_1 - Da x_1 e^{x_3} \\ \dot{x}_2 &= -x_2 + Da x_1 e^{x_3} - Da \gamma x_2 e^{x_3} \\ \dot{x}_3 &= -(1 + \delta)x_3 + Da \beta x_1 e^{x_3} + Da \beta \gamma \alpha e^{x_3}\end{aligned}$$

This three-dimensional autonomous system has five parameters ( $Da$ ,  $\gamma$ ,  $\delta$ ,  $\beta$ ,  $\alpha$ ). We choose  $\delta$  as the bifurcation parameter and fix  $Da = .26$ ,  $\gamma = .5$ ,  $\beta = 57.77$ ,  $\alpha = -.426$ . For this set of parameters, the bifurcation diagram has a unique steady state branch, with two Hopf-bifurcation points at  $\delta_1 = 6.94$  and  $\delta_2 = 19.34$ . The bifurcation at  $\delta_1$  is super-critical and at  $\delta_2$  it is sub-critical. The

complex part of the critical eigenvalue is 21.5. The period of the limit cycle near  $\delta_1 = 6.94$  is approximately given by .2917 and is numerically found to be 0.3. This system exhibits a period-doubling cascade of bifurcations. The limit cycle emanating at 6.94 is destabilised by a Floquet multiplier, leaving the unit circle through -1, and a period-two solution is born. At  $\delta = 7.76$ , the period of this period-two limit cycle is .4. The limit cycles corresponding to these parameters are shown in Figs. 12.5 and 12.7, respectively. The period-two solution is destabilised again by a period doubling bifurcation; this gives rise to a period-four, solution. At  $\delta = 7.89$ , this period-four solution is stable and has a period equalling .791. This attractor is shown in phase-space in Fig. 12.8. Its Poincare map has four points. This 4p-solution gives rise to an 8p-solution as we increase  $\delta$ . As we vary  $\delta$  further, we see a cascade of period-doubling bifurcations. The bifurcation points converge at approximately 7.965 and yield a chaotic attractor. The reader interested in simulating the system can verify his results on period-doubling cascade with those reported in Chemburkar et al. (1987). The period-two solution (after the first period doubling) exists for  $7.49 < \delta < 7.89$ . Here the period of this solution varies continuously. On the other hand, the Poincare map (if chosen appropriately) always yields two points. The coordinates of these points, however, depend on the  $\delta$  value as discussed earlier.

We would like to emphasise here that the Poincare map must be chosen appropriately. If the section is chosen at  $x_2$  as given by AA, we will see no intersection, and for  $x_2$  as given by BB, we will have only one intersection (Fig. 12.7). The chaotic attractor consists of a set of points in phase-space, in the neighbourhood of which every trajectory emanating from its basin of attraction is attracted. The chaotic system here also possesses the property of sensitivity to initial conditions. The evolution of a trajectory from a point in phase-space is determined uniquely. Two trajectories starting from nearby points in phase-space remain close to each other for a short time. For a long time, the two trajectories become completely uncorrelated. Uncertainties in specifying the initial condition accurately prevent precise prediction of the system state. It is precisely for this reason that the validity of weather forecasts is for a short period. The validity period can be increased by improving the accuracy of specifying the initial condition.

The sensitivity to initial condition is an inherent feature of the chaotic system. This property exists even when the chaotic attractor is the only attractor of the system. This is in contrast to the sensitivity a system exhibits across the boundaries of the different basins of attraction (Chapter 11). Here it is necessary that several stable states co-exist, each with its own basin of attraction. Jorgensen and Aris (1983) have observed a period-doubling cascade of bifurcations in a CSTR with consecutive reactions. Sparrow (1982) contains a good review of bifurcation features of the Lorenz system.

A two-dimensional autonomous dynamical system cannot exhibit the complex dynamic behaviour of quasi-periodic solution or chaotic behaviour. A dynamical system must be at least three-dimensional to exhibit such behaviour. This third dimension gives an extra degree of freedom. A unique trajectory goes through every point in phase-space here as well. Consequently, no two trajectories can intersect. The trajectory or the attractor appears to intersect itself in Fig. (12.7) because we have projected it on the two-dimensional plane of the paper. The three-dimensional dynamical system we have seen can be viewed as a two-dimensional invertible map. Hence, two-dimensional maps can exhibit such behaviour. How then does one explain that the one-dimensional logistic map can exhibit chaotic behaviour? This is possible because the logistic map is noninvertible.

### 12.3 RUELLE-TAKENS SCENARIO

The period-doubling route to chaos has been extensively investigated in the literature. In comparison

to this, the Ruelle-Takens scenario has caught very little attention. The transition sequence here can be succinctly written as

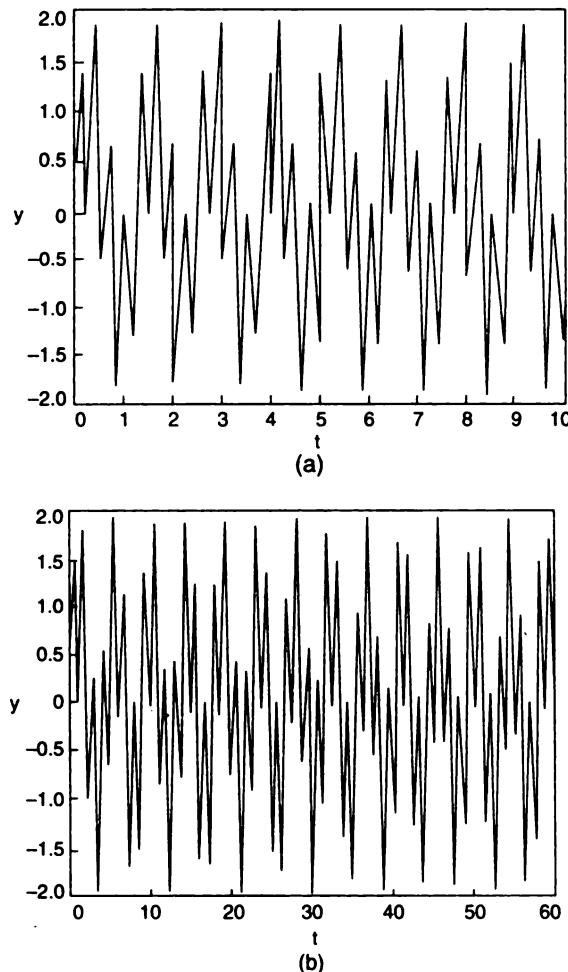
Periodic solution  $\rightarrow$  quasi-periodic solution  $\rightarrow$  chaotic solution

