

# DIFFERENTIAL EQUATIONS AND MATHEMATICAL BIOLOGY

SECOND EDITION

1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23

D. S. JONES, M. J. PLANK,  
AND B. D. SLEEMAN



CRC Press

Taylor & Francis Group

A CHAPMAN & HALL BOOK

**DIFFERENTIAL  
EQUATIONS AND  
MATHEMATICAL BIOLOGY**

**SECOND EDITION**

# CHAPMAN & HALL/CRC

## Mathematical and Computational Biology Series

### Aims and scope:

This series aims to capture new developments and summarize what is known over the whole spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical and computational methods into biology by publishing a broad range of textbooks, reference works and handbooks. The titles included in the series are meant to appeal to students, researchers and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

### Series Editors

Alison M. Etheridge

*Department of Statistics  
University of Oxford*

Louis J. Gross

*Department of Ecology and Evolutionary Biology  
University of Tennessee*

Suzanne Lenhart

*Department of Mathematics  
University of Tennessee*

Philip K. Maini

*Mathematical Institute  
University of Oxford*

Shoba Ranganathan

*Research Institute of Biotechnology  
Macquarie University*

Hersh M. Safer

*Weizmann Institute of Science  
Bioinformatics & Bio Computing*

Eberhard O. Voit

*The Wallace H. Coulter Department of Biomedical Engineering  
Georgia Tech and Emory University*

Proposals for the series should be submitted to one of the series editors above or directly to:

**CRC Press, Taylor & Francis Group**

4th, Floor, Albert House

1-4 Singer Street

London EC2A 4BQ

UK

# Published Titles

## Algorithms in Bioinformatics: A Practical Introduction

Wing-Kin Sung

## Bioinformatics: A Practical Approach

Shui Qing Ye

## Biological Sequence Analysis Using the SeqAn C++ Library

Andreas Gogol-Döring and Knut Reinert

## Cancer Modelling and Simulation

Luigi Preziosi

## Combinatorial Pattern Matching

## Algorithms in Computational Biology

## Using Perl and R

Gabriel Valiente

## Computational Biology: A Statistical

## Mechanics Perspective

Ralf Blossey

## Computational Neuroscience: A

## Comprehensive Approach

Jianfeng Feng

## Data Analysis Tools for DNA Microarrays

Sorin Draghici

## Differential Equations and Mathematical Biology, Second Edition

D.S. Jones, M.J. Plank, and B.D. Sleeman

## Engineering Genetic Circuits

Chris J. Myers

## Exactly Solvable Models of Biological Invasion

Sergei V. Petrovskii and Bai-Lian Li

## Gene Expression Studies Using Affymetrix Microarrays

Hinrich Göhlmann and Willem Talloen

## Glycome Informatics: Methods and Applications

Kiyoko F. Aoki-Kinoshita

## Handbook of Hidden Markov Models in Bioinformatics

Martin Gollery

## Introduction to Bioinformatics

Anna Tramontano

## An Introduction to Systems Biology: Design Principles of Biological Circuits

Uri Alon

## Kinetic Modelling in Systems Biology

Oleg Demin and Igor Goryanin

## Knowledge Discovery in Proteomics

Igor Jurisica and Dennis Wigle

## Meta-analysis and Combining Information in Genetics and Genomics

Rudy Guerra and Darlene R. Goldstein

## Modeling and Simulation of Capsules and Biological Cells

C. Pozrikidis

## Niche Modeling: Predictions from Statistical Distributions

David Stockwell

## Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems

Qiang Cui and Ivet Bahar

## Optimal Control Applied to Biological Models

Suzanne Lenhart and John T. Workman

## Pattern Discovery in Bioinformatics: Theory & Algorithms

Laxmi Parida

## Python for Bioinformatics

Sebastian Bassi

## Spatial Ecology

Stephen Cantrell, Chris Cosner, and Shigui Ruan

## Spatiotemporal Patterns in Ecology and Epidemiology: Theory, Models, and Simulation

Horst Malchow, Sergei V. Petrovskii, and Ezio Venturino

## Stochastic Modelling for Systems Biology

Darren J. Wilkinson

## Structural Bioinformatics: An Algorithmic Approach

Forbes J. Burkowski

## The Ten Most Wanted Solutions in Protein Bioinformatics

Anna Tramontano



Chapman & Hall/CRC Mathematical and Computational Biology Series

**DIFFERENTIAL  
EQUATIONS AND  
MATHEMATICAL BIOLOGY  
SECOND EDITION**

**D. S. JONES, M. J. PLANK,  
AND B.D. SLEEMAN**



**CRC Press**

Taylor & Francis Group

Boca Raton London New York

---

CRC Press is an imprint of the  
Taylor & Francis Group an **informa** business  
A CHAPMAN & HALL BOOK

MATLAB® and Simulink® are trademarks of The MathWorks, Inc. and are used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® and Simulink® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® and Simulink® software.

CRC Press  
Taylor & Francis Group  
6000 Broken Sound Parkway NW, Suite 300  
Boca Raton, FL 33487-2742

© 2009 by Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works  
Version Date: 20131120

International Standard Book Number-13: 978-1-4200-8358-3 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at  
<http://www.taylorandfrancis.com>

and the CRC Press Web site at  
<http://www.crcpress.com>

To Julie for her continuing support, encouragement and companionship.



---

# Contents

Preface to the First Edition	xiii
Preface to the Second Edition	xv
<b>1 Introduction</b>	<b>1</b>
1.1 Population growth . . . . .	1
1.2 Administration of drugs . . . . .	4
1.3 Cell division . . . . .	9
1.4 Differential equations with separable variables . . . . .	11
1.5 Equations of homogeneous type . . . . .	14
1.6 Linear differential equations of the first order . . . . .	16
1.7 Numerical solution of first-order equations . . . . .	19
1.8 Symbolic computation in MATLAB® . . . . .	24
1.9 Notes . . . . .	27
<b>2 Linear Ordinary Differential Equations with Constant Coefficients</b>	<b>33</b>
2.1 Introduction . . . . .	33
2.2 First-order linear differential equations . . . . .	35
2.3 Linear equations of the second order . . . . .	36
2.4 Finding the complementary function . . . . .	37
2.5 Determining a particular integral . . . . .	41
2.6 Forced oscillations . . . . .	50
2.7 Differential equations of order $n$ . . . . .	52
2.8 Uniqueness . . . . .	55
<b>3 Systems of Linear Ordinary Differential Equations</b>	<b>61</b>
3.1 First-order systems of equations with constant coefficients . . . . .	61
3.2 Replacement of one differential equation by a system . . . . .	64
3.3 The general system . . . . .	66
3.4 The fundamental system . . . . .	68
3.5 Matrix notation . . . . .	72
3.6 Initial and boundary value problems . . . . .	77
3.7 Solving the inhomogeneous differential equation . . . . .	82
3.8 Numerical solution of linear boundary value problems . . . . .	84

<b>4 Modelling Biological Phenomena</b>	<b>91</b>
4.1 Introduction . . . . .	91
4.2 Heartbeat . . . . .	91
4.3 Nerve impulse transmission . . . . .	94
4.4 Chemical reactions . . . . .	100
4.5 Predator-prey models . . . . .	106
4.6 Notes . . . . .	109
<b>5 First-Order Systems of Ordinary Differential Equations</b>	<b>115</b>
5.1 Existence and uniqueness . . . . .	115
5.2 Epidemics . . . . .	118
5.3 The phase plane and the Jacobian matrix . . . . .	119
5.4 Local stability . . . . .	121
5.5 Stability . . . . .	128
5.6 Limit cycles . . . . .	133
5.7 Forced oscillations . . . . .	139
5.8 Numerical solution of systems of equations . . . . .	143
5.9 Symbolic computation on first-order systems of equations and higher-order equations . . . . .	147
5.10 Numerical solution of nonlinear boundary value problems . . . . .	149
5.11 Appendix: existence theory . . . . .	153
<b>6 Mathematics of Heart Physiology</b>	<b>163</b>
6.1 The local model . . . . .	163
6.2 The threshold effect . . . . .	166
6.3 The phase plane analysis and the heartbeat model . . . . .	168
6.4 Physiological considerations of the heartbeat cycle . . . . .	171
6.5 A model of the cardiac pacemaker . . . . .	173
6.6 Notes . . . . .	175
<b>7 Mathematics of Nerve Impulse Transmission</b>	<b>177</b>
7.1 Excitability and repetitive firing . . . . .	177
7.2 Travelling waves . . . . .	185
7.3 Qualitative behaviour of travelling waves . . . . .	187
7.4 Piecewise linear model . . . . .	190
7.5 Notes . . . . .	194
<b>8 Chemical Reactions</b>	<b>197</b>
8.1 Wavefronts for the Belousov-Zhabotinskii reaction . . . . .	197
8.2 Phase plane analysis of Fisher's equation . . . . .	198
8.3 Qualitative behaviour in the general case . . . . .	199
8.4 Spiral waves and $\lambda - \omega$ systems . . . . .	204
8.5 Notes . . . . .	207

<b>9 Predator and Prey</b>	<b>211</b>
9.1 Catching fish . . . . .	211
9.2 The effect of fishing . . . . .	213
9.3 The Volterra–Lotka model . . . . .	215
<b>10 Partial Differential Equations</b>	<b>223</b>
10.1 Characteristics for equations of the first order . . . . .	223
10.2 Another view of characteristics . . . . .	230
10.3 Linear partial differential equations of the second order . . . . .	232
10.4 Elliptic partial differential equations . . . . .	235
10.5 Parabolic partial differential equations . . . . .	239
10.6 Hyperbolic partial differential equations . . . . .	239
10.7 The wave equation . . . . .	240
10.8 Typical problems for the hyperbolic equation . . . . .	245
10.9 The Euler–Darboux equation . . . . .	250
10.10 Visualisation of solutions . . . . .	251
<b>11 Evolutionary Equations</b>	<b>259</b>
11.1 The heat equation . . . . .	259
11.2 Separation of variables . . . . .	262
11.3 Simple evolutionary equations . . . . .	269
11.4 Comparison theorems . . . . .	277
11.5 Notes . . . . .	289
<b>12 Problems of Diffusion</b>	<b>293</b>
12.1 Diffusion through membranes . . . . .	293
12.2 Energy and energy estimates . . . . .	299
12.3 Global behaviour of nerve impulse transmissions . . . . .	304
12.4 Global behaviour in chemical reactions . . . . .	308
12.5 Turing diffusion driven instability and pattern formation . . . . .	311
12.6 Finite pattern forming domains . . . . .	321
12.7 Notes . . . . .	325
<b>13 Bifurcation and Chaos</b>	<b>329</b>
13.1 Bifurcation . . . . .	329
13.2 Bifurcation of a limit cycle . . . . .	334
13.3 Discrete bifurcation and period-doubling . . . . .	336
13.4 Chaos . . . . .	342
13.5 Stability of limit cycles . . . . .	346
13.6 The Poincaré plane . . . . .	350
13.7 Averaging . . . . .	355

<b>14 Numerical Bifurcation Analysis</b>	<b>367</b>
14.1 Fixed points and stability . . . . .	367
14.2 Path-following and bifurcation analysis . . . . .	370
14.3 Following stable limit cycles . . . . .	376
14.4 Bifurcation in discrete systems . . . . .	378
14.5 Strange attractors and chaos . . . . .	380
14.6 Stability analysis of partial differential equations . . . . .	384
14.7 Notes . . . . .	385
<b>15 Growth of Tumours</b>	<b>389</b>
15.1 Introduction . . . . .	389
15.2 Mathematical Model I of tumour growth . . . . .	392
15.3 Spherical tumour growth based on Model I . . . . .	395
15.4 Stability of tumour growth based on Model I . . . . .	399
15.5 Mathematical Model II of tumour growth . . . . .	401
15.6 Spherical tumour growth based on Model II . . . . .	404
15.7 Stability of tumour growth based on Model II . . . . .	406
15.8 Notes . . . . .	407
<b>16 Epidemics</b>	<b>411</b>
16.1 The Kermack–McKendrick model . . . . .	411
16.2 Vaccination . . . . .	413
16.3 An incubation model . . . . .	414
16.4 Spreading in space . . . . .	418
<b>Answers to Selected Exercises</b>	<b>427</b>
<b>Index</b>	<b>439</b>

---

## **Preface to the First Edition**

In recent years, mathematics has made a considerable impact as a tool with which to model and understand biological phenomena. In return biology has confronted the mathematician with a variety of challenging problems which have stimulated developments in the theory of nonlinear differential equations. This book is the outcome of the need to introduce undergraduates of mathematics, the physical and biological sciences to some of these developments. It is primarily directed to university students who are interested in modelling and the application of mathematics to biological and physical situations.

Chapter 1 is introductory, showing how the study of first-order ordinary differential equations may be used to model the growth of a population, monitoring the administration of drugs and the mechanism by which living cells divide. In Chapter 2, a fairly comprehensive account of a linear ordinary differential equation with constant coefficients is given while Chapter 3 extends the theory to systems of equations. Such equations arise frequently in the discussion of the biological models encountered throughout the text. Chapter 4 is devoted to modelling biological phenomena and in particular includes (i) physiology of the heartbeat cycle, (ii) blood flow, (iii) the transmission of electrochemical pulses in the nerve, (iv) the Belousov–Zhabotinskii chemical reaction and (v) predator–prey models.

Nearly all the biological models described in Chapter 4 have special solutions which arise as solutions to first-order autonomous systems of nonlinear differential equations. Chapter 5 gives an account of such systems through the use of the Poincaré phase plane.

With the knowledge of differential equations developed thus far, we are in a position to begin an analysis of the heartbeat, nerve impulse transmission, chemical reactions and predator–prey problems. These are the subjects of Chapters 6–9.

In order to gain a deeper insight into biological models, it is necessary to have knowledge of partial differential equations. These are the subject of Chapters 10 and 11. In particular, a number of the models discussed in Chapter 4 involve processes of diffusion (Chapter 12), and the evolutionary equations considered in Chapter 11 are basic for an understanding of these processes. A special feature of Chapter 12 is a treatment of pattern formation in developmental biology based on Turing’s famous idea of diffusion driven instabilities. The theory of bifurcation and chaotic behaviour is playing an increasing role in fundamental problems of biological modelling. An introduction to these topics is contained in Chapter 13. Chapter 14 models and studies

problems of growth of solid avascular tumours. Again differential equations play a fundamental part. However, a new feature here is that we encounter moving boundary problems. The book concludes in Chapter 15 with a discussion of epidemics and the spread of infectious diseases, modelled via various differential equations.

As an encouragement to further study, some of the chapters have notes indicating sources of material as well as references to additional literature. Each chapter has a set of exercises which either illustrate some of the ideas discussed or require readers to develop and test models of their own.

In writing the book, the authors have endeavoured to give it a multipurpose role. For example, it can be used (i) as a course in differential equations based on Chapters 1, 2, 3, 5, 10 and 11; (ii) as a course in biological modelling for students of mathematics and the physical sciences or (iii) as a course in differential equation models of biology for life science students based on Chapters 1, 2, 3, 5, 10, 11, 12 and 13, together with a selection of the remaining chapters depending on the students' interests. Throughout the stages of writing this book, the authors have benefited from discussions and advice from colleagues in the Department of Mathematics, University of Dundee and the School of Mathematics, University of Leeds. To them all, we express our appreciation.

Finally, it is a pleasure to thank Nick Hill, Mel Holmes, Michael Plank, Doreen Ross and David Sleeman for all their efforts in connection with the preparation of this book.

D. S. JONES and B. D. SLEEMAN  
Dundee and Leeds

---

## Preface to the Second Edition

Since the first appearance of *Differential Equations and Mathematical Biology*, published by Allen and Unwin in 1983, mathematical biology has become a mainstream branch of applied mathematics and many student programs in colleges and universities across the world now offer mathematical biology and systems biology courses. Over this past quarter of a century, theoretical research in biology has intensified leading to significant advances in many fields. The book grew out of the need to introduce students in the physical, mathematical and biological sciences to some of fundamental modelling and analytical techniques with which to deepen understanding of biological phenomena.

In the first edition published by CRC Press in 2003, and in light of biological developments, some reassessment of the material was undertaken. For example the chapter on catastrophes was replaced by a discussion of bifurcation and chaos. The chapter on problems of diffusion was expanded to include an introductory account of diffusion driven instability and Turing patterns. A further important addition was an introduction to the symbolic algebra package *Mathematica*®.

The current edition represents a continuation of this evolutionary process. Many of the chapters have been expanded to include new and topical material. For instance Chapter 8 now includes a section on spiral waves and Chapter 9 has been expanded to include some recent developments in tumour biology. The most significant part of the evolutionary process has been to include much more on the numerical solution of differential equations and numerical bifurcation analysis. An ideal tool for such computational approaches is the programming package MATLAB®, which is now widely used in the mathematical and systems biology communities.

Many more examples and exercises have been added which we hope will be useful to students.

In the course of writing this edition the authors have benefitted from discussions and cogent criticisms from colleagues and students both in the School of Mathematics, University of Leeds and the Department of Mathematics and Statistics, University of Canterbury. Thanks to them all.

All MATLAB® files found in the book are available for download from the publisher's Web site. MATLAB® is a registered trademark of The MathWorks, Inc. For product information, please contact:

The MathWorks, Inc.  
3 Apple Hill Drive  
Natick, MA 01760-2098 USA  
Tel: 508 647 7000  
Fax: 508 647 7001  
E-mail: [info@mathworks.com](mailto:info@mathworks.com)  
Web: [www.mathworks.com](http://www.mathworks.com)

D. S. JONES, M. J. PLANK and B. D. SLEEMAN  
Dundee, Canterbury and Leeds  
July 2009

---

## About the Authors

**Douglas S. Jones**, a Fellow of the Royal Society and of the Royal Society of Edinburgh, is Professor Emeritus, formerly Ivory Professor, of mathematics in the University of Dundee. His many contributions to applied mathematics and in particular to electromagnetic theory have been recognized by a number of awards including the Naylor Prize of the London Mathematical Society, the van der Pol Gold Medal of the International Union of Radio Science, the Keith Prize of the Royal Society of Edinburgh, the Marconi Prize of the Institute of Electrical Engineers and an Honorary Doctorate of the University of Strathclyde. Professor Jones is an Honorary Fellow of Corpus Christi College Oxford and past president of the Institute of Mathematics and its Applications.

**Michael J. Plank** is a senior lecturer in Mathematics at the University of Canterbury, Christchurch, New Zealand. He completed his PhD in applied mathematics under the supervision of Professor Sleeman at the University of Leeds in 2003. He moved to the University of Canterbury in 2004, first as a postdoctoral fellow, and subsequently as a lecturer from 2006. He is actively researching in a number of areas of mathematical biology and was recently awarded an Emerging Researcher Award by the University of Canterbury.

**Brian D. Sleeman**, a Fellow of the Royal Society of Edinburgh, is Professor Emeritus, formerly professor of applied mathematics, in the University of Leeds. In recognition of his contributions to applied analysis and in particular to mathematical biology he holds honorary professorships at the University of Dundee, the University of Abertay Dundee and the University of Wuhan, Peoples Republic of China. He is also a member of Clare Hall, University of Cambridge. Professor Sleeman is a former Erskine Visiting Fellow of the University of Canterbury, New Zealand and past president of the Edinburgh Mathematical Society. He is also founding Editor of the *Journal Computational and Mathematical Methods in Medicine*.



# Chapter 1

---

## Introduction

---

### 1.1 Population growth

To indicate why the study of differential equations can be useful some simple examples will be considered.

The way that the size of a population varies in time is a matter of interest in several contexts. Let the number of individuals in a given area at time  $t$  be  $p(t)$ . At time  $t+T$  the number of individuals is  $p(t+T)$  so that  $p(t+T) - p(t)$  must be the number of individuals that have been added to the population during the time interval  $T$ . The longer the interval the more individuals can be expected to arrive and the shorter the time the fewer additions can be expected. So write the change in the time interval  $T$  as  $NT$  and then

$$p(t+T) - p(t) = NT$$

or

$$\frac{p(t+T) - p(t)}{T} = N.$$

Letting  $T \rightarrow 0$  we see that the left-hand side becomes the derivative of  $p$  with respect to  $t$ . Consequently, we have

$$\frac{dp(t)}{dt} = N.$$

To gain an idea of the properties of  $N$  suppose that the change in size of the population is due entirely to individuals being born. As time progresses the fertility of parents may alter so that more or less offspring are born. Then the number born in the interval  $T$  may vary as time proceeds. In other words  $N$  may alter as  $t$  does. Another effect is introduced by the simple hypothesis that the more individuals there are at time  $t$  the more births are likely to occur. Then  $N$  will depend on  $p(t)$  also. Both possibilities can be allowed for by rewriting our equation as

$$\frac{dp(t)}{dt} = N\{t, p(t)\} \quad (1.1.1)$$

to show explicitly quantities on which  $N$  depends.

Sometimes

$$\frac{1}{p} \frac{dp}{dt}$$

is known as the **specific growth rate** or **per capita growth rate**. So another way of describing (1.1.1) is to say that the specific growth rate is  $N(t, p)/p$ .

It is plausible to assume that, in a short time interval, there will be about twice as many births as in a time interval of half its length. Thus, one could expect that the number of births would be proportional to  $p(t)T$  when  $T$  is small. If the fertility of the parents does not change, the actual number of births in the time interval  $T$  can be expressed as  $N_0 p(t)T$  with  $N_0$  a suitable constant. Then (1.1.1) becomes

$$\frac{dp(t)}{dt} = N_0 p(t) \quad (1.1.2)$$

which states that the specific growth rate is  $N_0$ , the same for all times and all sizes of population.

The solution of (1.1.2) is

$$p(t) = p_0 e^{N_0 t} \quad (1.1.3)$$

where  $p_0$  is any constant. This may be confirmed by taking a derivative of (1.1.3) with respect to  $t$ . The value of  $p_0$  can be fixed by putting  $t = 0$  in (1.1.3); evidently  $p_0$  is the size of the population at  $t = 0$ . By rearranging (1.1.3), it can be seen that

$$t = \frac{1}{N_0} \ln \frac{p(t)}{p_0}.$$

When  $p(t) = ep_0$ , we get  $t = 1/N_0$ , and hence  $1/N_0$  is the time taken for the population to increase by a factor of  $e$ .

Another way of verifying the result in (1.1.3) is to integrate (1.1.2) with respect to  $t$ . Thus

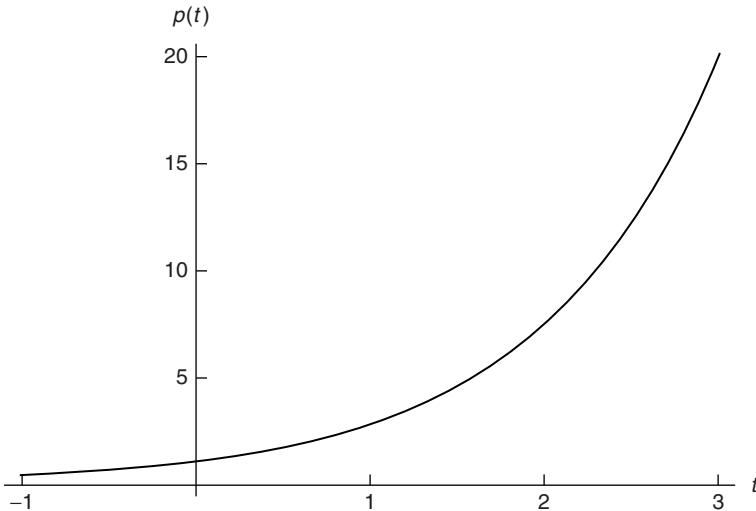
$$\begin{aligned} N_0 \int_0^t dt &= \int_0^t \frac{1}{p} \frac{dp}{dt} dt \\ &= \int_{p(0)}^{p(t)} \frac{dp}{p} \end{aligned}$$

on changing the variable of integration from  $t$  to  $p$ . Hence

$$N_0 t = \ln\{p(t)/p(0)\} \quad (1.1.4)$$

which agrees with (1.1.3).

The behaviour of the population as time increases according to (1.1.3) is displayed in Figure 1.1.1. The size grows steadily, and the increase becomes dramatic as time goes on. Of course, in any real situation, there will be a limit



**FIGURE 1.1.1:** Graph of exponential growth.

to the growth because of a shortage of essential supplies, insufficient food and self-pollution of the environment. Nevertheless, many organisms exhibit exponential growth in their initial stages. It is always easy to check whether a population is growing exponentially by plotting  $\ln p$  against time; a straight line should be obtained. The slope of the line is the specific growth rate  $N_0$  as is clear from (1.1.4).

Equation (1.1.2) has been derived on the assumption that only births occur. In the event that there are deaths but no births the same equation can be reached. However,  $N_0$  is now a negative number since the population decreases in the time interval  $T$ . It follows from (1.1.3) that the population decays exponentially with time from its size at  $t = 0$ .

More facets of the population problem can be incorporated. For instance, we may postulate that the number of deaths in the short time interval  $T$  is  $D_0 p(t)T$ . Similarly, individuals may enter the given area from outside, say  $I(t)T$  immigrants in the interval  $T$ . Likewise, some may depart from the area giving rise to  $E(t)T$  emigrants. Then

$$p(t + T) - p(t) = N_0 p(t)T - D_0 p(t)T + I(t)T - E(t)T$$

leading to

$$\frac{dp(t)}{dt} = (N_0 - D_0)p(t) + I(t) - E(t) \quad (1.1.5)$$

when  $T \rightarrow 0$ .

More generally,  $I$  and  $E$  could be made to depend on  $p$  so that (1.1.5), which is often called **Verhulst's differential equation**, can be difficult to

solve. In fact, (1.1.5) could be regarded as a particular case of (1.1.1) with an appropriate definition of  $N$  but that does not make it any easier to solve. Notwithstanding this, it is transparent that, if we hope to predict the size of a population at a given time, finding the solution of a differential equation will be an essential requirement.

---

## 1.2 Administration of drugs

When a drug is administered, it forms a concentration in the body fluids. This concentration diminishes in time through elimination, destruction or inactivation. The rate of reduction of the concentration is found, in most cases, to be proportional to the concentration. Therefore, if  $c(t)$  is the concentration at time  $t$ , and there is no external input of the drug, we have

$$\frac{d}{dt}c(t) = -\frac{c(t)}{\tau}, \quad (1.2.1)$$

where  $\tau$  is a constant which measures the rapidity with which the concentration falls. This simple model describes drug concentration in a single compartment (such as the blood plasma), which is assumed to be well mixed (i.e., the concentration throughout the compartment is assumed to be the same at any given point in time).

Exactly the same differential equation can be derived from (1.1.5) for a population which changes only by deaths which happen at a constant specific rate. So conclusions about drugs can be transferred easily to populations which alter by death alone.

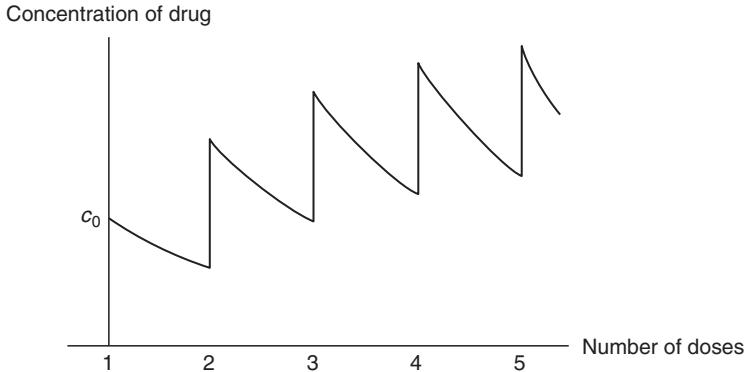
Analogous to (1.1.3) the solution of (1.2.1) is

$$c(t) = c_0 e^{-t/\tau}, \quad (1.2.2)$$

where  $c_0$  is the concentration at time  $t = 0$ . Notice that when  $t = \tau$  the concentration has dropped to  $c_0/e$ , so that in time  $\tau$  the concentration has been reduced to  $1/e$  of its initial value. This explains the significance of the time  $\tau$ ; the larger it is, the more slowly the drug disperses.

According to (1.2.2) the drug never disappears completely from the body except after infinite time. However, the residual concentration will usually be negligible when  $t \gg \tau$ . Notwithstanding this, the fact that some of the drug is always left is relevant when repeated doses are made, as is common practice. The level to which the drug accumulates is then of particular importance.

Suppose that a dose is administered regularly at the times  $t = 0, t_0, 2t_0, 3t_0, \dots$ . It is assumed that each dose raises the drug concentration by a fixed amount  $c_0$ , and that this change in concentration happens instantaneously



**FIGURE 1.2.1:** Concentration of drug initially.

(i.e., all of the drug administered is instantly absorbed into the fluid compartment). At time  $t_0$  the residual concentration  $r_1$  **just before the second dose is**

$$r_1 = c_0 e^{-t_0/\tau},$$

and then the second dose is given so that the total concentration  $c_1$  is given by

$$c_1 = c_0 + c_0 e^{-t_0/\tau}.$$

From  $t = t_0$  to  $t = 2t_0$  the concentration will fall exponentially so that the residual concentration  $r_2$  **just before the third dose at**  $t = 2t_0$  is

$$r_2 = c_1 e^{-t_0/\tau} = c_0 e^{-t_0/\tau} (1 + e^{-t_0/\tau}).$$

After the third dose the concentration  $c_2$  is

$$c_2 = c_0 + r_2 = c_0 (1 + e^{-t_0/\tau} + e^{-2t_0/\tau}).$$

Clearly, if we keep doing this, we shall discover that at  $t = (n - 1)t_0$ , with  $n$  a positive integer,

$$\begin{aligned} c_{n-1} &= c_0 (1 + e^{-t_0/\tau} + e^{-2t_0/\tau} + \dots + e^{-(n-1)t_0/\tau}) \\ &= c_0 \frac{1 - e^{-nt_0/\tau}}{1 - e^{-t_0/\tau}} \end{aligned} \tag{1.2.3}$$

on summing the geometric series. The residual concentration  $r_n$  at  $t = nt_0$  will be

$$r_n = c_{n-1} e^{-t_0/\tau} = c_0 e^{-t_0/\tau} \frac{1 - e^{-nt_0/\tau}}{1 - e^{-t_0/\tau}} \tag{1.2.4}$$

from (1.2.3).

The manner in which the drug builds up as the number of doses increases is illustrated in Figure 1.2.1. The level of concentration grows in an oscillatory

fashion. At first sight it looks as though there is no limit to the concentration. To check whether this is true examine (1.2.3). The only term involving  $n$  is  $e^{-nt_0/\tau}$  in the numerator. Since this term decreases as  $n$  increases,  $c_n$  does increase with  $n$ . However, the growth is not unlimited because the exponential decays to zero. Thus the concentration never exceeds  $c_M$  where

$$c_M = \frac{c_0}{1 - e^{-t_0/\tau}}. \quad (1.2.5)$$

The concentration becomes  $c_M$  only in the limit as  $n \rightarrow \infty$ , yet  $c_M$  is a good estimate of the concentration immediately after a dose when  $nt_0/\tau \gg 1$ . Indeed, if  $nt_0 > 5\tau$ ,  $c_{n-1}$  differs from  $c_M$  by less than 1% so that, unless  $t_0/\tau$  is small, the level of concentration will not be far from  $c_M$  after a few doses. To put it another way, if you want to reach the maximum concentration in about 5 doses, you should make the interval between doses larger than  $\tau$ . Naturally, the larger  $t_0/\tau$  is made the closer the maximum becomes to  $c_0$ , the concentration of a single dose.

Similarly, just before a dose the residue approaches  $r$  as  $n$  increases, where  $r$  is given by

$$r = c_M e^{-t_0/\tau} = \frac{c_0}{e^{t_0/\tau} - 1}. \quad (1.2.6)$$

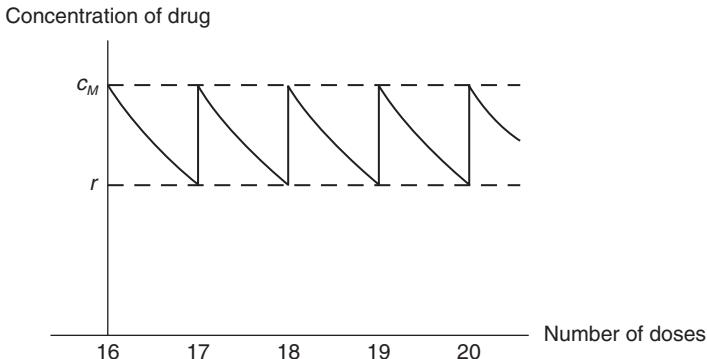
Observe that  $c_M = c_0 + r$ . Notice also that (1.2.6) implies that  $r$  becomes small when  $t_0/\tau$  is large. The larger  $t_0/\tau$  the more the level of concentration varies between doses. Thus, there is a trade-off between keeping the residue above a certain level and reaching  $c_M$  in a few doses.

When sufficient doses have been administered for (1.2.5) and (1.2.6) to be good approximations the concentration behaves as in Figure 1.2.2. It swings between  $c_M$  and  $r$ , never exceeding  $c_M$  nor falling below  $r$ .

The oscillatory build-up of Figure 1.2.1 may be undesirable. Several antibiotics can have harmful effects until their concentration has surpassed a certain threshold, since sub-optimal concentrations may induce resistance to the drug by the micro-organisms. The oscillatory growth can be avoided by taking advantage of the behaviour in Figure 1.2.2. An initial large dose of  $c_0 + r$  or  $c_M$  is given and thereafter doses of  $c_0$  are supplied at intervals of  $t_0$ . The first dose takes the concentration to  $c_M$  and from then on the level follows the curve of Figure 1.2.2. So long as  $r$  is above any threshold imposed the difficulty referred to has been surmounted.

### **Example 1.2.1**

A drug is available in doses of  $M$  mg, which raise the blood plasma concentration of an average adult by  $10 \text{ mg l}^{-1}$ . Once in the blood plasma, the drug concentration decays according to equation (1.2.1) with  $\tau = 4 \text{ h}$ . Health regulations require that the concentration never exceed  $15 \text{ mg l}^{-1}$ . What is shortest safe time interval at which doses can be given regularly? What is the minimum concentration if doses are given at this time interval? In general, what will be the minimum concentration of a drug that decays according to



**FIGURE 1.2.2:** Concentration of drug after many doses.

(1.2.1), and is given in doses that raise plasma concentration by  $c_0$  at regular intervals to ensure that its concentration never exceeds  $c_M$ ?

Rearranging equation (1.2.5) to find  $t_0$ , the time interval between doses in terms of maximum safe concentration  $c_M$ , the dose concentration  $c_0$  and the time constant  $\tau$  gives

$$t_0 = \tau \ln \left( \frac{c_M}{c_M - c_0} \right).$$

Substituting in the values  $c_M = 15 \text{ mg l}^{-1}$ ,  $c_0 = 10 \text{ mg l}^{-1}$  and  $\tau = 4 \text{ h}$  gives  $t_0 = 4.4 \text{ h}$ . If the drug is given regularly at this time interval, the lowest concentration will occur immediately before the second dose at  $t = t_0$ . The concentration at this time will be

$$c(t_0) = c_0 e^{-t_0/\tau} = 3.3 \text{ mg l}^{-1}.$$

In the general case, the minimum concentration always occurs just before the second dose at  $t = t_0$ . Using the expression above for  $t_0$  gives

$$c(t_0) = c_0 e^{-t_0/\tau} = c_0 \left( 1 - \frac{c_0}{c_M} \right).$$

In other words, the drug concentration will fall to a minimum of  $1 - c_0/c_M$  of its administered concentration.  $\square$

### Example 1.2.2

Suppose a virus enters the blood stream and develops at a rate proportional to its concentration. An antidote to the virus is administered at a time  $h$  and decays according to the law (1.2.1). Write down a simple model of virus-antidote interaction and estimate the dosage of antidote needed to eradicate the viral infection as quickly as possible.

Let the virus concentration be denoted by  $c(t)$  and the antidote concentration by  $a(t)$ . A basic model is represented by:

$$\begin{aligned}\frac{dc}{dt} &= kc - pa, \\ \frac{da}{dt} &= -\frac{a}{\tau},\end{aligned}\tag{1.2.7}$$

where  $k > 0$  is the rate of proportionality and  $p$  is the rate at which the antidote kills virus cells.

At  $t = 0$  suppose  $c(0) = c_0$ . For  $0 \leq t < h$  we assume  $a(t) = 0$  (which means that  $c(h) = c_0 e^{kh}$ ) and at  $t = h$ ,  $a(h) = a_0$ .

From (1.2.2) we find that  $a(t)$  is given by;

$$\begin{aligned}a(t) &= 0, \quad 0 \leq t < h, \\ &= a_0 e^{\frac{h-t}{\tau}}, \quad t \geq h.\end{aligned}$$

Substituting this result into the equation for  $c$  we have

$$\frac{dc}{dt} - kc = -pa_0 e^{\frac{h-t}{\tau}} - kt, \quad t \geq h.$$

This equation can be rearranged as

$$\frac{d(ce^{kt})}{dt} = -pa_0 e^{\frac{h-t}{\tau} - kt}$$

(a general method for doing this will be explained in Section 1.6), which on integrating both sides with respect to  $t$  gives

$$c(t)e^{-kt} - c_0 = -pa_0 e^{\frac{h}{\tau}} \int_h^t e^{-(k+1/\tau)x} dx,$$

or

$$c(t) = \left( c_0 - \frac{pa_0 e^{-kh}}{(k+1/\tau)} \right) e^{kt} + \frac{pa_0}{(k+1/\tau)} e^{\frac{(h-t)}{\tau}}.\tag{1.2.8}$$

For  $c(t)$  to decrease it is clear that the antidote administered at  $t = h$  should satisfy the condition

$$a_0 > \frac{c_0}{p} (k + 1/\tau) e^{kh}.$$

This shows that the virus will be eradicated at a time  $T > h$  satisfying the equation

$$\begin{aligned}e^{(k+1/\tau)(T-h)} &= \frac{(pa_0 - c_0(k+1/\tau)e^{kh})}{pa_0}, \\ &= 1 - \frac{c_0}{pa_0} (k + 1/\tau) e^{kh}.\end{aligned}\tag{1.2.9}$$

Clearly the larger the dose  $a_0$ , the quicker the virus  $c(t)$  is cleared from the blood stream. □

### 1.3 Cell division

When cells divide, their numbers grow by a process akin to that of Section 1.1. A new feature is that the multiplication in numbers is restricted by crowding effects. Biochemically, these may be due to lack of nutrient, shortage of oxygen, change in *pH* or the production of inhibitors, for example. Whatever the cause, the cells are interacting with one another. Since each cell can interact with  $p$  others there are  $p^2$  possibilities. This suggests that, in (1.1.1), we should put

$$N\{t, p(t)\} = N_0 p(t) - ap(t)^2 \quad (1.3.1)$$

where  $N_0$  and  $a$  are positive constants. The term involving  $N_0$  is the same as before and accounts for the increase due to division. The term containing  $a$  represents the inhibition on growth caused by crowding. With the substitution (1.3.1), (1.1.1) gives

$$\frac{dp}{dt} = N_0 p - ap^2 \quad (1.3.2)$$

which is called the **logistic equation**.

If we integrate (1.3.2) from 0 to  $t$ , as in Section 1.1, we obtain

$$\begin{aligned} \int_0^t dt &= \int_0^t \frac{1}{N_0 p - ap^2} \frac{dp}{dt} dt \\ &= \int_{p(0)}^{p(t)} \frac{dp}{N_0 p - ap^2}, \end{aligned}$$

when  $p$  is employed as the variable of integration. Now,

$$\int \frac{dp}{N_0 p - ap^2} = \frac{1}{N_0} \int \left( \frac{1}{p} - \frac{a}{ap - N_0} \right) dp = \frac{1}{N_0} \ln \left( \frac{p}{ap - N_0} \right).$$

Hence we have

$$t = \frac{1}{N_0} \ln \left( \frac{p(t)\{ap(0) - N_0\}}{\{ap(t) - N_0\}p(0)} \right)$$

whence

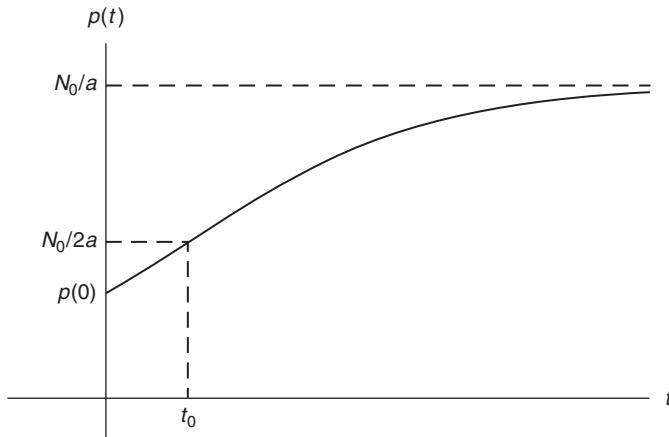
$$p(t)\{ap(0) - N_0\} = \{ap(t) - N_0\}p(0)e^{N_0 t}.$$

Consequently,

$$p(t) = \frac{N_0 p(0)}{ap(0) + \{N_0 - ap(0)\}e^{-N_0 t}} \quad (1.3.3)$$

which is known as the **logistic law of growth**.

The curve of logistic growth is shown in Figure 1.3.1, assuming that  $N_0 > ap(0)$ . The curve rises steadily from the value  $p(0)$  at  $t = 0$  to an eventual value of  $N_0/a$ , there being no maxima or minima in between. There may,



**FIGURE 1.3.1:** The curve of logistic growth.

however, be a point of inflexion where the curve crosses its tangent at  $t = t_0$  where

$$t_0 = \frac{1}{N_0} \ln \left( \frac{N_0}{ap(0)} - 1 \right)$$

and  $p(t_0) = N_0/2a$ .

Observe that the final value  $N_0/a$  of  $p$  does not involve  $p(0)$ , so that, no matter what the initial size of the population, its final size is always the same and does not depend on the starting size of the population. This final value imposes a limit on growing populations and is referred to as the **carrying capacity** of the environment.

### Example 1.3.1

Suppose a population of cells grows logically according to equation (1.3.1), with a per capita growth rate of  $N_0 = 0.1 \text{ h}^{-1}$  (i.e., a small population that is not approaching carrying capacity will take approximately 10 h to increase by a factor of  $e$ ). If the initial size of the population is 10% of the carrying capacity, how long will it take for the population to reach 95% of the carrying capacity?

The solution (1.3.3) to the logistic equation can be rewritten, by multiplying the numerator and denominator by  $1/a$  as:

$$p(t) = \frac{Kp(0)}{p(0) + (K - p(0)) e^{-N_0 t}},$$

where  $K = N_0/a$  is the carrying capacity. Rearranging to find  $t$  in terms of  $p(t)$  gives

$$t = \frac{1}{N_0} \ln \frac{p(t)(K - p(0))}{p(0)(K - p(t))}.$$

Substituting in  $N_0 = 0.1 \text{ h}^{-1}$  and the conditions  $p(0) = 0.1K$  and  $p(t) = 0.95K$ , we find  $t = 51.4 \text{ h}$ .  $\square$

The logistic law assumes that all the cells divide at the same rate. This is not always true. There are types in which some cells divide faster than others. Whether the logistic law can be applied still depends upon the differences between the various rates of division present. If the rates are not too far apart it is probably feasible to take  $N_0$  as their average. For greater deviations it may be necessary to adopt a model in which the statistics of the number of cells of a given age and type at a given time play a part.

---

## 1.4 Differential equations with separable variables

Having seen from the preceding sections that models lead naturally to differential equations, we shall devote the rest of this chapter to investigating some of the methods of solution for differential equations. In general, a differential equation picked out of a hat will be insoluble, at least in terms of elementary functions. However, progress can be made with particular types and attention will be concentrated on those which yield to analytical attack.

The first type to be considered is that which can be written

$$f(y) \frac{dy}{dt} = g(t). \quad (1.4.1)$$

It is called a **differential equation of the first order with separable variables**. The phrase “of the first order” signifies that the only derivative which is present, namely  $dy/dt$ , is of the first order. Integrate both sides of (1.4.1) to obtain

$$\int_{t_0}^t f(y) \frac{dy}{dt} dt = \int_{t_0}^t g(t) dt$$

where  $t_0$  is some constant. If  $t_1$  is another constant

$$\int_{t_0}^t g(t) dt = \int_{t_1}^t g(t) dt + \int_{t_0}^{t_1} g(t) dt$$

and the second term on the right-hand side is just a constant. Therefore we can write

$$\int^t f(y) \frac{dy}{dt} dt = \int^t g(t) dt + C \quad (1.4.2)$$

without indicating the lower limit of integration (other than that it is a constant) providing that the constant  $C$  is left at our disposal. On the left-hand side of (1.4.2) change the variable of integration to  $y$  and then

$$\int^y f(y) dy = \int^t g(t) dt + C. \quad (1.4.3)$$

If (1.4.1) were expressed formally as

$$f(y)dy = g(t)dt,$$

(1.4.3) could be obtained by placing an integral sign on each term and adding the constant  $C$ .

A value cannot be assigned to  $C$  unless  $y$  is prescribed for some value of  $t$ . Thus a general solution such as (1.4.3) is bound to involve an arbitrary constant.

### **Example 1.4.1**

Find the general solution of

$$(T^2 - t^2) \frac{dy}{dt} + ty = 0,$$

$T$  being a constant.

Using the formal approach described above, we rewrite the differential equation as

$$\frac{dy}{y} + \frac{tdt}{T^2 - t^2} = 0,$$

so that the general solution is

$$\int^y \frac{dy}{y} + \int^t \frac{tdt}{T^2 - t^2} = C.$$

By carrying out the integrations, we obtain

$$\ln y - \frac{1}{2} \ln |T^2 - t^2| = C$$

whence

$$y = e^C |T^2 - t^2|^{1/2}.$$

Put  $e^C = K$  and then the general solution is

$$y = K |T^2 - t^2|^{1/2}, \quad (1.4.4)$$

where now  $K$  is the arbitrary constant.

If it is known that  $y = 1$  when  $t = T/2$ , substitution in (1.4.4) gives

$$1 = KT\sqrt{3}/2$$

from which is deduced that  $K = 2/T\sqrt{3}$ . Thus

$$y = \frac{2|T^2 - t^2|^{1/2}}{T\sqrt{3}} \quad (1.4.5)$$

is the solution of the differential equation which takes the value 1 when  $t = T/2$ . □

**Example 1.4.2**

Find the general solution of

$$\frac{dy}{dt} = \frac{ky}{y+M},$$

where  $k$  and  $M$  are constants.

Rearranging the equation to separate the variables  $y$  and  $t$  gives

$$\frac{y+M}{y} dy = k dt$$

and integrating both side of this equation gives

$$y + M \ln y = kt + C,$$

where  $C$  is an arbitrary constant. This can be easily solved to give  $t$  in terms of  $y$ , but there is no explicit solution for  $y$  in terms of  $t$ . The value of  $C$  is again determined by some given initial condition. For example, if  $y = y_0$  when  $t = 0$  then

$$y_0 + M \ln y_0 = C,$$

which leads to

$$y - y_0 + M \ln(y/y_0) = kt.$$

□

**Example 1.4.3**

Find the general solution of the differential equation

$$\frac{dy}{dt} = y^2.$$

Rearrange the equation as in the previous example to get

$$\frac{1}{y^2} dy = dt$$

and integrate both sides to obtain

$$y = -\frac{1}{t+c}$$

where  $c$  is an arbitrary constant.

This example has the important feature that the solution may not exist for all time. For instance if at  $t = 0$ ,  $y = y_0 > 0$  then

$$y(t) = \frac{y_0}{(1-y_0 t)}$$

and so  $y(t)$  cannot exist beyond the time  $t = 1/y_0$ .

□

All the solutions derived in the first three sections were based on separable equations but the existence of Verhulst's equation (1.1.5) indicates that other types need to be discussed.

## 1.5 Equations of homogeneous type

A differential equation that has the form

$$\frac{dy}{dt} = f\left(\frac{y}{t}\right), \quad (1.5.1)$$

where  $f(y/t)$  is a function of the single variable  $y/t$ , is said to be of **homogeneous type**.

To solve (1.5.1), make the substitution

$$y = tz. \quad (1.5.2)$$

Then, by the product rule for taking a derivative,

$$\frac{dy}{dt} = z + t\frac{dz}{dt}$$

and so (1.5.1) becomes

$$z + t\frac{dz}{dt} = f(z)$$

or

$$t\frac{dz}{dt} = f(z) - z. \quad (1.5.3)$$

In (1.5.3) the variables are separable, and so it may be solved as in Section 1.4. All that the reader needs to remember, therefore, is to make the change of variable (1.5.2) when an equation of homogeneous type is met.

### **Example 1.5.1**

Find the general solution of

$$\frac{dy}{dt} = \left(\frac{y}{t}\right)^2.$$

Making the substitution (1.5.2) gives

$$z + t\frac{dz}{dt} = z^2.$$

We rearrange this equation to separate the variables  $z$  and  $t$ :

$$\frac{1}{z(z-1)} dz = \frac{1}{t} dt.$$

Using partial fractions to expand the left-hand side of this equation, and integrating, gives

$$\int \left( \frac{1}{z-1} - \frac{1}{z} \right) dz = \int \frac{1}{t} dt.$$

Carrying out the integration gives

$$\ln \left| \frac{z-1}{z} \right| = \ln t + C,$$

which leads to

$$z = \frac{1}{1 \pm Kt},$$

where  $K = e^C$ . Finally, substituting for  $z$  according to (1.5.2) gives the solution in the original variables  $y$  and  $t$ :

$$y = \frac{t}{1 \pm Kt}.$$

Note that, if we allow the constant  $K$  to be positive or negative, the  $\pm$  may be replaced by a  $+$ . □

### **Example 1.5.2**

Find the general solution of

$$(3t - y) \frac{dy}{dt} + t = 3y. \quad (1.5.4)$$

Since the differential equation can be written as

$$\frac{dy}{dt} = \frac{3y - t}{3t - y} = \frac{3(y/t) - 1}{3 - (y/t)},$$

it is of homogeneous type. With the substitution (1.5.2), we have

$$z + t \frac{dz}{dt} = \frac{3z - 1}{3 - z},$$

which implies that

$$t \frac{dz}{dt} = \frac{z^2 - 1}{3 - z}.$$

Hence the general solution is

$$\int \frac{3-z}{z^2-1} dz = \int \frac{dt}{t} + C.$$

Since

$$\frac{3}{z^2-1} = \frac{3}{2} \left( \frac{1}{z-1} - \frac{1}{z+1} \right),$$

we obtain

$$\frac{3}{2} \ln \left| \frac{z-1}{z+1} \right| - \frac{1}{2} \ln |z^2-1| = \ln t + C.$$

Consequently

$$\frac{|z - 1|}{(z + 1)^2} = Kt$$

where  $K = e^C$ . Substituting for  $z$  from (1.5.2), we obtain

$$y - t = K(y + t)^2 \quad (1.5.5)$$

as the general solution in the original variables. □

---

## 1.6 Linear differential equations of the first order

A differential equation that can be expressed as

$$\frac{dy}{dt} + f(t)y = g(t), \quad (1.6.1)$$

where  $y$  does not occur in either  $f(t)$  or  $g(t)$ , is called a **linear differential equation of the first order**.

Why the nomenclature “linear” is employed can be understood from consideration of (1.6.1) when  $g$  is replaced by zero so that the differential equation becomes

$$\frac{dy}{dt} + f(t)y = 0.$$

Suppose that  $y_1$  and  $y_2$  are solutions of this differential equation, i.e.,

$$\frac{dy_1}{dt} + f(t)y_1 = 0 \quad (1.6.2)$$

$$\frac{dy_2}{dt} + f(t)y_2 = 0. \quad (1.6.3)$$

The addition of (1.6.2) and (1.6.3) shows that  $y_1 + y_2$  is also a solution. Actually, if  $A$  and  $B$  are constants,  $Ay_1 + By_2$  is a solution. This construction of solutions by adding together constant multiples of solutions is the property of linearity. In nonlinear differential equations, the sum of two solutions cannot be asserted to be a solution without independent verification. Another linear property is that, if  $y_1 = 1$  when  $t = t_0$  and  $y_2 = C$  when  $t = t_0$ , then  $y_2(t) = Cy_1(t)$ . Thus, if the initial value of a solution is increased, all subsequent values are increased in the same proportion. The problem of cell division in Section 1.3 led to a solution which after a long time was the same for all initial populations and so the logistic growth equation is nonlinear, as is also evident from the structure of (1.3.2).

The technique for solving (1.6.1) is to multiply by  $h(t)$  to form

$$h(t) \frac{dy}{dt} + h(t)f(t)y = h(t)g(t).$$

The function  $h$  is to be chosen so that the left-hand side is the same as  $d(hy)/dt$ ; it is then known as an **integrating factor**.

Now

$$\frac{d}{dt}(hy) = h \frac{dy}{dt} + \frac{dh}{dt}y$$

which is the same as the left-hand side provided that

$$\frac{dh}{dt} = hf.$$

In this differential equation for  $h$  the variables are separable and so

$$\ln h(t) = \int^t f(u)du + C_1$$

where  $C_1$  is a constant, the lower limit of integration being omitted for the reasons given in deriving (1.4.2). Hence

$$h(t) = \exp\left(C_1 + \int^t f(u)du\right).$$

However, if  $h(t)$  is an integrating factor so is  $C_2h(t)$  when  $C_2$  is any constant. Therefore we can take for our integrating factor

$$h(t) = \exp\left(\int^t f(u)du\right). \quad (1.6.4)$$

The differential equation, after multiplication by the integrating factor, can be written as

$$\frac{d}{dt}(hy) = hg$$

with the general solution

$$h(t)y(t) = \int^t h(v)g(v)dv + C.$$

Substitution from (1.6.4) gives

$$y(t) = \left[ \int^t g(v) \exp\left(\int^v f(u)du\right) dv + C \right] \exp\left(-\int^t f(u)du\right). \quad (1.6.5)$$

The rule for the reader to remember for a linear equation of the first order is to multiply by the integrating factor defined by (1.6.4), making sure first

that the differential equation is in the form (1.6.1). It is not worth committing (1.6.5) to memory.

**Example 1.6.1**

Find the general solution of

$$(t^2 + 1) \frac{dy}{dt} + ty = \frac{1}{2}. \quad (1.6.6)$$

The differential equation is linear and can be rearranged into the form of (1.6.1) with

$$f(t) = \frac{t}{t^2 + 1}, \quad g(t) = \frac{1}{2} \frac{1}{t^2 + 1}.$$

Therefore

$$\int f(t) dt = \int \frac{t}{t^2 + 1} dt = \frac{1}{2} \ln(t^2 + 1)$$

and, according to (1.6.4), the integrating factor is

$$h(t) = \exp\left\{\frac{1}{2} \ln(t^2 + 1)\right\} = (t^2 + 1)^{1/2}.$$

Hence, multiply

$$\frac{dy}{dt} + \frac{t}{t^2 + 1} y = \frac{1}{2} \frac{1}{t^2 + 1}$$

by  $(t^2 + 1)^{1/2}$  with the result

$$(t^2 + 1)^{1/2} \frac{dy}{dt} + \frac{t}{(t^2 + 1)^{1/2}} y = \frac{1}{2} \frac{1}{(t^2 + 1)^{1/2}}.$$

Consequently

$$\frac{d}{dt}\{(t^2 + 1)^{1/2} y\} = \frac{1}{2} \frac{1}{(t^2 + 1)^{1/2}},$$

which gives on integration

$$(t^2 + 1)^{1/2} y = \frac{1}{2} \ln\{t + (t^2 + 1)^{1/2}\} + C \quad (1.6.7)$$

where  $C$  is an arbitrary constant. Equation (1.6.7) provides the general solution of (1.6.6).  $\square$

**Example 1.6.2**

Find the general solution of

$$\frac{dy}{dt} = 3 + t - \frac{2y}{t}. \quad (1.6.8)$$

The differential equation is linear with

$$f(t) = \frac{2}{t}, \quad g(t) = 3 + t.$$

According to (1.6.4), the integrating factor is

$$h(t) = \exp(2 \ln t) = t^2.$$

Multiplying both sides of the differential equation (1.6.8) by  $h(t)$ :

$$\frac{d}{dt}(t^2 y) = t^2(3 + t).$$

Integrating and rearranging to find  $y$  in terms of  $t$  gives the general solution of (1.6.6):

$$y = t + \frac{t^2}{4} + \frac{C}{t^2},$$

where  $C$  is an arbitrary constant. □

Observe that, if  $f(t)$  has the constant value  $f_0$ , the integrating factor is  $e^{f_0 t}$  so that the general solution of (1.6.1) is then

$$y(t) = e^{-f_0 t} \int^t g(x) e^{f_0 x} dx + C e^{-f_0 t}, \quad (1.6.9)$$

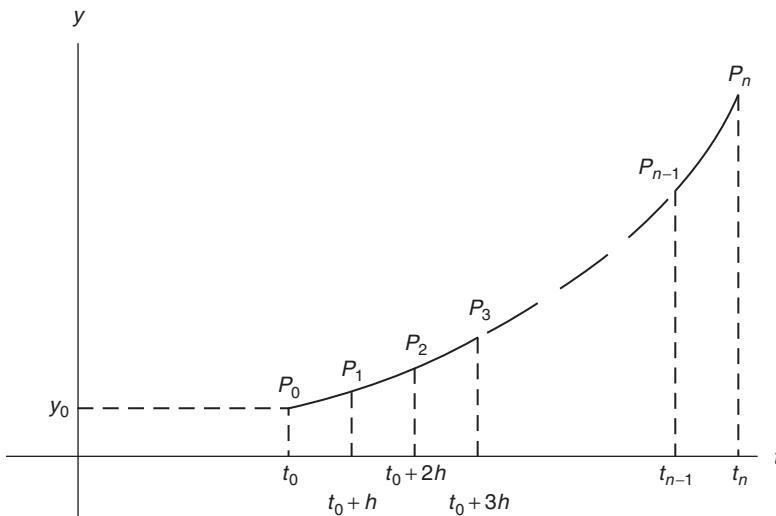
$C$  being an arbitrary constant.

## 1.7 Numerical solution of first-order equations

Many differential equations, particular nonlinear equations, are difficult or impossible to solve analytically. In such cases, it is necessary to turn to numerical techniques to find an **approximation** to the true solution. A typical differential equation of the first order can be expressed as

$$\frac{dy}{dt} = f(t, y). \quad (1.7.1)$$

Precise attributes of the function  $f$  that will guarantee that the differential equation does possess a solution are not of concern here. It will be sufficient to suppose that, corresponding to each pair of values  $(t, y)$ , there is a definite value of  $f$  and that small changes of  $t, y$  will be accompanied by only a small variation in  $f$ . Then, if a point  $(t, y)$  in the  $(t, y)$ -plane is chosen, (1.7.1) asserts that at  $(t, y)$  the derivative  $dy/dt$  has the value  $f(t, y)$ . In other words, (1.7.1) assigns a direction to a point of the  $(t, y)$ -plane. If a curve can be drawn so that, at each of its points, its gradient satisfies (1.7.1), then this curve will have an equation which is a solution of the differential equation. Any such curve may be called a **solution curve** of the differential equation.



**FIGURE 1.7.1:** Polygonal approximation to solution curve.

An approximate solution curve can be constructed graphically. Suppose it is required that  $y = y_0$  when  $t = t_0$  (this is the initial condition for the differential equation). Take the point  $(t_0, y_0)$  in the  $(t, y)$ -plane as  $P_0$  (Figure 1.7.1). Draw the straight line through  $P_0$  with slope  $f(t_0, y_0)$  and let it intersect  $t = t_0 + h$  at  $P_1$ . Let  $P_1$  be the point  $(t_0 + h, y_1)$ . At  $P_1$  draw the straight line of slope  $f(t_0 + h, y_1)$  and let it intersect  $t = t_0 + 2h$  at  $P_2$ , with ordinate  $y_2$ . Draw the straight line  $P_2P_3$  with slope  $f(t_0 + 2h, y_2)$ . Continuing in this way, we construct a polygonal curve  $P_0P_1P_2P_3\dots$  which has the slope prescribed by the differential equation at the points  $t = t_0, t_0 + h, t_0 + 2h, \dots$ . If  $h$  is kept small, the slope of the polygonal curve will never deviate by much from that demanded by the differential equation. It would therefore seem that, as  $h \rightarrow 0$ , the polygonal curve would tend to the solution curve of the differential equation which passes through  $(t_0, y_0)$ , i.e., provide the solution of (1.7.1) such that  $y(t_0) = y_0$ . Since only one slope can be drawn at each point, there is only one solution curve.

The procedure can be defined iteratively as follows

$$t_{n+1} = t_n + h, \quad y_{n+1} = y_n + h f(t_n, y_n).$$

We anticipate that  $y_1, y_2 \dots$  will be reasonable approximations to the values of the solution curve at  $t_1, t_2, \dots$ , though it must be stressed that we have not proved that this is so.

This approximate technique is known as **Euler's method** and is straightforward to implement in MATLAB® (or any other computer programming language). The following is an example of a MATLAB function called `euler` that uses Euler's method to solve a differential equation of the form (1.7.1).

```

function [t, y] = euler(dydt, tspan, y0, Nsteps)
% Function to solve a single first-order DE by Euler's method
% INPUTS: dydt - handle to a function f(t, y) that calculates
%          dy/dt
% tspan - 1x2 row vector containing the starting and
%         finishing values of t
% y0 - initial value of y
% Nsteps - number of steps of Euler's method to use
% OUTPUTS: t - vector of equally spaced t values
%           y - vector of corresponding y values
dt = (tspan(2)-tspan(1))/Nsteps;
t = (tspan(1):dt:tspan(2))'; % vector of equally spaced t values
y = zeros(Nsteps+1, 1); % set up vector to store y values
y(1) = y0; % store initial condition in y(1)
% Loop through t values calculating new y value each time:
for I = 1:Nsteps
    y(I+1) = y(I) + dt*dydt(t(I), y(I));
end

```

This program should be saved in a file called `euler.m`.

Although Euler's method is simple to use, it is not particularly efficient (it is often necessary to use a very small step size to obtain an accurate solution). However, it is the prototype for a suite of more advanced numerical methods of solving a differential equation by a step-by-step process that are capable of providing solutions to any required degree of accuracy. Indeed, there are methods that automatically adjust themselves so as to achieve the level of accuracy prescribed by the user at the outset. MATLAB has several built-in functions that make use of these more sophisticated methods. One of the most commonly used of these is `ode45` (see below for an example of usage), which uses adaptive step size to ensure an accurate solution, so there is no need for the user to specify the number of steps to take (or equivalently the step size  $h$ ).

The above MATLAB function, `euler`, can be thought of as a black box. Specify the starting and finishing  $t$  values, and the number of steps to take, and supply some means of calculating  $f(t, y)$  on the right-hand side of (1.7.1), and `euler` will calculate the solution. It remains to write a short function file that calculates  $f(t, y)$  for any given values of  $t$  and  $y$  (sometimes it is first necessary to rearrange the equation so that it is in the form (1.7.1)). This is illustrated in the following example.

### **Example 1.7.1**

$$\frac{dy}{dt} = -\frac{Qy^p}{K + y^p} \quad (1.7.2)$$

is called a Hill equation, and arises in enzymatic reactions (see Section 4.4). This equation is difficult to solve in general for  $y$  as an explicit function of  $t$ .

For parameter values of say  $Q = 1$ ,  $K = 5$  and  $p = 3.3$ , calculate and plot an approximate solution from  $t = 0$  to  $t = 10$ , with an initial condition  $y(0) = 5$ .

The function `euler` needs to know how to calculate  $dy/dt$  for any given values of  $t$  and  $y$ . We must therefore provide it with a MATLAB function that does this. The input arguments to this function are  $t$  and  $y$ , and the output is  $dy/dt$ , calculated according to equation (1.7.2). We must also give this MATLAB function a name, say `hill_fn` (which means that the file name should be `hill_fn.m`). The function should read as follows.

```
function f = hill_fn(t, y)
Q = 1; K = 5; p = 3.3;
f = -Q*y^p/(K+y^p);
```

To solve the differential equation using Euler's method with say 10 steps, the following commands may be entered at the MATLAB command prompt:

```
>> t0 = 0; t1 = 10; y0 = 5;
>> [t, y] = euler(@hill_fn, [t0, t1], y0, 10);
```

Note the four arguments (listed in the round brackets) to `euler` in the above command correspond to the four inputs in the `euler` program listed above. The first argument `@hill_fn` is called a function handle, and tells `euler` the name of the function to use to evaluate the right-hand side of the differential equation (this doesn't always have to be `@hill_fn` – it can be any appropriate function  $f(t, y)$ ). The second argument `[t0, t1]` tells `euler` to solve from  $t=t0$  to  $t=t1$ . The third and fourth arguments respectively specify the initial value of  $y$ , and the number of steps to use.

Alternatively, to use the built-in MATLAB function `ode45`, the command calling `euler` can be replaced by

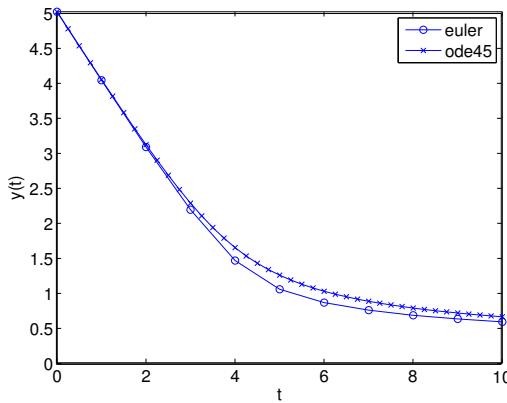
```
>> [t2, y2] = ode45(@myrhs, [t0, t1], y0);
```

Note that the arguments to `ode45` are the same as to `euler`, except that there is no fourth argument to `ode45` since there is no need to specify the number of steps to take. This time, the outputs from the function are stored in `t2` and `y2`, so the results from `euler` are not overwritten.

Now we would like to plot the results.

```
>> plot(t, y, 'o-', t2, y2, 'x-')
>> axis([t0 t1 0 y0])
>> legend('euler', 'ode45')
>> xlabel('t')
>> ylabel('y(t)')
```

The `plot` command draws two curves on the same graph: one of  $y$  against  $t$ , and one of  $y2$  against  $t2$  (you can include as many pairs of vectors in the graph as you like). The '`o-`' and '`x-`' arguments tell MATLAB to mark the positions of the points with symbols (`o` for the first curve and `x` for the



**FIGURE 1.7.2:** Numerical solution to equation (1.7.2) with  $Q = 1$ ,  $K = 5$ ,  $p = 3.3$  and  $y(0) = 5$  calculated using Euler's method (`euler`) and MATLAB's built-in function (`ode45`).

second), and connect them with solid lines (-). The `axis`, `legend` and `label` commands customize the plot, by adjusting the limits of the axes, and adding a legend and axis labels. Running the above commands gives the graph shown in Figure 1.7.2. The solution calculated by `ode45` is more accurate than the one calculated by `euler`.  $\square$

Sometimes we are less interested in the solution  $y$  as a function of the independent variable  $t$  than in the time at which the solution reaches a certain value. For instance, in the example above, we might want to know the time at which  $y$  first drops below 1. The following MATLAB function can be used to find all the times at which  $y$  is equal to a predefined value  $y_{\text{crit}}$  ( $y$  is assumed to be a continuous function of  $t$ ).

```
function th = hitting_times(t, y, y_crit)
% Function to find the times at which y = y_crit
% INPUTS: t - vector of increasing t values
%          y - vector of corresponding y values
%          y_crit - critical y value
% OUTPUTS: th - vector of times at which y = y_crit
N = length(t);
th = zeros(N,1);
hits = 0;
% Loop through the vector of t and y values
for I = 1:N-1
    % Test if y crosses y_crit between t(I) and t(I+1)
    if (y(I)-y_crit)*(y(I+1)-y_crit) <= 0
```

```

hits = hits+1;
% Interpolate between t(I) and t(I+1) to find the
% approximate hitting time
th(hits) = t(I)+(t(I+1)-t(I))*(y_crit-y(I))/(y(I+1)-y(I));
end
th = th(1:hits); % truncate th to the number of hits found

```

### **Example 1.7.2**

Find the time at which the solution from `ode45` shown in Figure 1.7.2 falls below  $y = 1$ .

This can be calculated by the following command.

```
>> th = hitting_times(t2, y2, 1)
th = 6.1240
```



## 1.8 Symbolic computation in MATLAB®

Although primarily designed for numerical calculations, MATLAB can also be implemented to perform computations using symbolic algebra. This feature can be used to obtain analytical solutions to some differential equations. A full exposition of MATLAB's symbolic computation power is beyond the scope of this book (for further details, see the MATLAB help files). Here, we take a brief look at the basic commands and syntax, and a few examples of how they are used to solve differential equations (see also Section 5.9 for solving higher-order equations and systems of equations).

MATLAB's symbolic representation of the derivative  $dy/dt$  is simply `Dy`. Similarly  $d^2y/dt^2$  is represented by `D2y`, etc. The command for solving a differential equation is `dsolve`. The basic syntax is

```
y_sol = dsolve('ode', 'ic1', 'ic2', ... , 'var')
```

where `ode` is the equation to be solved, `ic1`, `ic2` ... are the initial conditions, and `var` is the independent variable. The initial conditions may be omitted, in which case MATLAB will attempt to find the general solution. If the independent variable is omitted, it will be assumed that the independent variable is `t`.

**Example 1.8.1**

To solve the separable equation

$$\frac{dy}{dt} = \frac{2t}{(1+t^2)y}, \quad (1.8.1)$$

the appropriate MATLAB command is

```
>> y_sol = dsolve('Dy = 2*t/((1+t^2)*y)')
```

This produces the following output

```
y_sol =
(2*log(1+t^2)+C1)^(1/2)
-(2*log(1+t^2)+C1)^(1/2)
```

Here, `y_sol` is a vector containing the two branches of the solution. The symbol `C1` is an arbitrary constant. This tells us that the general solution to (1.8.1) is

$$y(t) = \pm (2 \ln(1 + t^2) + C_1)^{\frac{1}{2}}. \quad (1.8.2)$$

If we have a specific initial condition, say  $y(0) = -1$ , we can provide this information to MATLAB in the `dsolve` command:

```
>> y_sol = dsolve('Dy = 2*t/((1+t^2)*y)', 'y(0) = -1')
y_sol = -(2*log(1+t^2)+1)^(1/2)
```

Note that this corresponds to the negative branch of the general solution (1.8.2) with  $C_1 = 1$ .

In order to plot the solution, we need to generate vectors containing actual  $t$  and  $y$  values. In order to do this, we first declare `t` as a symbolic variable.

```
>> syms t
```

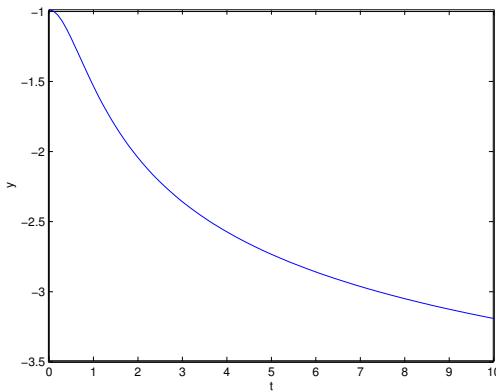
Now we define a vector of  $t$  values to plot (let's try  $0 \leq t \leq 10$ , at intervals of 0.1, and call it `t_plot`). Then we get MATLAB to calculate the corresponding  $y$  values, by substituting the values in `t_plot` into the symbolic variable `t`:

```
>> t_plot = [0:0.1:10];
>> y_plot = subs(y_sol, t, t_plot);
>> plot(t_plot, y_plot)
```

The resulting plot is shown in Figure 1.8.1.

**Example 1.8.2**

To solve equation (1.5.4), the appropriate MATLAB input and output is



**FIGURE 1.8.1:** Plot of the solution (1.8.2) to equation (1.8.1) with initial condition  $y(0) = -1$ .

```
>> y_sol = dsolve('(3*t-y)*Dy+t=3*y')
y_sol =
-1/2*(2*C1*t-1-(-8*C1*t+1)^(1/2))/C1
-1/2*(2*C1*t-1+(-8*C1*t+1)^(1/2))/C1
```

Hence, the general solution is

$$y = -\frac{1}{2C_1} \left( 2C_1 t - 1 \pm (1 - 8C_1 t)^{\frac{1}{2}} \right),$$

which you may recognise as the roots of the quadratic equation (1.5.5), which was found as the general solution in Example 1.5.2.  $\square$

### Example 1.8.3

If we try to solve the Hill equation (1.7.2) symbolically, this is what happens:

```
>> y_sol = dsolve('Dy=-Q*y^p/(K+y^p)')
Warning: Explicit solution could not be found; implicit solution
         returned.
> In dsolve at 315
y_sol =
y = 0
t+1/Q*y-1/Q/(p-1)*y^(1-p)*K+C1 = 0
```

MATLAB cannot find  $y$  as an explicit function of  $t$  (excluding the trivial solution  $y = 0$ ). Instead, it returns an implicit solution

$$t + \frac{y}{Q} - \frac{Ky^{1-p}}{Q(p-1)} + C_1 = 0. \quad (1.8.3)$$

Numerical techniques are needed to obtain explicit  $y$  values. This may be done either by solving equation (1.8.3) by a root finding method, such as Newton's method, for each value of  $t$  required, or by solving the original equation (1.7.2) numerically, as described in Section 1.7.

□

---

## 1.9 Notes

For a specialist text on numerical techniques, the reader is referred to J. D. Lambert, *Computational Methods in Ordinary Differential Equations*, John Wiley & Sons, New York, 1973 and *Numerical Methods for Ordinary Differential Systems*, John Wiley & Sons, New York, 1991. For a text on implementing numerical methods in MATLAB, see S. C. Chapra, *Applied Numerical Methods with MATLAB*, McGraw-Hill, New York, 2005. For a guide to the MATLAB programming language, see D. J. Higham and N. J. Higham, *MATLAB Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2005.

---

## Exercises

- 1.1 A spherical water drop loses volume by evaporation at a rate proportional to its surface area. Express its radius at time  $t$  in terms of the constant of proportionality and its radius  $r_0$  at  $t = 0$ .
- 1.2 The rate of increase of bacteria in a culture is proportional to the number present. The population multiplies by the factor  $n$  in the time interval  $T$ . Find the number of bacteria at time  $t$  when the initial population is  $p_0$ .
- 1.3 In Exercise 1.2 the population is found to increase by 2455 bacteria from  $t = 2$  to  $t = 3$  and by 4314 bacteria from  $t = 4$  to  $t = 5$ . Show that  $p_0 = 4291$  approximately and that, when  $T = 3$ ,  $n$  is about 2.33.
- 1.4 What changes take place in the curve of logistic growth in Figure 1.3.1 if  $N_0 < ap(0)$ ?
- 1.5 Observations on animal tumours indicate that their sizes obey the

### Gompertz growth law

$$\frac{ds}{dt} = ks \ln\left(\frac{S}{s}\right)$$

rather than the logistic law. Here  $k$  and  $S$  are positive constants. By putting  $y = \ln s$  prove that

$$s(t) = S \exp(-Ae^{-kt})$$

where  $A = \ln(S/s_0)$ ,  $s_0$  being the size at  $t = 0$ . Deduce that, in Gompertz growth, the size moves steadily from its initial value to an eventual value of  $S$  without passing through maxima or minima, though there is a point of inflection if  $s_0 < Se^{-1}$ .

- 1.6 Find the general solutions of the following equations and check your answer using MATLAB's symbolic computation.

$$\begin{aligned} (a) \quad & t \frac{dy}{dt} = (1+t)y, \\ (b) \quad & t(2y+3) \frac{dy}{dt} = y(3+y), \\ (c) \quad & 2ty(1+t) \frac{dy}{dt} = 1+y^2. \end{aligned}$$

- 1.7 Find the general solutions of the following equations and check your answer using MATLAB's symbolic computation.

$$\begin{aligned} (a) \quad & \frac{dy}{dt} = \frac{t}{y} + \frac{y}{t}, \\ (b) \quad & \frac{dy}{dt} = \left(\frac{t}{y}\right)e^{-y/t} + \frac{y}{t}, \\ (c) \quad & t^2 \frac{dy}{dt} = ty - y^2. \end{aligned}$$

- 1.8 (a) By means of the substitution  $w - 1 = y, u + 2 = t$  show that

$$\frac{dw}{du} = \frac{u+w+1}{u-w+3}$$

can be solved.

- (b) Generalise this result to

$$\frac{dw}{du} = \frac{au+bw+c}{a'u+b'w+c'}$$

when  $ab' \neq a'b$  by substituting  $w = y + h, u = t + k$  where

$$\begin{aligned} ak + bh + c &= 0, \\ a'k + b'h + c' &= 0. \end{aligned}$$

1.9 Find the general solutions of

- (a)  $(1-t^2)\frac{dy}{dt} - ty = (1-t^2)^{1/2}$ ,
- (b)  $\frac{dy}{dt} = (t-4)e^{4t} + ty$ ,
- (c)  $tw\frac{dw}{dt} = t^4 + w^2$  by putting  $w^2 = y$ .

In each case, write a MATLAB program to verify your solution by solving the differential equation numerically with an initial condition  $y(0) = 1$ , and plotting the numerical and the analytical solutions on the same graph.

1.10 **Bernoulli's differential equation** has the form

$$\frac{dy}{dt} + f(t)y = g(t)y^\nu.$$

Show that it can be made linear by the substitution  $w = y^{1-\nu}$ . Hence find the general solution of

$$\frac{dy}{dt} + ty = ty^2.$$

1.11 Use Exercise 1.10 to solve

- (a)  $(t+1)\left(y\frac{dy}{dt} - 1\right) = y^2$ ,
- (b)  $(t^2 - 2y + 1)\frac{dy}{dt} = t$ .

1.12 The function  $y_1(t)$  is known to be a solution of the **Riccati equation**

$$\frac{dy}{dt} + a(t)y + b(t)y^2 = c(t).$$

Show that the general solution can be found by putting  $y = y_1 + w$  and using Exercise 1.10. Hence solve

$$\frac{dy}{dt} + 2y^2 = \frac{6}{t^2}$$

given that  $y = 2/t$  is a solution.

1.13 Use Exercise 1.12 to find the general solutions of

- (a)  $t^2\frac{dy}{dt} + ty + t^2y^2 = 4$  given that  $y = 2/t$  is a solution,
- (b)  $t\frac{dy}{dt} - (2t+1)y + y^2 = -t^2$  given that  $y = t$  is a solution.

1.14 Find the general solutions of

- (a)  $(t^2 - 1)\frac{dy}{dt} + y = \frac{1}{2}(t^2 - 1)^{1/2}$ ,
- (b)  $t\frac{dy}{dt} = y + (\frac{1}{4}t^2 + y^2)^{1/2}$ ,
- (c)  $2\frac{dy}{dt} = e^{t+2y}$ ,
- (d)  $(2t + 2y + 5)\frac{dy}{dt} = 2y - 2t + 1$ ,
- (e)  $\frac{dy}{dt} = 2(2t - y)^2$ ,
- (f)  $(t + a)\frac{dy}{dt} - 3y = 2(t + a)^5$ ,
- (g)  $t\frac{dy}{dt} = y - \frac{1}{2}t \cos^2(2y/t)$ ,
- (h)  $(2y - 2t + 5)\frac{dy}{dt} = 2y - 2t + 1$ ,
- (i)  $\frac{dy}{dt} + (1 - 4y^2) \tan 2t = 0$ ,
- (j)  $(1 - t^2)\frac{dy}{dt} - \frac{1}{2}(1 + t)y = (1 - t^2)^{1/2}$ ,
- (k)  $(t + y)^2\frac{dy}{dt} = (t + y + 1)^2$ ,
- (l)  $1 + t\frac{dy}{dt} = e^{-y}\frac{dy}{dt} \sec^2 y$ .

1.15 According to Newton's law of cooling, the rate of decrease of temperature of a body is proportional to the difference between its temperature and that of its environment. If the temperature of the environment is 20°C and the body cools from 80°C to 60°C in 1 h, show that it will take somewhat over 4 h to cool to 30°C. Verify your answer by solving the differential equation numerically in MATLAB and plotting the temperature against time.

1.16 A body cools in 10 min from 100°C to 60°C when the environment is at 20°C. How long does it take to cool to 25°C?

1.17 After administration of a dose, the concentration of a drug decreases by 50% in 30 h. How long does it take to fall to 1% of its initial value?

1.18 The amount of light absorbed by a layer of material is proportional to the incident light and to the thickness of the layer. If a layer 35 cm thick absorbs half the light incident on its surface, what percentage of the incident light will be absorbed by a layer 200 cm thick?

- 1.19 When the drug theophylline is administered for asthma, a concentration below  $5 \text{ mg l}^{-1}$  has little effect and undesirable side-effects appear if the concentration exceeds  $20 \text{ mg l}^{-1}$ . For a body that weighs  $W \text{ kg}$ , the concentration when  $M \text{ mg}$  is present is  $2M/W \text{ mg l}^{-1}$ . If the constant that measures the rapidity at which the concentration falls is  $\tau = 6 \text{ h}$ , find the concentration at time  $t \text{ h}$  after an initial dose of  $D \text{ mg}$ .  
 If  $D = 500$  and  $W = 70$ , show that a second dose is necessary after about  $6 \text{ h}$  to prevent the concentration from becoming ineffective. What further time can elapse before a third dose is necessary?  
 What is the shortest safe time interval  $t_0$  at which doses of  $500 \text{ mg}$  can be given regularly?  
 Verify your results by solving the problem numerically.
- 1.20 If the drug in Exercise 1.19 is fed into the blood stream continuously by infusion at a rate of  $D_1 \text{ mg h}^{-1}$ , instead of being given by separate doses, show that the concentration approaches a steady level of  $12D_1/W \text{ mg l}^{-1}$ . What permissible range of  $D_1$  does this imply for a 60-kg patient?
- 1.21 In the reservoir model of the heart, it is imagined as a balloon. The balloon or heart is blown up by the influx of blood during the systole and, when the heart valve closes, the reservoir forces blood out during the diastole. At time  $t$ , the volume is  $v(t)$ , the inflow per unit time is  $I(t)$  and the outflow per unit time is  $F(t)$ . Interpret

$$dv/dt = I(t) - F(t).$$

If the pressure  $p(t)$  is such that

$$p(t) = K\{v(t) - v_0\}, \quad F(t) = p(t)/R,$$

where  $K, v_0$  and  $R$  are constants, find the differential equation satisfied by  $p$ .

In the diastole which lasts from  $t = t_0$  to  $t = T, I(t) = 0$ . Find  $p(t)$  in terms of  $p_0$ , the value of  $p(t_0)$ , during the time interval. In the systole, from  $t = 0$  to  $t = t_0, I(t) = I_0$  (a constant). Find  $p$ .