Here a periodic solution gives birth to a quasi-periodic solution through a torus bifurcation. On further varying the bifurcation parameter, the quasi-periodic solution gets destabilised and we see a chaotic solution. Our treatment of this problem will be at a very elementary and conceptual level. This is necessitated since this scenario can be demonstrated only computationally. Before we discuss this route in detail, let us see what is quasi-periodic behaviour.

Consider a signal  $y(t)$  with two frequencies as

$$y(t) = \sin 2t + \sin 7t \quad (12.8)$$

$\sin 2t$  is periodic with period  $\pi$ , and  $\sin 7t$  is periodic with period  $2\pi/7$  (see Fig. 12.10). The period of  $y(t)$  in (12.8) is  $2\pi$ . The period is defined as the minimum time interval over which the solution



**Fig. 12.10** Signal with two frequencies: (a) periodic; (b) quasi-periodic.

repeats itself. Thus a signal which repeats itself every 3 seconds also repeats itself every 6 seconds. In fact, it repeats itself every integer multiple of 3 seconds. The period of this signal is 3 and not  $3n$ , where  $n \geq 2$ . (Of course, we are assuming that the signal does not repeat itself for any time less than 3.)  $y(t)$  in (12.8) is a periodic signal, as it consists of two frequencies or periods whose ratio is a rational number.

Consider now the signal

$$y(t) = \sin 2t + \sin \sqrt{2}t \quad (12.9)$$

This signal is also composed of two frequencies 2 rad/s and  $\sqrt{2}$  rad/s, i.e. time periods  $\pi$  and  $\pi\sqrt{2}$ . The two frequencies are incommensurate, i.e. their ratio is an irrational number. Consequently,  $y(t)$  is not a periodic signal (Fig. 12.10b). The signal does not repeat itself over a finite interval of time. However, it is composed of two frequencies and the power spectra would yield two peaks as shown in Fig. 12.1b. Hence it is called a quasi-periodic signal. The concept of quasi-periodic behaviour can be best understood by considering the discretised circle map

$$\theta_{n+1} = (\theta_n + \Omega) \bmod 2\pi \quad (12.10)$$

Here  $\theta$  represents the angle made with the  $x$ -axis as shown in Fig. 12.11. The map can be viewed as generating points on the unit circle whose angular position is determined by  $\theta$ . Let  $\Omega$  be  $\pi$ . The

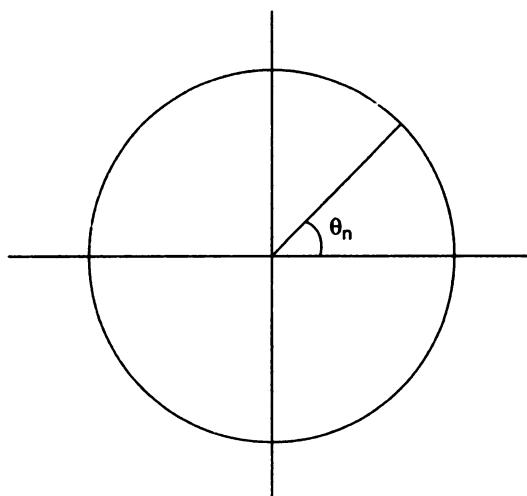
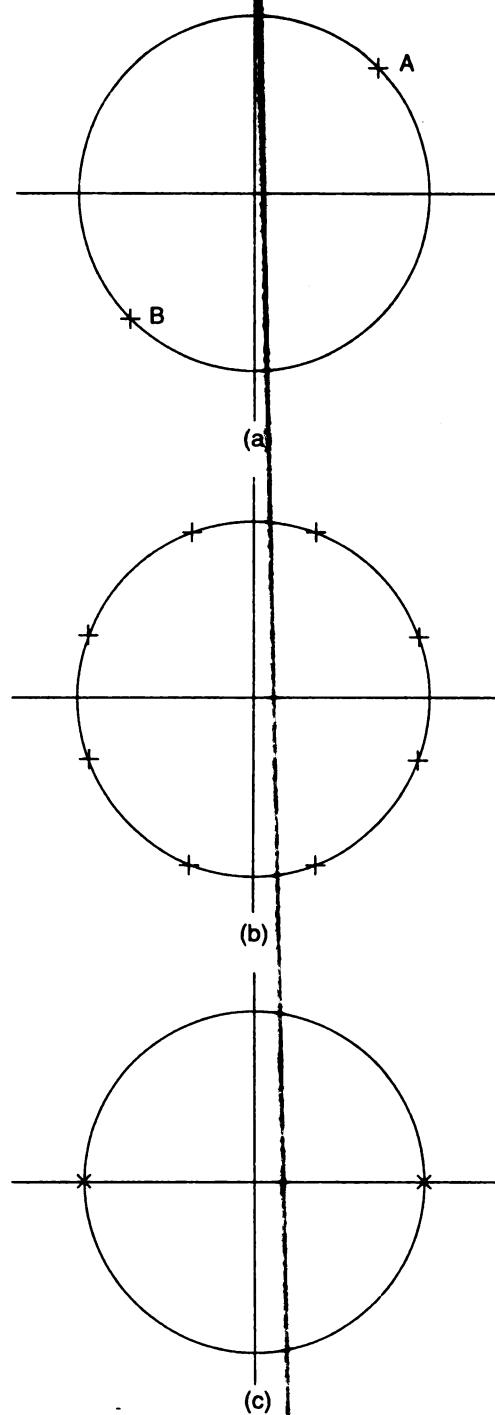