Since the heart is cyclic,  $p(0) = p(T)$ . Deduce that

$$R = \frac{p_0}{I_0} \frac{1 - \exp(-KT/R)}{1 - \exp(-Kt_0/R)}.$$



# Chapter 2

---

## Linear Ordinary Differential Equations with Constant Coefficients

---

### 2.1 Introduction

In Chapter 1, the solution of a first-order differential equation was considered. Higher derivatives can occur in some problems, and so we are led to the general ordinary differential equation

$$F\left(t, y, \frac{dy}{dt}, \frac{d^2y}{dt^2}, \dots, \frac{d^n y}{dt^n}\right) = 0, \quad (2.1.1)$$

where  $F$  is some function with  $n + 2$  arguments. It is called **ordinary** because it involves only the ordinary derivatives of  $y$  with respect to the single variable  $t$ . Later on we will study cases in which  $y$  is a function of more than one variable. In that case, partial derivatives of  $y$  can arise and a partial differential equation has to be solved.

The **order** of an ordinary differential equation is the order of the highest derivative appearing. Thus

$$\frac{d^4y}{dt^4} = \frac{dy}{dt} \left( \frac{d^3y}{dt^3} \right)^4 + y^2$$

is of order 4, whereas

$$\left( \frac{dy}{dt} \right)^2 = t^2 + y^2$$

is of order 1.

The two main categories into which ordinary differential equations are classified are **linear** and **nonlinear**. The form of the general linear ordinary differential equation of order  $n$  is

$$a_n(t) \frac{d^n y}{dt^n} + a_{n-1}(t) \frac{d^{n-1} y}{dt^{n-1}} + \dots + a_1(t) \frac{dy}{dt} + a_0(t)y = f(t), \quad (2.1.2)$$

where  $a_0(t), \dots, a_n(t)$  are known functions of  $t$ . If all of  $a_0(t), \dots, a_n(t)$  are constants, (2.1.2) is known as a **linear ordinary differential equation with**

**constant coefficients.** Any ordinary differential equation which does not have the structure of (2.1.2) is called nonlinear; it will contain products such as

$$y^2, \quad \frac{dy}{dt} \frac{d^2y}{dt^2}$$

or functions such as  $e^y$ . For example,

$$t \frac{d^2y}{dt^2} + e^t \frac{dy}{dt} + y \cos t = t^3 \tan t$$

is linear and of order 2,

$$5 \frac{d^2y}{dt^2} + 4 \frac{dy}{dt} + 3y = \ln^2 t$$

is linear with constant coefficients and of order 2, while

$$y^2 \frac{dy}{dt} = t$$

is nonlinear.

The solution of an ordinary differential equation is always sought on an interval  $(a, b)$  ( $a < b$ ) of  $t$ . It is a relation between  $y$  and  $t$  which satisfies the ordinary differential equation when  $t$  is any point of the interval and does not contain any derivatives or integrals of  $y$ . Integrals of functions of  $t$  may be involved but these should be evaluated when it is reasonable to do so.

The **general solution** (sometimes called the **complete primitive**) of an ordinary differential equation of order  $n$  must contain  $n$  arbitrary constants. Any solution that does not have  $n$  arbitrary constants is not the general solution. For instance, you can check that

$$y = e^t - 1/t \tag{2.1.3}$$

satisfies the ordinary differential equation

$$t^3 \left( \frac{d^2y}{dt^2} - y \right) = t^2 - 2, \tag{2.1.4}$$

but it is not the general solution because the general solution must contain two arbitrary constants whereas (2.1.3) has none. Similarly

$$y = Ce^{-t} - 1/t, \tag{2.1.5}$$

with  $C$  an arbitrary constant, is a solution of (2.1.4) but is not the general solution. On the other hand,

$$y = C_1 e^t + C_2 e^{-t} - 1/t, \tag{2.1.6}$$

with  $C_1$  and  $C_2$  arbitrary constants, is the general solution of (2.1.4).

When additional information is available, it may be possible to assign particular values to the arbitrary constants in the general solution. For example,

suppose that a solution of (2.1.4) is desired such that  $y = 1$  and  $dy/dt = 1$  when  $t = 1$ . From the general solution (2.1.6)

$$dy/dt = C_1e^t - C_2e^{-t} + 1/t^2,$$

and so the conditions at  $t = 1$  can be satisfied if

$$\begin{aligned}C_1e + C_2e^{-1} - 1 &= 1, \\C_1e - C_2e^{-1} + 1 &= 1.\end{aligned}$$

These require that  $C_1 = e^{-1}$  and  $C_2 = e$ ; hence

$$y = e^{t-1} + e^{1-t} - 1/t$$

is the solution of (2.1.4) which takes the correct values at  $t = 1$ .

An ordinary differential equation may not possess a solution. There are theorems, called **existence theorems**, which tell you that certain types of differential equations have a solution. If a solution exists and you can prove that it is the only one that satisfied any conditions imposed, then you have demonstrated a **uniqueness theorem**. Existence and uniqueness are beyond the scope of this chapter (for some information, see Section 5.11).

Even when existence and uniqueness theory is available, the actual finding of a solution may be a difficult task. For instance, there is an existence theorem for linear ordinary differential equations with variable coefficients, but the solution cannot always be written down easily. Again, the solution of

$$\frac{dy}{dt} = t^2 + y^2$$

cannot be expressed in terms of elementary functions although the solution is known to exist. Therefore, one must turn to numerical techniques, as discussed for example in Section 1.7, to find an approximation to the solution.

This chapter will be confined to discussing linear ordinary differential equations with constant coefficients. For these, not only is existence theory available but also the general solution can be determined explicitly.

## 2.2 First-order linear differential equations

The linear ordinary differential equation with constant coefficients of the first order is

$$a_1 \frac{dy}{dt} + a_0 y = f(t)$$

where  $a_1$  and  $a_0$  are constants. However,  $a_1 \neq 0$  otherwise there would not be a differential equation to solve. Therefore we can divide by  $a_1$  or, equivalently,

put  $a_1 = 1$  and take as the standard form

$$\frac{dy}{dt} + a_0 y = f(t). \quad (2.2.1)$$

This has been solved already in Section 1.6 and we may quote the general solution

$$y = e^{-a_0 t} \int^t e^{a_0 u} f(u) du + C e^{-a_0 t}, \quad (2.2.2)$$

where  $C$  is an arbitrary constant.

The term  $C e^{-a_0 t}$  satisfies

$$\left( \frac{d}{dt} + a_0 \right) C e^{-a_0 t} = 0.$$

Since  $C e^{-a_0 t}$  is the general solution of (2.2.1) with zero right-hand side (i.e., with  $f(t) = 0$ ) it is called the **complementary function**. The integral term in (2.2.2) satisfies (2.2.1) and any solution of (2.2.1) is called a **particular integral**. Consequently,

$$\begin{aligned} & \text{the general solution of (2.2.1)} \\ &= \text{particular integral} + \text{complementary function}. \end{aligned}$$

It does not matter which particular integral is chosen as we shall show now. Suppose  $h(t)$  is any function such that

$$\frac{dh}{dt} + a_0 h = f(t).$$

Put  $y = h + z$ ; then

$$\frac{dy}{dt} + a_0 y = \frac{dh}{dt} + \frac{dz}{dt} + a_0 h + a_0 z = f(t) + \frac{dz}{dt} + a_0 z$$

so that, if  $y$  satisfies (2.2.1), we must have

$$\frac{dz}{dt} + a_0 z = 0,$$

i.e.,  $z$  is the complementary function.

## 2.3 Linear equations of the second order

The linear ordinary differential equation of the second order with constant coefficients can be expressed as

$$\frac{d^2 y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = f(t), \quad (2.3.1)$$

after division by the coefficient of the second derivative (which must be non-zero if the differential equation is to be of the second order). With zero right-hand side, (2.3.1) is

$$\frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y = 0 \quad (2.3.2)$$

and is known as the **associated homogeneous differential equation**. Let the general solution of (2.3.2) be denoted by  $y_c$  and called the **complementary function**. Let  $y_p$  be any solution of (2.3.1) and designate it as a **particular integral**. Put  $y = y_p + z$ . Then

$$\begin{aligned} \frac{d^2y}{dt^2} + a_1 \frac{dy}{dt} + a_0 y &= \frac{d^2y_p}{dt^2} + a_1 \frac{dy_p}{dt} + a_0 y_p + \frac{d^2z}{dt^2} + a_1 \frac{dz}{dt} + a_0 z \\ &= f(t) + \frac{d^2z}{dt^2} + a_1 \frac{dz}{dt} + a_0 z. \end{aligned}$$

Therefore  $y$  satisfies (2.3.1) provided that

$$\frac{d^2z}{dt^2} + a_1 \frac{dz}{dt} + a_0 z = 0,$$

i.e.,  $z = y_c$ . Hence  $y = y_p + y_c$  and

general solution = particular integral + complementary function

as before.

Thus the general solution of both first- and second-order equations has the same structure, and we shall find that this is true for the linear ordinary differential equation with constant coefficients of order  $n$ .

## 2.4 Finding the complementary function

The determination of the complementary function requires the general solution of (2.3.2), i.e., of

$$\ddot{y} + a_1 \dot{y} + a_0 y = 0 \quad (2.4.1)$$

if we use the notation  $\dot{y} \equiv dy/dt$ ,  $\ddot{y} \equiv d^2y/dt^2$ .

Consider the equation for  $\lambda$

$$\lambda^2 + a_1 \lambda + a_0 = 0, \quad (2.4.2)$$

which will be called the **characteristic equation**. It has two roots  $\lambda_1$  and  $\lambda_2$  such that

$$\lambda_1 + \lambda_2 = -a_1$$

and

$$\lambda_1 \lambda_2 = a_0.$$

Now

$$\frac{d}{dt} \left( \frac{dy}{dt} - \lambda_1 y \right) = \frac{d^2 y}{dt^2} - \lambda_1 \frac{dy}{dt},$$

so that

$$\begin{aligned} \left( \frac{d}{dt} - \lambda_2 \right) \left( \frac{dy}{dt} - \lambda_1 y \right) &= \frac{d^2 y}{dt^2} - (\lambda_1 + \lambda_2) \frac{dy}{dt} + \lambda_1 \lambda_2 y \\ &= \ddot{y} + a_1 \dot{y} + a_0 y \end{aligned}$$

because of the properties of  $\lambda_1$  and  $\lambda_2$ . Consequently (2.4.1) can be written as

$$\left( \frac{d}{dt} - \lambda_2 \right) \left( \frac{dy}{dt} - \lambda_1 y \right) = 0.$$

Write

$$w = \frac{dy}{dt} - \lambda_1 y.$$

Then

$$\frac{dw}{dt} - \lambda_2 w = 0.$$

This is of the first order and has general solution

$$w = C e^{\lambda_2 t}.$$

Therefore

$$\frac{dy}{dt} - \lambda_1 y = C e^{\lambda_2 t}.$$

Since this is linear and of the first order, it can be solved by multiplying by the integrating factor  $e^{-\lambda_1 t}$ , which gives

$$\frac{d}{dt} (y e^{-\lambda_1 t}) = C e^{(\lambda_2 - \lambda_1)t}$$

whence

$$y e^{-\lambda_1 t} = C \int^t e^{(\lambda_2 - \lambda_1)u} du + D.$$

If  $\lambda_1 \neq \lambda_2$ ,

$$\int^t e^{(\lambda_2 - \lambda_1)u} du = \frac{e^{(\lambda_2 - \lambda_1)t}}{\lambda_2 - \lambda_1}$$

and

$$y e^{-\lambda_1 t} = \frac{C e^{(\lambda_2 - \lambda_1)t}}{\lambda_2 - \lambda_1} + D$$

or

$$y = C_2 e^{\lambda_2 t} + C_1 e^{\lambda_1 t},$$

where  $C_1$  and  $C_2$  are arbitrary constants.

If  $\lambda_1 = \lambda_2$ ,

$$\int^t e^{(\lambda_2 - \lambda_1)u} du = \int^t du = t$$

and

$$ye^{-\lambda_1 t} = Ct + D$$

or

$$y = (C_1 + C_2 t)e^{\lambda_1 t}.$$

The rule for finding the complementary function can be stated now as: *solve the characteristic equation (2.4.2):*

(a) *if the roots are different*

$$y_c = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t};$$

(b) *if the roots are the same*

$$y_c = (C_1 + C_2 t)e^{\lambda_1 t}.$$

The characteristic equation may be arrived at in the following way. Try  $y = e^{\lambda t}$  so that  $\dot{y} = \lambda e^{\lambda t}$ ,  $\ddot{y} = \lambda^2 e^{\lambda t}$ . Then (2.4.1) is satisfied if

$$(\lambda^2 + a_1 \lambda + a_0) e^{\lambda t} = 0$$

or

$$\lambda^2 + a_1 \lambda + a_0 = 0,$$

which is the characteristic equation.

### **Example 2.4.1**

Find the general solution of

$$16\ddot{y} - 8\dot{y} + y = 0.$$

The characteristic equation is

$$16\lambda^2 - 8\lambda + 1 = 0$$

so that  $\lambda_1 = \frac{1}{4} = \lambda_2$ . The rule now gives

$$y = (C_1 + C_2 t)e^{t/4}$$

as the general solution. □

It may happen that  $\lambda_1$  or  $\lambda_2$  is complex. No problem occurs because, if  $\lambda_1 = a + ib$  where  $a$  and  $b$  are real, it is known that (compare with De Moivre's theorem)

$$e^{\lambda_1 t} = e^{at} e^{ibt} = e^{at} (\cos bt + i \sin bt),$$

so that the solution can be expressed in terms of trigonometric functions if desired.

If  $a_0$  and  $a_1$  are real, we may be seeking a real solution of (2.4.1). In this case when  $\lambda_1 = a + ib$  the characteristic equation forces  $\lambda_2$  to be the complex conjugate, i.e.,  $\lambda_2 = a - ib$ . Since  $b \neq 0$ , because we are assuming a complex root of (2.4.2), the rule gives

$$y_c = C_1 e^{(a+ib)t} + C_2 e^{(a-ib)t}$$

where  $C_1$  and  $C_2$  are arbitrary (complex) constants. To make  $y_c$  real, we must make  $y_c^* = y_c$ , the asterisk indicating a complex conjugate. But, when  $a_0$  and  $a_1$  are real,  $y_c^*$  also satisfies (2.4.1) and so

$$y_c^* = C_1^* e^{(a-ib)t} + C_2^* e^{(a+ib)t}.$$

Thus  $y_c^* = y_c$  demands that  $C_2^* = C_1$ . This means that, if  $C_1 = A + iB$  with  $A$  and  $B$  real,  $C_2 = A - iB$ . It follows that

$$\begin{aligned} y_c &= (A + iB)e^{(a+ib)t} + (A - iB)e^{(a-ib)t} \\ &= 2e^{at}(A \cos bt - B \sin bt), \end{aligned}$$

which gives a real complementary function with real arbitrary constants  $A$  and  $B$ .

### **Example 2.4.2**

Find the solution of

$$\ddot{y} - 2\dot{y} + 2y = 0$$

satisfying the initial conditions,  $y(0) = 0$ ,  $\dot{y}(0) = 1$ .

The characteristic equation is

$$\lambda^2 - 2\lambda + 2 = 0,$$

so that  $\lambda_1 = 1+i$  and  $\lambda_2 = 1-i$ . From the above reasoning we can immediately write down the general solution in the form

$$y = e^t(A \cos t + B \sin t).$$

By imposing the initial conditions we find that  $A = 0$  and  $B = 1$ , which shows that the solution is

$$y = e^t \sin t.$$

□

## 2.5 Determining a particular integral

There are various devices for finding a particular integral, each of which has advantages and disadvantages. Three methods will be described here.

### 2.5.1 Undetermined coefficients

The method consists essentially of guessing an appropriate form for the answer and substituting it in the differential equation. It is suitable if the right-hand side is a polynomial, exponential, sine, cosine or a product of these. The technique will be illustrated by means of examples.

#### **Example 2.5.1**

Find a particular integral of

$$\ddot{y} - 2\dot{y} - 3y = 3t^2.$$

The idea is to try to find a polynomial solution, and we show how this can be done by starting from the simplest and gradually making it more complicated.

Try  $y = C$ . Then  $\ddot{y} - 2\dot{y} - 3y = -3C$ . Since this can never agree with  $3t^2$ , it must be rejected.

Try  $y = Bt$ . Then  $\ddot{y} - 2\dot{y} - 3y = -2B - 3Bt$ , and again the attempt is unsatisfactory.

Try  $y = At^2$ . Then  $\ddot{y} - 2\dot{y} - 3y = 2A - 4At - 3At^2$ , which again is not suitable but does at least contain a term involving  $t^2$ .

This suggests that we should try  $y = At^2 + Bt + C$ . Then

$$\ddot{y} - 2\dot{y} - 3y = 2A - 2B - 3C - (4A + 3B)t - 3At^2.$$

We can make this the same as  $3t^2$  if

$$\begin{aligned} -3A &= 3, \\ 4A + 3B &= 0, \\ 2A - 2B - 3C &= 0. \end{aligned}$$

These equations are satisfied if  $A = -1$ ,  $B = \frac{4}{3}$  and  $C = -\frac{14}{9}$ . Therefore

$$y_p = -t^2 + \frac{4}{3}t - \frac{14}{9}$$

supplies a particular integral. □

In general this suggests, since the derivative of a polynomial of degree  $n$  is a polynomial of degree  $n - 1$ , that *when the right-hand side of (2.3.1) is a*

polynomial of degree  $n$  and  $a_0 \neq 0$  try a polynomial of degree  $n$  with arbitrary coefficients. If  $a_0 = 0$  but  $a_1 \neq 0$  try a polynomial of degree  $n + 1$ .

### Example 2.5.2

Find a particular integral of

$$\ddot{y} - 2\dot{y} - 3y = e^{2t}.$$

In this case we try  $y = Ae^{2t}$ . Then the left-hand side is

$$4Ae^{2t} - 4Ae^{2t} - 3Ae^{2t} = -3Ae^{2t}$$

which is the same as  $e^{2t}$  if  $A = -\frac{1}{3}$ . Therefore

$$y_p = -\frac{1}{3}e^{2t}. \quad \square$$

### Example 2.5.3

Find a particular integral of

$$\ddot{y} - 2\dot{y} - 3y = e^{-t}.$$

Again we try  $y = Ae^{-t}$ . The left-hand side becomes

$$Ae^{-t} + 2Ae^{-t} - 3Ae^{-t} = 0$$

so that  $A$  cannot be found. The try fails in this case because  $e^{-t}$  satisfies the associated homogeneous differential equation and so is part of the complementary function.

The lesson to be learned is that the complementary function should always be found first. The characteristic equation is

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

so that  $\lambda_1 = 3, \lambda_2 = -1$  and

$$y_c = C_1e^{3t} + C_2e^{-t}.$$

If the right-hand side is part of the complementary function, it cannot be a suitable particular integral.

Try instead  $y = Ate^{-t}$  so that  $\dot{y} = A(1-t)e^{-t}, \ddot{y} = A(t-2)e^{-t}$ . Then

$$\begin{aligned}\ddot{y} - 2\dot{y} - 3y &= \{t-2-2(1-t)-3t\}Ae^{-t} \\ &= -4Ae^{-t}\end{aligned}$$

which agrees with  $e^{-t}$  if  $A = -\frac{1}{4}$ . Therefore

$$y_p = -\frac{1}{4}te^{-t}. \quad \square$$

**Example 2.5.4**

Find a particular integral of

$$\ddot{y} - 4\dot{y} + 4y = e^{2t}.$$

To determine the complementary function, solve the characteristic equation

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

which gives  $\lambda_1 = 2 = \lambda_2$ . Therefore

$$y_c = (C_1 + C_2 t)e^{2t}.$$

In this case, both  $e^{2t}$  and  $te^{2t}$  will fail as particular integrals. So try  $At^2e^{2t}$ , which will be found to work. □

**Example 2.5.5**

Find a particular integral of

$$\ddot{y} - 2\dot{y} - 3y = -9te^{2t}.$$

The right-hand side is not part of the complementary function and is a product of a polynomial and an exponential. This suggests that we try  $(At + B)e^{2t}$ . The left-hand side becomes then

$$(2A - 3B - 3At)e^{2t},$$

which agrees with the right-hand side if  $A = 3$  and  $B = 2$ . Therefore

$$y_p = (3t + 2)e^{2t}.$$

If  $e^{2t}$  had been part of the complementary function but not  $te^{2t}$ , we would have tried  $(At + B)te^{2t}$ ; if  $te^{2t}$  had been part of the complementary function also, the trial function would have been  $(At + B)t^2e^{2t}$ . □

From these examples we can construct the following prescription for a particular integral. *If*

$$\ddot{y} + a_1\dot{y} + a_0y = t^m e^{at} \tag{2.5.1}$$

where  $a_0 \neq 0$  and  $m$  is a nonnegative integer,

(I) when  $a$  is not a root of the characteristic equation  $\lambda^2 + a_1\lambda + a_0 = 0$ , try

$$y = e^{at}(A_m t^m + A_{m-1} t^{m-1} + \cdots + A_0);$$

(II) when  $a$  is a single root of the characteristic equation, multiply the expansion in (I) by  $t$ ;

- (III) when  $a$  is a double root of the characteristic equation, multiply the expansion in (I) by  $t^2$ .

Substitution of the proposed expansion in the differential equation will confirm that the right-hand side can be obtained by an appropriate choice of  $A_0, \dots, A_m$ .

A right-hand side that is composed of a sum of the type in (2.5.1) can be handled because, if

$$\ddot{y} + a_1\dot{y} + a_0y = f_1(t) + f_2(t) + \dots + f_n(t) \quad (2.5.2)$$

and  $y_i$  is a particular integral of

$$\ddot{y} + a_1\dot{y} + a_0y = f_i(t),$$

then

$$y_p = y_1 + y_2 + \dots + y_n$$

is a particular integral of (2.5.2) as may be confirmed by substitution in the differential equation.

Trigonometric functions are also covered by the rule given because we can write

$$\cos t = \frac{1}{2}(e^{it} + e^{-it}), \quad \sin t = \frac{1}{2i}(e^{it} - e^{-it}) \quad (2.5.3)$$

and use the fact that  $a$  can be complex in (2.5.1).

### **Example 2.5.6**

Find a particular integral of

$$\ddot{y} - 2\dot{y} - 3y = 2 \sin t.$$

First consider the right-hand side of  $e^{it}$ . This is not part of the complementary function as can be seen from Example 2.5.3. Therefore try  $y = Ae^{it}$ . The left-hand side is  $(-4 - 2i)Ae^{it}$ , which agrees with  $e^{it}$  if  $A = -1/(4 + 2i)$ . Similarly, the particular integral corresponding to  $e^{-it}$  is  $-e^{-it}/(4 - 2i)$ . Hence, from (2.5.3),

$$\begin{aligned} yp &= \frac{1}{i} \left( -\frac{e^{it}}{4+2i} + \frac{e^{-it}}{4-2i} \right) \\ &= \frac{1}{5}(\cos t - 2 \sin t). \end{aligned} \quad \square$$

### **Example 2.5.7**

Find a particular integral of

$$\ddot{y} + 4y = 32t \cos 2t - 8 \sin 2t.$$

The characteristic equation is  $\lambda^2 + 4 = 0$ , which implies that  $\lambda_1 = 2i, \lambda_2 = -2i$ . The complementary function may be expressed in terms of  $e^{2it}$  and  $e^{-2it}$  or, in real form, by

$$y_c = C_1 \cos 2t + C_2 \sin 2t.$$

For the right-hand side we consider first  $e^{2it}$ . Since  $2i$  is a single root of the characteristic equation, case (II) above applies and we try  $Ate^{2it}$ . Similarly for  $te^{2it}$  we would try  $(Bt + C)te^{2it}$ . However, it is more economical to combine the two and make  $(Bt + C)te^{2it}$  reproduce the term  $(16t + 4i)e^{2it}$  required on the right-hand side. We find  $B = -2i$  and  $C = 2$ , so that

$$\begin{aligned} y_p &= (-2it + 2)te^{2it} + (2it + 2)te^{-2it} \\ &= 4t \cos 2t + 4t^2 \sin 2t. \end{aligned}$$

The general solution is

$$y = C_1 \cos 2t + C_2 \sin 2t + 4t(\cos 2t + t \sin 2t). \quad \square$$

### Example 2.5.8

Find a particular integral of

$$\ddot{y} - \dot{y} - 2y = \cosh t.$$

The characteristic equation is  $\lambda^2 - \lambda - 2 = 0$ , so that  $\lambda_1 = 2, \lambda_2 = -1$  and the complementary function is

$$y_c = C_1 e^{2t} + C_2 e^{-t}.$$

The right-hand side can be converted to standard form by using  $\cosh t = \frac{1}{2}(e^t + e^{-t})$ . For  $e^t$  case (I) applies because 1 is not a root of the characteristic equation, but for  $e^{-t}$  case (II) is relevant because  $-1$  is a single root of the characteristic equation. Therefore, try  $Ae^t + Bte^{-t}$ ; it is found that  $A = -\frac{1}{4}$ ,  $B = -\frac{1}{6}$  and

$$y_p = -\frac{1}{4}e^t - \frac{1}{6}te^{-t}$$

is a particular integral.  $\square$

The differential equation

$$\ddot{y} + 2\dot{y} + y = te^{-t}$$

provides an illustration of case (III). A particular integral is

$$y_p = \frac{1}{6}t^3 e^{-t}.$$

### 2.5.2 Factorisation of the operator

In this method the differential equation is split into two first-order differential equations in the same way as was employed for discovering the complementary function in Section 2.4. An example will demonstrate the technique.

#### **Example 2.5.9**

Reconsider

$$\ddot{y} - \dot{y} - 2y = \cosh t$$

which was discussed in Example 2.5.8.

The differential equation can be expressed as

$$\left(\frac{d}{dt} + 1\right)\left(\frac{dy}{dt} - 2y\right) = \cosh t.$$

Put  $u = \dot{y} - 2y$ . Then

$$\frac{du}{dt} + u = \cosh t = \frac{1}{2}(e^t + e^{-t}).$$

The integrating factor is  $e^t$  and

$$\frac{d}{dt}(ue^t) = \frac{1}{2}(e^{2t} + 1)$$

so that

$$ue^t = \frac{1}{4}e^{2t} + \frac{1}{2}t + C.$$

Therefore

$$\dot{y} - 2y = \frac{1}{4}e^t + \frac{1}{2}te^{-t} + Ce^{-t}.$$

The integrating factor is  $e^{-2t}$  and

$$\frac{d}{dt}(ye^{-2t}) = \frac{1}{4}e^{-t} + \frac{1}{2}te^{-3t} + Ce^{-3t}.$$

Since

$$\begin{aligned} \int^t xe^{-3x}dx &= -\frac{1}{3}te^{-3t} + \frac{1}{3}\int e^{-3x}dx = -\frac{1}{3}te^{-3t} - \frac{1}{9}e^{-3t}, \\ ye^{-2t} &= -\frac{1}{4}e^{-t} - \frac{1}{6}te^{-3t} - \frac{1}{18}e^{-3t} - \frac{1}{3}Ce^{-3t} + D \end{aligned}$$

whence

$$y = -\frac{1}{4}e^t - \frac{1}{6}te^{-t} - \frac{1}{18}e^{-t} - \frac{1}{3}Ce^{-t} + De^{2t}.$$

Actually, this analysis has given the general solution, but we could have obtained a particular integral by leaving  $C$  and  $D$  out when they arose. The resulting particular integral differs from that of Example 2.5.8 by  $-\frac{1}{18}e^{-t}$ .

but, since this is part of the complementary function, there is no change to the general solution.  $\square$

The advantages of this method are that it can give the general solution directly and always works even when the right-hand side is not a polynomial, exponential, sine or cosine. On the other hand, it often involves more labour than undetermined coefficients when the right-hand side is such that either method is applicable. It can also be awkward to implement when the roots of the characteristic equation are complex.

### 2.5.3 Variation of parameters

Another method that can be adopted for any right-hand side is variation of parameters.

Suppose that the complementary function of

$$\ddot{y} + a_1\dot{y} + a_0y = f(t) \quad (2.5.4)$$

is

$$y_c = C_1y_1(t) + C_2y_2(t).$$

We seek a solution of (2.5.4) in the form

$$y = u(t)y_1(t) + v(t)y_2(t), \quad (2.5.5)$$

i.e., we allow the parameters  $C_1$  and  $C_2$  to vary—which explains the nomenclature. There are two unknown functions  $u$  and  $v$  so that two conditions are needed to determine them. One condition is obtained by substituting (2.5.5) in (2.5.4). The other we can pick for ourselves and we want to do it so as to avoid second derivatives of  $u$  and  $v$  if possible; otherwise we are no better off. Now, from (2.5.5),

$$\dot{y} = u\dot{y}_1 + v\dot{y}_2 + \dot{u}y_1 + \dot{v}y_2$$

and, if second derivatives of  $u$  and  $v$  are not to occur in  $\ddot{y}$ , we must insist that

$$\dot{u}y_1 + \dot{v}y_2 = 0. \quad (2.5.6)$$

Then

$$\dot{y} = u\dot{y}_1 + v\dot{y}_2$$

and

$$\ddot{y} = u\ddot{y}_1 + v\ddot{y}_2 + \dot{u}\dot{y}_1 + \dot{v}\dot{y}_2.$$

Now (2.5.4) can be satisfied if

$$u\ddot{y}_1 + v\ddot{y}_2 + \dot{u}\dot{y}_1 + \dot{v}\dot{y}_2 + a_1(u\dot{y}_1 + v\dot{y}_2) + a_0(uy_1 + vy_2) = f(t)$$

or

$$(\ddot{y}_1 + a_1\dot{y}_1 + a_0y_1)u + (\ddot{y}_2 + a_1\dot{y}_2 + a_0y_2)v + \dot{u}\dot{y}_1 + \dot{v}\dot{y}_2 = f(t).$$

But  $y_1$  and  $y_2$  both satisfy the associated homogeneous differential equation, and so the terms in the two sets of parentheses vanish. Consequently

$$\dot{u}y_1 + \dot{v}y_2 = f(t). \quad (2.5.7)$$

The equations (2.5.6) and (2.5.7) are now solved for  $\dot{u}$  and  $\dot{v}$ . Integration then supplies  $u, v$  and hence a particular integral.

It should be remarked that it is possible to solve (2.5.6) and (2.5.7) provided that  $y_1\dot{y}_2 - \dot{y}_1y_2 \neq 0$ . Suppose that this is not true and that  $y_1\dot{y}_2 - \dot{y}_1y_2 = 0$ . Then  $\dot{y}_2/y_2 = \dot{y}_1/y_1$  which implies that  $\ln y_2 = \ln y_1 + C$  or  $y_2 = Ay_1$ . Thus  $C_1y_1 + C_2y_2 = (C_1 + AC_2)y_1$  and this cannot be the complementary function since it contains only one arbitrary constant. Therefore, so long as the genuine complementary function has been found, the solution of (2.5.6) and (2.5.7) can always be carried out.

### Example 2.5.10

Solve

$$\ddot{y} - \dot{y} - 2y = \cosh t$$

by means of variation of parameters.

The complementary function is

$$y_c = C_1e^{2t} + C_2e^{-t}.$$

Therefore  $y_1 = e^{2t}, y_2 = e^{-t}$  and (2.5.6) gives

$$\dot{u}e^{2t} + \dot{v}e^{-t} = 0$$

whereas (2.5.7) becomes

$$2\dot{u}e^{2t} - \dot{v}e^{-t} = \cosh t.$$

Hence

$$3\dot{u}e^{2t} = \cosh t, \quad \dot{v}e^{-t} = -\frac{1}{3}\cosh t$$

or

$$\dot{u} = \frac{1}{6}(e^{-t} + e^{-3t}), \quad \dot{v} = -\frac{1}{6}(1 + e^{2t}).$$

Consequently

$$u = -\frac{1}{6}e^{-t} - \frac{1}{18}e^{-3t} + C, \quad v = -\frac{1}{6}t - \frac{1}{12}e^{2t} + D$$

and

$$\begin{aligned} y &= \left(-\frac{1}{6}e^{-t} - \frac{1}{18}e^{-3t} + C\right)e^{2t} + \left(-\frac{1}{6}t - \frac{1}{12}e^{2t} + D\right)e^{-t} \\ &= Ce^{2t} + \left(D - \frac{1}{18}\right)e^{-t} - \frac{1}{4}e^t - \frac{1}{6}te^{-t}. \end{aligned}$$

This is the same general solution as derived in Examples 2.5.8 and 2.5.9. If we had placed  $C = 0, D = 0$  earlier we would have obtained a particular integral but not the general solution.  $\square$

**Example 2.5.11**

Solve Example 2.5.7 by the variation of parameters method. In this example it is more convenient to replace the trigonometric functions by their exponential representation. That is we consider the differential equation

$$\ddot{y} + 4y = 16t(e^{2it} + e^{-2it}) + 4i(e^{2it} - e^{-2it}).$$

The complementary function is

$$y_c = C_1e^{2it} + C_2e^{-2it}.$$

Consequently (2.5.6) and (2.5.7) become

$$\dot{u}e^{2it} + \dot{v}e^{-2it} = 0$$

and

$$\dot{u}e^{2it} - \dot{v}e^{-2it} = -8it(e^{2it} + e^{-2it}) + 2(e^{2it} - e^{-2it}).$$

This pair of equations is easily solved to give

$$\dot{u} = -4it(1 + e^{-4it}) + (1 - e^{-4it}),$$

$$\dot{v} = 4it(1 + e^{4it}) + (1 - e^{4it}),$$

which on integration by parts and ignoring the arbitrary constants of integration gives the results

$$u = t(1 - 2it) + e^{-4it}(t - \frac{3i}{8}),$$

$$v = t(1 + 2it) + e^{+4it}(t + \frac{3i}{8}).$$

Substituting these expressions for  $u$  and  $v$  into (2.5.5) and writing the exponential forms back into trigonometric functions we obtain the particular integral

$$y_p = 4t \cos 2t + 4t^2 \sin 2t - \frac{3}{4} \sin 2t,$$

which when combined with the complementary function gives the solution as found before. □

The method of variation of parameters is applicable for any right-hand side and works even if the characteristic equation has complex roots. It can also be generalised to other types of differential equations. Its disadvantage is that it often requires a lot of effort to carry through.

In summary, the strategy suggested is that the method of undetermined coefficients should be employed whenever the right-hand side has the right form. If it does not, try either factorisation of the operator or variation of parameters with preference for variation of parameters when the roots of the characteristic equation are complex.

## 2.6 Forced oscillations

A differential equation that arises frequently in practice is

$$\ddot{y} + 2b\Omega\dot{y} + \Omega^2 y = F \cos(\omega t + \beta), \quad (2.6.1)$$

where  $b, \Omega, F, \omega$  and  $\beta$  are real constants with  $\Omega > 0$ . It represents an oscillatory system subject to damping, when  $b > 0$  as we shall assume, being vibrated by external means. In one application  $y$  represents the displacement of a particle subject to a restoring force  $\Omega^2 y$  per unit mass and viscous damping  $2b\Omega\dot{y}$  per unit mass acted on by a force of magnitude  $F$  per unit mass and circular frequency  $\omega$ .

If  $b \neq 1$ , the complementary function is

$$y_c = C_1 e^{\delta_1 t} + C_2 e^{\delta_2 t}, \quad (2.6.2)$$

where  $\delta_1 = -b\Omega + \Omega(b^2 - 1)^{1/2}$  and  $\delta_2 = -b\Omega - \Omega(b^2 - 1)^{1/2}$ . The quantities  $\delta_1$  and  $\delta_2$  are real if  $b > 1$  and complex if  $b < 1$ . In either case, the real parts of  $\delta_1$  and  $\delta_2$  are negative.

If  $b = 1$ , the characteristic equation has the double root  $-\Omega$  and

$$y_c = (A + Bt)e^{-\Omega t}. \quad (2.6.3)$$

To find a particular integral, we consider the right-hand side  $F e^{i(\omega t + \beta)}$ . Since  $i\omega$  is not a root of the characteristic equation, we try  $C e^{i(\omega t + \beta)}$  which leads to

$$(-\omega^2 + 2ib\Omega\omega + \Omega^2)C = F.$$

Therefore

$$\begin{aligned} 2y_p &= \frac{F e^{i(\omega t + \beta)}}{\Omega^2 - \omega^2 + 2ib\Omega\omega} + \frac{F e^{-i(\omega t + \beta)}}{\Omega^2 - \omega^2 - 2ib\Omega\omega} \\ &= \frac{2F\{(\Omega^2 - \omega^2)\cos(\omega t + \beta) + 2b\Omega\omega\sin(\omega t + \beta)\}}{(\Omega^2 - \omega^2)^2 + 4b^2\Omega^2\omega^2}. \end{aligned}$$

Consequently, the general solution of (2.6.1) is

$$y = y_c + \frac{F \cos(\omega t + \beta - \phi)}{\{(\Omega^2 - \omega^2)^2 + 4b^2\Omega^2\omega^2\}^{1/2}}, \quad (2.6.4)$$

where  $y_c$  is given by (2.6.2) or (2.6.3) depending on the value of  $b$ , and

$$\begin{aligned} \cos \phi &= \frac{\Omega^2 - \omega^2}{\{(\Omega^2 - \omega^2)^2 + 4b^2\Omega^2\omega^2\}^{1/2}}, \\ \sin \phi &= \frac{2b\Omega\omega}{\{(\Omega^2 - \omega^2)^2 + 4b^2\Omega^2\omega^2\}^{1/2}}. \end{aligned} \quad (2.6.5)$$

The first term on the right-hand side of (2.6.4) is always present whether  $F = 0$  or not. Its value depends upon what conditions are set at  $t = 0$ , but wherever it starts it will diminish to zero as  $t$  becomes large enough because of the form of (2.6.2) and (2.6.3). For this reason, the first term of (2.6.4) is often known as a **transient**.

The second term of (2.6.4) only occurs when  $F \neq 0$  and varies at the same rate as the forcing device. Its amplitude is independent of the conditions at  $t = 0$  and there is no decay as  $t \rightarrow \infty$ . It is known as the **forced oscillation**.

When one is concerned only with what happens for large  $t$ , the forced oscillation alone survives. Its amplitude  $Y$  can be expressed as

$$Y = \frac{F/\Omega^2}{[\{1 - (\omega^2/\Omega^2)\}^2 + 4b^2\omega^2/\Omega^2]^{1/2}}.$$

As  $\omega$  varies, changes in  $Y$  occur only in the denominator and are dictated by the behaviour of

$$(1 - z^2)^2 + 4b^2z^2$$

with  $z^2 = \omega^2/\Omega^2$ . The derivative of this is

$$-4(1 - z^2)z + 8b^2z$$

which vanishes when  $z = 0$  or  $z^2 = 1 - 2b^2$ . The second derivative has the value  $4(2b^2 - 1)$  when  $z = 0$  and the value  $8(1 - 2b^2)$  when  $z^2 = 1 - 2b^2$ . Since  $z^2$  cannot be negative, the denominator of  $Y$  has a minimum at  $z = 0$  and no other stationary point if  $2b^2 > 1$ , but a maximum at  $z = 0$  and a minimum at  $z^2 = 1 - 2b^2$  if  $2b^2 < 1$ . Hence (i) if  $2b^2 > 1$ ,  $Y$  has a maximum at  $\omega = 0$  and (ii) if  $2b^2 < 1$ ,  $Y$  has a minimum at  $\omega = 0$  and a maximum at  $\omega^2/\Omega^2 = 1 - 2b^2$ . The value of  $Y$  at  $\omega^2/\Omega^2 = 1 - 2b^2$  is  $F/2\Omega^2b(1 - b^2)^{1/2}$ . So the maximum of the forced oscillation is larger the smaller  $b$  is, i.e., *with low damping it is possible to excite large vibrations provided that  $\omega$  is chosen appropriately*.

It can be seen from (2.6.5) that  $\phi = \frac{1}{2}\pi$  when  $\omega = \Omega$ , that  $\phi \approx 0$  when  $\omega/\Omega \ll 1$  and that  $\phi \rightarrow \pi$  as  $\omega/\Omega \rightarrow \infty$ .

### Example 2.6.1

Solve the differential equation

$$\ddot{x} + \epsilon k\dot{x} + x = \epsilon F \cos t,$$

governing simple harmonic motion with weak forcing and weak damping. The characteristic equation is

$$\lambda^2 + \epsilon k\lambda + 1 = 0$$

so that  $\lambda_1, \lambda_2 = -\frac{\epsilon k}{2} \pm i\sqrt{1 - \epsilon^2 k^2/4}$ . For  $\epsilon$  sufficiently small we can approximate the complementary function  $y_c$  by

$$y_c \simeq e^{-\epsilon kt/2}(A \cos t + B \sin t)$$

for arbitrary constants  $A, B$ .

To determine the particular solution we try

$$y_p = C \cos t + D \sin t$$

from which it is found that  $C = 0$  and  $D = F/k$ . Consequently we have the approximate solution

$$y_c \simeq e^{-\epsilon kt/2} (A \cos t + B \sin t) + \frac{F}{k} \sin t.$$

We conclude that as  $t \rightarrow \infty$  the solution becomes one of simple harmonic motion with amplitude  $F/k$ . □

---

## 2.7 Differential equations of order n

Much of the discussion for the equation of second order carries over to the linear differential equation with constant coefficients of order  $n$ , namely

$$\frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = f(t). \quad (2.7.1)$$

Exactly as in Section 2.3, one may show that the general solution can be written

$$y = y_c + y_p$$

where  $y_p$  is a particular integral of (2.7.1) and  $y_c$  is the complementary function, i.e., the general solution of the associated homogeneous differential equation.