Fig. 12.11 Illustration of the circle map.

map generates two points on the unit circle, i.e. starting with  $A$ , we generate point  $B$ , and  $B$  gives back point  $A$  (see Fig. 12.12a). To come back to point  $A$ , we have to traverse two points ( $A$  and  $B$ ) and rotate once around the unit circle. Defining the **rotation number**  $R$  as the number of rotations of the unit circle per point traversed, we obtain for  $\Omega = \pi$ ,  $R = 1/2$ .

Consider now the case when  $\Omega$  is  $\pi/4$ . The map generates eight points on the unit circle as illustrated in Fig. 12.12b. Here,  $R = 1/8$ . For  $\Omega = 3\pi$ , we generate two points on the unit circle. However, now we go through three rotations of the unit circle before we come back to the same point. This yields  $R = 3/2$  as shown in Fig. 12.12c.

The map generates a discrete set of points in all the above cases and the rotation number  $R$  is always rational. This corresponds to the case where the solution is periodic in nature when the Poincaré map shows a finite and a discrete number of points.  $R$  as we have defined, is  $(\Omega/2\pi)$ .

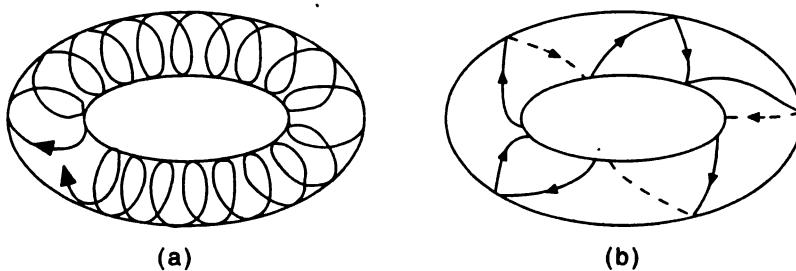


**Fig. 12.12** Circle map for: (a)  $\Omega = \pi$ ; (b)  $\Omega = \pi/8$ ; (c)  $\Omega = 3\pi$ .

Consider now the case where the ratio of  $\Omega$  and  $2\pi$  is an irrational number. Let  $\Omega = 1$ . The map now generates points on the unit circle, which completely fill it. The map does not generate a discrete set of points anymore but the entire unit circle or a closed curve. Take a specific point on the unit circle. The points generated by the map come close to any point an infinite number of times but never reach the specific point exactly. The rotation number here is  $R = 1/2\pi$ . This is defined as the number of rotations of the unit circle per point traversed in the limit  $n \rightarrow \infty$ . There is an interplay of two frequencies in the circle map (12.10): (a) the frequency with which the points are generated on the unit circle; and (b) the number of rotations around the unit circle. The ratio of these frequencies ( $R$ ) determines the nature of the system behaviour. This idea in maps helps us understand how the interplay of two frequencies gives rise to quasi-periodic behaviour (see Schuster, 1984).

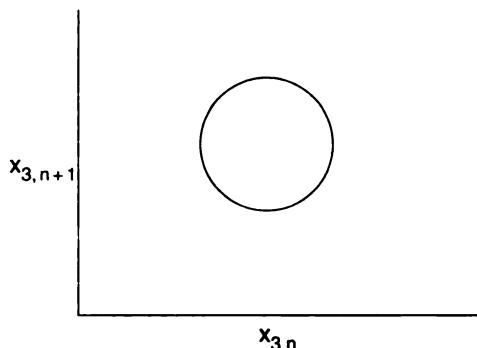
Let us see now how a quasi-periodic solution can be visualised in a dynamical system. Consider for the sake of simplicity the three-dimensional dynamical system. A periodic solution is a closed curve in phase-space. The Poincare section of this trajectory is a discrete set of points.

A quasi-periodic solution here can be thought of as a motion on a two-dimensional toroidal surface. For the sake of clarity, this is being depicted as the motion on a torus (see Fig. 12.13). A torus can be obtained by bending a cylinder around its axis so that the two flat faces meet. Consider a limit cycle composed of two frequencies  $\omega_1, \omega_2$ . This is a closed curve in phase-space and can be visualised as being a closed curve on the toroidal surface. It is similar to a snake lying coiled around this surface such that it is nibbling its tail.  $\omega_1$  can be thought of as the frequency of motion around the axis of the cross-sectional area of the torus and  $\omega_2$  as the frequency of motion along the axis of the cylinder. When  $\omega_1/\omega_2$  is a rational number, we have a closed curve on the torus (Fig. 12.13b). A Poincare section of this attractor yields a discrete set of points. When  $\omega_1/\omega_2$  is incommensurate, i.e. it is irrational, the trajectory completely covers the surface of the torus as illustrated in Fig. 12.13a. The trajectory comes close to a point on the surface an infinite number of times without ever reaching the same point. The surface gets fully covered and the Poincare section of such a trajectory is a closed curve.



**Fig. 12.13** Conceptual visualisation of a torus: (a) quasi-periodic motion; (b) periodic motion.

Quasi-periodic solutions usually arise in the study of periodically forced dynamic systems (see Scott, 1991). Here the period of forcing  $\omega_f$  and the natural frequency of the system  $\omega_n$  interact, giving rise to periodic or quasi-periodic behaviour (Fig. 12.14). The motion on the terms is called a  $T^2$  solution, to signify a two-dimensional toroidal surface. In the Ruelle-Takens route to chaotic behaviour, the  $T^2$  solution gives rise to chaotic behaviour abruptly. Here we do not observe an infinite cascade of bifurcations as in the periodic-doubling route. Such a route to chaotic behaviour has been observed by Konnur and Pushpavanam (1994) in a periodically forced operation of a batch reactor.