The complementary function is determined by solving the characteristic equation. Recalling from Section 2.4 that the characteristic equation could be reached by trying  $y = e^{\lambda t}$ , we do the same here and obtain

$$\lambda^n + a_{n-1} \lambda^{n-1} + \cdots + a_0 = 0. \quad (2.7.2)$$

Equation (2.7.2) has  $n$  roots  $\lambda_1, \lambda_2, \dots, \lambda_n$ . For any  $\lambda_i$  which is different from all the rest, there is a contribution to the complementary function of  $C_i e^{\lambda_i t}$ . If, however,  $\lambda_j$  occurs  $m$  times, the complementary function acquires a term

$$(D_1 + D_2 t + \cdots + D_m t^{m-1}) e^{\lambda_j t}.$$

### Example 2.7.1

Find the general solution of

$$\frac{d^5 y}{dt^5} + 6 \frac{d^4 y}{dt^4} + 15 \frac{d^3 y}{dt^3} + 26 \frac{d^2 y}{dt^2} + 36 \frac{dy}{dt} + 24y = 0.$$

The characteristic equation is

$$\lambda^5 + 6\lambda^4 + 15\lambda^3 + 26\lambda^2 + 36\lambda + 24 = 0.$$

Now

$$\begin{aligned} \lambda^5 + 6\lambda^4 + 15\lambda^3 + 26\lambda^2 + 36\lambda + 24 &= (\lambda + 2)(\lambda^4 + 4\lambda^3 + 7\lambda^2 + 12\lambda + 12) \\ &= (\lambda + 2)^2(\lambda^3 + 2\lambda^2 + 3\lambda + 6) \\ &= (\lambda + 2)^3(\lambda^2 + 3). \end{aligned}$$

Thus the roots of the characteristic equation are  $-2$  (three times) and  $\pm i\sqrt{3}$ , once each. Therefore the general solution is

$$\begin{aligned} y &= (C_1 + C_2t + C_3t^2)e^{-2t} + C_4e^{it\sqrt{3}} + C_5e^{-it\sqrt{3}} \\ &= (C_1 + C_2t + C_3t^2)e^{-2t} + A \cos \sqrt{3}t + B \sin \sqrt{3}t \end{aligned}$$

for a real solution. □

Any of the three methods described for deriving a particular integral for the second-order differential equation may be employed for the general order in the right circumstances. For undetermined coefficients, the rules (I), (II) and (III) of Section 2.5(a) need to be supplemented because a root of the characteristic equation can be repeated more than twice. It is, however, clear from (I), (II) and (III) that the pertinent change is that *if a appears m times as a root of the characteristic equation, multiply by t<sup>m</sup> the expansion that would have been tried for a solitary root*. For instance, if the right-hand side in Example 2.7.1 were  $te^{-2t}$ , the trial function would be  $t^3(A + Bt)e^{-2t}$ , leading to  $A = 1/168, B = 2/147$ .

There is nothing new to add to the method of factorisation of the operator, but now  $n$  equations of the first order have to be solved and the method becomes increasingly cumbersome as the size of  $n$  grows.

The principle of variation of parameters is unaltered but further detail is necessary. Let

$$y_c = C_1y_1(t) + C_2y_2(t) + \cdots + C_ny_n(t)$$

be the complementary function. Then a solution of (2.7.1) is sought in the form

$$y = u_1(t)y_1(t) + u_2(t)y_2(t) + \cdots + u_n(t)y_n(t) \quad (2.7.3)$$

with  $u_1, \dots, u_n$  subject to the  $n - 1$  conditions

$$\begin{aligned} \frac{du_1}{dt}y_1 + \frac{du_2}{dt}y_2 + \cdots + \frac{du_n}{dt}y_n &= 0, \\ \frac{du_1}{dt}\frac{dy_1}{dt} + \frac{du_2}{dt}\frac{dy_2}{dt} + \cdots + \frac{du_n}{dt}\frac{dy_n}{dt} &= 0, \end{aligned} \quad (2.7.4)$$

$$\begin{array}{ccccccc} & \cdots & & \cdots & & \cdots & \\ & \cdots & & \cdots & & \cdots & \\ \frac{du_1}{dt} \frac{d^{n-2}y_1}{dt^{n-2}} + \frac{du_2}{dt} \frac{d^{n-2}y_2}{dt^{n-2}} + \cdots + \frac{du_n}{dt} \frac{d^{n-2}y_n}{dt^{n-2}} & = 0. \end{array}$$

If (2.7.3) is inserted in (2.7.1) and the conditions (2.7.4) imposed, the additional equation

$$\frac{du_1}{dt} \frac{d^{n-1}y_1}{dt^{n-1}} + \frac{du_2}{dt} \frac{d^{n-1}y_2}{dt^{n-1}} + \cdots + \frac{du_n}{dt} \frac{d^{n-1}y_n}{dt^{n-1}} = f(t) \quad (2.7.5)$$

is obtained. The equations (2.7.4) and (2.7.5) constitute  $n$  linear equations for the unknowns  $\dot{u}_1, \dots, \dot{u}_n$ . Solving this system for  $\dot{u}_1, \dots, \dot{u}_n$ , integrating and substituting in (2.7.3) then leads to the desired solution.

### Example 2.7.2

Solve, by variation of parameters,

$$\frac{d^3y}{dt^3} + 5\frac{d^2y}{dt^2} - 6\frac{dy}{dt} = 9e^{3t}.$$

The characteristic equation is

$$\lambda^3 + 5\lambda^2 - 6\lambda = 0$$

with roots 0, 1,  $-6$ , so that

$$y_c = C_1 + C_2 e^t + C_3 e^{-6t}.$$

Now try

$$y = u_1(t) + u_2(t)e^t + u_3(t)e^{-6t}.$$

Equations (2.7.4) become

$$\begin{aligned} \dot{u}_1 + \dot{u}_2 e^t + \dot{u}_3 e^{-6t} &= 0, \\ \dot{u}_2 e^t - 6\dot{u}_3 e^{-6t} &= 0, \end{aligned}$$

while (2.7.5) is

$$\dot{u}_2 e^t + 36\dot{u}_3 e^{-6t} = 9e^{3t}.$$

Therefore  $\dot{u}_1 = -\frac{3}{2}e^{3t}$ ,  $\dot{u}_2 = \frac{9}{7}e^{2t}$ ,  $\dot{u}_3 = \frac{3}{14}e^{9t}$  and

$$u_1 = -\frac{1}{2}e^{3t} + D_1, \quad u_2 = \frac{9}{14}e^{2t} + D_2, \quad u_3 = \frac{1}{42}e^{9t} + D_3.$$

Consequently

$$y = \frac{1}{6}e^{3t} + D_1 + D_2 e^t + D_3 e^{-6t}$$

is the required general solution. □

## 2.8 Uniqueness

The general solutions that have been constructed contain arbitrary constants and are therefore capable of assuming different values by assigning the constants in different ways. What we shall show now is that, if certain conditions are imposed, the values of the constants are fixed once and for all.

The general solution of

$$\dot{y} + a_0 y = 0 \quad (2.8.1)$$

is  $y(t) = Ce^{-a_0 t}$ . Suppose now that the extra condition  $y(0) = 0$  is imposed. Then, there is no alternative to taking  $C = 0$  and  $y(t) \equiv 0$  throughout the interval where the differential equation holds.

The second order differential equation

$$\ddot{y} + a_1 \dot{y} + a_0 y = 0 \quad (2.8.2)$$

can be rewritten, according to Section 2.4, as

$$\dot{w} - \lambda_2 w = 0 \quad (2.8.3)$$

where

$$w = \dot{y} - \lambda_1 y. \quad (2.8.4)$$

Now require that  $y = 0$  and  $\dot{y} = 0$  at  $t = 0$ . It follows from (2.8.4) that  $w = 0$  at  $t = 0$ . But (2.8.3) is of the same form as (2.8.1) and we conclude from the preceding paragraph that  $w \equiv 0$ . That makes (2.8.4) of the same type as (2.8.1) and since  $y = 0$  at  $t = 0$ ,  $y \equiv 0$  is the only possibility. In other words, the solution of (2.8.2) such that  $y = 0$  and  $\dot{y} = 0$  at  $t = 0$  vanishes throughout the interval.

Evidently, by factorising the operator in

$$\frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = 0 \quad (2.8.5)$$

and proceeding step by step as above we may deduce that *the solution of (2.8.5) such that*

$$y = 0, \quad \frac{dy}{dt} = 0, \quad \dots, \quad \frac{d^{n-1} y}{dt^{n-1}} = 0 \quad (2.8.6)$$

*at  $t = 0$  vanishes identically.*

There is an important consequence for the differential equation (2.7.1). Let it be desired to find a solution such that

$$y = D_0, \quad \frac{dy}{dt} = D_1, \quad \dots, \quad \frac{d^{n-1} y}{dt^{n-1}} = D_{n-1} \quad (2.8.7)$$

at  $t = 0$  with  $D_0, \dots, D_{n-1}$  some given constants. There could be two or more solutions which satisfy these conditions. Let  $y_1$  and  $y_2$  be two of them. Since  $y_1$  and  $y_2$  both comply with (2.8.7) their difference  $y_1 - y_2$  satisfies (2.8.6) at  $t = 0$ . Also, since both are solutions of (2.7.1), their difference  $y_1 - y_2$  is a solution of (2.8.5). By what has been established in the preceding paragraph  $y_1 - y_2$  vanishes identically, i.e., there is no difference between  $y_1$  and  $y_2$ .

This result may be stated as: *there is one, and only one, solution of*

$$\frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = f(t) \quad (2.8.8)$$

*which meets the conditions (2.8.7) at  $t = 0$ .*

To put it another way, the solution of (2.8.8) subject to (2.8.7) at  $t = 0$  is *unique*.

Of course, it is not necessary to specify the conditions at the origin—that was chosen to fix ideas. If (2.8.7) held at  $t = a$  the same argument would carry through and uniqueness would be valid still. However, it is essential that all the conditions are imposed at the same value of  $t$ . If some are enforced at one value of  $t$  and others at another, the situation is changed totally as will be seen in the next chapter.

---

## Exercises

2.1 Find the general solutions of

- (a)  $\ddot{y} - 8\dot{y} + 15y = 0$ ,
- (b)  $\ddot{y} - 8\dot{y} + 16y = 0$ ,
- (c)  $\ddot{y} + 2\dot{y} + 5y = 5t^2 + 5t$ ,
- (d)  $\ddot{y} - \dot{y} - 2y = 30e^{2t}$ ,
- (e)  $\ddot{y} + \omega^2 y = \cos \Omega t$  ( $\Omega^2 \neq \omega^2$ ),
- (f)  $\ddot{y} + \omega^2 y = \cos \omega t$ ,
- (g)  $\ddot{y} + \dot{y} + y = 3 \sin^2 t$ .

2.2 Find the general solutions of

- (a)  $\ddot{y} + 3\dot{y} + 2y = 4t^2 - 2t$ ,
- (b)  $\ddot{y} + 2\dot{y} - 2y = \frac{3}{5}t^5$ ,
- (c)  $\ddot{y} + 3\dot{y} = 6t^3 + 3t + 3$ ,
- (d)  $\ddot{y} + \dot{y} + y = 26e^t \sin t$ ,
- (e)  $\ddot{y} + 3\dot{y} = 15 \sin t + 5$ ,
- (f)  $\ddot{y} - 2\dot{y} + y = 6e^t + 2 \sin t$ ,
- (g)  $\ddot{y} + 9y = 3 \sin 3t + 6 \cos 3t$ .

2.3 Give a suitable real form, but do not evaluate the coefficients, for a particular integral of

- (a)  $\ddot{y} + y = 2t + t \sin t,$
- (b)  $\ddot{y} - 4\dot{y} + 4y = t^2 + 3te^{2t} + 4t \sin 2t,$
- (c)  $\ddot{y} + 3\dot{y} + 2y = 2(1 + t^2)e^t \sin 2t - e^t \cos t + 3e^t,$
- (d)  $\ddot{y} + 4y = 2te^t + 3t \sin 2t.$

Do you think it would be simpler to employ exponentials with complex exponents in any of these cases?

2.4 Show that

$$t^2\ddot{y} + b_1t\dot{y} + b_0y = f(t),$$

where  $b_0$  and  $b_1$  are constants, can be converted to a linear ordinary differential equation with constant coefficients by the substitution  $t = e^x$ . Hence find the general solution of

$$t^2\ddot{y} - t\dot{y} + y = 2t.$$

2.5 Use factorisation of the operator to find the general solution of

$$\ddot{y} - \dot{y} - 2y = 15e^{2t}.$$

2.6 Find the general solution of  $\ddot{y} + y = 3t$  by (a) undetermined coefficients and (b) factorisation of the operator.

2.7 By means of (a) undetermined coefficients and (b) variation of parameters find a particular integral of

$$\ddot{y} + 4y = e^t + \sin 2t$$

explaining any difference between the answers.

2.8 (a) Find a particular integral of

$$\ddot{y} + 2\dot{y} + y = 5e^{-t}(1+t)^{1/2}$$

by (i) factorisation of the operator and (ii) variation of parameters.

(b) Find the general solution of

$$\ddot{y} + y = 1/\sin t$$

by variation of parameters.

Would the method of undetermined coefficients be suitable for (a) or (b)? Would it be feasible to use factorisation of the operator for (b)?

2.9 The general solution of

$$\ddot{y} + p_1(t)\dot{y} + p_0(t)y = 0$$

is  $y = C_1y_1(t) + C_2y_2(t)$ . Show, by variation of parameters, that the general solution of

$$\ddot{y} + p_1(t)\dot{y} + p_0(t)y = f(t)$$

can be expressed as

$$y = -y_1(t) \int^t \frac{y_2(u)f(u)}{W(u)} du + y_2(t) \int^t \frac{y_1(u)f(u)}{W(u)} du$$

where  $W(t) = y_1(t)\dot{y}_2(t) - \dot{y}_1(t)y_2(t)$ .

2.10 Find the general solutions of

- (a)  $\frac{d^4y}{dt^4} + 5\frac{d^2y}{dt^2} + 4y = 0$ ,
- (b)  $\frac{d^3y}{dt^3} + 7\frac{d^2y}{dt^2} + 16\frac{dy}{dt} + 12y = 0$ ,
- (c)  $\frac{d^4y}{dt^4} + 2\frac{d^3y}{dt^3} + 10\frac{d^2y}{dt^2} = 0$ ,
- (d)  $\frac{d^4y}{dt^4} - 4\frac{d^3y}{dt^3} + 6\frac{d^2y}{dt^2} - 4\frac{dy}{dt} + y = 0$ .

2.11 Prove that, if  $p$  and  $q$  are nonnegative integers,

$$\left[ \frac{d}{dt} - a \right]^p t^q e^{at} = \begin{cases} 0 & (p > q), \\ q(q-1)\cdots(q-p+1)t^{q-p}e^{at} & (p \leq q). \end{cases}$$

2.12 Use the method of variation of parameters to solve

- (a)  $\frac{d^3y}{dt^3} - \frac{dy}{dt} = \sin t$ ,
- (b)  $\frac{d^4y}{dt^4} - y = 1 + t^2$ ,
- (c)  $\frac{d^4y}{dt^4} - \frac{d^2y}{dt^2} = 4te^t$ .

2.13 (a) Show that, if  $y_0(t)$  is a solution of the associated homogeneous differential equation of

$$\frac{d^n y}{dt^n} + p_{n-1}(t)\frac{d^{n-1}y}{dt^{n-1}} + \cdots + p_0(t)y = f(t),$$

the substitution  $y = u(t)y_0(t)$  leads to a differential equation of order  $n-1$  for  $du/dt$ .

- (b) Given that  $e^{-t^2}$  is a solution of

$$\ddot{y} + 4t\dot{y} + (4t^2 + 2)y = 0,$$

find the general solution.

- (c) Given that  $1/t$  satisfies the associated homogeneous differential equation of

$$t^2\ddot{y} + 4t\dot{y} + 2y = t \sin t,$$

find  $y$  so that  $y = 1, \dot{y} = 0$  at  $t = 1$ .



# Chapter 3

---

## Systems of Linear Ordinary Differential Equations

---

### 3.1 First-order systems of equations with constant coefficients

In studying natural phenomena, we are often interested in more than one quantity and the several quantities may well be connected by differential equations. We are therefore led to consider what happens when more than one differential equation has to be solved at a time. Suppose

$$a_1\dot{x} + b_1\dot{y} + c_1x + d_1y = f_1(t), \quad (3.1.1)$$

$$a_2\dot{x} + b_2\dot{y} + c_2x + d_2y = f_2(t) \quad (3.1.2)$$

where  $a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2$  are constants and  $x, y$  are to be found. In other words, two simultaneous differential equations of the first order have to be solved.

Multiply (3.1.1) by  $b_2$  and (3.1.2) by  $b_1$ . Then subtraction gives

$$(a_1b_2 - a_2b_1)\dot{x} + \alpha x + \beta y = F(t) \quad (3.1.3)$$

where  $\alpha = c_1b_2 - c_2b_1$ ,  $\beta = d_1b_2 - d_2b_1$  and  $F(t) = b_2f_1(t) - b_1f_2(t)$ .

There are two distinct cases to discuss according as  $a_1b_2 - a_2b_1$  is or is not zero. We call  $a_1b_2 - a_2b_1$  the **test determinant**.

#### 3.1.1 Test determinant is nonzero

If  $\beta \neq 0$ , (3.1.3) can be solved to give  $y$  in terms of  $x$  and  $\dot{x}$ . If this expression is substituted in (3.1.1) or (3.1.2) a linear differential equation with constant coefficients of order 2 is obtained for  $x$ . This differential equation can be solved by techniques already described and its general solution will involve two arbitrary constants. Having found  $x$  we can determine  $y$  from (3.1.3). No further arbitrary constants are introduced and so the whole solution contains two arbitrary constants.

If  $\beta = 0$ , (3.1.3) is a differential equation of the first order for  $x$  which can be resolved by means of an integrating factor. Its general solution will possess

one arbitrary constant. Once  $x$  is known, it can be substituted in (3.1.1) or (3.1.2) resulting in a differential equation of the first order for  $y$ . Its general solution will bring in another arbitrary constant so again the whole solution contains two arbitrary constants.

### 3.1.2 Test determinant is zero

When  $a_1b_2 = a_2b_1$ , (3.1.3) reduces to

$$\alpha x + \beta y = F(t). \quad (3.1.4)$$

If  $\beta \neq 0$ , solve (3.1.4) for  $y$  and substitute in (3.1.1) or (3.1.2). The consequent differential equation for  $x$  is of the first order and so its general solution has one arbitrary constant. With  $x$  known,  $y$  is given by (3.1.4). Since no additional arbitrary constant is entailed, the whole solution possesses one arbitrary constant.

If  $\beta = 0$ , (3.1.4) immediately furnishes  $x$  provided that  $\alpha \neq 0$ . Then (3.1.1) or (3.1.2) supplies  $y$  with one arbitrary constant. Again the whole solution contains one arbitrary constant.

If  $\beta = 0$  and  $\alpha = 0$ , (3.1.4) becomes  $0 = F(t)$ . There are now two possibilities. Either  $F(t)$  is not zero over the interval of  $t$  under consideration when (3.1.4) cannot be satisfied and the original differential equations are inconsistent, or  $F(t)$  is zero over the interval and then (3.1.2) is a constant multiple of (3.1.1).

To sum up, when the test determinant is nonzero, the whole solution can be found and contains two arbitrary constants. When the test determinant vanishes, either there is a solution and it includes one arbitrary constant, or there is no solution, or the two differential equations are not different.

Although  $\dot{y}$  was eliminated to arrive at (3.1.3), a similar equation could be obtained by eliminating  $\dot{x}$ . The general conclusion concerning the role of the test determinant would remain unaltered. In practice, it is a matter of convenience whether  $\dot{x}$  or  $\dot{y}$  is eliminated.

#### **Example 3.1.1**

Solve

$$\dot{x} + 2\dot{y} + x - y = t, \quad (3.1.5)$$

$$\dot{x} - \dot{y} - x - 2y = 1. \quad (3.1.6)$$

The test determinant is  $-1 - 2$ , which is nonzero; so the whole solution should possess two arbitrary constants.

Eliminating  $\dot{y}$  from (3.1.5) and (3.1.6) we obtain

$$3\dot{x} - x - 5y = t + 2$$

whence

$$y = \frac{1}{5}(3\dot{x} - x - t - 2). \quad (3.1.7)$$

Insertion of (3.1.7) in (3.1.6) gives

$$-\frac{3}{5}\ddot{x} - \frac{3}{5}x = -\frac{2}{5}t$$

or

$$\ddot{x} + x = \frac{2}{3}t.$$

It follows that

$$x = C_1 \cos t + C_2 \sin t + \frac{2}{3}t. \quad (3.1.8)$$

From (3.1.7)

$$y = \frac{1}{5}(3C_2 - C_1) \cos t - \frac{1}{5}(3C_1 + C_2) \sin t - \frac{1}{3}t.$$

Only two arbitrary constants  $C_1$  and  $C_2$  appear in the solution in accordance with earlier observations.  $\square$

Having found  $x$  in (3.1.8), one might have tried to find  $y$  from (3.1.6) instead of (3.1.7). Not only would this require more effort but also it would display another feature. Substitution of (3.1.8) in (3.1.6) gives

$$\dot{y} + 2y = (C_2 - C_1) \cos t - (C_1 + C_2) \sin t - \frac{2}{3}t - \frac{1}{3}$$

from which can be deduced

$$y = C_3 e^{-2t} + \frac{1}{5}(3C_2 - C_1) \cos t - \frac{1}{5}(3C_1 + C_2) \sin t - \frac{1}{3}t. \quad (3.1.9)$$

An extra arbitrary constant  $C_3$  has made its presence known. However, because we did not find  $y$  from (3.1.7), there is no guarantee that (3.1.5) is satisfied. If (3.1.8) and (3.1.9) are put in (3.1.5) it will be discovered that  $x$  and  $y$  do not satisfy (3.1.5) unless  $C_3 = 0$ . It is therefore essential to minimise labour, to work through (3.1.7) or (3.1.3) in the general case.

There is another technique which suggests itself and which should be avoided, namely to try to treat (3.1.5) and (3.1.6) like algebraic equations and remove both  $\dot{x}$  and  $x$  (or  $\dot{y}$  and  $y$ ) at the same time. Apply  $(d/dt) - 2$  to (3.1.5) and  $(d/dt) - 1$  to (3.1.6). There results

$$\begin{aligned} \ddot{x} + 3\dot{x} + 2x + 2\ddot{y} + 3\dot{y} - 2y &= 1 + 2t, \\ 2\ddot{x} - 3\dot{x} + x - 2\ddot{y} - 3\dot{y} + 2y &= -1. \end{aligned}$$

By addition

$$3(\ddot{x} + x) = 2t,$$

which supplies the same  $x$  as in (3.1.8). But, unless we re-derive (3.1.7), we have to determine  $y$  by substituting for  $x$  in one of the original differential

equations. This will run into the same trouble as in the preceding paragraph and recover (3.1.9).

**Example 3.1.2**

Find the solution of

$$\dot{x} - 2\dot{y} - x = e^t, \quad (3.1.10)$$

$$-2\dot{x} + 4\dot{y} + y = 1. \quad (3.1.11)$$

Here the test determinant vanishes and one of three possibilities may occur. Elimination of  $\dot{y}$  provides

$$2x + y = 2e^t + 1. \quad (3.1.12)$$

Thus the equations are consistent and different; consequently, there will be a general solution with one arbitrary constant only.

Substitute for  $y$  from (3.1.12) in (3.1.10). Then

$$5\dot{x} + x = 5e^t$$

whence

$$x = Ce^{-t/5} + \frac{5}{6}e^t.$$

From (3.1.12)

$$y = \frac{1}{3}e^t + 1 - 2Ce^{-t/5}$$

and the whole solution contains the single arbitrary constant  $C$ . □

Finally, note that, when the test determinant does not vanish, (3.1.3) can be divided by  $a_1b_2 - a_2b_1$  with the result

$$\dot{x} = \alpha_1x + \beta_1y + F_1(t). \quad (3.1.13)$$

Similarly, by getting rid of  $\dot{x}$  from (3.1.1) and (3.1.2), we obtain

$$\dot{y} = \alpha_2x + \beta_2y + F_2(t). \quad (3.1.14)$$

Thus (3.1.1) and (3.1.2) could be replaced by (3.1.13) and (3.1.14), if desired, so long as the test determinant is nonzero.

### 3.2 Replacement of one differential equation by a system

The second-order differential equation

$$\ddot{y} + a_1\dot{y} + a_0y = f(t) \quad (3.2.1)$$

can be represented as a system of first-order equations. Put

$$y(t) = x_1(t), \quad \dot{y}(t) = x_2(t). \quad (3.2.2)$$

From the first of (3.2.2),  $\dot{y} = \dot{x}_1$  and so, from the second of (3.2.2),

$$\dot{x}_1 = x_2. \quad (3.2.3)$$

Furthermore, (3.2.1) can be written as

$$\dot{x}_2 = f(t) - a_1 x_2 - a_0 x_1. \quad (3.2.4)$$

Thus (3.2.1) gives rise to the system of first-order equations (3.2.3) and (3.2.4). Any solution  $y$  of (3.2.1) provides a solution of the system via the identification (3.2.2). Conversely, given a solution of (3.2.3) and (3.2.4), we can substitute from (3.2.3) in (3.2.4) to obtain

$$\ddot{x}_1 + a_1 \dot{x}_1 + a_0 x_1 = f(t)$$

so that a solution of (3.2.1) is obtained by putting  $y = x_1$ . Therefore the differential equation (3.2.1) and the system (3.2.3)–(3.2.4) are equivalent. It is thereby possible to deduce properties of a second-order differential equation from those of a first-order system or vice versa.

These notions can be extended to the differential equation of order  $n$

$$\frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = f(t) \quad (3.2.5)$$

by placing

$$y(t) = x_1(t), \quad \dot{y} = x_2(t), \quad \dots, \quad \frac{d^{n-1}}{dt^{n-1}} y(t) = x_n(t).$$

It is evident from these last relations that

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_3, \\ &\vdots \\ \dot{x}_{n-1} &= x_n. \end{aligned}$$

Moreover, (3.2.5) can be expressed as

$$\dot{x}_n = f(t) - a_0 x_1 - \cdots - a_{n-1} x_n.$$

Again a system (of  $n$  first-order equations) has been produced. The equivalence of the system and (3.2.5) can be demonstrated in the same manner as for the second-order differential equation.

The systems derived here have the same structure as (3.1.13) and (3.1.14) but the latter are more general than the former. From now on it will be assumed that the test determinant of our system is nonzero so that (3.1.13) and (3.1.14) are valid. Their generalisation will be investigated in the next section.

### 3.3 The general system

A system of differential equations,

$$\dot{x}_i = \sum_{j=1}^n a_{ij}x_j + f_i(t) \quad (i = 1, \dots, n), \quad (3.3.1)$$

where every  $a_{ij}$  is a constant, is known as a **system of linear differential equations of the first order with constant coefficients**. The system

$$\dot{x}_i = \sum_{j=1}^n a_{ij}x_j \quad (i = 1, \dots, n) \quad (3.3.2)$$

is called the **associated homogeneous system**. A general solution of (3.3.1) or of (3.3.2) must determine the  $n$  quantities  $x_1, x_2, \dots, x_n$ . It is convenient to use the abbreviated notation  $\mathbf{x}$  for the  $n$  quantities  $x_1, x_2, \dots, x_n$  and say that  $\mathbf{x}$  is a solution of (3.3.2) when the  $x_1, x_2, \dots, x_n$  satisfy (3.3.2).

Suppose now that  $\mathbf{x}^{(k)}$  is a solution of (3.3.2) for  $k = 1, \dots, n$ . Consider

$$x_i = \sum_{k=1}^n C_k x_i^{(k)} \quad (i = 1, \dots, n) \quad (3.3.3)$$

where the  $C_k$  are constants. Then

$$\begin{aligned} \dot{x}_i &= \sum_{k=1}^n C_k \dot{x}_i^{(k)} = \sum_{k=1}^n C_k \sum_{j=1}^n a_{ij} x_j^{(k)} \\ &= \sum_{j=1}^n a_{ij} \sum_{k=1}^n C_k x_j^{(k)} \\ &= \sum_{j=1}^n a_{ij} x_j \end{aligned}$$

so that (3.3.3) also furnishes a solution of the associated homogeneous system. This explains why the system is called linear.

The formula (3.3.3) is a candidate for the general solution. It will be satisfactory if we can choose  $C_1, \dots, C_n$  so that  $x_i(t_0) = x_{i0}$  ( $i = 1, \dots, n$ ) for any  $t_0$  in the interval under consideration and for any selection of the constants  $x_{i0}$  that we care to make. The choice is possible if

$$\sum_{k=1}^n C_k x_i^{(k)}(t_0) = x_{i0} \quad (i = 1, \dots, n).$$

These constitute  $n$  linear algebraic equations for the unknowns  $C_1, \dots, C_n$ . They can be solved for arbitrary right-hand sides if, and only if, the determinant of the coefficients is nonzero, i.e.,

$$\begin{vmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(n)} \\ \vdots & \vdots & & \vdots \\ x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(n)} \end{vmatrix} \neq 0.$$

The determinant is known as the **Wronskian** and written  $W(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)})$  for brevity. If the Wronskian is nonzero throughout the interval, the solutions  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  are said to form a **fundamental system**. We conclude that if  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  form a fundamental system, the general solution of (3.3.2) can be expressed as

$$x_i = \sum_{k=1}^n C_k x_i^{(k)} \quad (i = 1, \dots, n)$$

where  $C_1, \dots, C_n$  are arbitrary constants.

Given the general solution of the associated homogeneous system, the general solution of (3.3.1) is

$$x_i = \sum_{k=1}^n C_k x_i^{(k)} + X_i \quad (i = 1, \dots, n) \quad (3.3.4)$$

where  $X_1, \dots, X_n$  is some particular solution of (3.3.1). The proof is the same as for the second-order differential equation, the first term on the right of (3.3.4) corresponding to the complementary function and the second to the particular integral.

The most reliable way of finding  $\mathbf{X}$  for general values of  $n$  is the method of variation of parameters. With the solutions  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  forming a fundamental system, we look for a solution of (3.3.1) of the type

$$x_i(t) = \sum_{k=1}^n u_k(t) x_i^{(k)}(t) \quad (i = 1, \dots, n). \quad (3.3.5)$$

Then

$$\begin{aligned} \dot{x}_i &= \sum_{k=1}^n (\dot{u}_k x_i^{(k)} + u_k \dot{x}_i^{(k)}) \\ &= \sum_{k=1}^n \dot{u}_k x_i^{(k)} + \sum_{k=1}^n u_k \sum_{j=1}^n a_{ij} x_j^{(k)} \end{aligned}$$

because  $\mathbf{x}^{(k)}$  satisfies (3.3.2). Thus (3.3.5) implies that

$$\dot{x}_i - \sum_{j=1}^n a_{ij} x_j = \sum_{k=1}^n \dot{u}_k x_i^{(k)},$$

and (3.3.1) is satisfied if

$$\sum_{k=1}^n \dot{u}_k x_i^{(k)} = f_i \quad (i = 1, \dots, n). \quad (3.3.6)$$

There are  $n$  equations in (3.3.6) for the  $n$  unknowns  $\dot{u}_1, \dots, \dot{u}_n$ . They can always be solved because the determinant of the coefficients is the same as the Wronskian, which is nonzero because  $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$  make a fundamental system. Integration supplies  $u_1, \dots, u_n$  and a particular integral has been found.

The reader should verify that, for a single differential equation of order  $n$ , the equations (3.3.6) do go over to (2.7.4) and (2.7.5) when the substitutions of Section 3.2 are made.

---

### 3.4 The fundamental system

It is evident from the preceding section that the general solution of a system can be elicited if a fundamental system can be unearthed. For the single differential equation, searching for solutions proportional to  $e^{\lambda t}$  was profitable and so the same device may be effective for a system. Therefore try  $x_i = c_i e^{\lambda t}$  for  $i = 1, \dots, n$ . The system (3.3.2) will be satisfied if

$$\lambda c_i e^{\lambda t} = \sum_{j=1}^n a_{ij} c_j e^{\lambda t}$$

or

$$\sum_{j=1}^n a_{ij} c_j = \lambda c_i \quad (i = 1, \dots, n). \quad (3.4.1)$$

The linear equations (3.4.1) will force every  $c_i$  to be zero unless the determinant of the coefficients vanishes. To obtain a nonzero solution, at least one  $c_i$  must be different from zero. Hence the determinant of the coefficients must be made to vanish, i.e.,

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{vmatrix} = 0. \quad (3.4.2)$$

When expanded, the determinant becomes a polynomial of degree  $n$  in  $\lambda$ , which may be expressed as

$$\lambda^n + p_{n-1}\lambda^{n-1} + \dots + p_0 = 0. \quad (3.4.3)$$

This polynomial has a special property (needed later) which is usually designated as the Cayley-Hamilton Theorem. Denote the  $n \times n$ -matrix  $(a_{ij})$ , with entries as in (3.4.2) when  $\lambda = 0$ , by  $A$ . Then, if  $I$  is the unit  $n \times n$ -matrix,

$$A^n + p_{n-1}A^{n-1} + \cdots + p_0I = 0. \quad (3.4.4)$$

There are  $n$  values of  $\lambda$  which satisfy (3.4.2). Let these roots be denoted by  $\lambda_1, \dots, \lambda_n$ ; they are called the **eigenvalues** of the matrix  $(a_{ij})$ . Put  $\lambda = \lambda_k$  in (3.4.1) and solve for  $c_i$  to obtain  $c_i^{(k)}$ , say. Then  $x_i = c_i^{(k)}e^{\lambda_k t}$  is a solution of the associated homogeneous system.

We must now investigate whether the identification  $x_i^{(k)} = c_i^{(k)}e^{\lambda_k t}$  for  $k = 1, \dots, n$  will construct a fundamental system. Suppose that  $\lambda_1, \dots, \lambda_n$  are all different. Let there be  $n$  constants  $B_1, \dots, B_n$  such that

$$\sum_{k=1}^n B_k c_i^{(k)} = 0 \quad (i = 1, \dots, n). \quad (3.4.5)$$

Then

$$\sum_{j=1}^n a_{ij} \sum_{k=1}^n B_k c_j^{(k)} = 0 \quad (i = 1, \dots, n)$$

or, from (3.4.1),

$$\sum_{k=1}^n \lambda_k B_k c_i^{(k)} = 0 \quad (i = 1, \dots, n). \quad (3.4.6)$$

Multiply (3.4.5) by  $\lambda_n$  and subtract from (3.4.6). Then

$$\sum_{k=1}^{n-1} (\lambda_k - \lambda_n) B_k c_i^{(k)} = 0. \quad (3.4.7)$$

Since  $\lambda_k \neq \lambda_n$  for  $k \neq n$ , we have equations of the same form as (3.4.6) except that  $c_i^{(n)}$  has been removed. Starting from (3.4.7) we can repeat the process and strike off  $c_i^{(n-1)}$ . Continuing in this way we arrive at

$$B'_1 c_i^{(1)} = 0 \quad (i = 1, \dots, n) \quad (3.4.8)$$

where  $B'_1$  is a nonzero multiple of  $B_1$ . At least one of  $c_1^{(1)}, \dots, c_n^{(1)}$  is not zero so that  $B'_1 = 0$ , which implies that  $B_1 = 0$ . But now the stage before (3.4.8) will enforce  $B'_2 c_i^{(2)} = 0$ , which entails  $B_2 = 0$ . Repetition of the procedure leads to the conclusion that, if (3.4.5) holds,  $B_k = 0$  ( $k = 1, \dots, n$ ). However, that is possible only if the determinant of the coefficients is nonzero, i.e.,

$$\begin{vmatrix} c_1^{(1)} & c_1^{(2)} & \cdots & c_1^{(n)} \\ c_2^{(1)} & c_2^{(2)} & \cdots & c_2^{(n)} \\ \vdots & \vdots & & \vdots \\ c_n^{(1)} & c_n^{(2)} & \cdots & c_n^{(n)} \end{vmatrix} \neq 0.$$

Remembering that the Wronskian is  $e^{(\lambda_1 + \dots + \lambda_n)t}$  times this determinant, we see that the Wronskian does not vanish. Thus, in this case, a fundamental system has been obtained.

What has been shown is that, if all the eigenvalues are distinct, the general solution of the associated homogeneous system is

$$x_i = \sum_{k=1}^n c_i^{(k)} e^{\lambda_k t} \quad (i = 1, \dots, n). \quad (3.4.9)$$

### Example 3.4.1

Find the general solution of

$$\begin{aligned}\dot{x} &= 4x + y, \\ \dot{y} &= 3x + 2y.\end{aligned}$$

The determinantal equation (3.4.2) is

$$\begin{vmatrix} 4 - \lambda & 1 \\ 3 & 2 - \lambda \end{vmatrix} = 0$$

or

$$\lambda^2 - 6\lambda + 5 = 0.$$

Thus we can take  $\lambda_1 = 1$  and  $\lambda_2 = 5$ .

For  $\lambda = 1$ , (3.4.1) becomes

$$3c_1^{(1)} + c_2^{(1)} = 0$$

twice. Consequently  $c_1^{(1)} = C_1$ ,  $c_2^{(1)} = -3C_1$  where  $C_1$  is an arbitrary constant.

For  $\lambda = 5$ , (3.4.1) goes over to

$$-c_1^{(2)} + c_2^{(2)} = 0$$

so that  $c_1^{(2)} = C_2$ ,  $c_2^{(2)} = C_2$  where  $C_2$  is an arbitrary constant.

The desired general solution is

$$\begin{aligned}x &= C_1 e^t + C_2 e^{5t}, \\ y &= -3C_1 e^t + C_2 e^{5t}.\end{aligned}$$
□

When some of the eigenvalues are repeated, the situation is much more complicated. Assume that  $\lambda_1$  occurs  $p$  times. When  $\lambda$  is placed equal to  $\lambda_1$  in (3.4.1) there may be  $p$  solutions  $c_i^{(1)}, \dots, c_i^{(p)}$  such that

$$\sum_{k=1}^p B_k c_i^{(k)} = 0 \quad (i = 1, \dots, n)$$

enforces  $B_k = 0$  ( $k = 1, \dots, p$ ). In that case our earlier analysis ensures that the contribution of this eigenvalue has the same form as in (3.4.9).

It may happen that  $p$  solutions cannot be found. Additional solutions must now be generated. Drawing on our experience with the single differential equation we try

$$x_i = (d_i + c_i t)e^{\lambda t}.$$

This will satisfy (3.3.2) if

$$(\lambda d_i + c_i + \lambda c_i t)e^{\lambda t} = \sum_{j=1}^n a_{ij}(d_j + c_j t)e^{\lambda t} \quad (i = 1, \dots, n).$$

These can be true for an interval of  $t$  only if

$$\sum_{j=1}^n a_{ij}c_j = \lambda c_i, \quad (3.4.10)$$

$$\sum_{j=1}^n a_{ij}d_j = \lambda d_i + c_i \quad (3.4.11)$$

for  $i = 1, \dots, n$ . Equation (3.4.10) is the same as (3.4.1) and may be solved in the same way as before. Once  $c_i$  has been determined, we solve (3.4.11) for  $d_i$ . In this manner, extra solutions of the differential system may be created.

There may still not be enough to fill the  $p$  slots available. If so, quadratic and possibly higher powers of  $t$  can be added into the expression for  $x_i$ . In fact, it can be asserted that the trial solution

$$x_i = (r_i + \dots + d_i t^{p-2} + c_i t^{p-1})e^{\lambda t}$$

is bound to produce enough solutions corresponding to the eigenvalue  $\lambda_1$ .

An alternative method for manufacturing solutions is discussed in the next section.

### **Example 3.4.2**

Find the general solution of

$$\begin{aligned} \dot{x} &= 5x + 3y, \\ \dot{y} &= -3x - y. \end{aligned}$$

In this case (3.4.2) is

$$\begin{vmatrix} 5 - \lambda & 3 \\ -3 & -1 - \lambda \end{vmatrix} = 0$$

or

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

The eigenvalue 2 occurs twice. With  $\lambda = 2$  in (3.4.1)

$$\begin{aligned} 3c_1^{(1)} + 3c_2^{(1)} &= 0, \\ -3c_1^{(1)} - 3c_2^{(1)} &= 0. \end{aligned}$$

Now  $c_1^{(1)} = C_1, c_2^{(1)} = -C_1$  is the only possible solution and so is insufficient for our purposes. Invoking (3.4.11) we have

$$\begin{aligned} 3d_1 + 3d_2 &= c_1, \\ -3d_1 - 3d_2 &= -c_1 \end{aligned}$$

since  $c_2 = -c_1$ . A solution is  $d_1 = C_2, d_2 = 0, c_1 = 3C_2$ . No other solution is necessary since it will differ from this only by a solution of (3.4.10). Two solutions are known now and

$$\begin{aligned} x &= C_1 e^{2t} + C_2 (1 + 3t) e^{2t}, \\ y &= -C_1 e^{2t} - 3C_2 t e^{2t}. \end{aligned}$$

The reader should confirm that the Wronskian is nonzero, consistent with the derivation of a fundamental system. □

---

### 3.5 Matrix notation

The system (3.3.1) can be expressed in terms of matrices by introducing the column vector  $\mathbf{x}$  with components  $x_1, \dots, x_n$ , the column vector  $\mathbf{f}$  with components  $f_1, \dots, f_n$  and the matrix  $A = (a_{ij})$ . Then

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f} \quad (3.5.1)$$

and the associated homogeneous system is

$$\dot{\mathbf{x}} = A\mathbf{x}. \quad (3.5.2)$$

The similarity of (3.5.2) to the single first-order differential equation suggests that it ought to be possible to write the general solution as

$$\mathbf{x} = e^{tA}\mathbf{C} \quad (3.5.3)$$

where  $\mathbf{C}$  is an arbitrary column vector. However, (3.5.3) has no significance until a meaning is attributed to the matrix  $e^{tA}$ . A suitable definition is

$$e^{tA} = I + tA + \frac{1}{2!}t^2 A^2 + \frac{1}{3!}t^3 A^3 + \dots \quad (3.5.4)$$

where  $I$  is the unit  $n \times n$ -matrix. The presence of an infinite series in the definition means that (3.5.3) has a deceptive air of simplicity. It conceals the fact that it may be quite difficult to calculate  $e^{tA}$ .