**Fig. 12.14** Poincaré section of a quasi-periodic solution—a closed curve.

A third route to chaotic behaviour is called the transition to chaos via intermittency. Here, an apparently ‘steady state’ exhibits excursions from the steady state by showing intermittent bursts before relaxing to the steady state. This route is called the *Pomeau-Manneville Scenario*. It has been observed in some systems and is not as well understood as the two other routes to chaotic behaviour we have discussed. Konnur and Pushpavanam (1994) have observed this in their periodically forced batch reactor system.

## 12.4 CHARACTERISATION OF TRAJECTORY

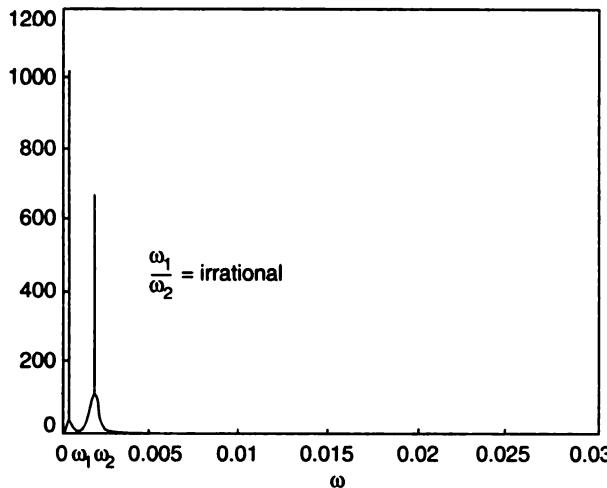
So far we have seen the mechanisms and origin of the different complex dynamic behaviour in systems. We now discuss methods by which we can quantify the long-time system behaviour. The system behaviour is obtained by measuring the different state-space variables as a function of time. This data is usually in discrete form, i.e. as a time-series. Since we are interested in determining the presence of different frequencies in a time signal, the fast Fourier transform (FFT) can be used effectively to characterise system behaviour.

### 12.4.1 Fast Fourier Transform

The FFT generates the power spectra of a signal. The power spectrum represents the square of the amplitude of a particular frequency component in a signal. It signifies the strength of a particular frequency in a signal. Many standard programs are available to compute the FFT efficiently. The FFT is the application of Fourier series (which we defined for continuous periodic functions) to discrete representations of continuous signals.

A periodic signal (of infinite length) with a single frequency  $\omega_0$  has an FFT as shown in Fig. 12.1a. It is like a Dirac-delta function at  $\omega = \omega_0$  and vanishes for  $\omega \neq \omega_0$ . The power spectrum (obtained from FFT) for a periodic signal with two frequencies  $\omega_1, \omega_2$  (where  $\omega_1/\omega_2 = \text{rational}$ ) has two peaks at  $\omega_1$  and  $\omega_2$  respectively (see Fig. 12.1b). A quasi-periodic solution with two frequencies  $\omega_3$  and  $\omega_4$  also shows two peaks at  $\omega_3$  and  $\omega_4$ . These peaks occur at the frequencies which are incommensurate ( $\omega_3/\omega_4 = \text{irrational}$ ) as depicted in Fig. 12.15. It is therefore difficult to use the FFT to distinguish between a periodic and a quasi-periodic signal. A chaotic signal, on the other hand, has a broad-band power spectrum and does not have discrete peaks. It has a low frequency component and the strength of the signal vanishes above a cut-off frequency. This is in contrast to a system with noise which has a high frequency component in the power spectra.

A problem with using the FFT is that, in addition to the actual frequency of the signal, other



**Fig. 12.15** Power spectrum of quasiperiodic signal.

harmonics also manifest in the power spectra. Besides errors in computation in experimental measurements, and finite length of data can generate many spurious frequencies, thereby confusing the investigator. Extreme care and experience is the only way in which an FFT can be used to characterise the system (see Scott, 1991).

#### 12.4.2 Poincare Maps

We have already discussed the Poincare maps in detail. This is an effective tool which can be used to classify system behaviour.

The Poincare map of a periodic state consists of a discrete set of points, and that of a quasi-periodic solution shows a continuous closed curve as already discussed (see Fig. 12.14). A chaotic solution, on the other hand, has a Poincare section which can be thought of as segments of continuous curves.

The Poincare map can be used as a tool to characterise a dynamical system by exploiting these characteristics. It represents a dynamical system by a lower dimensional map. Thus for a three-dimensional system, instead of visualising trajectories in three-dimensional space, we can observe points in the two-dimensional plane and follow the system evolution (see Scott, 1991).

#### 12.4.3 Dimension of an Attractor

A steady state of a system is a point in  $n$ -dimensional space. It can be thought of geometrically as being zero-dimensional. A limit cycle, on the other hand, is a closed curve in phase-space. This curve can be cut at a point and stretched to give a line. This line is one dimensional and we can define the dimension of a limit cycle as being unity (see Nicolis, 1986).

A quasi-periodic trajectory is represented as the motion on a torus. The surface of this torus is two-dimensional. We can cut open a torus twice, first along the cross-section to get the cylinder and then along the axis of the cylinder. It can be stretched open now and gives a two-dimensional plane. The quasi-periodic attractor has a dimension of two.

A characteristic of a chaotic attractor is that its dimension is not an integer; it is fractional. Such an object having a dimension which is not an integer is called a *fractal*.