By putting  $t = 0$  in (3.5.4) we see at once that

$$e^0 = I. \quad (3.5.5)$$

Furthermore, by taking derivatives of (3.5.4) with respect to  $t$  term-by-term, without worrying about the legitimacy, we obtain

$$\frac{d}{dt} e^{tA} = A e^{tA} = e^{tA} A. \quad (3.5.6)$$

Next, notice that the polynomial (3.4.3) would occur when considering solutions of the differential equation

$$\frac{d^n w}{dt^n} + p_{n-1} \frac{d^{n-1} w}{dt^{n-1}} + \cdots + p_0 w = 0. \quad (3.5.7)$$

This suggests that the solutions of (3.5.2) are related in some way to those of (3.5.7). In fact, a derivative of (3.5.2) gives

$$\ddot{\mathbf{x}} = A\dot{\mathbf{x}} = A^2\mathbf{x}$$

by virtue of (3.5.2). Clearly

$$\frac{d^m \mathbf{x}}{dt^m} = A^m \mathbf{x} \quad (3.5.8)$$

in general. Observe that this is consistent with (3.5.3) as can be seen by invoking (3.5.6). From (3.5.8)

$$\frac{d^n \mathbf{x}}{dt^n} + p_{n-1} \frac{d^{n-1} \mathbf{x}}{dt^{n-1}} + \cdots + p_0 \mathbf{x} = (A^n + p_{n-1} A^{n-1} + \cdots + p_0 I) \mathbf{x} = 0 \quad (3.5.9)$$

on account of (3.4.4). Comparison of (3.5.9) with (3.5.7) reveals that *each element of  $\mathbf{x}$  is a solution of the differential equation (3.5.7)*.

Suppose now that  $\mathbf{x} = \mathbf{0}$  at  $t = 0$ . Then (3.5.8) implies that the derivatives of  $\mathbf{x}$  are also zero at  $t = 0$ . Hence each element of  $\mathbf{x}$  is a solution of (3.5.7) which, together with its derivatives, vanishes at  $t = 0$ . The theory of Section 2.8 tells us that such a solution must be identically zero. Hence, *if  $\mathbf{x} = \mathbf{0}$  at  $t = 0$ , the solution of (3.5.2) is zero throughout the interval under consideration*. It follows, as in Section 2.8, that the solution of (3.5.1) *such that  $\mathbf{x} = \mathbf{C}$  at  $t = 0$  is unique*. Naturally, the point  $t = 0$  can be replaced by some other if desired.

Now consider some special solutions  $w_1, w_2, \dots, w_n$  of (3.5.7). We choose  $w_1$  so that

$$w_1 = 1, \quad dw_1/dt = 0, \quad d^2 w_1/dt^2 = 0, \quad \dots, \quad d^{n-1} w_1/dt^{n-1} = 0$$

at  $t = 0$ . For  $w_2$  we take

$$w_2 = 0, \quad dw_2/dt = 1, \quad d^2w_2/dt^2 = 0, \quad \dots, \quad d^{n-1}w_2/dt^{n-1} = 0$$

at  $t = 0$  and generally we select  $d^m w_i / dt^m$  to vanish at  $t = 0$  except for  $m = i - 1$  when the value is to be unity. The functions  $w_1, w_2, \dots, w_n$  are uniquely defined (Section 2.8) and any solution of (3.5.7) can be expressed in terms of them by adding appropriate multiples to reproduce values specified for the solution and its derivatives at  $t = 0$ . Accordingly, there are constant vectors  $\mathbf{c}_1, \dots, \mathbf{c}_n$  such that

$$\mathbf{x} = w_1 \mathbf{c}_1 + w_2 \mathbf{c}_2 + \dots + w_n \mathbf{c}_n. \quad (3.5.10)$$

Let  $\mathbf{x} = \mathbf{C}$  at  $t = 0$ . Putting  $t = 0$  in (3.5.10) we have  $\mathbf{c}_1 = \mathbf{C}$ . Take a derivative of (3.5.10) and put  $t = 0$ . Then  $\dot{\mathbf{x}} = \mathbf{c}_2 = A\mathbf{C}$  by virtue of (3.5.8). Repeating the process we obtain  $\mathbf{c}_3 = A^2\mathbf{C}$  and generally  $\mathbf{c}_m = A^{m-1}\mathbf{C}$ . Consequently, the solution of (3.5.2) such that  $\mathbf{x} = \mathbf{C}$  at  $t = 0$  is

$$\mathbf{x} = (w_1 I + w_2 A + \dots + w_n A^{n-1})\mathbf{C}.$$

On the other hand, (3.5.5) indicates that (3.5.3) is the solution that is  $\mathbf{C}$  at  $t = 0$ . By the uniqueness property, which has been demonstrated already, the two solutions must be the same, i.e.,

$$e^{tA}\mathbf{C} = (w_1 I + w_2 A + \dots + w_n A^{n-1})\mathbf{C}.$$

But  $\mathbf{C}$  is arbitrary since no particular values have been assigned to it and so

$$e^{tA} = w_1 I + w_2 A + \dots + w_n A^{n-1}. \quad (3.5.11)$$

The formula (3.5.11) has the advantage over (3.5.4) of being a finite series rather than an infinite one. However, it does entail the determination of  $n$  solutions of (3.5.7). Whether it is more effective in practice than the method described in earlier sections is more difficult to assess. Probably, the earlier method is best when all the eigenvalues of  $A$  are distinct. With repeated eigenvalues the scales will tend to tilt towards (3.5.11). To aid the reader in forming an assessment some of the preceding examples will be tackled by the method of this section.

### **Example 3.5.1**

Find the general solution of

$$\begin{aligned} \dot{x} &= 4x + y, \\ \dot{y} &= 3x + 2y. \end{aligned}$$

The analogue of (3.5.7) is, from Example 3.4.1,

$$\ddot{w} - 6\dot{w} + 5 = 0.$$

Consequently,  $w_1 = (5e^t - e^{5t})/4$  and  $w_2 = (e^{5t} - e^t)/4$ . Substitute in (3.5.11) with  $n = 2$ . Then

$$\begin{aligned} 4e^{tA} &= (5e^t - e^{5t}) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + (e^{5t} - e^t) \begin{pmatrix} 4 & 1 \\ 3 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 3e^{5t} + e^t & e^{5t} - e^t \\ 3e^{5t} - 3e^t & e^{5t} + 3e^t \end{pmatrix}. \end{aligned}$$

Consequently, if  $\mathbf{C}$  has elements  $C_1, C_2$

$$\begin{aligned} x &= (3C_1 + C_2)e^{5t}/4 + (C_1 - C_2)e^t/4, \\ y &= (3C_1 + C_2)e^{5t}/4 - 3(C_1 - C_2)e^t/4. \end{aligned}$$

This has the same structure as in Example 3.4.1 although here  $C_1$  and  $C_2$  are the values of  $x$  and  $y$  at  $t = 0$ .  $\square$

### **Example 3.5.2**

Find the general solution of

$$\begin{aligned} \dot{x} &= 5x + 3y, \\ \dot{y} &= -3x - y. \end{aligned}$$

The differential equation to be solved is, by Example 3.4.2,

$$\ddot{w} - 4\dot{w} + 4w = 0.$$

Therefore  $w_1 = (1 - 2t)e^{2t}$  and  $w_2 = te^{2t}$ . Hence

$$e^{tA} = e^{2t} \begin{pmatrix} 1 + 3t & 3t \\ -3t & 1 - 3t \end{pmatrix}$$

resulting in

$$\begin{aligned} x &= \{C_1 + 3(C_1 + C_2)t\}e^{2t}, \\ y &= \{C_2 - 3(C_1 + C_2)t\}e^{2t}. \end{aligned}$$

Again there is consistency with Example 3.4.2 on adjusting the constants.  $\square$

### **Example 3.5.3**

Find the general solution of

$$\begin{aligned} \dot{x} &= 3x + y, \\ \dot{y} &= 3y + z, \\ \dot{z} &= 3z. \end{aligned}$$

In this case  $A = \begin{pmatrix} 3 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{pmatrix}$  and  $A^2 = \begin{pmatrix} 9 & 6 & 1 \\ 0 & 9 & 6 \\ 0 & 0 & 0 \end{pmatrix}$ . The relevant differential equation is

$$\frac{d^3w}{dt^3} - 9\frac{d^2w}{dt^2} + 27\frac{dw}{dt} - 27w = 0.$$

Hence  $w_1 = (1 - 3t + 9t^2/2)e^{3t}$ ,  $w_2 = (t - 3t^2)e^{3t}$ ,  $w_3 = t^2e^{3t}/2$  leading to

$$e^{tA} = e^{3t} \begin{pmatrix} 1 & t & t^2/2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\begin{aligned} x &= (C_1 + C_2t + C_3t^2/2)e^{3t}, \\ y &= (C_2 + C_3t)e^{3t}, \\ z &= C_3e^{3t}. \end{aligned}$$
□

Matrix notation offers a neat way of representing a particular integral of the system

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f}(t). \quad (3.5.12)$$

When  $\mathbf{f}$  is absent we know that a solution of (3.5.12) can be expressed as  $\mathbf{x} = e^{tA}\mathbf{x}_0$ . Therefore, as in the method of variation of parameters, try

$$\mathbf{x} = e^{tA}\mathbf{y}(t) \quad (3.5.13)$$

as a solution of (3.5.12). Since, from (3.5.6),

$$\dot{\mathbf{x}} = Ae^{tA}\mathbf{y}(t) + e^{tA}\dot{\mathbf{y}}(t) = A\mathbf{x} + e^{tA}\dot{\mathbf{y}}(t)$$

(3.5.12) is satisfied provided that

$$e^{tA}\dot{\mathbf{y}}(t) = \mathbf{f}(t).$$

Hence  $\dot{\mathbf{y}}(t) = e^{-tA}\mathbf{f}(t)$  and

$$\mathbf{y}(t) = \int^t e^{-uA}\mathbf{f}(u)du.$$

It follows from (3.5.13) that a particular integral of (3.5.12) is

$$\mathbf{x}(t) = e^{tA} \int^t e^{-uA}\mathbf{f}(u)du. \quad (3.5.14)$$

The meaning of  $\int \mathbf{g}(u)du$  is specified by

$$\int \mathbf{g}(u)du = \begin{pmatrix} \int g_1(u)du \\ \int g_2(u)du \\ \vdots \\ \int g_n(u)du \end{pmatrix}$$

when  $\mathbf{g}^T = (g_1, g_2, \dots, g_n)$  where  $\mathbf{g}^T$  is the transpose of  $\mathbf{g}$ .

Observe that  $e^{-tA}$  differs from  $e^{tA}$  only in the sign of  $t$ . Consequently,  $e^{-tA}$  can be written down as soon as  $e^{tA}$  has been calculated.

### Example 3.5.4

Find a particular integral for

$$\begin{aligned}\dot{x} &= 5x + 3y + 2te^{2t}, \\ \dot{y} &= -3x - y + 4.\end{aligned}$$

We know from Example 3.5.2 that

$$e^{tA} = e^{2t} \begin{pmatrix} 1+3t & 3t \\ -3t & 1-3t \end{pmatrix}$$

and so

$$e^{-tA} = e^{-2t} \begin{pmatrix} 1-3t & -3t \\ 3t & 1+3t \end{pmatrix}.$$

Hence, with  $\mathbf{f}^T = (2te^{2t}, 4)$ ,

$$\int^t e^{-uA} \mathbf{f}(u) du = \begin{pmatrix} t^2 - 2t^3 + 3(2t+1)e^{-2t} \\ 2t^3 - (6t+5)e^{-2t} \end{pmatrix}.$$

We do not need to include arbitrary constants since they add only multiples of the complementary function.

After insertion in (3.5.14) we obtain the particular integral

$$\begin{aligned}x &= (t^2 + t^3)e^{2t} + 3, \\ y &= -t^3e^{2t} - 5.\end{aligned}$$
□

## 3.6 Initial and boundary value problems

It has been mentioned from time to time that solutions of differential equations are often subject to extra conditions. This section will be devoted to a discussion of two types of conditions that are of frequent occurrence in practice.

Suppose a solution of

$$\frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = f(t) \quad (3.6.1)$$

is required such that

$$y = y^{(0)}, \quad \frac{dy}{dt} = y^{(1)}, \quad \dots, \quad \frac{d^{n-1}y}{dt^{n-1}} = y^{(n-1)}$$

at  $t = t_0$ , with  $y^{(0)}, \dots, y^{(n-1)}$  prescribed constants. This is termed an **initial value problem**.

The general solution of (3.6.1) is

$$y = C_1 y_1 + \dots + C_n y_n + y_p$$

where  $y_p$  is a particular integral and the remaining terms represent the complementary function. Then the imposed conditions can be complied with if

$$C_1 \frac{d^m y_1}{dt^m} + \dots + C_n \frac{d^m y_n}{dt^m} = y^{(m)} - \frac{d^m y_p}{dt^m} \quad (m = 0, 1, \dots, n-1)$$

when  $t = t_0$ . These  $n$  equations for  $C_1, \dots, C_n$  can always be solved if

$$\begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ \frac{dy_1}{dt} & \frac{dy_2}{dt} & \cdots & \frac{dy_n}{dt} \\ \vdots & \vdots & & \vdots \\ \frac{d^{n-1}y_1}{dt^{n-1}} & \frac{d^{n-1}y_2}{dt^{n-1}} & \cdots & \frac{d^{n-1}y_n}{dt^{n-1}} \end{vmatrix} \neq 0.$$

This determinant is the same as that derived from the Wronskian of Section 3.3 when the differential equation is converted to a first-order system and, accordingly, is also known as a Wronskian. The nonvanishing of the Wronskian warrants the statement that

$$B_1 y_1(t) + \dots + B_n y_n(t) = 0 \quad (3.6.2)$$

for an interval of  $t$ , which necessitates the constants  $B_1, \dots, B_n$  all being zero. For  $n-1$  derivatives of (3.6.2) give a set of equations for  $B_1, \dots, B_n$  with nonzero determinant and zero right-hand side. Expressed in other words, the nonvanishing of the Wronskian makes  $y_1, \dots, y_n$  **linearly independent** over the interval. To put it another way, it makes certain that the complementary function has been determined correctly.

Since there is only one set of  $C_1, \dots, C_n$  that satisfies the equations for nonzero Wronskian, it has been demonstrated that *the initial value problem always possesses a solution and there is only one which satisfies the imposed conditions*. This constitutes another verification of the uniqueness property.

The initial value problem is characterised by all the restrictions being applied at a single value of  $t$ . In some instances the conditions refer to more than one value of  $t$ —we then have a **boundary value problem**. In contrast

to the initial value problem, it is by no means certain that a boundary value problem has a solution. Consider

$$\ddot{y} + y = 0$$

of which the general solution is

$$y = C_1 \cos t + C_2 \sin t.$$

Let the conditions be  $y(0) = 0, y(1) = 0$ . The first requires  $C_1 = 0$  and the second  $C_2 \sin 1 = 0$ . Since  $\sin 1 \neq 0$  we must have  $C_2 = 0$  and the only solution is the trivial one which vanishes everywhere. Now change the conditions to  $y(0) = 0, y(\pi) = 0$ . In this event,  $y = C_2 \sin t$  is a solution with  $C_2$  arbitrary. Thus, boundary value problems may have many solutions or none (if the trivial one is discounted). It is also obvious that the interval of  $t$  has a critical role to play.

Instead of varying the interval, it is usual to fix it and incorporate a parameter in the differential equation. A typical problem might be to solve

$$\frac{d}{dt} \left( p(t) \frac{dy}{dt} \right) + \{q(t) + \lambda\}y = 0 \quad (3.6.3)$$

subject to  $y(a) = 0, y(b) = 0$ . This is an example of a **Sturm–Liouville problem**. The values of  $\lambda$ , which is independent of  $t$ , are crucial. For some there will be only the trivial solution and for others there will be many solutions. Those  $\lambda$  for which nontrivial solutions exist are called **eigenvalues** and the corresponding solutions **eigenfunctions**.

### Example 3.6.1

Consider

$$\ddot{y} + \lambda y = 0$$

under the conditions  $y(0) = 0, y(\pi) = 0$ .

If  $\lambda = 0$ , the general solution is  $y = A + Bt$ , which satisfies the boundary conditions only if  $A = 0$  and  $B = 0$ . Therefore  $\lambda = 0$  is not an eigenvalue.

If  $\lambda \neq 0$ , the general solution is

$$y = C_1 \cos \sqrt{\lambda}t + C_2 \sin \sqrt{\lambda}t.$$

To comply with the boundary conditions we must have  $C_1 = 0$  and  $C_2 \sin \sqrt{\lambda}\pi = 0$ . For a nontrivial solution  $C_2 \neq 0$  and  $\lambda = m^2$  where  $m$  is a positive integer. The eigenvalues are real and infinite in number. They may be designated  $\lambda_1, \lambda_2, \dots$  where  $\lambda_m = m^2$ . The eigenfunction corresponding to  $\lambda_m$  is  $C_m \sin mt$  where  $C_m$  is arbitrary.  $\square$

The discussion of (3.6.3) will assume that  $p(t)$  is a continuously differentiable real function that does not change sign for any  $t$  in  $(a, b)$ . No loss of

generality is incurred in taking it to be positive. The function  $q$  will be assumed to be real and continuous in  $(a, b)$ . We shall also suppose that there is an infinite set of eigenvalues  $\lambda_1, \lambda_2, \dots$  with associated eigenfunctions  $Y_1, Y_2, \dots$

With these assumptions the first thing to be shown is that *the eigenvalues are real*.  $Y_m$  satisfies

$$\frac{d}{dt} \left( p(t) \frac{dY_m}{dt} \right) + \{q(t) + \lambda_m\} Y_m = 0 \quad (3.6.4)$$

and  $Y_m(a) = 0, Y_m(b) = 0$ . By taking a complex conjugate

$$\frac{d}{dt} \left( p(t) \frac{dY_m^*}{dt} \right) + \{q(t) + \lambda_m^*\} Y_m^* = 0 \quad (3.6.5)$$

and  $Y_m^*(a) = 0, Y_m^*(b) = 0$ . Multiply (3.6.4) by  $Y_m^*$ , (3.6.5) by  $Y_m$  and subtract. There results

$$Y_m^* \frac{d}{dt} \left( p(t) \frac{dY_m}{dt} \right) - Y_m \frac{d}{dt} \left( p(t) \frac{dY_m^*}{dt} \right) + (\lambda_m - \lambda_m^*) |Y_m|^2 = 0.$$

Hence

$$\begin{aligned} (\lambda_m^* - \lambda_m) \int_a^b |Y_m|^2 dt &= \int_a^b \left\{ Y_m^* \frac{d}{dt} \left( p(t) \frac{dY_m}{dt} \right) - Y_m \frac{d}{dt} \left( p(t) \frac{dY_m^*}{dt} \right) \right\} dt \\ &= \left[ Y_m^* p(t) \frac{dY_m}{dt} - Y_m p(t) \frac{dY_m^*}{dt} \right]_a^b \end{aligned} \quad (3.6.6)$$

by integration by parts. The right-hand side of (3.6.6) is zero because of the conditions on  $Y, Y^*$  at  $t = a, t = b$ . The integral on the left is positive because  $Y_m$  is a nontrivial solution. Consequently,  $\lambda_m = \lambda_m^*$  and  $\lambda_m$  is real.

The reality of the eigenvalues means that there is no loss of generality in taking the eigenfunctions to be real. With this understood, note that

$$\frac{d}{dt} \left( p(t) \frac{dY_n}{dt} \right) + \{q(t) + \lambda_n\} Y_n = 0 \quad (3.6.7)$$

and  $Y_n(a) = 0, Y_n(b) = 0$ . Multiply (3.6.4) by  $Y_n$ , (3.6.7) by  $Y_m$ , subtract and proceed as above. Then

$$(\lambda_n - \lambda_m) \int_a^b Y_m Y_n dt = \left[ p(t) \left( Y_n \frac{dY_m}{dt} - Y_m \frac{dY_n}{dt} \right) \right]_a^b. \quad (3.6.8)$$

The right-hand side is zero on account of the values of  $Y_m, Y_n$  at the endpoints. Therefore, if  $\lambda_m \neq \lambda_n$ ,

$$\int_a^b Y_m Y_n dt = 0. \quad (3.6.9)$$

Functions that satisfy (3.6.9) are said to be *orthogonal*, i.e., the eigenfunctions of distinct eigenvalues are orthogonal. If, in addition, the eigenfunctions are **normalised** so that  $\int_a^b Y_m^2 dt = 1$ , the eigenfunctions are called **orthonormal**.

It will not have escaped the reader's notice that the right-hand sides of (3.6.6) and (3.6.8) can vanish for conditions other than those delineated. For example, if  $y(a) = 0$  is replaced by

$$\alpha_1 y(a) + \alpha_2 \dot{y}(a) = 0, \quad (3.6.10)$$

where at least one of the real  $\alpha_1, \alpha_2$  is nonzero, the right-hand sides are still zero. A similar remark is true if  $y(b) = 0$  is changed to

$$\beta_1 y(b) + \beta_2 \dot{y}(b) = 0 \quad (3.6.11)$$

where  $\beta_1, \beta_2$  are real with at least one nonzero. Equations (3.6.10) and (3.6.11) can be deemed **standard boundary conditions** (they include the previous ones by putting  $\alpha_2 = 0, \beta_2 = 0$ ). What has been shown is that the *eigenvalues are real and the eigenfunctions orthogonal for standard boundary conditions*.

### Example 3.6.2

Find the eigenfunctions of  $\ddot{y} + \lambda y = 0$  subject to  $\dot{y}(0) = 0, \dot{y}(\pi) = 0$ .

For  $\lambda = 0$ , the solution  $y = A + Bt$  meets the boundary conditions if  $B = 0$ . The eigenfunction is  $y = A$ .

For  $\lambda \neq 0$ , proceed as in Example 3.6.1 to show that there is an eigenvalue  $n^2$  with eigenfunction  $C_n \cos nt$ .

The first eigenfunction can be subsumed in the second group by allowing  $n = 0$ . Thus the eigenfunctions are  $C_n \cos nt$  for  $n = 0, 1, 2, \dots$ . □

Eigenvalues can also occur for **periodic boundary conditions** where  $p(a) = p(b), y(a) = y(b), \dot{y}(a) = \dot{y}(b)$ . Again the eigenvalues are real and the eigenfunctions orthogonal.

### Example 3.6.3

Find the eigenfunctions of  $\ddot{y} + \lambda y = 0$  subject to  $y(0) = y(\pi)$  and  $\dot{y}(0) = \dot{y}(\pi)$ . If  $\lambda = 0$ , the general solution is  $y = A + Bt$  which satisfies the boundary conditions if  $B = 0$  and so  $\lambda = 0$  is an eigenvalue with a corresponding constant eigenfunction.

If  $\lambda \neq 0$  then the general solution is

$$y = A \sin \sqrt{\lambda} t + B \cos \sqrt{\lambda} t.$$

On applying the boundary conditions we find that the following set of homogeneous equations must be satisfied:

$$A \sin \sqrt{\lambda} \pi + B(\cos \sqrt{\lambda} \pi - 1) = 0$$

$$A(\cos \sqrt{\lambda} \pi - 1) - B \sin \sqrt{\lambda} \pi = 0.$$

This set of equations has a nontrivial solution only if  $\cos \sqrt{\lambda}\pi = 1$ , that is  $\lambda = 4n^2$ ,  $n = 1, 2, \dots$  is an eigenvalue with corresponding eigenfunction  $y_n = A_n \sin 2nt + B_n \cos 2nt$ , where  $A_n$  and  $B_n$  are arbitrary constants. These constants can be chosen so that the eigenfunctions are normalised to give  $y_n = \sqrt{2/\pi} \sin(2nt + \alpha_n)$  with arbitrary constant  $\alpha_n$ .  $\blacksquare$

---

### 3.7 Solving the inhomogeneous differential equation

This section is concerned with the boundary value problem in which

$$\frac{d}{dt} \left( p(t) \frac{dy}{dt} \right) + q(t)y = f(t). \quad (3.7.1)$$

For simplicity, the conditions  $y(a) = 0, y(b) = 0$  will be imposed, although it will be clear that the technique is equally valid for the standard boundary conditions.

Let  $\lambda_1, \lambda_2, \dots$  and  $Y_1, Y_2, \dots$  be the eigenvalues and eigenfunctions determined in the section before. Assume that we can write

$$f(t) = \sum_{m=1}^{\infty} b_m Y_m(t).$$

Multiply by  $Y_n$  and integrate from  $a$  to  $b$ . Then, by virtue of the orthogonality of the eigenfunctions,

$$b_n \int_a^b Y_n^2 dt = \int_a^b f(t) Y_n(t) dt$$

which specifies the coefficient  $b_n$ . Putting

$$y = \sum_{m=1}^{\infty} a_m Y_m(t) \quad (3.7.2)$$

in (3.7.1) we obtain, provided that derivatives can be taken term-by-term,

$$-\sum_{m=1}^{\infty} a_m \lambda_m Y_m(t) = \sum_{m=1}^{\infty} b_m Y_m(t),$$

which suggests that (3.7.2) is the desired solution of the boundary value problem when  $a_m = -b_m/\lambda_m$ .

#### **Example 3.7.1**

Find the solution of

$$\ddot{y} = t$$

such that  $y(0) = 0, y(\pi) = 0$ .

It should first be pointed out that this problem can be solved easily without eigenfunctions but for many problems there is no option to finding the solution as a series of eigenfunctions.

The eigenfunctions are  $\sin nt$  and so

$$b_n = \int_0^\pi t \sin nt dt \Big/ \int_0^\pi \sin^2 nt dt = (-1)^{n+1} 2/n.$$

Since  $\lambda_n = n^2$  our solution is

$$y = 2 \sum_{m=1}^{\infty} \frac{(-1)^m}{m^3} \sin mt.$$

□

The derivation of the series solution rested on a number of assumptions, namely

- (a)  $f$  can be expanded in a series of eigenfunctions,
- (b)  $a_m = -b_m/\lambda_m$  for  $m = 1, 2, \dots$ ,
- (c)  $\sum_{m=1}^{\infty} b_m Y_m / \lambda_m$  is a continuous function that possesses two derivatives which can be calculated by taking derivatives of the series term-by-term.

There is no difficulty about (b) when  $\lambda_m \neq 0$  for  $m = 1, 2, \dots$  and a unique solution is obtained. If, however, one of the eigenvalues, say  $\lambda_1$ , is zero, the boundary value problem has no solution when  $b_1 \neq 0$  and an infinite number of solutions when  $b_1 = 0$ .

Both (a) and (c) raise delicate matters because they require knowledge of the properties of expansions of functions in terms of eigenfunctions. It would take us too far afield to derive these properties; so we content ourselves with a few observations without proof. Generally speaking, the smoother  $f$  is, i.e., the more derivatives it has, the more likely is the process to be legitimate. In fact, for the standard boundary conditions, the series for a piecewise smooth function converges uniformly and absolutely to the function on any closed interval in which the function is continuous. Nevertheless, even in the absence of a specific theorem, there is nothing to prevent one from carrying out the process formally and then attempting to confirm that the resulting series has the desired properties.

### 3.8 Numerical solution of linear boundary value problems

The solution of boundary value problems by numerical methods is quite different from the solution of initial value problems (see Section 1.7). The reason is that, for initial value problems, the solution is completely known at some initial time, and can be systematically stepped forward to obtain an approximation to the solution at any future time. In contrast, a boundary value problem has partial information specified at two (or more) different values of the independent variable ( $t$  or  $x$ ), so simple marching methods, such as Euler's method, are not applicable. Instead, a system of equations for the solution at each value of  $x$  must be set up and solved simultaneously. This is done by dividing the domain up into a set of grid points and replacing derivatives in the differential equation by corresponding finite difference approximations.

As an example, consider the second-order linear differential equation

$$\frac{d^2y}{dx^2} + p(x)\frac{dy}{dx} + q(x)y = g(x), \quad (3.8.1)$$

subject to the boundary conditions

$$y(a) = \alpha, \quad y(b) = \beta. \quad (3.8.2)$$

Note that the coefficients  $p$  and  $q$ , and the right-hand side  $g$ , are permitted to be functions of the independent variable  $x$ . The basic idea is to approximate derivatives of  $y$  in (3.8.1) using **central difference approximations**. The approximations for the first two derivatives are

$$\begin{aligned}\frac{dy}{dx} &= \frac{y(x+h) - y(x-h)}{2h}, \\ \frac{d^2y}{dx^2} &= \frac{y(x+h) + y(x-h) - 2y(x)}{h^2}.\end{aligned}$$

These may be derived from the Taylor expansion of  $y(x \pm h)$  about  $x$ . We now define a grid of equally spaced  $x$  values  $x_0, \dots, x_n$  (where  $x_i = a + ih$  and the step size is  $h = (b - a)/n$ ), with corresponding approximate  $y$  values  $y_0, \dots, y_n$ . Equation (3.8.1) now reads

$$\frac{y_{i+1} + y_{i-1} - 2y_i}{h^2} + p(x_i)\frac{y_{i+1} - y_{i-1}}{2h} + q(x_i)y_i = g(x_i). \quad (3.8.3)$$

By setting  $i = 1, \dots, n-1$ , this gives an equation for each of the  $n-1$  interior lattice points. The boundary conditions  $y_0 = \alpha$  and  $y_n = \beta$  provide two further equations, giving a system of  $n+1$  equations in the  $n+1$  unknowns  $y_0, \dots, y_n$ . Since the differential equation (3.8.1) is linear, the resulting system

of finite difference equations is also linear. It may, therefore, be written in the form of a matrix–vector equation  $A\mathbf{y} = \mathbf{b}$ , where

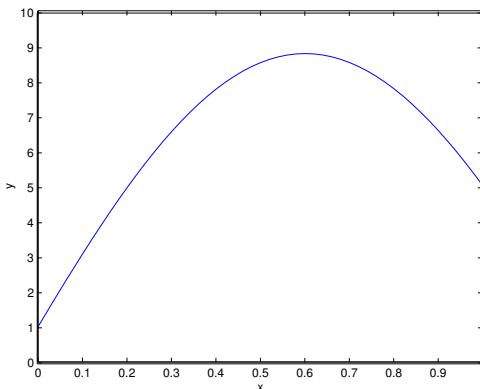
$$A = \begin{bmatrix} 1 & & & & \\ \frac{1}{h^2} - \frac{p(x_1)}{2h} & -\frac{2}{h^2} + q(x_1) & \frac{1}{h^2} + \frac{p(x_1)}{2h} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{h^2} - \frac{p(x_1)}{2h} & -\frac{2}{h^2} + q(x_1) & \frac{1}{h^2} + \frac{p(x_1)}{2h} \\ & & & & 1 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} \alpha \\ g(x_1) \\ \vdots \\ g(x_n) \\ \beta \end{bmatrix}.$$

This system can be readily solved by standard methods. Indeed, MATLAB has a built-in command for solving linear systems: the solution of the matrix equation  $A\mathbf{y} = \mathbf{b}$  may be found via the command  $\mathbf{y} = A \backslash \mathbf{b}$ . Note that, because each of the equations (3.8.3) only contains three of the unknowns ( $y_{i-1}$ ,  $y_i$  and  $y_{i+1}$ ), the matrix  $A$  has a tridiagonal structure (i.e., all entries that are not on the main diagonal or the first subdiagonal or superdiagonal are zero). Moreover, there is a clear pattern followed by the entries on these three diagonals. These properties can be used to define the matrix  $A$  efficiently. The following MATLAB function file calculates the approximate solution to the boundary value problem (3.8.1) and (3.8.2) using a specified number of lattice points  $N$ .

```
function [x, y] = bvp(coeffs, xspan, BCs, N);
% Function to solve a linear second-order BVP
% INPUTS: coeffs - handle to a function that returns the three
%          coefficients p(x), q(x) and g(x)
%          xspan - 1x2 vector containing the min and max values
%          of x
%          BCs - 1x2 vector containing the corresponding values
%          of y
%          N - number of lattice points to use
% OUTPUTS: x - vector of equally spaced x values
%          y - vector of corresponding y values
dx = (xspan(2)-xspan(1))/N; % spacing between lattice points
x = (xspan(1):dx:xspan(2))'; % vector of equally spaced values
A = zeros(N+1, N+1); % initialise A as a matrix of zeros
b = zeros(N+1, 1); % initialise b as a vector of zeros
A(1, 1) = 1; % row 1 of matrix A
b(1) = BCs(1); % row 1 of vector b
for I = 2:N % loop through rows 2 to N of A and b
    % Tridiagonal matrix construction
    A(I, I-1) = -2/dx^2 + coeffs(2)(x(I));
    A(I, I) = 1 + coeffs(1)(x(I));
    A(I, I+1) = -2/dx^2 + coeffs(2)(x(I+1));
    % Boundary conditions
    if I == 1
        b(I) = coeffs(3)(x(I));
    else
        b(I) = 0;
    end
end
% Solve the linear system
y = A \ b;
x = xspan;

```



**FIGURE 3.8.1:** Numerical solution of the boundary value problem (3.8.1) and (3.8.2) with  $p(x) = 0$ ,  $q(x) = 6$ ,  $g(x) = 2x$ ,  $a = 0$ ,  $b = 1$ ,  $\alpha = 1$ ,  $\beta = 5$ , using  $n = 50$  lattice points.

```
% get values of coefficients at x(I):
[p, q, g] = coeffs(x(I));
A(I, I-1) = 1/dx^2-p/(2*dx);
A(I, I) = -2/dx^2+q;
A(I, I+1) = 1/dx^2+p/(2*dx);
b(I) = g;
end
A(N+1, N+1) = 1; % row N+1 of A
b(N+1) = BCs(2); % row N+1 of b
y = A\b; % solution to matrix equation Ay=b
```

The coefficients  $p(x)$ ,  $q(x)$  and  $g(x)$  are calculated by a simple user-supplied function. For example, if  $p(x) = 0$ ,  $q(x) = 6$  and  $g(x) = 2x$ , we would write

```
function [p, q, g] = mycoeffs(x)
p = 0; q = 6; g = 2*x;
```

The boundary value problem with boundary conditions of, for example,  $y(0) = 1$  and  $y(1) = 5$  may be solved, using 50 lattice points, and plotted (see Figure 3.8) with the following commands:

```
>> a = 0; b = 1; alpha = 1; beta = 5;
>> [x, y] = bvp(@mycoeffs, [a, b], [alpha, beta], 50);
>> plot(x, y)
```

In the above example, both boundary conditions are in the form  $y(x_b) = y_b$ , i.e., they specify the value of the dependent variable  $y$  at the boundary values of  $x$ . This is known as a **Dirichlet boundary condition**. Another

common form of boundary condition is a **Neumann boundary condition**, which specifies the value of the **derivative**  $dy/dx$  (also written  $y'(x)$ ) at the boundary value of  $x$ . For example, the Neumann boundary conditions

$$y'(a) = \alpha, \quad y'(b) = \beta$$

are equally valid conditions for the differential equations such as (3.8.1). Furthermore, any boundary condition can contain a mixture of Dirichlet and Neumann terms, e.g., (3.6.10).

Boundary conditions containing derivatives are slightly more complicated than Dirichlet conditions to implement numerically, but the basic approach is the same: replace derivatives of  $y$  with finite difference approximations. The difficulty is, in order to use a central difference at the left-hand boundary point  $x = x_0$ , we need  $x_0 + h$  and  $x_0 - h$ . The first of these is fine as  $x_0 + h = x_1$ , but for  $x_0 - h$ , we must introduce an additional lattice point  $x_{-1}$  (with corresponding  $y$  value  $y_{-1}$ ) that is outside the domain. Then, we can use the central difference approximation:

$$y'(a) \approx \frac{y_1 - y_{-1}}{2h}.$$

In doing this, we have introduced an additional unknown  $y_{-1}$ , so we need an additional equation. This is achieved simply by taking the difference equation (3.8.3) with  $i = 0$ , as well as  $i = 1, \dots, n$ . If the right-hand boundary condition also contains a derivative term, the same procedure can be carried out at  $x = x_n$ . Here, an additional lattice point is introduced at  $x_{n+1}$ , and we get an additional equation (3.8.3) with  $i = n$ .

## Exercises

3.1 Find the general solution of

- (a)  $\dot{x} + 2x + 3y = 0,$   
 $\dot{y} + 3x + 2y = 2e^{2t},$
- (b)  $\dot{x} + \dot{y} - 5x + 3y = 15t^{1/2}e^t,$   
 $\dot{x} - 2\dot{y} + x = -30t^{1/2}e^t.$

3.2 Find the general solution of

$$\begin{aligned} x - \dot{x} + \dot{y} + y &= 1, \\ 2\dot{x} + x - 2\dot{y} - y &= 2 \end{aligned}$$

by eliminating (a)  $y$ , (b)  $\dot{y}$ .

3.3 Find the general solution of

- (a)  $\dot{x} = 9x - 8y,$   
 $\dot{y} = 24x - 19y;$
- (b)  $\dot{x} = 2y - x,$   
 $\dot{y} = -2x - y;$
- (c)  $\dot{x} = x + y,$   
 $\dot{y} = y;$
- (d)  $\dot{x} = x - 2y - z,$   
 $\dot{y} = -x + y + z,$   
 $\dot{z} = x - z;$
- (e)  $\dot{x} = x - y + z,$   
 $\dot{y} = x + y - z,$   
 $\dot{z} = 2z - y.$

3.4 Show that  $x = e^t$ ,  $y = 2e^t$  and  $x = te^t$ ,  $y = (2t - 1)e^t$  form a fundamental system for

$$\dot{x} = 3x - y + 1, \quad \dot{y} = 4x - y + t$$

and hence find the solution of the system such that  $x = 1$ ,  $y = 0$  at  $t = 0$ .

3.5 Find the solutions of the initial value problems

- (a)  $\dot{x} + 2x + 3y = 0$ ,  $\dot{y} + 3x + 2y = 2e^{2t}$  with  $x = 0$ ,  $y = 0$  at  $t = 0$ ;
- (b)  $\dot{x} + 3x + 2y = 3t - 1$ ,  $\dot{y} + 3x - 2y = 3t - 10$  with  $x = 1$ ,  $y = 0$  at  $t = 0$ .

3.6 Find a particular integral by matrix methods of

$$\begin{aligned}\dot{x} &= 3x + y + 3t^2 e^{3t}, \\ \dot{y} &= 3y + z + 9t, \\ \dot{z} &= 3z + 18.\end{aligned}$$

3.7 The function  $f(t)$  is defined by

$$f(t) = \begin{cases} t & (0 \leq t \leq \pi), \\ \pi e^{\pi-t} & (t \geq \pi). \end{cases}$$

Find the solution of the initial value problem

$$\ddot{y} + y = f(t)$$

which is continuous, with a continuous derivative, for all  $t \geq 0$  and such that  $y = 0$ ,  $\dot{y} = 1$  at  $t = 0$ .

- 3.8 If  $f(t)$  is defined as in Exercise 3.7, show that the boundary value problem with  $y(0) = 0$ ,  $y(2\pi) = a$  has no continuous solution with continuous derivative on  $0 \leq t \leq 2\pi$  if  $a \neq \frac{1}{2}\pi(e^{-\pi} - 1)$ , but has infinitely many such solutions if  $a = \frac{1}{2}\pi(e^{-\pi} - 1)$ .
- 3.9 Find the eigenvalues and eigenfunctions of  $\ddot{y} + \lambda y = 0$  subject to the boundary conditions
- $\dot{y}(-\pi) = 0$ ,  $\dot{y}(\pi) = 0$ ;
  - $y(0) = 0$ ,  $ay(b) + \dot{y}(b) = 0$  ( $a > 0, b > 0$ ).

- 3.10 Show that

$$\frac{d^4y}{dt^4} - \mu^4 y = 0$$

has nontrivial solutions satisfying the boundary conditions  $y(0) = 0$ ,  $\dot{y}(0) = 0$ ,  $y(1) = 0$ ,  $\dot{y}(1) = 0$  if, and only if,  $\cos \mu \cosh \mu = 1$  ( $\mu \neq 0$ ).

- 3.11 The steady-state temperature  $T$  of a one-dimensional object in a medium of ambient temperature  $T_a$  satisfies the steady-state equation

$$\frac{d^2T}{dx^2} + c(T_a - T) = 0,$$

where  $c > 0$  is a heat transfer coefficient. The endpoints of the object are held at a constant temperature of  $T = \alpha$  at  $x = 0$  and  $T = \beta$  at  $x = 1$ . Find the solution to this boundary value problem analytically. Confirm your answer by solving the problem numerically when  $c = 4$ ,  $T_a = 0$ ,  $\alpha = 20$  and  $\beta = 40$ , and plotting the analytical and numerical solution on the same graph.

If the left-hand endpoint is insulated so that there is no heat flow in or out of the object, the boundary condition becomes a Neumann condition  $T'(0) = 0$ . How does this change the solution to the problem?

- 3.12 Find series solutions in terms of eigenfunctions of

- $\ddot{y} = t(t - 2\pi)$  subject to  $y(0) = 0$ ,  $\dot{y}(\pi) = 0$ ;
- $\ddot{y} = \sin(\pi t/b)$  subject to  $y(0) = 0$ ,  $\dot{y}(b) = 0$ ;
- $t \frac{d}{dt} \left( t \frac{dy}{dt} \right) + 5y = 3 \sin(5 \ln t)$  subject to  $y(1) = 0$ ,  $y(e^\pi) = 0$ .

- 3.13 Show that, for

$$\frac{d}{dt} \left( p(t) \frac{dy}{dt} \right) + \{q(t) + \lambda r(t)\}y = 0$$

subject to the standard boundary conditions and  $r$  positive, the eigenvalues are real and the eigenfunctions are orthogonal with respect to the weight function  $r$ , i.e.,

$$\int_a^b r Y_m Y_n dt = 0$$

when  $Y_m$  and  $Y_n$  correspond to distinct eigenvalues.

3.14 The eigenfunctions  $Y_1(t), Y_2(t), \dots$  satisfy

$$t^2 \frac{d^2 Y_m}{dt^2} - t \frac{d Y_m}{dt} + \left( \frac{3}{4} + \lambda_m t^2 \right) Y_m = 0$$

on  $\pi < t < 2\pi$  under the boundary conditions  $Y_m(\pi) = 0$ ,  $Y_m(2\pi) = 0$ . Show that the eigenfunctions are orthogonal with respect to a suitable weight function.

By means of the substitution  $y(t) = t^{1/2} u(t)$ , obtain as a series of eigenfunctions the solution of

$$t^2 \ddot{y} - t \dot{y} + \left( \frac{3}{4} - t^2 \right) y = t^{5/2}$$

such that  $y(\pi) = 0$ ,  $y(2\pi) = 0$ .

# *Chapter 4*

---

## *Modelling Biological Phenomena*

---

### **4.1 Introduction**

Mathematical modelling of physical phenomena, such as the dynamics of a rigid body, the deformation of an elastic material or the propagation of electromagnetic waves in the atmosphere, is based, to the best of our present scientific knowledge, on sound physical laws. Thus to describe the dynamical behaviour of a rigid body undergoing the influence of external forces we have the fundamental Newtonian laws of motion at our disposal. The behaviour of deforming elastic materials is governed by the constitutive equations of continuum mechanics; Maxwell's equations are the fundamental postulates that govern electromagnetic waves.

These basic laws have been the result of centuries of experiment, observation and inspiration of mathematicians and scientists including Sir Isaac Newton, Leonard Euler and James Clerk Maxwell.

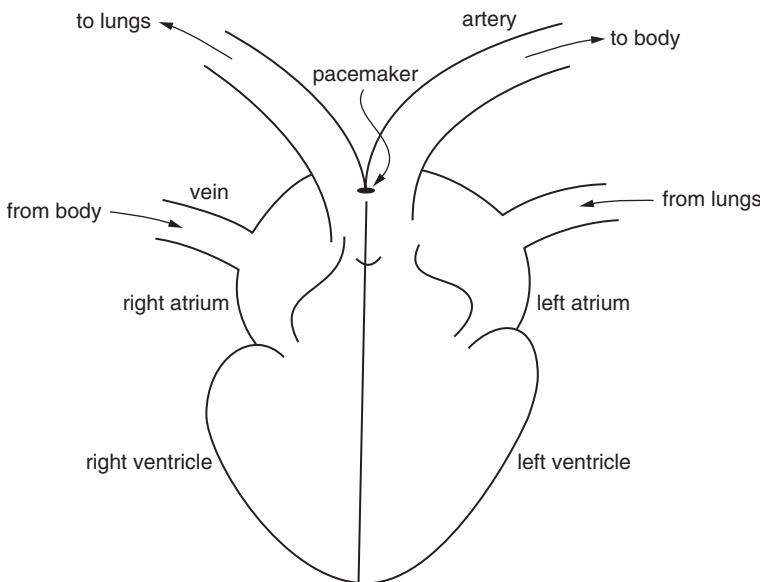
In biology and the life sciences in general, this interplay between the observed phenomenon and its mathematical description is still in the early stages of development and, apart from the **Hardy–Weinberg law** associated with the Mendelian theory of genetics, there are few sound postulates to guide us. Instead the philosophy is to develop mathematical models that in the first instance describe in only a qualitative way the observed biological process. As in all scientific endeavours, the real test of the model is that it not only agrees qualitatively with the biological process but has the ability to suggest new experiments and bring deeper insight to the biological situation. If enough experience is gained by this philosophy then hopefully, together with a better understanding of the life sciences, sound postulates will emerge upon which a mathematical theory can be developed.

This philosophy of qualitative description will be exploited in the topics treated in this chapter and throughout a major portion of this book.

---

### **4.2 Heartbeat**

The heart is a complex but robust pump (see the simplified illustration in Figure 4.2.1). It consists of four chambers and four valves. There are essentially



**FIGURE 4.2.1:** Schematic description of the heart as viewed from the front of the body.

two circuits for the blood, one which spreads through the lungs to pick up oxygen and the other which spreads through the body to deliver the oxygenated blood. The first circuit is a low-pressure circuit so as not to damage the delicate membrane in the lungs, whereas the second is a high-pressure circuit in order for the blood to get down to the feet and up again. From Figure 4.2.1, it is apparent that the right side of the heart is the low-pressure pump to the lungs while the left side is the high-pressure pump to the body.

Each pump has a main pumping chamber called the ventricle with an inlet and an outlet valve. The purpose of the inlet valve is to prevent flow back up the veins while pumping, and the outlet valve is to prevent flow back from the arteries while filling. Since the heart is made of nonrigid tissue it only has the power to push out and no power to suck in. Thus to get a good pump of blood it is necessary to fill the ventricle completely, and to aid this there is a small chamber called the atrium whose job is to pump gently beforehand, just enough to fill the ventricle but not enough to cause any flow back.

During the heartbeat cycle there are two extreme equilibrium states, namely **diastole** which is the relaxed state and **systole** which is the contracted state. What makes the heart beat is the presence of a pacemaker which is located on the top of the atrium. The pacemaker causes the heart to contract into systole. That is, it triggers off an electrochemical wave which spreads slowly over the atria causing the muscle fibres to contract and push blood into the ventricles and then spreads rapidly over the ventricles causing the whole ventricle to

contract into systole and deliver a big pump of blood down the arteries. The muscle fibres then rapidly relax and return the heart to diastole; the process is then repeated.

In order to develop a mathematical model that reflects the behaviour of the heartbeat action described above, we choose to single out the following features. First of all, the model should exhibit an equilibrium state corresponding to diastole. Secondly, there must be a threshold for triggering the electrochemical wave emanating from the pacemaker causing the heart to contract into systole. Thirdly, the model must reflect the rapid return to the equilibrium state.

We begin by doing a little mathematical experimentation. Suppose we let  $x$  denote muscle fibre length referred to some convenient origin, say  $x = 0$ , which corresponds to the equilibrium state. Let  $b$  be an electrical control variable which governs the electrochemical wave. As far as the muscle fibres are concerned, we look for a differential equation which has  $x = 0$  as an equilibrium state and at least for small times has a rapidly decreasing solution. An appropriate equation exhibiting these features is

$$\epsilon \frac{dx}{dt} = -x, \quad (4.2.1)$$

where  $\epsilon$  is a small positive parameter. When  $\frac{dx}{dt}$ , the velocity of the fibre, is zero we have the equilibrium state  $x = 0$ . Furthermore we know (Chapter 1) that (4.2.1) has the general solution

$$x = A \exp(-t/\epsilon) \quad (4.2.2)$$

which is rapidly decreasing in time. Thus (4.2.1) seems to be a good candidate to represent the behaviour initially of the muscle fibres causing contraction into systole.

Turning now to the electrochemical wave, we need the control  $b$  to represent initially the relatively slow spread of this wave over the atria. A simple model which does this is

$$\frac{db}{dt} = -b. \quad (4.2.3)$$

Here  $b = 0$  is an equilibrium state and (4.2.3) has the solution

$$b = B \exp(-t), \quad (4.2.4)$$

which, in comparison with (4.2.2), represents a relatively slow decay time.

The features that are not covered by this simple model obtained from (4.2.1) and (4.2.3) are (i) the threshold or trigger and (ii) the rapid return to equilibrium. At this stage, our knowledge of differential equations is insufficient to include these features and the discussion must be deferred until Chapter 6. The model that incorporates the desired features is the coupled nonlinear

first-order system

$$\begin{aligned}\epsilon \frac{dx}{dt} &= -(x^3 + ax + b), \\ \frac{db}{dt} &= x - x_a.\end{aligned}\tag{4.2.5}$$

Here  $x$  represents the length of the muscle fibre,  $-a$  represents tension,  $b$  represents the chemical control and  $x_a$  represents a typical fibre length when the heart is in diastole. The model (4.2.5) is due to E.C. Zeeman.

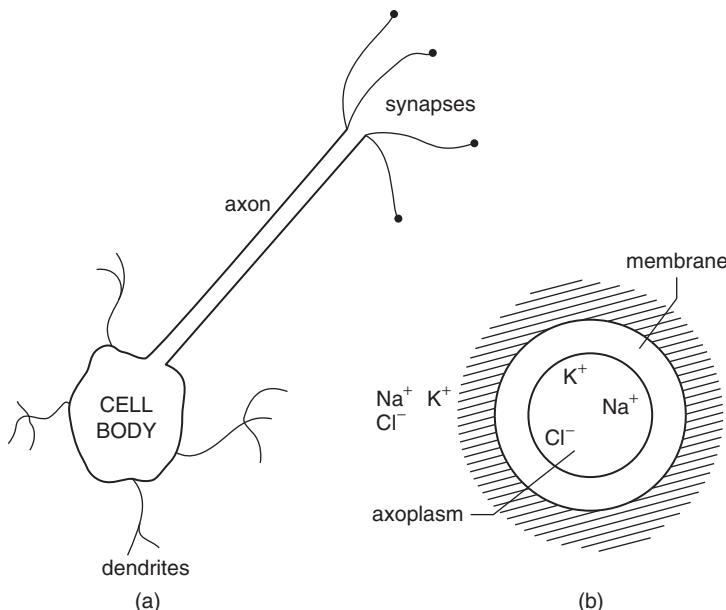
That Zeeman chose to single out the above three qualities of the heartbeat cycle and to attempt to model them through the system of equations (4.2.5) should not lead the reader to assume that such a description is the only one. Indeed, the model can only be considered a reasonable one if it reflects the basic features of the heartbeat cycle well.

It is appropriate to remark, however, that the model (4.2.5) has been quite successful in distinguishing between some extreme forms of heartbeat behaviour, for example, the effects of high blood pressure or an excess of adrenalin in the bloodstream due to rage or vigorous exercise. Likewise there is the situation when the heart beats in a feeble manner and does not contract into systole.

### 4.3 Nerve impulse transmission

The axon portion of a nerve cell (see Figure 4.3.1) is made up of a conducting material called axoplasm which is contained in a roughly cylindrical membrane between 50 and 70 Å thick. The membrane is permeable to potassium ions,  $K^+$ , concentrated in the interior, and to sodium ions,  $Na^+$ , concentrated in the exterior. Also present, but to a much lesser extent, are other ions such as chlorine  $Cl^-$ . In nature, the high concentration of sodium ions is maintained by the organism in the fluid medium exterior to the nerve cell. In its resting state there is a potential difference across the axon membrane of between  $-50$  and  $-70$  millivolts (mV).

Suppose we place a segment of axon in a bath containing a sodium concentration similar to the one usually present in the exterior fluid medium and apply a potential difference across the membrane. In the laboratory, this is usually done by inserting a fine micropipette into the axon and injecting sodium ions. The induced sodium gives rise to an applied current. It is observed that if a small potential difference, which is positive relative to the resting potential, is applied across the membrane, the sodium and potassium ionic currents are briefly disturbed but quickly return to their zero resting state and the membrane settles back to the resting potential. If a much larger



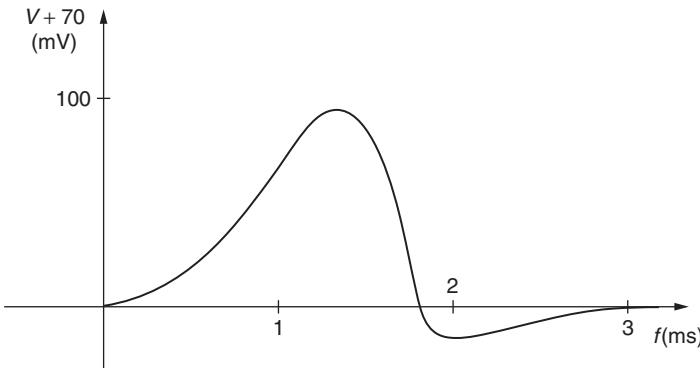
**FIGURE 4.3.1:** Schematic description of the nerve axon.

positive membrane potential is applied (between 7 and 10 mV) the equilibrium state is exceeded and the sodium currents become active. What happens now is that the axon membrane becomes permeable to positive sodium ions which flow inwards, making the membrane potential increase and this causes the membrane to become even more permeable to sodium ions. If circumstances are just right in that the inflow of sodium ions keeps the membrane potential increasing, there is a critical level in membrane permeability which we call the **threshold**, which results in a rapid impulsive rise in potential difference across the membrane to about 100 mV relative to the negative resting potential.

Following this “firing” of the axon two things happen. First the sodium ion permeability begins to decrease slowly and secondly the potassium ion permeability rapidly rises. Potassium ions thus flow outwards and eventually restore the membrane potential once again to its resting state after an overshoot of about 5 mV. The impulse lasts about 3 to 5 ms (see Figure 4.3.2).

The question to ask now is how an impulse is transmitted along the axon during this process. Near the point of stimulation, an impulse is created, which is shot off down the axon along the membrane, being renewed at each point, to approximately 100 mV as the membrane potential at each point achieves a value that initiates the active phase of sodium. As we mentioned above, a threshold is involved, that is, either this voltage is not achieved and no impulse is propagated or it is achieved and at least one impulse is propagated.

Except at a discrete set of points called **nodes of Ranvier**, which are about 1 mm apart, every vertebrate nerve axon is covered with a sheath

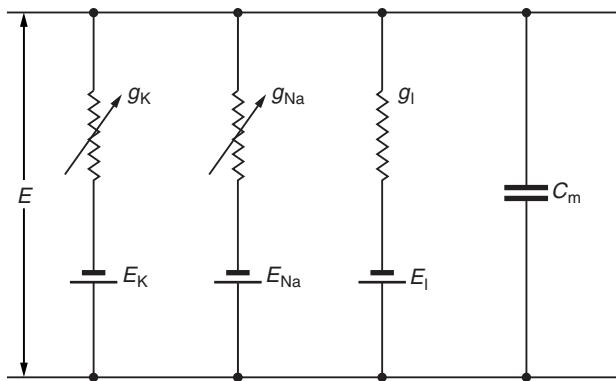


**FIGURE 4.3.2:** Membrane potential.

that electrically insulates the axoplasm and surrounding membrane from the exterior medium. A vertebrate nerve fibre is said to be **myelinated** and, although current in a vertebrate nerve fibre can easily pass freely along the axoplasm or the exterior fluid as happens in an **invertebrate unmyelinated nerve**, it can pass through the membrane only at certain points. Currents circulate on paths around the boundary of a section of arbitrary location for the membrane of an unmyelinated axon, but only on paths that pass through the nodes of Ranvier for a myelinated axon.

Our understanding of the mechanism governing the action potential is due chiefly to the inspired and carefully executed experiments of the physiologists A.L. Hodgkin and A.F. Huxley in 1952. This work, which led to the award of a Nobel prize, was largely performed on the large axon to be found in the squid *Loligo*, and culminated in the development of a mathematical model. This model not only agrees in a qualitative way with the experimental results but gives remarkably accurate quantitative results. Since the development of this model, others have been subsequently formulated which reflect current experimental findings with good agreement. However the so-called **Hodgkin–Huxley** model is still regarded as the fundamental model governing nerve impulse transmissions. The experiments performed by Hodgkin and Huxley were set up in the same way as described above, except that the induced current was achieved not by injection of sodium ions but by inserting a fine current-carrying wire into the axon. This induced current, which we denote by  $I$ , is found to give rise to a membrane potential  $E$  which is the same at each point of the segment of axon and depends only on the time  $t$ . That is,  $E$  is independent of  $x$ , the position of a point along the axon relative to some convenient origin. This configuration is known as **space clamp**.

The basic assumption of the Hodgkin–Huxley theory is that there are separate channels for the sodium, potassium and other ions like chlorine. We can envisage these channels in terms of the electrical circuit shown in Figure 4.3.3. Thus, each channel is described in terms of a voltaic cell  $E$  in series with a



**FIGURE 4.3.3:** Conductance model of the nerve membrane.

conductance  $g$  together with a capacitance  $C_m$  across the whole ensemble.

The transmembrane current is then given by

$$I = C_m \frac{dE}{dt} + g_{Na}(E - E_{Na}) + g_K(E - E_K) + g_l(E - E_l), \quad (4.3.1)$$

where  $I$  is the current density,  $E$  is the membrane potential,  $C_m$  is the membrane capacity,  $g_{Na}$  is the sodium conductance,  $g_K$  is the potassium conductance,  $g_l$  is the leakage conductance,  $E_{Na}$  is the sodium equilibrium potential,  $E_K$  is the potassium equilibrium potential and  $E_l$  is the leakage equilibrium potential. The conductances  $g_{Na}$  and  $g_K$  are assumed to vary with time and the potential  $E$ , while  $g_l$  is assumed constant. To describe the variation in  $g_{Na}$  and  $g_K$  Hodgkin and Huxley assumed that  $g_K$  is described by

$$g_K = \bar{g}_K n^4, \quad (4.3.2)$$

and that  $g_{Na}$  is determined by

$$g_{Na} = \bar{g}_{Na} m^3 h. \quad (4.3.3)$$

In these expressions,  $\bar{g}_K$  is the maximum potassium conductance and  $\bar{g}_{Na}$  is the maximum sodium conductance. The quantities  $n$ ,  $m$  and  $h$  are dimensionless quantities that vary between 0 and 1 and are functions of  $E$  and  $t$ . Their precise forms are determined as solutions of the system of ordinary differential equations:

$$\frac{dm}{dt} = \alpha_m(E)(1 - m) + \beta_m(E)m, \quad (4.3.4)$$

$$\frac{dh}{dt} = \alpha_h(E)(1 - h) + \beta_h(E)h, \quad (4.3.5)$$

$$\frac{dn}{dt} = \alpha_n(E)(1 - h) + \beta_n(E)n. \quad (4.3.6)$$

In these equations the coefficient functions  $\alpha_j, \beta_j, j = m, h, n$ , are functions of the membrane potential  $E$ , and were found by careful empirical fitting with the experimental results. The exact forms of these coefficient functions are complicated formulae involving  $\exp E$ . The mathematical model developed from the Hodgkin–Huxley theory is the formidable system of linked differential equations given by (4.3.1)–(4.3.6). That Hodgkin and Huxley were able to solve these equations at all is quite remarkable, especially if one appreciates the limited computational facilities available in 1952. It should be remarked, however, that although this system admits to numerical solution, the underlying analytical structure is by no means fully understood.

If the space clamp is removed in the sense that the membrane potential is allowed to vary with position  $x$  along the axon, then Hodgkin and Huxley assumed Kelvin's cable theory to assert that the current  $I$  is given by

$$I = \frac{a}{2R} \frac{\partial^2 E}{\partial x^2} \quad (4.3.7)$$

where  $a$  is the radius of the axon and  $R$  is the specific resistivity of the axoplasm. Since  $E$  now depends on both  $x$  and  $t$  all derivatives must be replaced by partial derivatives, and so by incorporating (4.3.7) in (4.3.1) we are led to consider the partial differential equation

$$\frac{a}{2R} \frac{\partial^2 E}{\partial x^2} = C_m \frac{\partial E}{\partial t} + g_{Na}(E - E_{Na}) + g_K(E - E_K) + g_l(E - E_l), \quad (4.3.8)$$

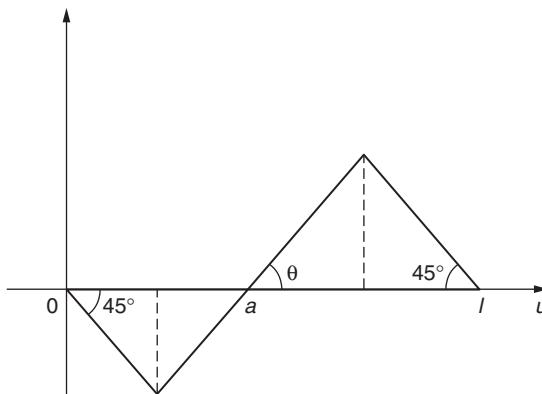
together with the system (4.3.2)–(4.3.6).

As we have mentioned before, several alternative mathematical models of nerve impulse transmission have been developed since 1952. Some of these are of the same complexity as the Hodgkin–Huxley model while others, by making certain additional assumptions about the behaviour of the various ionic conductances, are considerably simpler. These simplified models nevertheless retain the main features characteristic of the Hodgkin–Huxley theory. One such model that has attracted much interest is the **FitzHugh–Nagumo** model, originally proposed by R. FitzHugh in 1961 and subsequently developed by J. Nagumo and his co-workers in 1962. This model is developed in analogy with the **van der Pol oscillator**, well known to electrical engineers and physicists, and takes the form

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial t} - u(1-u)(u-a) + w, \quad (4.3.9)$$

$$\frac{\partial w}{\partial t} = bu - \gamma w, \quad (4.3.10)$$

where  $a, b$  and  $\gamma$  are positive constants and  $0 < a < 1$ . In this simplified model,  $u$  represents the membrane potential  $E$  and, as before,  $x$  measures distance along the axon and  $t$  is time. The cubic term  $u(1-u)(u-a)$  in (4.3.9) is analogous to an instantaneous turning on of sodium permeability



**FIGURE 4.3.4:** The piecewise linear term.

and can be thought of as playing the role of the variable  $m$  in the Hodgkin–Huxley equations.  $w$  is a recovery variable and is analogous to the turning on of potassium permeability and so behaves like the variable  $n$  in the Hodgkin–Huxley model. There is no counterpart to inactivation of sodium permeability.

There is a further simplification which can be adopted and which comes about from the following observation. The simplified model cannot be expected to give quantitative comparisons with experiment, as the Hodgkin–Huxley model does. Thus we can only expect a qualitative comparison and, consequently, if we could simplify the nonlinearity in (4.3.9), (4.3.10) without destroying the desired behaviour of solutions, then the methods of solution could be greatly simplified. To this end H.P. McKean in 1970 proposed that the term  $u(1 - u)(u - a)$  in (4.3.9) could be replaced, for example, by the piecewise linear term illustrated in Figure 4.3.4, where the angle  $\theta$  can take any value in the semi-open interval  $0 < \theta \leq \pi/2$ . The benefit of incorporating this simplification is that, along each line segment, the system of equations (4.3.9) and (4.3.10) is linear and in general linear equations are much easier to solve than nonlinear ones.

The mathematical models of nerve impulse transmission are incomplete without some notion of the solutions to expect or the appropriate initial and boundary conditions. If we consider the Hodgkin–Huxley model as applied to the giant axon of the squid, we know that the length of the axon is large compared with its radius  $a$ . Thus if we put

$$X = x/\sqrt{a} \quad (4.3.11)$$

in (4.3.7) and (4.3.8) we can approximate the range of  $X$  to be infinitely large. In this way we take the range of  $x$  in the system (4.3.9), (4.3.10) also to be  $0 \leq x < \infty$ .

Both of the models we have discussed have solutions that depend only on  $x + ct$  and are called **travelling waves**. That is, there are membrane

potentials  $E(x, t) = E(x+ct)$  that move along the axon with no loss of strength and in a direction determined by the sign of  $c$ . The rate of propagation is  $|c|$  and the problem of finding this value is very important.

Suppose the axon is stimulated from one end defined, for example, by  $x = 0$ . Then the appropriate conditions are:

$$\begin{aligned} E(0, t) &= P(t), \quad t > 0, \\ E(x, 0) &= 0, \quad x \geq 0. \end{aligned}$$

Here  $P(t)$  is a function of  $t$  which defines the stimulus emanating at  $x = 0$ . An important and largely unresolved problem is to classify those stimuli  $P(t)$  so that  $E(x, t)$  approaches a travelling wave  $E(x + ct)$  as  $t \rightarrow \infty$ . The study of travelling waves will be taken up in Chapter 7, while some idea of the behaviour of stimulated membrane potentials will be considered in Chapter 12.

---

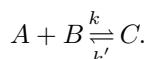
## 4.4 Chemical reactions

### Chemical notation and the law of mass action

As is standard in Chemistry, we write



to indicate that a molecule of substance  $A$  and a molecule of substance  $B$  combine to form a molecule of substance  $C$ . If the reaction is reversible (i.e., a molecule of  $C$  can also decompose into a molecules each of  $A$  and  $B$ ), we write



The quantities  $k$  and  $k'$  are rate constants.

The law of mass action states that the rate of the reaction is proportional to the product of the concentrations of the reactants. Thus, for reaction (4.4.1) (and denoting the concentration of substance  $A$  by  $[A]$ , etc.), the law of mass action gives

$$\frac{d[C]}{dt} = k[A][B] = -\frac{d[A]}{dt} = -\frac{d[B]}{dt}.$$

### The Belousov–Zhabotinskii reaction

The chemical reaction described here is not part of any living system and is certainly not biological. However, its interest and importance lies in its ability to “oscillate”, forming patterns such as parallel bands, concentric rings and

cell-like structures. Such phenomena are well known in biology: for example, in morphogenesis (discussed later in Chapter 12), the regulating processes involved in living cells and organisms, and in information transmission. A further reason for studying pattern formation in nonliving systems is that the constituent forces are more limited than those of even the simplest biological systems where electrical forces, surface tensions, colloid properties and crystallising forces, and complex chemical reactions may play a significant role.

The reaction we shall consider is the usually dramatic oscillatory reaction discovered by Belousov in 1958 and Zhabotinskii in 1964. The **Belousov–Zhabotinskii** reaction, as it is now called, is essentially the oxidation of malonic acid by bromate in a sulphuric acid medium, in the presence of a cerium catalyst. In the reaction, two overall processes I and II can be identified, and the chemistry involved can be described as follows. In the reaction when the bromide ion ( $\text{Br}^-$ ) is above some critical concentration, process I occurs. Here the bromate ion ( $\text{BrO}_3^-$ ) is reduced to bromine ( $\text{Br}_2$ ), with bromous acid ( $\text{HBrO}_2$ ) as an intermediary, and the malonic acid,  $\text{CH}_2(\text{COOH})_2$ , is brominated. During this process there is little oxidation of the cerium ion  $\text{Ce(III)}$ . Process I thus uses up the bromide. When the concentration of bromide becomes sufficiently low, process II takes over. In this the bromous acid and the bromate ion produce a radical bromate species ( $\text{BrO}_2^{\cdot}$ ), which oxidises the cerium ion  $\text{Ce(III)}$  to the  $\text{Ce(IV)}$  form with bromous acid generated autocatalytically. When all of the  $\text{Ce(III)}$  has been oxidised to  $\text{Ce(IV)}$  and the bromide ion concentration is low, the  $\text{Ce(IV)}$  then reacts with the bromomalonic acid to produce the cerium ion  $\text{Ce(III)}$  and bromide again. When the bromide passes a critical concentration, process I takes over again and the cycle is repeated. During the process, one sees the reagent oscillate in colour, turning alternatively bright blue and reddish purple if ferroin is used as an indicator.

On the bases of the mechanism described above, Field and Noyes have suggested the following mathematical model which involves three intermediaries which oscillate and which are associated with the bromide ion ( $\text{Br}^-$ ), bromous acid ( $\text{HBrO}_2$ ) and the cerium ion  $\text{Ce(IV)}$ .

If  $X = \text{HBrO}_2$ ,  $Y = \text{Br}^-$  and  $Z = \text{Ce(IV)}$ , the Field–Noyes system is



In this system,  $A$  is the reactant  $\text{BrO}_3^-$ , which is maintained at a constant concentration,  $P$  and  $Q$  are products and  $f$  is a constant stoichiometric factor. To arrive at the mathematical model, we invoke the law of mass action. For

the reaction (4.4.2), the law of mass action says that  $X$  is produced at the rate  $k_1[A][Y]$ . However, at the same time,  $X$  is being used up in reaction (4.4.3) to produce  $P$  at rate  $k_2[X][Y]$ , and by reaction (4.4.5) to produce  $Q$  at rate  $k_4[X]^2$ . Finally, reaction (4.4.4) produces  $X$  at the rate  $k_3[A][X]$ . Thus the total rate of change of  $X$  is

$$\frac{d[X]}{dt} = k_1[A][Y] - k_2[X][Y] + k_3[A][X] - k_4[X]^2. \quad (4.4.7)$$

If we analyse the rates of change of  $Y$  and  $Z$  in the same way, the law of mass action gives

$$\frac{d[Y]}{dt} = -k_1[A][Y] - k_2[X][Y] + fk_5[Z], \quad (4.4.8)$$

and

$$\frac{d[Z]}{dt} = k_3[A][X] - k_5[Z]. \quad (4.4.9)$$

The system of coupled first-order differential equations (4.4.7)–(4.4.9) is the mathematical model of Field, Körös and Noyes representing the Belousov–Zhabotinskii (B-Z) reaction.

Before we proceed further with the mathematical models of the B-Z reaction, let us describe one or two of the experiments performed on this reaction and the patterns that arise.

Under certain conditions, the reagent is capable of organising itself into spatially inhomogeneous structures that are seen as coloured patterns. The physical reasons for the pattern formations as well as the patterns themselves are different according to the experimental setup. If the reagent is placed in a thin layer and allowed to convect, say, by heating the fluid from below or cooking it from above by evaporation, then a reaction–diffusion process resembling the Bénard phenomenon of cellular convection occurs. Here the fluid, when viewed from above, organises itself into hexagonal or rectangular cells, which are outlined in reddish purple. It seems that the boundaries of the convection cells appear to contain most of their cerium in the reduced state, while the rest of each cell contains more cerium in an oxidised state. However, the precise behaviour of this hydrodynamic and chemical phenomenon is not fully understood.

An interesting pattern formation is observed when the reagent is placed in a vertical container and a gradient of temperature or of one of the concentrations is imposed on the fluid. In the latter case, sulphuric acid is carefully added after the other ingredients have been mixed. Under these circumstances, horizontal bands form and propagate vertically through the container. In this case there is no fluid motion at all; it is only the lines of constant phase of the oscillation that are moving through the fluid. The explanation of this phenomenon is as follows: the concentration gradient or temperature gradient produces a vertical gradient in the frequency of the oscillation. This frequency gradient can account in detail for the space-time behaviour of the patterns that emerge.

Diffusion plays a negligible role unless the pattern has a very small spatial scale.

Perhaps the most striking pattern formations that can be observed are those found in the experiments of Zaikin and Zhabotinskii using ferroin for the catalyst and malonic acid in the Belousov reaction. Here the reagent is spread thinly, about 2 mm thick on a petri dish. Circular chemical waves are observed propagating outwards. These waves, essentially oxidation bands, are blue and they propagate through the reddish background fluid. When two waves collide both disappear. When a faster one catches up with a slower one the latter is entrained. These waves have been designated “trigger” waves since diffusion combines with the chemical reaction to trigger the waves. To include diffusion effects in our model (4.4.7)–(4.4.9) we suppose the intermediaries  $X, Y$  and  $Z$  can diffuse with diffusion coefficients  $D_X, D_Y, D_Z$  and hence are functions of the space variables  $x, y, z$  and time  $t$ . The system (4.4.7)–(4.4.9) is modified to read

$$\frac{\partial[X]}{\partial t} = k_1[A][Y] - k_2[X][Y] + k_3[A][X] - k_4[X]^2 + D_X \nabla^2[X], \quad (4.4.10)$$

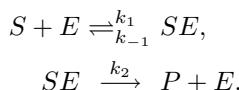
$$\frac{\partial[Y]}{\partial t} = -k_1[A][Y] - k_2[X][Y] + f k_5[Z] + D_Y \nabla^2[Y], \quad (4.4.11)$$

$$\frac{\partial[Z]}{\partial t} = k_3[A][X] - k_5[Z] + D_Z \nabla^2[Z]. \quad (4.4.12)$$

This model is a **reaction–diffusion** system of coupled partial differential equations and, under certain circumstances, admits travelling wave solutions. For instance, if we take  $[Z] = 0$  and assume  $[X] = \xi(x+ct)$  and  $[Y] = \eta(x+ct)$ , then the reduced system (4.4.10), (4.4.11) can be studied in the same way as the travelling wave solutions of nerve axon equations are considered. Of course boundary and initial conditions are somewhat different. These considerations will be taken up in Chapters 8 and 12.

## Enzymatic catalysis

A reaction that is extremely common in biological systems is **enzymatic catalysis**. This covers the broad class of reactions where a substrate ( $S$ ) reversibly reacts with an enzyme (or catalyst) ( $E$ ) to form a compound substance (or complex) ( $SE$ ). This compound then decomposes into a product ( $P$ ), and the original enzyme ( $E$ ):



Typically, the enzyme is present at very low concentration, which limits the rate at which the substrate can be converted to the product. Hence, this type of reaction is often referred to as a rate-limited reaction.

Applying the law of mass action to this reaction gives the following differential equations for the four reactants:

$$\begin{aligned}\frac{d[S]}{dt} &= -k_1[S][E] + k_{-1}[SE], \\ \frac{d[E]}{dt} &= -k_1[S][E] + k_{-1}[SE] + k_2[SE], \\ \frac{d[SE]}{dt} &= k_1[S][E] - k_{-1}[SE] - k_2[SE], \\ \frac{d[P]}{dt} &= k_2[SE].\end{aligned}$$

The appropriate initial conditions are

$$[S](0) = S_0, \quad [E](0) = E_0, \quad [SE](0) = 0, \quad [P](0) = 0,$$

where  $S_0$  and  $E_0$  represent the initial concentrations of substrate and enzyme respectively. This system of four differential equations can be simplified by noting that the first three equations do not depend on  $[P]$  and that  $d/dt([E](t) + [SE](t)) = 0$ , so  $[E](t) + [SE](t) = E_0$ . We hence consider the reduced system

$$\frac{d[S]}{dt} = -k_1 E_0 [S] + (k_1 [S] + k_{-1}) [SE], \quad (4.4.13)$$

$$\frac{d[SE]}{dt} = k_1 E_0 [S] - (k_1 [S] + k_{-1} + k_2) [SE]. \quad (4.4.14)$$

It is useful to nondimensionalise all variables by making the following transformations:

$$\begin{aligned}T &= k_1 E_0 t, & u(T) &= [S](t)/S_0, & v(T) &= [SE](t)/E_0 \\ \lambda &= \frac{k_2}{k_1 S_0}, & K &= \frac{k_{-1} + k_2}{k_1 S_0}, & \epsilon &= \frac{E_0}{S_0}.\end{aligned}$$

Then, the system (4.4.13), (4.4.14) reads

$$\frac{du}{dT} = -u + (u + K - \lambda)v, \quad (4.4.15)$$

$$\epsilon \frac{dv}{dT} = u - (u + K)v, \quad (4.4.16)$$

with initial conditions  $u(0) = 1$  and  $v(0) = 0$ . Methods for dealing with nonlinear systems of this type are dealt with in Chapter 5 and, in general, finding an explicit solution is not trivial. However, the fact that the initial concentration of enzyme  $E_0$  is typically much smaller than that of the substrate  $S_0$  means that  $\epsilon$  is a small parameter ( $\epsilon \ll 1$ ). This can be exploited by seeking a solution as a power series in  $\epsilon$ :

$$u(T) = u_0(T) + \epsilon u_1(T) + \epsilon^2 u_2(T) + \dots,$$

$$v(T) = v_0(T) + \epsilon v_1(T) + \epsilon^2 v_2(T) + \dots.$$

We now substitute these forms into (4.4.15) and (4.4.16) and equate coefficients of powers of  $\epsilon$ . At  $O(\epsilon^0)$ , we have

$$\frac{du_0}{dT} = -u_0 + (u_0 + K - \lambda)v_0, \quad (4.4.17)$$

$$0 = u_0 - (u_0 + K)v_0, \quad (4.4.18)$$

with initial conditions  $u_0(0) = 1$  and  $v_0(0) = 0$ . Equation (4.4.18) implies that

$$v_0(T) = \frac{u_0(T)}{u_0(T) + K}. \quad (4.4.19)$$

Substituting this into (4.4.17), solving by separation of variables (see Section 1.4), and using the initial condition on  $u_0$  gives

$$u_0(T) + K \ln u_0(T) = 1 - \lambda T. \quad (4.4.20)$$

The problem now is that the solution for  $v_0(T)$  does not satisfy the initial condition  $v_0(0) = 0$ . This problem arises because we have essentially assumed that  $\epsilon dv/dT \ll 1$ . However, at  $T = 0$ , the initial conditions state that  $u = 1$  and  $v = 0$ , and hence from equation (4.4.16)  $dv/dT$  is of the order  $\epsilon^{-1}$ , so our assumption is incorrect there. This problem may be elegantly solved by introducing a *fast time variable*,  $S = T/\epsilon$ , transforming equations (4.4.15) and (4.4.16) to this variable and again seeking solutions as a power series in  $\epsilon$ . This gives a solution (called the inner solution) that is valid for very small times (i.e., values of  $S$  of the order 1). By matching with the solution found above (called the outer solution), which is valid for  $T \gg \epsilon$ , it is possible to obtain a matched solution that is valid for all times.

The fact that  $dv/dT$  is very large near  $T = 0$  indicates that the  $v$  reaction is so fast that it is close to equilibrium at all times. Hence, in practice,  $v$  may be assumed to be in the **pseudo-steady state** defined by (4.4.19). This was the hypothesis made by Michaelis and Menten in their 1913 work on chemical kinetics. Typically, it is the reaction rate  $du/dT$  that is of primary interest, rather than the concentration  $u(T)$ . Implicit differentiation of (4.4.20) gives

$$\frac{du_0}{dT} = -\frac{\lambda u_0}{K + u_0}.$$

Neglecting terms of order  $\epsilon$  and higher, and returning to the original dimensional variables, we have

$$\frac{d[S]}{dt} = -\frac{k_2 E_0 [S]}{K_m + S}, \quad (4.4.21)$$

where  $K_m = (k_{-1} + k_2)/k_1$  is called the Michaelis constant. We hence see that the rate of uptake of the substrate  $S$  is proportional to the enzyme concentration  $E_0$ , but is nonlinearly related to the concentration of substrate  $[S]$ . When  $[S] \ll K_m$ , the increase in reaction rate is approximately linear in  $[S]$ , but as  $[S]$  approaches  $K_m$ , the reaction rate levels off (saturates), and has a maximum value of  $k_2 E_0$ . The reaction rate defined as (4.4.21) is sometimes referred to as **Michaelis–Menten kinetics**.

## 4.5 Predator-prey models

The problems we shall consider here are of fundamental importance in ecology, i.e., the study of the interactions between living organisms and their environment. Let us consider two organisms or species characterised, for example, by their respective population densities, say  $X$  and  $Y$ . Thus  $X$  may be the population density of a carnivore occupying a certain habitat and  $Y$  may be the population density of a herbivore occupying the same habitat as  $X$  and considered as a food source for  $X$ . Alternatively,  $X$  and  $Y$  may be used to represent parasite and host, or herbivore and plant.

It is convenient to classify the direct interaction between a pair of species into the following categories:

- (a) **Competition:** each species has an inhibiting effect on the growth of the other.
- (b) **Commensalism:** each species has an accelerating effect on the growth of the other.
- (c) **Predation:** one species, the “predator”, has an inhibiting effect on the growth of the other; the “prey” has an accelerating effect on the growth of the predator.

Throughout this section, we consider the interaction between  $X$  and  $Y$  to be that of predation. Furthermore, we shall make the following simplifying assumptions.

- (a) The density of a species – that is, the number of individuals per unit area – can be represented as a function of a single variable, time. Thus we ignore possible age differences and differences of sex or genotype.
- (b) Changes in density are deterministic; that is, we assume there are no random effects in the environment influencing the interaction between  $X$  and  $Y$ . Obviously this is a severe limiting assumption in many realisable situations.
- (c) The effects of interactions within and between species are instantaneous. In the predator-prey interaction, this means that the delay between the moment a predator eats a prey, and the moment when the ingested material is converted into part of a new predator is ignored.

We begin by considering some simple models of predator-prey interaction. Let  $X$  be the predator density and  $Y$  the prey density. In the absence of predators, we expect no inhibition in the growth of the prey. A simple growth relation  $Y$  could be

$$\frac{dY}{dt} = kY, \quad (4.5.1)$$

where  $t$  is time and  $k$  is a positive rate constant. We know that the ordinary differential equation (4.5.1) has the solution

$$Y = Y_0 e^{kt}, \quad (4.5.2)$$

where  $Y_0$  is the initial population density. Thus if (4.5.1) is used to describe the population growth of  $Y$ , then  $Y$  will increase exponentially in time. Such a growth behaviour is reasonable for a limited time, but ultimately an increasing population will exhaust its resources. Consequently we expect in practice that  $Y$  will either settle down to some steady state value, fluctuate between various levels or decline. If the first possibility arises then we could replace (4.5.1) by the logistic equation

$$\frac{dY}{dt} = aY - bY^2, \quad (4.5.3)$$

which has the general solution

$$Y = \frac{aY_0}{bY_0 + (a - bY_0)e^{-at}}. \quad (4.5.4)$$

The justification for this type of growth is that

- (a) when  $Y$  is small, (4.5.3) formally reduces to (4.5.1) and the growth is exponential;
- (b) as  $t$  increases  $Y$  approaches the value  $a/b$  steadily and without oscillation.

In (4.5.3), it is standard to call  $a$  the intrinsic rate of increase and  $k = a/b$  the carrying capacity.

Before we introduce into either (4.5.1) or (4.5.3) the effects of predation by  $X$ , let us consider the behaviour of the predator species  $X$  in the absence of prey. Without prey, the predators  $X$  are expected to decrease and so their decline could be represented in terms of the exponentially decaying solution  $X_0 \exp(-et)$  of the differential equation

$$\frac{dX}{dt} = -eX, \quad (4.5.5)$$

where  $e$  is a positive constant. Again, like (4.5.1), the exponentially declining population of predators  $X$  is only reasonable for a limited period of time, since one would expect  $X$  to become extinct in a finite rather than an infinite amount of time when starved of prey  $Y$ .