We have already seen that a chaotic attractor exists for a three-dimensional system. The dimensions of a steady state, limit cycle, and a quasi-periodic solution are 0, 1, and 2, respectively. Since we are dealing with a three-dimensional system, the dimension of a chaotic attractor has to be less than 3, and it cannot be an integer.

So far we have defined the dimension of attractors using our physical intuition. We would now like to define the dimension of an attractor mathematically and rigorously so that it can be used to analyse and quantify system behaviour. This definition should be such that it reduces to the physical (geometric) concept of dimension. We will then be in a position to employ this definition to obtain the dimension of a chaotic attractor.

Consider the three-dimensional dynamical system:

$$\left. \begin{array}{l} \dot{x} = f_1(x_1, x_2, x_3) \\ \dot{x} = f_2(x_1, x_2, x_3) \\ \dot{x} = f_3(x_1, x_2, x_3) \end{array} \right\} \quad (12.11)$$

This autonomous differential system possesses different kinds of attracting states like steady states, periodic solutions, etc. We would like to quantify these various states and we do this by assigning them a dimension. The dimension of an attractor can be defined in several ways. These definitions should be such that they reduce to the geometric (Euclidean) dimension. Thus we would like the dimension of a steady state to be zero and that of a limit cycle to be one, and so on.

We will now see the two definitions of dimension, viz. those of Hausdorff dimension, and Correlation dimension.

The former is used to introduce the notion of dimension and the latter as a computational tool to compute it (see Schuster (1984) and Nicolis (1986)).

**Hausdorff dimension.** Consider the three-dimensional phase-space  $x_1, x_2, x_3$ . This space can be filled by infinitesimal cubes of edge  $\epsilon$ , and volume  $\epsilon^3$ . Let  $N(\epsilon)$  be the number of cubes which fill the attractor, i.e. the points in the attractor that lie in  $N(\epsilon)$  cubes. The Hausdorff dimension  $d_H$  of the attractor is then defined as

$$d_H = \lim_{\epsilon \rightarrow 0} \frac{\ln(N(\epsilon))}{\ln(1/\epsilon)} \quad (12.12)$$

Consider an attractor lying on a curve of dimension  $P$ , where  $P$  is an integer. There is a contribution to  $N(\epsilon)$  only if the attractor passes through a box. In this case we obtain

$$N(\epsilon) \propto \epsilon^{-P}$$

and (12.12) yields  $d_H = P$ .

For a steady state represented by a point in phase-space,  $N(\epsilon) = 1$ ; then  $d_H = 0$ , as expected. Consider a limit cycle or a one-dimensional curve. Here as we decrease  $\epsilon$ ,  $N(\epsilon)$  varies as  $\epsilon^{-1}$ , i.e.  $N(\epsilon) = k\epsilon^{-1}$ , where  $k$  is some constant. This yields for the limit cycle  $d_H = 1$ . Similarly, the quasi-periodic attractor has  $d_H = 2$ . The dynamic systems of interest to engineers are dissipative in nature. This means that points in  $m$ -dimensional phase-space get attracted to attractors of dimension  $< m$ . Since the 0, 1, 2 dimensional attractors are steady state, the limit cycle, quasi-periodic solutions, and the chaotic attractors must necessarily have a noninteger or fractional dimension. Such attractors are also called strange attractors and are fractals. A technique to determine  $d_H$  is by plotting  $\ln(N(\epsilon))$  against  $\ln(1/\epsilon)$ , over a range of  $\epsilon$  values, usually such that the entire attractor is covered. The slope of this curve yields  $d_H$ .

The concept of Hausdorff dimension is relatively simple to define, but not very convenient to calculate. Data on the state of a system is obtained in the form of time-series. The correlation dimension is defined so that the time-series data can be used directly to get an idea of the attractor dimension.

**Correlation dimension.** Consider an attractor of an  $m$ -dimensional system. Let the vector  $x(t)$  represent the variables  $(x_1(t), x_2(t), \dots, x_m(t))$ . Consider the attractor to be described by  $N$  such points in phase-space,  $x^i(t)$ ,  $i = 1, N$ . The correlation function  $C^j(\Delta)$  is defined as

$$C^j(\Delta) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i \neq j}^N \text{(number of points that lie within an } m\text{-dimensional sphere of radius } \Delta \text{ around } x^j) \quad (12.13)$$

The correlation sum  $C(\Delta)$  is defined as the average of the correlation functions over all points on the attractor, i.e.,

$$C(\Delta) = \frac{1}{N} \sum_{j=1}^N C^j(\Delta) \quad (12.14a)$$

$$= \frac{1}{N^2} \sum_{j=1}^N \sum_{i \neq j}^N H(\Delta - d_2(x^i, x^j)) \quad (12.14b)$$

where  $H$  is the Heaviside function

$$\begin{aligned} H(x) &= 1 \text{ for } x > 0 \\ &= 0 \text{ for } x < 0 \end{aligned}$$

The correlation dimension  $d_c$  is defined such that in the  $\lim \Delta \rightarrow 0$ ,

$$C(\Delta) = \Delta^{d_c} \quad (12.15a)$$

or

$$d_c = \lim_{\Delta \rightarrow 0} \frac{\ln C(\Delta)}{\ln (\Delta)} \quad (12.15b)$$

$C(\Delta)$  represents an average density of points of an  $m$ -dimensional sphere of radius  $\Delta$ .

We conclude by remarking that the correlation dimension is an effective tool in characterising an attractor from time-series data. Many other definitions of attractor dimension exist. Their discussion is beyond the scope of this text.

## 12.5 CONCLUDING REMARKS ON NONLINEAR DYNAMICAL SYSTEMS

The late seventies and early eighties saw a spurt in research activity in the area of nonlinear dynamical systems. In chemical engineering the systems investigated were primarily in chemical reaction engineering. Lumped-parameter models of continuous stirred-tank reactors consisting of three-coupled ordinary differential equations were studied. This is the minimum number of equations required to generate complex behaviour, as already discussed. It may appear that the engineer would be interested only in simulating the open-loop behaviour of the system. However, many important issues exist in this area which we now discuss.