If we assume, as Volterra and Lotka did, that in the absence of predation,  $Y$  follows a logistic growth curve and that the rate at which prey are eaten is proportional to the product of the densities of predator and prey, then we arrive at the model

$$\frac{dX}{dt} = -eX + fXY, \quad (4.5.6)$$

$$\frac{dY}{dt} = aY - bY^2 - cXY. \quad (4.5.7)$$

The assumptions of Volterra and Lotka are valid under the following conditions:

- (a) one or both species move at random;
- (b) when they meet, there is a constant probability that the predator will kill the prey;
- (c) the time taken by the predator in consuming the prey is negligible.

There are a number of variants of the Volterra–Lotka model that have been proposed and that attempt to take into account other effects that may influence the predator–prey interaction. For example, when a constant number  $\bar{Y}$  of the prey can find some cover or refuge, which makes them inaccessible to the predator, then the system (4.5.6), (4.5.7) is modified to the form

$$\frac{dX}{dt} = -eX + fX(Y - \bar{Y}), \quad (4.5.8)$$

$$\frac{dY}{dt} = aY - bY^2 - cX(Y - \bar{Y}). \quad (4.5.9)$$

In a much more general way, we can follow the model of Rosenzweig and MacArthur (1963) and assume a general growth rate  $f(Y)$  for the prey in the absence of predators and assume that prey are eaten at a rate proportional to some function  $h(X, Y)$  (referred to as the **functional response**) by the predators. In this case we obtain the general model:

$$\frac{dX}{dt} = -eX + kh(X, Y), \quad (4.5.10)$$

$$\frac{dY}{dt} = f(Y) - h(X, Y). \quad (4.5.11)$$

If  $h(X, Y) = cXY$  (often called a **type-I** functional response) then (4.5.10) and (4.5.11) reduce to the basic Volterra–Lotka model (4.5.6), (4.5.7) with general prey growth rate  $f(Y)$  instead of logistic growth. A more realistic choice for  $h(X, Y)$  is a function that increases linearly with predator density  $X$ , but saturates with respect to prey density  $Y$ . This reflects the fact that there is typically a maximum amount of food than one predator can consume and, if there is sufficient prey available for predators to obtain their maximum intake, further increasing the prey density will not significantly alter the overall rate of prey consumption. A common choice is the function

$$h(X, Y) = \frac{cXY}{K + Y}.$$

This type of function was first used in a predator–prey model by Holling in 1959 and is often referred to as a **type-II** response. Note that this function is strikingly similar to the Michaelis–Menten reaction rate (see Section 4.4),

and shares the same characteristic of saturating at some maximum level as prey density (or, in the case of Michaelis–Menten, substrate concentration) becomes very large. A **type-III** functional response is of the form

$$h(X, Y) = \frac{cXY^2}{K + Y^2}.$$

Like the type-II response, the type-III response saturates for large  $Y$ ; unlike type-II, it is not linear in  $Y$  for small  $Y$  but increases as a sigmoidal (S-shaped) curve. This models the situation where the predators cannot find prey as effectively when the prey are very rare.

The Rosenzweig–MacArthur model (4.5.10), (4.5.11) will be examined in some depth in Chapter 9, under certain simplifying but nevertheless realistic assumptions regarding the functions  $f(Y)$  and  $h(X, Y)$ .

To conclude this section, we introduce a somewhat different but also interesting ecological interacting two-species model. Suppose we have two species  $X$  and  $Y$  competing to exist in the same habitat. We assume that, in the absence of  $Y$ ,  $X$  grows according to a logistic law and, similarly in the absence of  $X$ , that  $Y$  does the same. However, when  $X$  and  $Y$  are both present, it is natural to assume that each has an inhibiting effect on the growth of the other. Suppose the inhibition is the same as that in the Volterra–Lotka model. Then we have the system

$$\frac{dX}{dt} = X(e - fY - gX), \quad (4.5.12)$$

$$\frac{dY}{dt} = Y(a - bY - cX). \quad (4.5.13)$$

This is called a **competition model**.

## 4.6 Notes

### Heartbeat

Much of the material here is based on the work of E.C. Zeeman and can be found in his article “Differential equations for the heart beat and nerve impulse”, which appeared in *Towards a Theoretical Biology*, vol. 4, C.H. Waddington, Ed., Edinburgh University Press, Edinburgh, 1972.

### Nerve impulse transmission

It is highly recommended that the reader consult the fundamental papers of A.L. Hodgkin and A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, *J. Physiol.*,

**117**, 500–544. There are several sources of good background material relating to nerve modelling. See, for example, B. Katz, *Nerve, Muscle and Synapse*, McGraw-Hill, New York, 1966; D. Junge, *Nerve and Muscle Excitation*, Sinauer Associates, Sunderland, MA, 1976; H.C. Tuckwell, *Introduction to Theoretical Neurobiology*, vols. 1, 2, Cambridge University Press, London, 1988; and J. Cronin, *Mathematical Aspects of Hodgkin-Huxley Neural Theory*, Cambridge University Press, London, 1987.

## Chemical reactions

A comprehensive treatment of the Belousov-Zhabotinskii reaction and enzymatic catalysis is to be found in J.D. Murray, *Mathematical Biology*, Springer-Verlag, Heidelberg, 1993. For a dedicated text on methods for solving problems involving a small parameter, see E.J. Hinch, *Perturbation Methods*, Cambridge University Press, 1991.

## Predator-prey models

The books by J. Maynard Smith, *Models in Ecology*, Cambridge University Press, London, 1974 and J.D. Murray, *Mathematical Biology*, Springer-Verlag, Heidelberg, 1993 are good sources of information relating to predator-prey models as well as some general models of species interaction.

## Further reading

- Belousov, B. P., An oscillating reaction and its mechanism, *Sborn. Referat. Radiat. Med. Medgiz.*, Moscow, 145, 1959.
- Field, R. J., Körös, E., and Noyes, R. M., Oscillations in chemical systems. II. Thorough analysis of temporal oscillation in bromate-cerium malonic acid systems, *J. Am. Chem. Soc.*, **94**, 8649–8664, 1972.
- Field, R. J. and Noyes, R. M., Oscillations in chemical systems. IV. Limit cycle behaviour in a model of a real chemical reaction, *J. Chem. Phys.*, **60**, 1877–1884, 1974.
- FitzHugh, R., Impulses and physiological states in theoretical models of nerve membrane, *Biophys. J.*, **1**, 445–466, 1961.
- Holling, C. S., The components of predation as revealed by a study of small mammal predation on the European pine sawfly, *Can. Ent.* **91**, 293–320, 1959.
- Kolmogorov, A., Petrovskii, I., and Piskounov, N., Etude de l'équation de la diffusion avec croissance de la quantité de matière et son application à une problème biologique, *Bull. Univ. Moscow. Ser. Int.*, A, I, **6**, 1–25, 1937.
- Lotka, A. J., *Elements of Physical Biology*, Williams & Wilkins, Baltimore, 1925.

- McKean, H. P., Nagumo's equation, *Adv. Math.*, **4**, 209–223, 1970.
- Michaelis, L. and Menten, M.I., Die Kinetik der Invertinwirkung. *Biochem. Z* **49**, 333–369, 1913.
- Nagumo, J., Arimoto, S., and Yoshizawa, S., An active pulse transmission line simulating nerve axon, *Proc. IRE*, **50**, 2061–2071, 1962.
- Rosenzweig, M. L. and MacArthur, R. H., Graphical representation and stability conditions of predator-prey interactions, *Am. Nat.*, **97**, 209–223, 1963.
- Rybak, B. and Béchet, J. J., Recherches sur l'électromécanique cardiaque, *Path. Biol.*, **9**, 2035–2054, 1961.
- Volterra, V., Variazione e fluttuazioni del numero d'individui in specie animali conviventi, *Mem. Accad. Nazionale Lincei* (ser. 6) **2**, 31–113, 1926.
- Zaikin, A. N. and Zhabotinskii, A. M., Concentration wave propagation in two-dimensional liquid-phase self-oscillating systems, *Nature*, **225**, 535–537, 1970.
- Zhabotinskii, A. M., Periodic process of the oxidation of malonic acid in solution (Study of the kinetics of Belousov's), *Biofizika*, **9**, 306–311, 1964.

## Exercises

- 4.1 Show that Zeeman's heartbeat equations have a unique resting state  $x = x_a, b = -(x_a^3 + ax_a)$  and derive a single differential equation satisfied by the muscle fibre length  $x$ .
- 4.2 In the differential equation satisfied by the muscle fibre length  $x$  of Exercise 4.1, let  $x = x_a + y$  and assume that  $y$  is small and that  $y \frac{dy}{dt}$  and  $y^2 \frac{dy}{dt}$  can be neglected. Show that, if  $x_a < \sqrt{-a/3}$ ,  $y$  grows exponentially with time but decays exponentially with time if  $x_a > \sqrt{-a/3}$ . What happens when  $x_a = \sqrt{-a/3}$ ? Give possible interpretations of these three cases.
- 4.3 In the Hodgkin–Huxley model of nerve impulse transmission, assume the potential  $E(x, t)$  has the form of a travelling wave as well as the

conductances  $m, h$  and  $n$  and show that such solutions satisfy the system

$$\begin{aligned} \frac{a}{2R} \frac{d^2 E}{d\xi^2} &= cC_m \frac{dE}{d\xi} + g_{Na}(E - E_{Na}) + g_K(E - E_K) + g_l(E - E_l), \\ c \frac{dm}{d\xi} &= \alpha_m(E)(1 - m) + \beta_m(E)m, \\ c \frac{dh}{d\xi} &= \alpha_h(E)(1 - h) + \beta_h(E)h, \\ c \frac{dn}{d\xi} &= \alpha_n(E)(1 - h) + \beta_n(E)n, \end{aligned}$$

where  $\xi = x + ct$  and  $c$  is the wave number.

- 4.4 Assume the FitzHugh–Nagumo equations (4.3.9) and (4.3.10) admit travelling wave solutions  $u(x, t) = \phi(x + ct)$ ,  $v(x, t) = \psi(x + ct)$ . Deduce the equations to be satisfied by  $\phi$  and  $\psi$ .
- 4.5 Verify that if  $b = 0 = \gamma$  and (4.3.9) and (4.3.10) have travelling wave solutions then

$$\phi(x + ct) \equiv \phi(\xi) = \frac{1}{1 + e^{-\xi/\sqrt{2}}}$$

is a solution provided  $c = \sqrt{2(\frac{1}{2} - a)}$ ,  $0 < a < \frac{1}{2}$ .

- 4.6 For what values of the wave speed  $c$  is

$$\begin{aligned} \phi(x + ct) &= 3 \left( \sqrt{\left[ (2 - a) \left( \frac{1}{2} - a \right) \right]} \cosh \sqrt{ax} + (1 + 1/a) \right)^{-1}, \\ 0 < a &< \frac{1}{2} \end{aligned}$$

a solution of (4.3.9) and (4.3.10) when  $b = \gamma = 0$ ?

- 4.7 If the FitzHugh–Nagumo equations have travelling wave solutions, show that the system governing these solutions has a unique rest state if and only if

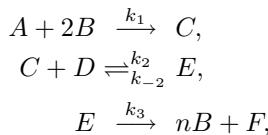
$$(1 - a)^2 < 4b/\gamma.$$

- 4.8 In the simplified McKean model of nerve impulse transmission (see Figure 4.3.4), let  $b = 0 = \gamma$  and determine the forms of travelling waves as functions of  $\xi = x + ct$ ,  $-\infty < \xi < \infty$  and which satisfy the conditions  $\phi(\xi) \rightarrow 0$  as  $\xi \rightarrow -\infty$ ,  $\phi(\xi) \rightarrow 1$  as  $\xi \rightarrow \infty$ .

Determine the wave speed  $c$  in each of the cases  $\theta = \frac{\pi}{2}$  and  $\tan \theta = \frac{1}{8}$ .

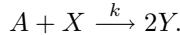
- 4.9 Determine the rest states in the Belousov–Zhabotinskii reaction governed by the model (4.4.7)–(4.4.9).

- 4.10 Using the law of mass action, derive a system of equations modelling the chemical reaction



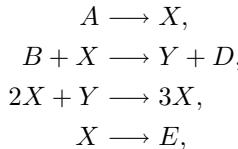
where  $n > 0$  is a constant.

- 4.11 Consider the following chemical reaction.



If the reactant  $A$  is held at a constant concentration  $a$ , use the law of mass action to derive a system of equations for the concentrations of  $X$  and  $Y$ . Suppose the initial concentrations of  $X$  and  $Y$  are  $X_0$  and  $Y_0$  respectively. Solve the system of equations to obtain  $X(t)$  and  $Y(t)$  (find  $X(t)$  first, then substitute into the equation for  $Y$  to find  $Y(t)$ ).

- 4.12 Using the law of mass action, derive a mathematical model governed by the intermediaries  $X$  and  $Y$  in the trimolecular reaction:



where  $A, B, D$  and  $E$  are initial and final products, and all rate constants are equal to 1.

- 4.13 Consider an enzymatic reaction in which the concentration of the enzyme  $E$  is not constant, as in equation (4.4.21), but decays according to law  $E(t) = E_0 e^{-ct}$ . Suppose the initial concentration of the substrate  $S$  is  $S(0) = S_0$ . Solve the differential equation

$$\frac{dS}{dt} = -\frac{kES}{K_m + S}$$

to obtain an equation for  $S(t)$ . Show that  $S(t)$  tends to some positive value as  $t \rightarrow \infty$ .

- 4.14 Determine the rest states in the Volterra–Lotka model of predator–prey interaction (4.5.6), (4.5.7). If predator and prey are present in the steady state, show that the coefficients  $a, b, c, e$  and  $f$  must satisfy the constraints

$$\frac{a}{b} > \frac{e}{f}, \quad c \neq 0.$$

- 4.15 Discuss the same problem as in Exercise 4.16 in relation to the models (4.5.8), (4.5.9) and (4.5.12), (4.5.13).



# Chapter 5

---

## First-Order Systems of Ordinary Differential Equations

---

### 5.1 Existence and uniqueness

In Chapters 2 and 3 certain types of differential equations have been discussed and methods for deriving their solutions have been described. When more general differential equations are considered it is not by any means obvious that they possess solutions. Spending a lot of time trying to solve a differential equation, which does not have a solution, can be very frustrating to say the least. Therefore, we shall give one theorem, which guarantees that a differential equation that satisfies its conditions possesses a solution, and say something about its ramifications. The proof of the theorem is given in the Appendix to this chapter as well as a method for finding the solution.

#### EXISTENCE THEOREM I

Let  $f(t, y)$  be a single-valued continuous function of  $t$  and  $y$  in  $t_0 \leq t \leq t_0 + h$ ,  $|y - y_0| \leq k$  that satisfies:

- (a)  $|f(t, y)| < M$ ,
- (b)  $|f(t, y) - f(t, y')| < K|y - y'|$

for any  $(t, y)$  and  $(t, y')$  that comply with the above inequalities. Then, for  $h < k/M$ , the differential equation

$$\dot{y} = f(t, y) \tag{5.1.1}$$

possesses one, and only one, continuous solution  $y(t)$  in  $t_0 \leq t \leq t_0 + h$  such that  $y(t_0) = y_0$ .

The constant  $h$  determines the range of  $t$  for which the solution is valid while the constant  $k$  sets a limit on how far  $y(t)$  deviates from its initial value. It would be ideal if  $h$  could be made as large as desired. However, the restriction  $h < k/M$  means that  $h$  cannot be increased beyond a certain point without a corresponding increase in  $k$ . But, larger  $h$  and  $k$  may entail an increase in the

bound  $M$  to meet condition (a) and this increase may be sufficient to prevent any improvement in  $k/M$ .

Nevertheless, it may be possible to extend the solution to larger  $t$  by taking  $y(t_0 + h)$  as the initial value at  $t = t_0 + h$  provided that suitable new  $h, k, M$  can be found with this starting point.

Suppose that  $f$  satisfies the conditions of the theorem and, by some means, two continuous solutions  $y_1(t)$ ,  $y_2(t)$  of (5.1.1) have been found such that  $y_1(t_0) = y_0$  and  $y_2(t_0) = y_0$ . Suppose, further, it is known that  $y_1(t)$  is valid for  $t_0 \leq t \leq t_0 + h_1$  whereas  $y_2(t)$  holds for  $t_0 \leq t \leq t_0 + h_2$  with  $h_2 > h_1$ . The uniqueness part of the theorem then says that  $y_1(t) = y_2(t)$  for  $t_0 \leq t \leq t_0 + h_1$ . The same assertion cannot be made for larger values of  $t$  unless it can be demonstrated that  $y_1(t)$  can be continued beyond  $t = t_0 + h_1$ . For example,  $y_1(t) = 1 - t + t^2 - \dots$  and  $y_2(t) = 1/(1+t)$  are solutions of  $(1+t)\dot{y} = -y$ , which are unity at  $t = 0$  so long as  $h_1 < 1$ , but the series in  $y_1$  is not valid in  $t > 1$ .

Generally, it is not difficult to recognise when  $f$  is continuous and to assess  $M$ . Checking (b) can require more effort but there is one case when (b) holds for sure and that is when

$$\left| \frac{\partial}{\partial y} f(t, y) \right| \leq N \quad (5.1.2)$$

for  $t_0 \leq t \leq y_0 + h$ ,  $|y - y_0| \leq k$  and  $N$  finite.

For then

$$\begin{aligned} |f(t, y) - f(t, y')| &= \left| \int_{y'}^y \frac{\partial}{\partial u} f(t, u) du \right| \\ &= \left| \int_0^{y-y'} \frac{\partial}{\partial u} f(t, u - y') du \right| \\ &\leq \int_0^{|y-y'|} \left| \frac{\partial}{\partial u} f(t, u - y') \right| du \\ &\leq N |y - y'|. \end{aligned}$$

### **Example 5.1.1**

The differential equation

$$\dot{y} = g(t)y^2$$

in which  $g$  is continuous clearly has  $f(t, y)$  continuous and, because

$$\frac{\partial}{\partial y} g(t)y^2 = 2g(t)y,$$

satisfies the condition (5.1.2) so long as  $t$  and  $y$  are bounded. Therefore the differential equation has one and only one continuous solution such that  $y(t_0) = y_0$ . It remains valid as  $t$  increases as long as  $t$  and  $y$  remain finite.

Consider, in particular,  $\dot{y} = y^2$ . The solution of this such that  $y(t_0) = 0$  is  $y(t) = 0$  for all  $t$ . On the other hand, if  $y(t_0) = y_0$  with  $y_0 \neq 0$ ,  $y(t) = y_0 / \{1 + (t_0 - t)y_0\}$ . If  $y_0 < 0$  this solution holds for all  $t \geq t_0$ . In contrast, if  $y_0$  is positive,  $y(t)$  becomes unbounded as  $t$  approaches  $t_0 + 1/y_0$ ; in this case the region of validity is confined to  $t_0 \leq t < t_0 + 1/y_0$ .  $\blacksquare$

The conditions of Existence Theorem I are sufficient but not necessary. There are differential equations that do not satisfy the conditions but which possess a unique continuous solution. For example,

$$\dot{y} = \begin{cases} (1 - 2t)y & (t > 0) \\ (2t - 1)y & (t < 0) \end{cases} \quad (5.1.3)$$

subject to  $y = y_0$  ( $\neq 0$ ) at  $t = 0$ . In this case,  $f$  is discontinuous at  $t = 0$  when  $y_0 \neq 0$  so the conditions of Existence Theorem I are not met. Nevertheless there is a unique continuous solution, namely

$$y(t) = \begin{cases} y_0 e^{t-t^2} & (t \geq 0) \\ y_0 e^{t^2-t} & (t \leq 0). \end{cases}$$

There may, however, be values of  $t_0$  or  $y_0$  for which the initial value problem

- (i) has no solution;
- (ii) has a discontinuous solution;
- (iii) has more than one continuous solution.

For instance, the differential equation

$$y\dot{y} = -t \quad (5.1.4)$$

has solution  $y^2 + t^2 = C$  where  $C$  is a constant. An example when there is no solution is to take  $y = 0$  at  $t = 0$ . Then  $C = 0$  and  $y^2 + t^2 = 0$ . This forces  $t = 0$  and there is no solution for  $t > 0$ .

An illustration of more than one solution is provided by  $y = 0$  at  $t = t_0 \neq 0$ . Then  $C = t_0^2$  and  $y = \pm(t_0^2 - t^2)^{1/2}$  giving two solutions while  $t^2 \leq t_0^2$ .

Thus the initial value  $y = 0$  originates difficulties for (5.1.4). For it,  $f(t, y)$  in (5.1.1) is  $-t/y$ , which is infinite at  $y = 0$  for any nonzero  $t$ , and so the conditions of Existence Theorem I cannot be met. Thus there is no warranty of a unique continuous solution. Yet, there is no problem if  $y(t_0) = y_0$  with  $y_0 \neq 0$ . Now Existence Theorem I applies and there is the unique continuous solution  $y = (t_0^2 + y_0^2 - t^2)^{1/2}$  so long as  $t^2 \leq t_0^2 + y_0^2$ .

Any point  $(t_0, y_0)$  at which (i), (ii) or (iii) is true is known as a **singular point**. For example, any point  $(t_0, 0)$  is a singular point of (5.1.4). At a singular point Existence Theorem I must fail, but the converse is false as (5.1.3) shows. Thus, while places where  $f$  does not abide by the conditions of Existence

Theorem I are candidates for singular points, a special investigation has to be undertaken to check whether or not they are actually singular points.

The same nomenclature of singular points is used in connection with systems and with the differential equation of order  $n$  when (i), (ii) or (iii) occurs. Existence Theorems II and III (given in the Appendix) are invalid at singular points but the points where their conditions are unsatisfied are not necessarily singular points.

The linear system

$$\dot{y}_i = \sum_{j=1}^n a_{ij}(t)y_j + f_i(t) \quad (i = 1, \dots, n) \quad (5.1.5)$$

conforms to Existence Theorem II except at those values of  $t$  where  $a_{ij}$  or  $f_i$  are discontinuous. Apart from these values, the initial value problem has a unique continuous solution. In particular, the linear system with constant coefficients possesses a unique continuous solution except, perhaps, for those  $t$  where  $f_i$  is not continuous.

---

## 5.2 Epidemics

The simplest model of the spread of an epidemic in a population stipulates that at time  $t$  there are  $x$  susceptible individuals and  $y$  infected who may transmit the disease. It is assumed that the mixing of these two groups passes on the illness and that, in the short time  $\delta t$ ,  $\mu xy\delta t$  new infections occur. Also some of those infected will die, or stop mixing, or recover and become immune; suppose that  $\nu y\delta t$  disappear in this way. Then, if  $\rho\delta t$  new susceptibles arrive in the interval  $\delta t$ ,

$$\dot{x} = -\mu xy + \rho, \quad \dot{y} = \mu xy - \nu y. \quad (5.2.1)$$

Normally  $\mu$ ,  $\nu$  and  $\rho$  are taken as nonnegative constants and it is convenient to assume that they are positive. The right-hand sides of (5.2.1) vanish for  $x = x_0$ ,  $y = y_0$  where

$$-\mu x_0 y_0 + \rho = 0, \quad \mu x_0 y_0 - \nu y_0 = 0 \quad (5.2.2)$$

or  $x_0 = \nu/\mu$ ,  $y_0 = \rho/\nu$ . If  $x(t) = x_0$ ,  $y(t) = y_0$  then  $\dot{x} = 0$ ,  $\dot{y} = 0$  and the differential equations (5.2.1) are satisfied. In other words,  $(x_0, y_0)$  is an *equilibrium state* in which the numbers of susceptibles and infected do not vary.

Let us now address the question of whether equilibrium is approached from a nearby state and, if so, in what manner. Put

$$x = x_0(1 + \xi), \quad y = y_0(1 + \eta)$$

where  $\xi$  and  $\eta$  are so small that their products may be neglected. Then (5.2.1) becomes

$$\dot{\xi} = -\sigma(\xi + \eta), \quad \dot{\eta} = \nu\xi$$

where  $\sigma = \mu\rho/\nu$ . Substituting for  $\xi$  from the second equation we obtain

$$\ddot{\eta} + \sigma\dot{\eta} + \sigma\nu\eta = 0$$

of which the solution such that  $\eta = \eta_0$ ,  $\dot{\eta} = \nu\xi_0$  at  $t = 0$  is

$$\eta = e^{-\sigma t/2} \left\{ \eta_0 \cos \omega t + \frac{1}{\omega} (\nu\xi_0 + \frac{1}{2}\sigma\eta_0) \sin \omega t \right\}$$

where  $\omega^2 = \sigma\nu - \sigma^2/4$ . Hence

$$\xi = e^{-\sigma t/2} \left\{ \xi_0 \cos \omega t - \frac{\sigma}{2\omega} (\xi_0 + 2\eta_0) \sin \omega t \right\}.$$

When  $\omega^2 > 0$ , i.e.,  $4\nu > \sigma$  or  $4\nu^2 > \mu\rho$ , the population, after a small departure from equilibrium, returns to equilibrium in an oscillatory fashion with exponential decay. If  $\omega^2 < 0$  or  $4\nu^2 < \mu\rho$  the fact that  $|\omega| < \sigma/2$  ensures exponential decay again but there is no accompanying oscillation. In either case the population returns to equilibrium, the approach being more rapid when oscillations are present.

### 5.3 The phase plane and the Jacobian matrix

If less specific assumptions are made about the mechanism of propagation of epidemics, the most that can be said is that a system of the type

$$\dot{x} = f(x, y), \quad \dot{y} = g(x, y) \tag{5.3.1}$$

will need to be solved. Any point  $(x_0, y_0)$  such that

$$f(x_0, y_0) = 0, \quad g(x_0, y_0) = 0 \tag{5.3.2}$$

is called a *critical point* or *fixed point* or *equilibrium point*. A solution that starts at an equilibrium point never leaves it because  $\dot{x}$  and  $\dot{y}$  both vanish there provided that  $f$  and  $g$  satisfy the conditions of Existence Theorem II.

When the solution of (5.3.1) has been found, say  $x = h_1(t)$ ,  $y = h_2(t)$ , the point  $(x, y)$  can be plotted in the  $(x, y)$ -plane at time  $t$ . As  $t$  varies,  $(x, y)$  will trace a curve in the  $(x, y)$ -plane. This curve is known as a *trajectory* and the  $(x, y)$ -plane is called the **phase plane**. By attaching an arrow to each trajectory the direction in which  $(x, y)$  moves as  $t$  increases can be indicated. The phase plane then contains all the information in (5.3.1) except the rate

at which the trajectory is traversed. The slope of a trajectory is given from (5.3.1) by

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{g(x, y)}{f(x, y)}. \quad (5.3.3)$$

A trajectory is vertical at any  $(x, y)$  where  $g(x, y) \neq 0$ ,  $f(x, y) = 0$  and horizontal where  $g(x, y) = 0$ ,  $f(x, y) \neq 0$ . The trajectory corresponding to an equilibrium point reduces to a single point.

When  $f$  and  $g$  satisfy the conditions of Existence Theorem II, the initial value problem has a single continuous solution in the neighbourhood of  $t = t_0$ . Therefore, in this case, only one trajectory passes through a given point of the phase plane, i.e., under the conditions of Existence Theorem II, *two trajectories do not intersect in general*.

All these notions can be generalised to a system of  $n$  equations. A solution still describes a trajectory, which is now a curve in a space of  $n$  dimensions, and we talk of a **phase space** rather than a phase plane. Diagrams are, however, much more difficult to draw.

The same question about behaviour near equilibrium that was asked for epidemics can be raised here. Put

$$x = x_0 + \xi, \quad y = y_0 + \eta$$

where  $(x_0, y_0)$  is in conformity with (5.3.2). Because of the smallness of  $\xi$  and  $\eta$ , it will be assumed that  $f(x, y)$  can be approximated by the first terms in its Taylor expansion, namely

$$\xi \left( \frac{\partial f}{\partial x} \right)_0 + \eta \left( \frac{\partial f}{\partial y} \right)_0$$

where  $(\cdot)_0$  means calculate the value at  $x = x_0$ ,  $y = y_0$ . The approximation to (5.3.1) is then

$$\dot{\xi} = \xi \left( \frac{\partial f}{\partial x} \right)_0 + \eta \left( \frac{\partial f}{\partial y} \right)_0, \quad (5.3.4)$$

$$\dot{\eta} = \xi \left( \frac{\partial g}{\partial x} \right)_0 + \eta \left( \frac{\partial g}{\partial y} \right)_0, \quad (5.3.5)$$

a linear system with constant coefficients. The behaviour of such systems in the phase plane will be examined in succeeding sections. It is worth noting that this linear system can be written in matrix–vector notation (see Section 3.5):

$$\dot{\mathbf{u}} = (\mathbf{J})_0 \mathbf{u},$$

where  $\mathbf{u} = (\xi, \eta)$  and  $\mathbf{J}$  is a matrix called the **Jacobian matrix**. The entries of  $\mathbf{J}$  are the four partial derivatives of  $f$  and  $g$  with respect to  $x$  and  $y$ :

$$\mathbf{J}(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{bmatrix}.$$

The behaviour near equilibrium is a matter of **local stability** and, as we shall see in the next section, can be determined by the eigenvalues of the Jacobian matrix. The larger question of what happens when the initial state is not near equilibrium is one of **global stability**. This will be tackled in Section 5.5 and, in the meantime, we merely remark that for systems of three or more equations global behaviour is very varied and imperfectly understood.

---

## 5.4 Local stability

It has been discovered in the preceding section that local stability reduces to a discussion of

$$\dot{x} = ax + by, \quad \dot{y} = cx + dy \quad (5.4.1)$$

where  $a, b, c$  and  $d$  are real constants (corresponding to the entries of the Jacobian matrix, evaluated at the fixed point). The goal of this section is to determine the trajectories of this system and, in doing so, the stability of the fixed point at  $(x, y) = (0, 0)$ .

Let us first remark that, if  $\alpha$  is a real constant,  $x(t + \alpha), y(t + \alpha)$  occupies the same points in the phase plane as  $t$  varies as  $x(t), y(t)$  though at a time  $\alpha$  earlier. So both points describe the same trajectory despite being different solutions. More than one solution can lie on one trajectory.

In finding the trajectories we shall ignore the degenerate case in which  $ad = bc$ ; should it arise, the equations can be integrated directly without trouble. It would, in any case, be necessary to reconsider the validity of (5.3.4) and (5.3.5) as an adequate prescription for local stability in these circumstances. Therefore, from now on, it will be assumed that

$$ad \neq bc \quad (5.4.2)$$

so that there is a single fixed point at the origin.

According to Section 3.4, the first attempt at a solution is  $x = \alpha e^{\lambda t}$ ,  $y = \beta e^{\lambda t}$ . For the satisfaction of (5.4.1) we require

$$(a - \lambda)\alpha + b\beta = 0, \quad (5.4.3)$$

$$c\alpha + (d - \lambda)\beta = 0. \quad (5.4.4)$$

These give nonzero  $\alpha, \beta$  only if

$$(a - \lambda)(d - \lambda) - bc = 0$$

or

$$\lambda^2 - (a + d)\lambda + ad - bc = 0.$$

The roots are  $\lambda_1, \lambda_2$  where

$$\begin{aligned} 2\lambda_1 &= a + d + \{(a - d)^2 + 4bc\}^{1/2}, \\ 2\lambda_2 &= a + d - \{(a - d)^2 + 4bc\}^{1/2}. \end{aligned}$$

Note that  $\lambda_1$  and  $\lambda_2$  are the **eigenvalues** of the Jacobian matrix

$$\mathbf{J} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Several cases have to be studied.

$$(a) (\mathbf{a} - \mathbf{d})^2 + 4\mathbf{b}\mathbf{c} > 0.$$

In this case the values of  $\lambda_1$  and  $\lambda_2$  are real and distinct; the special procedure for multiple roots does not have to be called on.

Assume first that  $b \neq 0$ . Then, from (5.4.3), we can choose  $\alpha = b$ ,  $\beta = \lambda_1 - a$  corresponding to  $\lambda_1$  and  $\alpha = b$ ,  $\beta = \lambda_2 - a$  corresponding to  $\lambda_2$ . Consequently

$$x = b(C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}), \quad (5.4.5)$$

$$y = (\lambda_1 - a)C_1 e^{\lambda_1 t} + (\lambda_2 - a)C_2 e^{\lambda_2 t} \quad (5.4.6)$$

where  $C_1$  and  $C_2$  are arbitrary constants. These equations may be rearranged to give

$$(\lambda_2 - a)x - by = b(\lambda_2 - \lambda_1)C_1 e^{\lambda_1 t}, \quad (5.4.7)$$

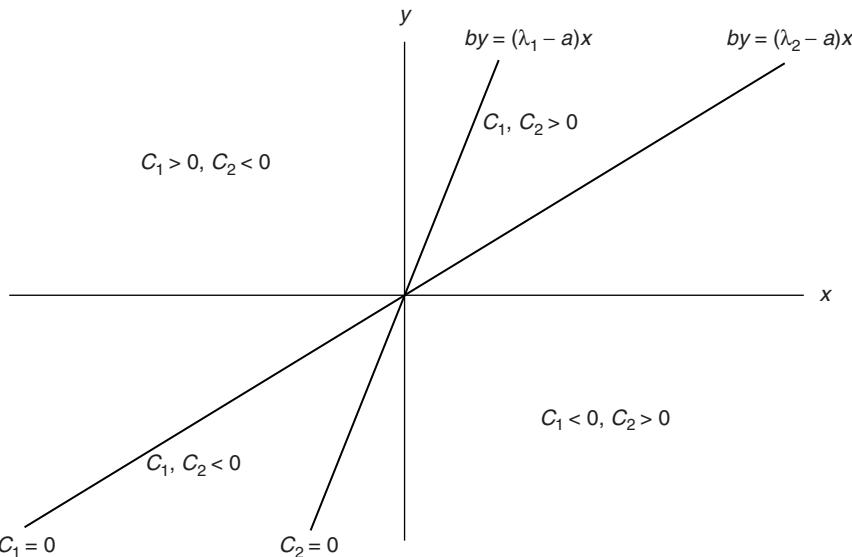
$$(\lambda_1 - a)x - by = b(\lambda_1 - \lambda_2)C_2 e^{\lambda_2 t}. \quad (5.4.8)$$

From (5.4.7),  $(\lambda_2 - a)x - by$  cannot change sign as  $t$  varies. Therefore the trajectory cannot go over the line  $(\lambda_2 - a)x = by$ . Similarly, from (5.4.8), the trajectory cannot trespass across  $(\lambda_1 - a)x = by$ . These lines are displayed in Figure 5.4.1 as well as the regions to which the trajectory is confined for various choices of  $C_1, C_2$  when  $b > 0$  and both  $\lambda_1, \lambda_2$  are negative.

Suppose now that both  $\lambda_1$  and  $\lambda_2$  are negative. It is evident that  $x \rightarrow 0$ ,  $y \rightarrow 0$  as  $t \rightarrow \infty$ . Also  $\lambda_1 > \lambda_2$  so that  $(\lambda_1 - a)x \sim by$  as  $t \rightarrow \infty$ , so long as  $C_1 \neq 0$ . Moreover, as  $t \rightarrow -\infty$ ,  $|x|$  and  $|y|$  become large and  $(\lambda_2 - a)x \sim by$  provided  $C_2 \neq 0$ ,  $x$  approaching  $-\infty$  when  $b > 0$ ,  $C_2 < 0$ . The trajectories, therefore, have the shape depicted in Figure 5.4.2 for  $b > 0$ . The exclusion so far of  $C_1 = 0$  or  $C_2 = 0$  can be remedied immediately because their trajectories are the dividing straight lines on account of (5.4.7) and (5.4.8). The arrows on the curves indicate the direction in which  $(x, y)$  moves as  $t$  increases. A fixed point of this type is known as a **stable node**.

When  $\lambda_1$  and  $\lambda_2$  are both positive, the curves are similar in character to those in Figure 5.4.2 but the directions of the arrows are reversed because  $(x, y)$  moves away from the origin as  $t$  increases. We have an **unstable node**.

The remaining possibility is that  $\lambda_1$  and  $\lambda_2$  have opposite signs so that  $\lambda_1 > 0$  and  $\lambda_2 < 0$ . From (5.4.7), the magnitude of  $(\lambda_2 - a)x - by$  increases



**FIGURE 5.4.1:** Lines that cannot be crossed by trajectories.

with time whereas that of  $(\lambda_1 - a)x - by$  diminishes. The origin can never be reached unless  $C_1 = 0$  when the trajectory is a straight line. The trajectory is also a straight line when  $C_2 = 0$ , but now the origin is departed from. The behaviour of the trajectories is displayed in Figure 5.4.3. The fixed point is called a **saddle-point**. Clearly a point  $(x, y)$  started near the origin cannot stay near it in general and there is no stability.

So far we have assumed that  $b \neq 0$ . If  $b = 0$  we see at once from (5.4.1) that  $x = C_1 e^{at}$  and then

$$y = C_2 e^{dt} + cC_1 e^{at}/(a - d).$$

We remark that  $a \neq d$  because  $(a - d)^2$  must be positive when  $b = 0$ . In this case, the dividing lines are  $x = 0$  and  $(a - d)y = cx$ . Apart from this change the pictures are practically unaltered. There is a node if  $a$  and  $d$  have the same sign (stable if  $a < 0$ , unstable if  $a > 0$ ) and a saddle-point if  $a$  and  $d$  have opposite signs.

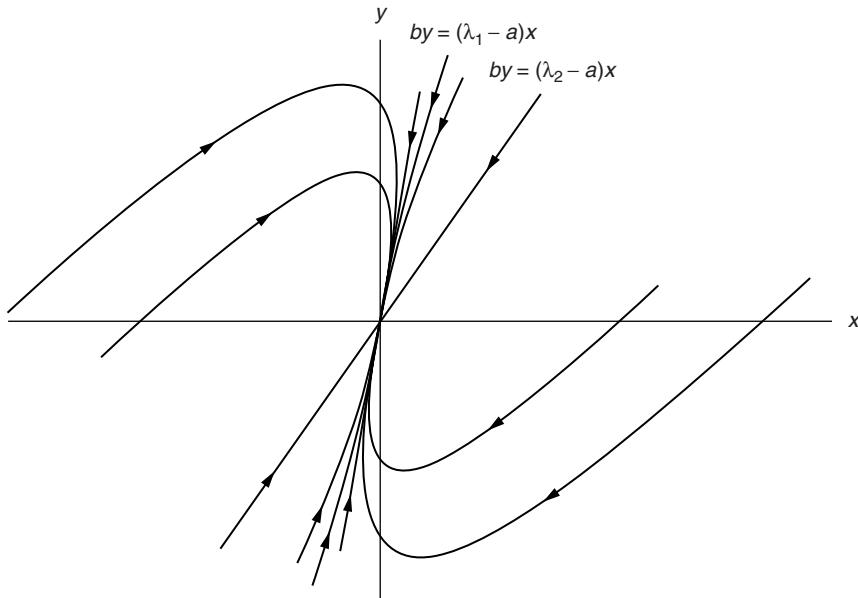
$$(b) (a - d)^2 + 4bc < 0$$

This possibility can occur only when  $bc < 0$ ; so neither  $b$  nor  $c$  vanishes and they have opposite signs. The roots  $\lambda_1$  and  $\lambda_2$  are still distinct but they are now complex conjugates. Write  $\lambda_1 = \frac{1}{2}(a + d) + i\omega$ ,  $\lambda_2 = \frac{1}{2}(a + d) - i\omega$  where

$$\omega^2 = -\frac{1}{4}(a - d)^2 - bc.$$

Then

$$x = Ae^{\frac{1}{2}(a+d)t} \cos(\omega t - \alpha) \quad (5.4.9)$$



**FIGURE 5.4.2:** The stable node.

where  $A$  and  $\alpha$  are arbitrary constants. From (5.4.1)

$$y = (A/b)e^{\frac{1}{2}(a+d)t} \left\{ \frac{1}{2}(d-a) \cos(\omega t - \alpha) - \omega \sin(\omega t - \alpha) \right\}. \quad (5.4.10)$$

The formulae (5.4.9) and (5.4.10) can be combined to give

$$cx^2 + (d-a)xy - by^2 = -(\omega^2 A^2/b)e^{(a+d)t}. \quad (5.4.11)$$

Suppose that  $a+d = 0$  so that  $\lambda_1$  and  $\lambda_2$  are purely imaginary. The equation of a trajectory is given by (5.4.11) as

$$cx^2 + (d-a)xy - by^2 = -\omega^2 A^2/b.$$

Rotate the axes by means of the transformation

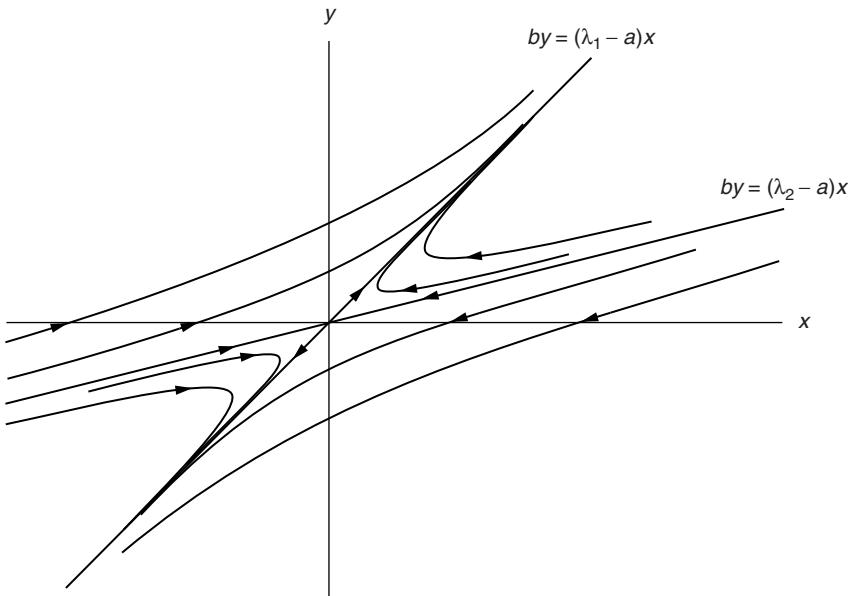
$$x = X \cos \theta - Y \sin \theta, \quad y = X \sin \theta + Y \cos \theta$$

where

$$\tan 2\theta = \frac{d-a}{b+c}.$$

The equation of the curve goes over to

$$A'X^2 + C'Y^2 = -\omega^2 A^2/b$$



**FIGURE 5.4.3:** The saddle-point.

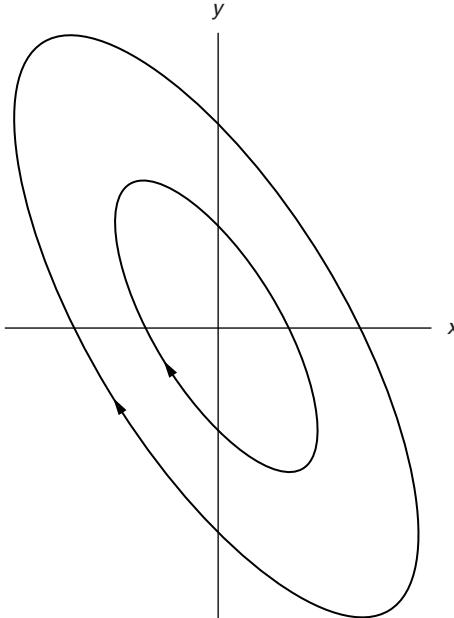
where

$$\begin{aligned} A' &= \frac{1}{2}(c - b) + \frac{1}{2}\{(b + c)^2 + (d - a)^2\}^{1/2}, \\ C' &= \frac{1}{2}(c - b) - \frac{1}{2}\{(b + c)^2 + (d - a)^2\}^{1/2}. \end{aligned}$$

Since  $A'C' = \omega^2$ ,  $A'$  and  $C'$  have the same sign. Also  $A' + C' = c - b$  so that if  $b < 0$ , which implies  $c > 0$ ,  $A'$  and  $C'$  are positive, whereas if  $b > 0$ , which makes  $c < 0$ ,  $A'$  and  $C'$  are negative. Thus  $A'$  and  $C'$  have the opposite sign to  $b$  and the trajectory is an ellipse with semi-axes  $\omega|A|/(-bA')^{1/2}$  and  $\omega|A|/(-bC')^{1/2}$ . Typical trajectories are drawn in Figure 5.4.4; the fixed point is known as a **centre**. With regard to the direction of motion on a trajectory, we see from (5.4.1) that when  $x = 0$ ,  $\dot{x} = by$ . Hence, when  $b > 0$ ,  $x$  must be increasing at positive  $y$  and so the direction is as shown in Figure 5.4.4; if  $b < 0$  the arrows have to be reversed.