The dynamic behaviour induced by nonlinear interactions is usually thought of as being undesirable or detrimental to the system performance. Chaotic oscillations in the reactor effluent

imply that the product composition varies with time and is not preferred when product quality is important. This also affects downstream processing units like separators or distillation columns and units upstream when a recycle is present. In hydrodynamic systems, the velocity in a channel can oscillate and this can induce mechanical stresses and cause fatigue. Chaotic behaviour is preferred in some systems, i.e. where mixing is important. The theory of nonlinear dynamics enables us to predict system behaviour. It provides an important tool by which the engineer can determine operating points in a system (i.e. parameter values to have the desired performance). He can avoid regions of pathological behaviour, like chaotic states, if they exist.

The phenomena discussed so far have been observed experimentally in most systems. Several examples can be found in Scott (1991). An important feature is that most of the secondary bifurcations are observed in very narrow regions of parameter space. Thus it is experimentally difficult to maintain parameter values in these intervals to detect high period states like the 16-period state. The patience and efforts of experimentalists definitely must be appreciated in these circumstances. The small regions of instability therefore imply that most plants operate at a stable steady state. Determining when a system exhibits some of the features detailed needs careful planning using the insight obtained from the theory. The emphasis so far can be viewed as studying the open-loop response of a system as done in process control. Classical control theory has restricted itself to "linear control" aspects only. Nonlinear control theory, i.e. the implications of bifurcations and chaos on the design of control systems, is not very well understood and promises to be an important area for future research.

Recent work in control of chaotic systems involves exploiting features of chaotic attractors like the sensitivity to parameters. The chaotic attractor arising from a period-doubling cascade contains many embedded period attractors. In this approach of control, a system is allowed to evolve from an initial condition, till it reaches the periodic state on which we intend to control the system. Small perturbations in the control parameter are now determined to maintain the system close to the periodic state. The advantage here is that since the system naturally comes close to the periodic state, a proportional law is sufficient to achieve control.

In many systems the processes going on are not very clearly understood. A reactor, for example, can sustain many side reactions. Here it would be difficult to write the detailed modelling equations. It is again not feasible to measure all variables of a system (this may just be uneconomical). Usually, only a single variable is measured. In a nonisothermal reactor, for example, on-line measurements of temperature are possible, but those of concentrations would be difficult. The experimentalist has only a single time-series to characterise system behaviour. Mathematical techniques to reconstruct attractors using this information have been developed. These techniques have been successfully implemented numerically for some systems. However, the theory has to be put on a firm footing.

The chaotic attractors of most systems have been found to be less than 3. This is even true of systems which are infinite dimensional (i.e. governed by partial differential equations). This implies that the system is confined effectively to a low three-dimensional subspace in phase-space. Lorenz (1963) was able to reduce his system of partial differential equations to a system of ordinary differential equations by doing a Fourier series expansion. The dimension of attractors can act as a guide in proposing phenomenological models for systems. We conclude by remarking that the area of nonlinear dynamics plays a valuable role in the modelling and simulation of systems.

# Appendix

## Floquet Theory

In this section we describe the basics of Floquet theory. The concepts of the stability of a steady state will be extended to determine the stability of a limit cycle. The conditions for a limit cycle to be stable will be obtained.

Consider the autonomous  $n$ -dimensional system

$$\dot{x} = F(x) \quad (12.16)$$

Let  $x_\Omega$  be an  $\Omega$ -periodic solution of this equation. Assuming that this solution (limit cycle) can be found, we want to determine its stability. Does it attract points in its immediate neighbourhood or not? Linearising the nonlinear system (12.16) around this periodic-state, we obtain

$$\dot{x}^* = L_\Omega(t)x^* \quad (12.17)$$

This system of equations, though linear, has periodic coefficients with period  $\Omega$  (Ince, 1956). The elements of the matrix  $L_\Omega$  have period  $\Omega$  as opposed to the Jacobian matrix  $A$  around a steady state which has constant coefficients.

Let  $x_0^i(t)$  ( $i = 1, \dots, n$ ) be a fundamental set of solutions to (12.17). Then  $x_0^i(t + \Omega)$  ( $i = 1, \dots, n$ ) is also a set of solutions. This set can be expressed as a linear combination of the fundamental set ( $x_0^i(t)$ ):

$$x_0^i(t + \Omega) = \sum_{k=1}^n a_{ik} x_0^k(t) \quad (12.18)$$

We now constitute a matrix  $X(t)$  from the fundamental set  $\{x_0^i(t) | i = 1, n\}$ . It follows from (12.18) that

$$X(t + \Omega) = AX(t) \quad (12.19)$$

Consider a solution vector  $y^i(t)$ . This is called a normal solution if

$$y^i(t + \Omega) = sy^i(t) \quad (12.20)$$

The set of normal solutions  $y^i(t)$  form the normal solution matrix  $Y(t)$ , so

$$Y(t + \Omega) = sY(t)$$

$Y(t)$  can also be expressed in terms of the fundamental set  $X(t)$  as

$$Y(t) = BX(t) \quad (12.21)$$

Since  $B$  is a constant matrix

$$Y(t + \Omega) = BX(t + \Omega)$$

$$sBX(t) = BAX(t)$$

It therefore follows that the scalars  $s$  defining the normal solutions are eigenvalues of the matrix  $A$ , defined in (12.19).

To see the role of the normal solutions, let us write the normal solution matrix as

$$Y(t) = e^{At} U(t) \quad (12.22)$$

where  $U(t)$  is a periodic function of  $t$  with period  $\Omega$ . Then

$$\begin{aligned} Y(t + \Omega) &= e^{\Lambda(t + \Omega)} U(t + \Omega) \\ &= e^{\Lambda(t + \Omega)} U(t) \end{aligned} \quad (12.23)$$

Also, as  $Y(t + \Omega) = sY(t)$ , we have

$$se^{\Lambda t} U(t) = e^{\Lambda(t + \Omega)} U(t) \quad (12.24)$$

or

$$sI = e^{\Lambda\Omega}$$

So, when  $|s_i| > 1$  or the eigenvalues of  $\Lambda$  are positive, the normal solution diverges from the base solution and the limit cycle is unstable, otherwise it is stable.

The matrix  $A$  is determined by integrating the set (12.16)–(12.17) numerically over a time period with the initial condition  $X(0) = I$ . The value of  $X(\Omega)$  then yields the matrix  $A$ , whose eigenvalues give ‘ $s$ ’. The stability condition of the limit cycle is identical to that of a map.

This analogy between the stability of a fixed point of a map and the limit cycle of a dynamical system can be easily seen using Poincare maps as explained in the text.

Differentiating (12.16) with respect to  $t$  yields

$$\ddot{x}(t) = L(x)\dot{x}(t) \quad (12.25)$$

It follows from this that  $\dot{x}(t)$  is a solution to the linearised system (12.17). At any instant of time the solution  $\dot{x}(t)$  is such that its direction is tangential to the base limit cycle which we are studying. Along this direction the perturbations neither decay nor grow. There is thus one eigenvalue that is always on the unit circle. This eigenvalue is  $+1$  and is called the *critical Floquet multiplier (CFM)*. This is frequently used as a check of the shooting method.

## REFERENCES

- Chemburkar, N., Kahlerts, O. and Varma, A., Dynamics of Consecutive Reactions in a CSTR—A case study, *Chemical Engineering Science*, **42**, 1507 (1987).
- Feigenbaum, M.J., Quantitative Universality for a Class of Non-linear Transformations, *J. Statistical Physics*, **19**, 25 (1978).
- Gleick, J., *Chaos Making a New Science*, Heinemann, London (1988).
- Haken, H., *Advanced Synergetics: Instability hierarchies of Self-organising Systems and Devices*, Springer-Verlag, Berlin (1983).
- Ince, E.L., *Ordinary Differential Equations*, Dover, New York (1956).
- Jorgensen, D.V. and Aris, R., On the dynamics of a stirred tank with conservative reactions, *Chemical Engineering Science*, **38**, 45 (1983).
- Konnur, R. and Pushpavanam, S., Dynamics of a fed-batch reactor, Transition from the batch to the CSTR, *Chemical Engineering Science*, **49**, 383 (1994).
- Kuramoto, Y., *Chemical Oscillations, Waves and Turbulence*, Springer-Verlag, New York (1984).
- Landau, L.D. and Lifshitz, E.M., *Fluid Mechanics*, Pergamon Press, Oxford (1959).

- Lorenz, E.N., Deterministic Non-periodic Flow, *J. Atmos. Sci.*, **20**, 130 (1963).
- Marsden, J.E. and McCracken, M., The Hopf Bifurcation and Its Applications, Springer-Verlag, New York (1976).
- Nicolis, J.S., Dynamics of Hierarchical Systems: An evolutionary approach, Springer-Verlag, Berlin (1986).
- Schuster, H.G., Deterministic Chaos—An Introduction Physics, Springer-Verlag, Weinherm (1984).
- Scott, S.K., Chemical Chaos, Clarendon Press, Oxford (1991).
- Sparrow, C., The Lorenz Equations, Bifurcations' Chaos and Strange Attractors, Springer-Verlag New York (1982).
- Vidyasagar, N., Non-linear Systems Analysis, Prentice-Hall, Englewood Cliffs, New Jersey (1979).



# Index

---

---

- Adjoint matrix, 49  
Adjoint operator, 119  
Adjugate matrix, 49  
Angle between vectors, 23  
Attractors, 262
- Banach space, 117  
Basis of a vector space, 37  
Bessel functions, 130  
Bessel's inequality, 140  
Bifurcation  
    diagram, 279  
    of maps, 290  
    theory, 273  
Biorthogonal set, 54  
Boundary conditions  
    Cauchy, 102  
    Dirichlet, 99, 101  
    Neumann, 100, 101  
    Robin or mixed, 100, 101  
Boundary value problems, 113, 118, 128
- Cauchy sequence, 234  
Chaotic solutions, 300  
Characteristic equation, 51, 271  
Closure, 22  
Complete set, 164  
Complete space, 116  
Conservative systems, 280  
Continuation across parametrically sensitive regions, 252  
Continuation methods, 248  
    obtaining a good first initial guess, 250  
Contraction map, 234  
Correlation dimension, 319  
CSTR stability features—case study, 275
- Determinant, 47  
Diagonalisation of matrix, 79
- Dimension of  
    attractor, 317  
    vector space, 28, 32, 115  
Dirac delta function, 182  
Discrete maps, 240  
Dissipative systems, 280  
Distance between vectors, 22  
Dynamic instability, 277  
Dynamic systems, 240, 243
- Eigen-values  
    cylindrical coordinates, 128  
    spherical coordinates, 135  
    theorems for differential operators, 118  
    theorems for matrices, 49
- Eigen-vectors, 50  
Elliptic boundary condition problem, 105  
Energy methods  
    elliptic problems, 223  
    parabolic problems, 224
- Field, 22  
Finite Fourier transform, 162  
Floquet theory, 320  
Focus, 268  
Fourier  
    coefficients, 140  
    cosine series, 164  
    cosine transform, 170  
    inversion theorem, 170  
    series, 138  
    sine series, 164  
    sine transform, 170  
    transform, 168
- Fredholm  
    alternative, 58  
    alternative elliptic problems, 225
- Frobenius method, 129
- Function  
    domain of, 45  
    range of, 45

- Generating set, 33  
 Gram Schmidt orthonormalisation, 39  
 Greens function  
     adjoint, 182, 195, 204  
     calculation of full eigenfunction expansion method, 196  
     calculation of partial eigenfunction expansion method, 200  
     calculation using Laplace transforms, 210  
     causal, 182  
     different coordinate systems, 193  
     ordinary differential equation, 182, 184  
     partial differential equations, 192, 203  
     physical significance of, 192  
     unbounded domains, 210
- Hard oscillations, 288  
 Hausdorff dimension, 318  
 Hilbert space, 117  
 Homogeneous equation, 2, 97  
 Hopf bifurcation, 274  
 Hyperbolic initial condition problem, 106
- Infinite dimensional space, 113  
 Initial value problem, 7  
 Inner-product space, 28, 115  
 Inversion theorem, 170
- Kronecker delta, 52
- Landau-Hopf scenario, 301  
 Laplace transform, 174  
 Lebesgue integration, 117  
 Legendre functions, 137  
 Length of vector, 23  
 Limit cycle, 273  
 Linear algebraic equations  
     examples of, 3  
     Fredholm alternative, 58  
     method of solution of, 66  
     solvability condition of, 58  
 Linear dependence, 30  
 Linear operator, 47, 102  
 Linear ordinary differential equations  
     applications in process control, 80  
     examples of, 7  
     method of solution of, 68  
 Linear partial differential equations  
     classification of, 97  
     degree of, 94  
     elliptic, 98, 105  
     examples of, 10  
     hyperbolic, 98, 105  
     order of, 94  
     parabolic, 98, 105  
 Linear space, 22  
 Linear stability  
     dynamical systems of, 261, 264  
     maps of, 290  
 Linearity and superposition, 103  
 Local stability, 263  
 Logistic map, 302  
 Lower solution, 228, 230
- Maps, 227  
 Matrices, 46  
 Maximal solution, 229  
 Maximum principles  
     elliptic nonlinear systems, 221  
     elliptic partial differential equations, 216  
     ordinary differential equations, 215  
     parabolic systems, 218  
     physical basis of, 220  
 Metric space, 26, 114  
 Minimal solution, 229  
 Modelling, 1  
 Monotone iteration methods  
     algebraic systems, 227  
     elliptic systems, 230  
     uniqueness conditions, 232, 234
- Necessary conditions, 264, 272  
 Newton-Raphson method, 246  
 Node, 263, 267  
 Nonhomogeneous equation, 2, 97  
     problem, 106  
 Nonlinear equations, examples of  
     algebraic equations, 12  
     ordinary differential equations, 13  
     partial differential equations, 13  
 Normed linear space, 27, 115
- Orthogonal vectors, 30, 124  
 Orthonormal vectors, 30, 52
- Parabolic initial condition problem, 105  
 Parseva's equation, 140  
 Period doubling bifurcation, 291  
 Period doubling cascade, 302  
 Phase-plane, 267  
 Poincare map, 288  
 Power spectrum, 316

Quasi-periodic solutions, 281, 300

Rank, 48

Rayleigh's quotient, 60, 141

Repellers, 262

Riemann integration, 117

Rotation number, 313

Routh-Hurwitz criterion, 273

Ruelle-Takens scenario, 311

Saddle, 268

node bifurcation, 274, 291

Scalar multiplication, 22

Self-adjoint operator, 49

Sensitivity to initial conditions, 305

Separation of variables

cylindrical coordinates, 155

rectangular coordinates, 147

spherical coordinates, 159

Shadowing property, 306

Shooting method, 294

Simulation, 1, 2

Singular matrix, 48

Soft oscillations, 288

Solvability condition, 58

Space time cylinder, 204

Stability condition, 265

Static instability, 277

Steady states numerical evaluation, 245

Stretch and fold of logistic map, 306

Sturm-Liouville problem, 121

Subcritical bifurcation, 281

Sub-harmonic solutions, 300

Subspace, 33

Sufficient conditions, 264, 272

Supercritical bifurcation, 281

Superposition, 102, 103

Torus bifurcation, 291

Transpose, 48

Uniqueness conditions

elliptic problems, 217

maps, 234

parabolic problems, 218

Unstable limit cycle determination, 294

Upper solution, 228, 230

Vector addition, 22

Vector space, 22

Vectors

algebraic representation of, 21

geometric representation of, 21

Wei and Prater problem, 82





# **Mathematical Methods in Chemical Engineering**

**S. PUSHPAVANAM**

This comprehensive, well organized and easy to read book presents concepts in a unified framework to establish a similarity in the methods of solutions and analysis of such diverse systems as algebraic equations, ordinary differential equations and partial differential equations. The distinguishing feature of the book is the clear focus on analytical methods of solving equations. The text explains how the methods meant to elucidate linear problems can be extended to analyse nonlinear problems. The book also discusses in detail modern concepts like bifurcation theory and chaos.

To attract engineering students to applied mathematics, the author explains the concepts in a clear, concise and straightforward manner, with the help of examples and analysis. The significance of analytical methods and concepts for the engineer/scientist interested in numerical applications is clearly brought out.

Intended as a textbook for the postgraduate students in engineering, the book could also be of great help to the research students.

S. PUSHPAVANAM (Ph.D., University of Florida) is Associate Professor, Department of Chemical Engineering, Indian Institute of Technology Madras. Earlier, he taught at Indian Institute of Technology Kanpur. A recipient of the prestigious Fullbright Fellowship, Dr. Pushpavanam pursues research in areas such as applied mathematics, mathematical modelling, chemical engineering and multiphase flow.

**Rs. 275.00**

**Prentice-Hall of India**

New Delhi

[www.phindia.com](http://www.phindia.com)

ISBN 81-203-1262-7

9 788120 312623

A standard barcode is displayed vertically, corresponding to the ISBN number 81-203-1262-7.