The equations (5.4.9) and (5.4.10) make it evident that  $x$  and  $y$  vary harmonically when  $a + d = 0$ . The point  $(x, y)$  therefore makes continual circuits round the origin and is forever retracing its path. A trajectory started from near the origin never leaves the neighbourhood but never swings into the origin. Therefore, there is stability in the sense that  $(x, y)$  remains in the vicinity of the fixed point, if it is initially near there, but it never attains the fixed point.

It should be observed that any closed trajectory implies periodic motion because it entails there being a fixed  $T$  such that  $x(t+T) = x(t)$ ,  $y(t+T) = y(t)$



**FIGURE 5.4.4:** The centre.

for all  $t$ . The motion need not, however, have the simple harmonic character mentioned above when the system is more general than (5.4.1).

Turning now to the case when  $a + d \neq 0$ , we note that the only difference is the exponential factor in (5.4.11). The trajectory may be thought of instantaneously as an ellipse whose axes are changing exponentially. The trajectory therefore spirals about the origin as shown in Figure 5.4.5. If  $a + d < 0$ , the point  $(x, y)$  must approach the origin as  $t \rightarrow \infty$ . The direction of motion along a trajectory is then that of Figure 5.4.5 and the fixed point is called a **stable focus**. When  $a + d > 0$ ,  $(x, y)$  departs from the origin, the arrows are reversed and we have an **unstable focus**.

$$(c) (a - d)^2 + 4bc = 0$$

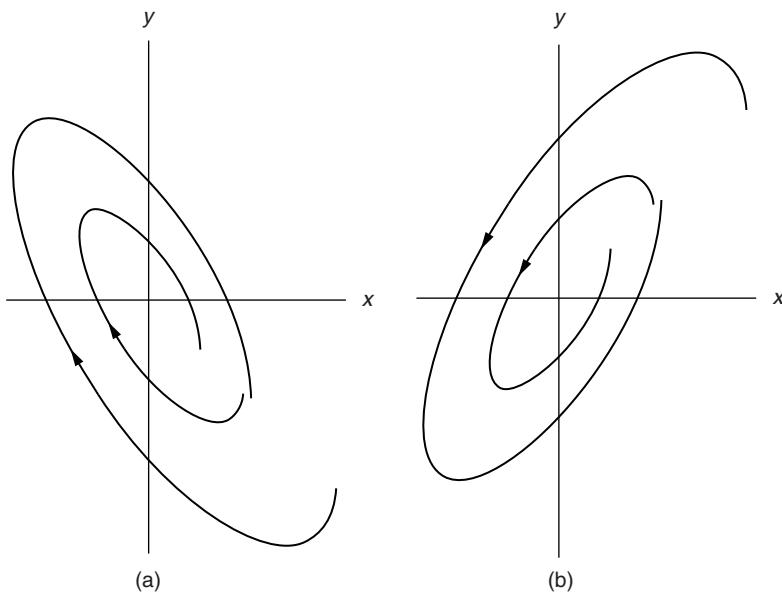
In this case  $\lambda_1 = \lambda_2 = \frac{1}{2}(a + d)$ . However,

$$(a + d)^2 = (a - d)^2 + 4ad = -4bc + 4ad \neq 0$$

by (5.4.2) so  $\lambda_1$  and  $\lambda_2$  are nonzero.

If  $b \neq 0$ , (5.4.3) and (5.4.4) supply only the single solution  $\alpha = b$ ,  $\beta = \frac{1}{2}(d - a)$ . To find a second solution we try, according to Section 3.4,  $x = (\gamma + \alpha t)e^{\frac{1}{2}(a+d)t}$ ,  $y = (\delta + \beta t)e^{\frac{1}{2}(a+d)t}$  with the result that  $\gamma = 0$ ,  $\delta = 1$ . The technique of Section 3.5 leads to the same answer. Consequently

$$\begin{aligned} x &= b(C_1 + C_2 t)e^{\frac{1}{2}(a+d)t}, \\ y &= \left\{ C_2 + \frac{1}{2}(d - a)(C_1 + C_2 t) \right\} e^{\frac{1}{2}(a+d)t}. \end{aligned}$$



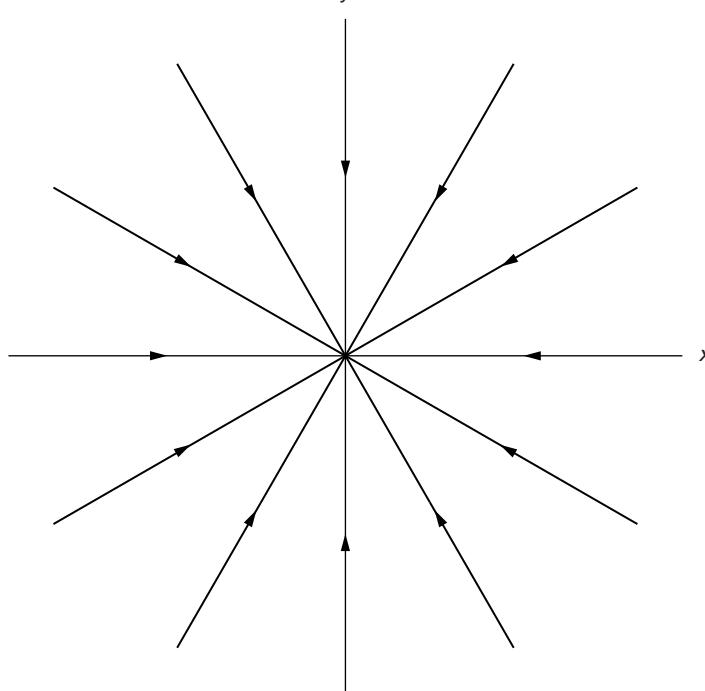
**FIGURE 5.4.5:** The focus when  $a + d < 0$ ; (a)  $b > 0$ , (b)  $b < 0$ .

The line  $by = \frac{1}{2}(d-a)x$  cannot be crossed and the structure of the trajectories is similar to that of Figure 5.4.2 when the dividing lines coalesce. This fixed point is therefore also termed a node; it is stable if  $a + d < 0$  and unstable if  $a + d > 0$ .

If  $b = 0$ , then  $a = d$  and  $x = C_1 e^{at}$ ,  $y = (C_2 + cC_1 t)e^{at}$  and the trajectories are not much changed in shape if  $c \neq 0$ . If, in addition,  $c = 0$  the trajectories are the straight lines  $y/x = \text{constant}$  (see Figure 5.4.6). The fixed point is still designated a node, stable if  $a < 0$  and unstable if  $a > 0$ .

These results can be summarised as follows. If  $ad \neq bc$ , the fixed point of (5.4.1) is:

1. **Stable** if  $\text{Re}(\lambda_1) < 0$  and  $\text{Re}(\lambda_2) < 0$  ( $\text{Re}(\lambda)$  denotes the real part of  $\lambda$ ).
  - (a) If  $\lambda_1$  and  $\lambda_2$  are real, the trajectories form a stable node.
  - (b) If  $\lambda_1$  and  $\lambda_2$  are a complex conjugate pair, the trajectories form a stable focus.
2. **A saddle point** if  $\lambda_1$  and  $\lambda_2$  are real and have opposite signs.
3. **A centre** if  $\lambda_1$  and  $\lambda_2$  are purely imaginary.
4. **Unstable** if  $\text{Re}(\lambda_1) > 0$  and  $\text{Re}(\lambda_2) > 0$ .
  - (a) If  $\lambda_1$  and  $\lambda_2$  are real, the trajectories form an unstable node.



**FIGURE 5.4.6:** The case  $b = c = 0$ ,  $a < 0$ .

- (b) If  $\lambda_1$  and  $\lambda_2$  are a complex conjugate pair, the trajectories form an unstable focus.

The reader will note that *the stability of the fixed point is determined solely by the sign of the real part of the eigenvalues  $\lambda_1$  and  $\lambda_2$* . The fixed point is stable if and only if both  $\lambda_1$  and  $\lambda_2$  have negative real part. In addition, there can be no oscillation if  $\lambda_1$  and  $\lambda_2$  are real; in particular, the system will not be oscillatory if  $bc > 0$ .

## 5.5 Stability

We now want to investigate what general conclusions can be drawn about the behaviour of solutions to (5.3.1) on the basis of the model of (5.3.4), (5.3.5) and the trajectories determined in Section 5.4. To fix ideas, we consider a somewhat generalised model of epidemics in which the number of susceptibles  $x$  and of infected  $y$  satisfy

$$\dot{x} = h(x, y)x, \quad \dot{y} = k(x, y)y.$$

The behaviour near an equilibrium point in which neither  $x$  nor  $y$  is zero is of interest. So  $x_0$  and  $y_0$  are taken to satisfy

$$h(x_0, y_0) = 0, \quad k(x_0, y_0) = 0.$$

It then follows that, in the notation of Section 5.3,

$$\begin{aligned} (\partial f / \partial x)_0 &= h_1 x_0, & (\partial f / \partial y)_0 &= h_2 x_0, \\ (\partial g / \partial x)_0 &= k_1 y_0, & (\partial g / \partial y)_0 &= k_2 y_0 \end{aligned}$$

where  $h_1$ ,  $h_2$ ,  $k_1$  and  $k_2$  are the values of  $\partial h / \partial x$ ,  $\partial h / \partial y$ ,  $\partial k / \partial x$  and  $\partial k / \partial y$ , respectively, at  $(x_0, y_0)$ . Thus, in the theory of Section 5.4,  $a = h_1 x_0$ ,  $b = h_2 x_0$ ,  $c = k_1 y_0$ ,  $d = k_2 y_0$ .

Since the presence of the infected tends to reduce the number of susceptibles by infection, we expect  $h_2 < 0$ . As the number of infected increases there will be less opportunity to affect the susceptibles and so  $k_2 < 0$ ,  $k_1 > 0$ . If there is a birth rate of susceptibles, we can suppose  $h_1 > 0$ , though  $h_1$  will be rather small in comparison with other partial derivatives in most epidemics, because they tend to spread much faster than susceptibles are created.

Since  $ad - bc = (h_1 k_2 - h_2 k_1)x_0 y_0$  we can be sure that  $ad \neq bc$  when  $h_1$  is small, as suggested above, and the theory of Section 5.4 can be applied. In order that there can be any kind of oscillation, we must have case (b) of Section 5.4, i.e.,

$$(h_1 x_0 - k_2 y_0)^2 + 4h_2 k_1 x_0 y_0 < 0.$$

The second term on the right-hand side is negative; so the inequality is feasible if the first term is not too large. Since  $h_1$  is small, this will be true if  $k_2$  is not too large. The oscillations are likely to remain near equilibrium because  $a + d = h_1 x_0 + k_2 y_0$  is negative on account of the smallness of  $h_1$ . The fixed point will be a stable focus or, possibly, a centre.

In the absence of oscillations,  $\lambda_1$  and  $\lambda_2$  will both be negative because  $ad - bc$  is positive and the equilibrium will be a stable node.

Quite a lot of qualitative information about the behaviour of the solution has been obtained without too specific assumptions about  $h$  and  $k$ . Of course, further conclusions could be drawn if more was known about  $h$  and  $k$ . In other problems, the signs of the partial derivatives might be different and the behaviour near equilibrium changed thereby. Some of the unstable fixed points might occur. However, such instability merely means departure from equilibrium and, once this exceeds a certain amount, the model of (5.3.4) and (5.3.5) loses its validity because it assumed motion near equilibrium. We then enter the arena of global stability, with the possibility of some kind of stable behaviour away from equilibrium, a matter to be discussed in the next section.

In the foregoing the presence of an equilibrium point has been assumed and it will not be amiss to say a word or two about how the existence of an equilibrium point is verified. Often it will be done most simply by graphical means

but sometimes an analytical argument is helpful. With  $\partial k/\partial x > 0$ ,  $k(x, y) = 0$  can be solved for  $x$  to give a unique function  $x(y)$  of  $y$ . Because

$$\frac{dx}{dy} = -\frac{\partial k/\partial y}{\partial k/\partial x}$$

we have  $dx/dy > 0$  when  $\partial k/\partial y < 0$ . There can be no infected if there is no population, and so  $x(y) > 0$  but less than some bound. Thus  $h(x(y), y)$  is such that

$$\frac{dh}{dy} = \frac{\partial h}{\partial x} \frac{dx}{dy} + \frac{\partial h}{\partial y},$$

which will be negative when both  $\partial h/\partial x$  and  $\partial h/\partial y$  are. Then  $h(x(y), y)$  decreases as  $y$  increases so that if  $h(x(y), y)$  is positive for small  $y$  and negative for large  $y$  there will be one and only one equilibrium point.

As a final illustration we consider a case in which the trajectories can be traced completely, namely

$$\dot{x} = y, \quad \dot{y} = \frac{1}{2}(1 - x^2)$$

or, equivalently,

$$\ddot{x} + \frac{1}{2}(x^2 - 1) = 0.$$

We have

$$y \frac{dy}{dx} = \frac{1}{2}(1 - x^2)$$

which gives, on integration,

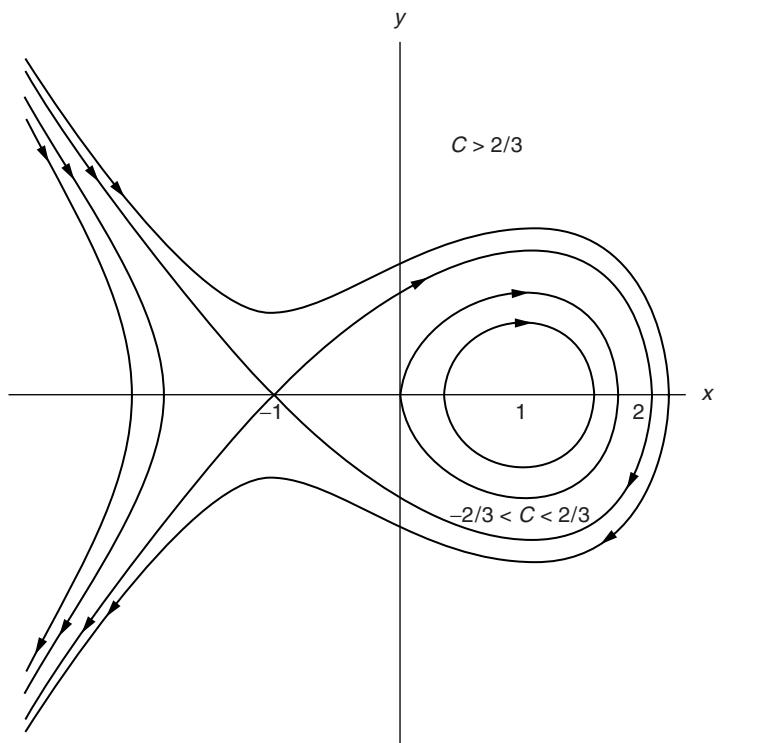
$$y^2 = x - \frac{1}{3}x^3 + C.$$

These trajectories are displayed graphically in Figure 5.5.1. The fixed point  $x = 1, y = 0$  is a centre and  $x = -1, y = 0$  is a saddle-point. If  $-\frac{2}{3} < C < \frac{2}{3}$  the closed curves surrounding  $x = 1, y = 0$  show that periodic motion is possible with proper initial conditions. If  $C = \frac{2}{3}$  no oscillations are allowed but  $x = -1, y = 0$  can be tended to if the initial conditions are appropriate. If  $C > \frac{2}{3}$ ,  $x$  and  $y$  always approach negative infinity as  $t \rightarrow \infty$ .

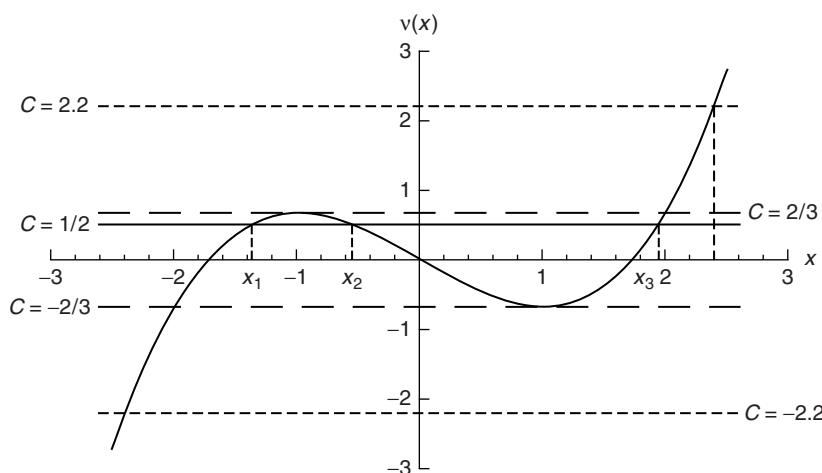
A quick idea of the shapes of the trajectories can be obtained as follows. Put

$$y^2 = C - v(x).$$

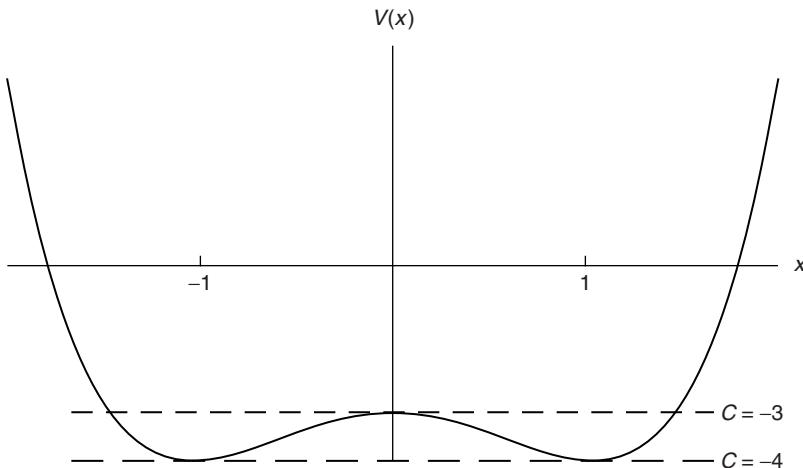
Since  $y^2$  cannot be negative the only values of  $x$  that can occur are those that satisfy  $v(x) \leq C$ . Such values can be seen easily from a graph of  $v(x)$ . In Figure 5.5.2 a graph is displayed together with various possibilities for  $C$ . The line on which  $C = 2.2$  intersects the curve at  $x = 2.4$  approximately and is above the curve for  $x < 2.4$ . Hence  $v(x)$  is below 2.2 for  $x < 2.4$ . Thus the trajectory starts at  $x = -\infty$  with  $y$  positive (because  $\dot{x}$  must be positive at the start) and moves to the right until it crosses the  $x$ -axis at  $x = 2.4$  approximately. Thereafter it returns to  $x = -\infty$  with  $y$  negative. Likewise,



**FIGURE 5.5.1:** Nonlinear conservative system.



**FIGURE 5.5.2:** Qualitative determination of trajectories.



**FIGURE 5.5.3:** Graph of  $x^4 - 2x^2 - 3$ .

when  $C = -2.2$ ,  $x$  is restricted to  $x < -2.4$ ; the trajectory goes from  $-\infty$  and back again, crossing the  $x$ -axis at  $x = -2.4$  approximately. On the other hand, when  $C = \frac{1}{2}$ , there are three intersections with the curve at  $x_1$ ,  $x_2$  and  $x_3$ . Two trajectories are possible now. On one  $x \leq x_1$  and it is similar to the one when  $C = -2.2$ . On the other trajectory  $x$  is confined to the interval  $x_2 \leq x \leq x_3$  so that the trajectory is a closed curve. It is clear from Figure 5.5.2 that trajectories that are closed curves are possible only for  $-\frac{2}{3} < C < \frac{2}{3}$ .

This is an example of a *nonlinear conservative system*. A more general version is

$$\dot{x} = y, \quad \dot{y} = -\frac{d}{dx}V(x)$$

or

$$\ddot{x} + \frac{dV(x)}{dx} = 0.$$

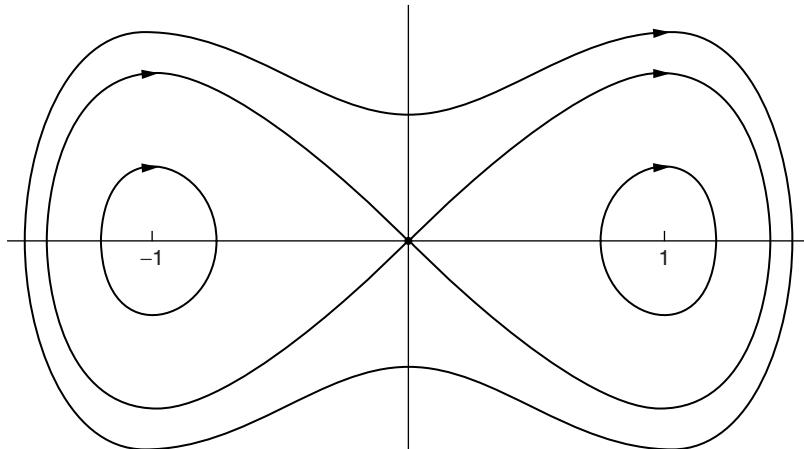
The trajectories can be determined as above and are

$$\frac{1}{2}y^2 + V(x) = C.$$

The graphical method will give an indication of the shapes of the trajectories. As an illustration take

$$V(x) = x^4 - 2x^2 - 3.$$

Its graph is drawn in Figure 5.5.3. Evidently, when  $C > -3$  the trajectory consists of a single closed curve. When  $-3 > C > -4$ , however, the trajectory has two parts. One is a closed curve surrounding  $x = -1$  while the other is a closed curve around  $x = 1$ . There is no trajectory for  $C$  below  $-4$ . Since the equilibrium points are  $x = 0, \pm 1$  the trajectories are as shown in Figure 5.5.4. The points  $x = 1$  and  $x = -1$  are centres whereas  $x = 0$  is a saddle-point.



**FIGURE 5.5.4:** The trajectories corresponding to Figure 5.5.3.

## 5.6 Limit cycles

As a beginning to the discussion of global stability, let us examine the system

$$\dot{x} = y, \quad \dot{y} = \epsilon(1 - x^2)y - x$$

where  $\epsilon$  is a nonnegative constant. Eliminating  $y$  we obtain

$$\ddot{x} - \epsilon(1 - x^2)\dot{x} + x = 0 \quad (5.6.1)$$

which is known as **van der Pol's equation**.

Assume that  $\epsilon$  is small. As a first approximation one would ignore the term involving  $\epsilon$ . The general solution is then

$$x = C \cos(t - \alpha)$$

where  $C$  and  $\alpha$  are arbitrary. Let us study how the presence of the term in  $\epsilon$  affects the solution  $x = C \cos t$ . Since this is expected to be a reasonable first approximation, we try  $x = A \cos \omega t$  where  $\omega$  is a constant that is nearly 1. Substitution in (5.6.1) leads to

$$\begin{aligned} A(1 - \omega^2) \cos \omega t &= -\epsilon(1 - A^2 \cos^2 \omega t)\omega A \sin \omega t \\ &= \epsilon \omega A \left\{ \frac{A^2}{4} \sin 3\omega t - \left( 1 - \frac{1}{4}A^2 \right) \sin \omega t \right\}. \end{aligned}$$

This equation can be satisfied for all  $t$  only if the coefficients of different sinusoidal terms vanish. The term in  $\cos \omega t$  disappears if  $\omega = 1$ , and the term in  $\sin \omega t$  is removed by taking  $A = 2$ . Thus we choose

$$\omega = 1, \quad A = 2.$$

That still leaves the term containing  $\sin 3\omega t$  unaccounted for because we have now fixed  $\omega$  and  $A$ . To get rid of the extra term we need to try a higher approximation such as

$$x = A \cos \omega t + B_1 \cos 2\omega t + C_1 \sin 2\omega t$$

and repeat the process. We shall expect  $\omega$  to be nearly 1,  $A$  to be nearly 2 and  $B_1, C_1$  small. Even then, there will still be extra terms in the equation which we cannot dispose of. To tackle these we could contemplate adding further terms to the expression for  $x$  by introducing sinusoidal functions of  $3\omega t, 4\omega t, \dots$ . However, we shall not go into this complication but stay with our first approximation in the belief that the extra terms will represent a small correction. According to our first approximation, there is a periodic motion

$$x = 2 \cos t$$

in which the amplitude 2 is in error by order  $\epsilon$  and the error in the argument of the cosine is of order  $\epsilon^2$ .

The influence of a small nonlinearity has been radical. Instead of a simple harmonic motion in which the amplitude  $C$  can take any value the periodic motion has been restricted to the single amplitude 2.

Of course, we do not know whether this periodic state can ever be reached and the theory of local stability is no aid because the motion is nowhere near the fixed point at  $x = 0, y = 0$ . Nevertheless, some progress can be made. Multiply (5.6.1) by  $\dot{x}$  and integrate with respect to  $t$  from  $t = \tau$  to  $t = \tau + 2\pi$ . Then

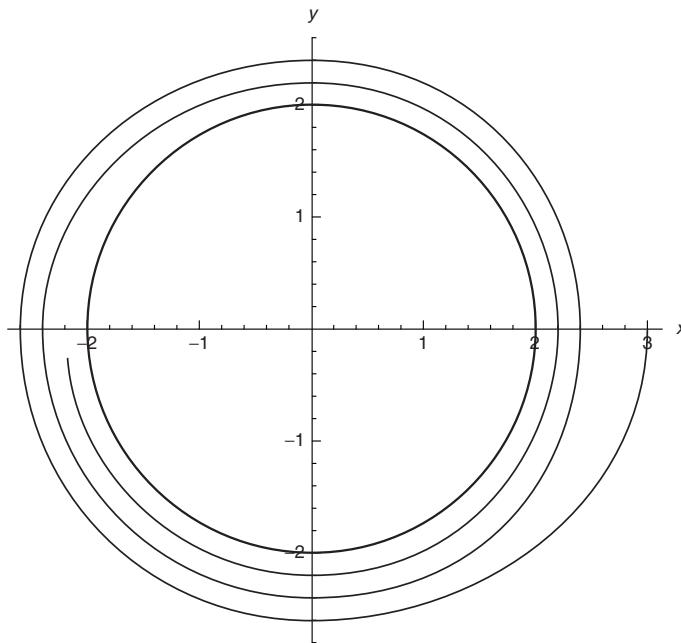
$$\frac{1}{2} [x^2 + y^2]_{\tau}^{\tau+2\pi} = \int_{\tau}^{\tau+2\pi} \epsilon(1 - x^2) \dot{x}^2 dt.$$

Although  $x$  is not known precisely, it can be expected to be substantially of the form  $B \cos t$  for a time interval of  $2\pi$ . The error in calculating the integral by this formula should not be more than order  $\epsilon^2$ . Now

$$\begin{aligned} \int_{\tau}^{\tau+2\pi} \epsilon(1 - x^2) \dot{x}^2 dt &= \int_{\tau}^{\tau+2\pi} \epsilon(1 - B^2 \cos^2 t) B^2 \sin^2 t dt \\ &= \frac{1}{4} \pi \epsilon B^2 (4 - B^2). \end{aligned}$$

Thus, if  $B > 2$ ,  $x^2 + y^2$  is reduced after one period. Repeating the argument for each consecutive period we conclude that if the motion starts with  $B > 2$  it must eventually arrive at the position where  $B = 2$ . Similarly, for a motion that begins with  $B < 2$ ,  $x^2 + y^2$  increases after each period and continues to do so until  $B = 2$ . Consequently, the system ends up in the specified periodic motion whatever the initial point.

A typical trajectory can be seen in Figure 5.6.1. Any closed trajectory that is eventually reached by a system is called a **limit cycle**. For van der Pol's equation  $x^2 + y^2 = 4$  is a limit cycle, according to our first approximation.



**FIGURE 5.6.1:** The limit cycle.

The theory can be arranged to cover the more general

$$\ddot{x} + x = -\epsilon g(x, \dot{x}) \quad (5.6.2)$$

and the corresponding system. Put  $x = A \cos \omega t$  and let

$$g(A \cos \omega t, -A\omega \sin \omega t) = a_1 \cos \omega t + b_1 \sin \omega t + a_2 \cos 2\omega t + b_2 \sin 2\omega t + \dots$$

Taking advantage of the orthogonality of the trigonometric functions and Section 3.7 we have

$$\begin{aligned} \int_0^{2\pi/\omega} g(A \cos \omega t, -A\omega \sin \omega t) \cos \omega t dt &= \pi a_1 / \omega, \\ \int_0^{2\pi/\omega} g(A \cos \omega t, -A\omega \sin \omega t) \sin \omega t dt &= \pi b_1 / \omega. \end{aligned}$$

Since the left-hand side of (5.6.2) is  $A(1 - \omega^2) \cos \omega t$  the coefficients of  $\cos \omega t$  and  $\sin \omega t$  can be made to agree by requiring that

$$1 - \omega^2 = -\epsilon a_1 / A = -\frac{\epsilon \omega}{\pi A} \int_0^{2\pi/\omega} g(A \cos \omega t, -A\omega \sin \omega t) \cos \omega t dt, \quad (5.6.3)$$

$$0 = b_1 \pi / \omega = \int_0^{2\pi/\omega} g(A \cos \omega t, -A\omega \sin \omega t) \sin \omega t dt. \quad (5.6.4)$$

Equation (5.6.3) tells us that  $\omega = 1 + O(\epsilon)$  and, taking advantage of this in (5.6.4), we obtain

$$\int_0^{2\pi} g(A \cos t, -A \sin t) \sin t dt = 0 \quad (5.6.5)$$

to determine  $A$ . Substitution back into (5.6.3) then leads to a more accurate determination of  $\omega$ . Knowing  $\omega$  and  $A$  enables us to locate the positions of any limit cycles.

To establish whether the system tends to a limit cycle we note that when  $x = B \cos t$

$$\frac{1}{2} [x^2 + y^2]_0^{2\pi} = \epsilon B \int_0^{2\pi} g(B \cos t, -B \sin t) \sin t dt. \quad (5.6.6)$$

It is too much to expect to be able to estimate the sign of the right-hand side of (5.6.6) in general, but it is possible to form an opinion of what happens near a limit cycle, i.e., when  $B \approx A$  where  $A$  satisfies (5.6.5). We want the left-hand side of (5.6.6) to decrease if  $B > A$  and to increase when  $B < A$  if the trajectory is to approach the limit cycle from either side, i.e., we need the right-hand side of (5.6.6) to go from positive to negative values as  $B$  increases through  $A$ . This demands that the derivative of the right-hand side of (5.6.6) with respect to  $B$  shall be negative when  $B = A$ , i.e.,

$$\epsilon \int_0^{2\pi} \{g + [g_x]A \cos t - [g_y]A \sin t\} \sin t dt < 0$$

where

$$[g_x] = \left[ \frac{\partial}{\partial x} g(x, y) \right], \quad [g_y] = \left[ \frac{\partial}{\partial y} g(x, y) \right]$$

and the substitution  $x = A \cos t$ ,  $y = -A \sin t$  is made after the derivatives have been performed. The first integrand gives a zero contribution by virtue of (5.6.5) but is retained because, by integration by parts,

$$\begin{aligned} \int_0^{2\pi} g \sin t dt &= \int_0^{2\pi} \frac{\partial g}{\partial t} \cos t dt \\ &= - \int_0^{2\pi} A \cos t \{[g_x] \sin t + [g_y] \cos t\} dt. \end{aligned}$$

Hence the inequality becomes

$$\int_0^{2\pi} [g_y] dt > 0 \quad (5.6.7)$$

since  $\epsilon$  and  $A$  are both positive. The inequality (5.6.7) is the condition under which the system tends to go into the periodic motion of the limit cycle, i.e.,

the limit cycle is **stable**. If the inequality in (5.6.7) is reversed, i.e., the left-hand side is negative, the system would tend to depart from the limit cycle, which would then be **unstable**.

The number of possible limit cycles is fixed by the number of distinct positive values of  $A$  which satisfy (5.6.5). Their stability can be assessed by means of (5.6.7). One can thereby obtain an idea of which limit cycle (if any) will be attained under given initial conditions.

The theory expounded above has been based on  $\epsilon$  being small and is plausible rather than rigorous. Notwithstanding this, it indicates what can happen. To deal with cases when  $\epsilon$  is not small, much more elaborate analysis is necessary and we shall confine ourselves to quoting two theorems.

### LIMIT CYCLE CRITERION

*In (5.6.2) let  $g(x, \dot{x}) = \dot{x}G(x)$  where  $G(x)$  is an even function of  $x$  such that  $G(x) < 0$  for  $|x| < 1$  and  $G(x) > 0$  for  $|x| > 1$ . Suppose further that for some finite  $x_0$  there is  $G_0 > 0$  such that  $G(x) \geq G_0$  for  $|x| \geq x_0$ . Then, for any  $\epsilon > 0$ , (5.6.2) has just one limit cycle and this limit cycle is stable.*

### POINCARÉ-BENDIXSON THEOREM

*If there is a bounded region  $D$  in the  $(x, y)$ -plane such that any solution of the system*

$$\dot{x} = f(x, y), \quad \dot{y} = g(x, y)$$

*that starts in  $D$  remains in  $D$ , then  $D$  contains either a stable fixed point or a limit cycle.*

Note that  $D$  is often referred to a **trapping region**, or an **invariant set**, due to the property that it “traps” trajectories: once a trajectory has entered  $D$ , it can never escape.

The Poincaré–Bendixson theorem deals with a more general situation than the limit cycle criterion but provides less specific information. The reader should not, however, attempt to apply the Poincaré–Bendixson theorem to a system of more than two differential equations.

Limit cycles make their presence felt only with nonlinear differential equations; they do not occur when the governing equations are linear. Caution should therefore be exercised in introducing a linear model for a natural phenomenon where the behaviour is essentially nonlinear. At best it will describe local stability characteristics but it may give no clue as to what takes place globally or, worse, may suggest misleading conclusions.

The Poincaré–Bendixson theorem does not distinguish between the cases of a fixed point and a limit cycle. It is sometimes helpful in deciding between them to observe that *a limit cycle must contain a fixed point*. For example, take the limit cycle as the boundary of  $D$  in the Poincaré–Bendixson theorem. Then there must be a limit cycle or fixed point inside. If there is no limit cycle

the assertion is proved and if there is one we repeat the argument for that one.

**Example 5.6.1**

Use the Poincaré–Bendixson theorem to show that the system

$$\dot{x} = x - y - x^3, \quad (5.6.8)$$

$$\dot{y} = x + y - y^3, \quad (5.6.9)$$

has a stable periodic orbit.

In examples like this one, it is often useful to convert to polar coordinates  $r$  and  $\phi$ , where

$$x = r \cos \phi, \quad y = r \sin \phi. \quad (5.6.10)$$

It can easily be seen geometrically from Figure 5.6.2 that  $r$  and  $\phi$  satisfy

$$r^2 = x^2 + y^2, \quad \tan \phi = \frac{y}{x}. \quad (5.6.11)$$

Note that, in this transformation,  $r$  can never be negative: it represents the length of the radius vector from the origin to the point  $(x, y)$ . Differentiating equations (5.6.11) with respect to  $t$  shows that

$$r\dot{r} = x\dot{x} + y\dot{y}, \quad r^2\dot{\phi} = x\dot{y} - y\dot{x}. \quad (5.6.12)$$

These formulae are very useful for transforming systems of two differential equations to polar coordinates. Substituting equations (5.6.8) and (5.6.9) into (5.6.12) gives

$$\dot{r} = r - r^3 (\cos^4 \phi + \sin^4 \phi), \quad (5.6.13)$$

$$\dot{\phi} = 1. \quad (5.6.14)$$

We immediately see that there is a constant angular velocity ( $\dot{\phi}$ ) of 1, so the radius vector in Figure 5.6.2 sweeps steadily counterclockwise, and there is no possibility of a fixed point other than possibly at the origin. If  $r = 0$  then  $\dot{r} = 0$ , so the system has a unique fixed point at the origin. To check its stability, we look at the Jacobian matrix for original system of differential equations in  $(x, y)$ :

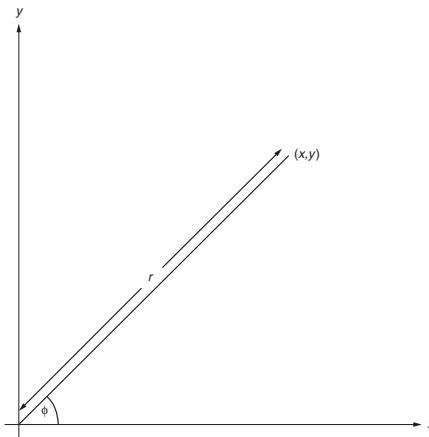
$$\mathbf{J} = \begin{bmatrix} 1 - 3x^2 & -1 \\ 1 & 1 - 3y^2 \end{bmatrix}.$$

At the fixed point  $(x, y) = (0, 0)$ ,

$$\mathbf{J} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

The eigenvalues  $\lambda$  of this matrix satisfy the equation

$$(1 - \lambda)(1 - \lambda) + 1 = 0,$$



**FIGURE 5.6.2:** Polar coordinates.

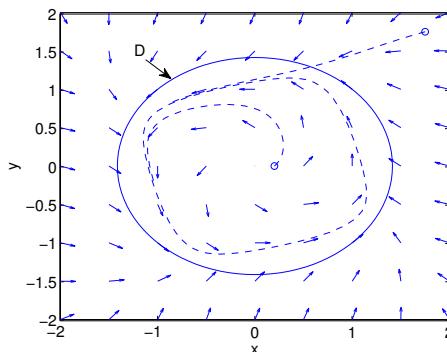
which has roots  $\lambda = 2 \pm 2i$ . These eigenvalues are a complex conjugate pair with positive real part. Hence we are in case 4(b) of Section 5.4 and the fixed point is an unstable focus. We now observe that  $\cos^4 \phi + \sin^4 \phi$  always lies between  $1/2$  and  $1$ . The largest possible value for  $\dot{r}$  in equation (5.6.13) therefore occurs when  $\cos^4 \phi + \sin^4 \phi = 1/2$ , and this gives an upper bound on  $\dot{r}$ :

$$\dot{r} \leq r - \frac{r^3}{2} = r \left(1 - \frac{r^2}{2}\right).$$

Hence we can guarantee that  $\dot{r} < 0$ , regardless of the value of  $\phi$ , provided  $r > \sqrt{2}$ . Therefore, the circle  $D$  of radius  $\sqrt{2}$  is a **trapping region**, meaning that any trajectory that starts within  $D$  will always remain within  $D$  (see Figure 5.6.3). Thus  $D$  satisfies the conditions of the Poincaré–Bendixson theorem and, since we have shown that  $D$  does not contain a stable fixed point (the only fixed point is the one at the origin, which is unstable),  $D$  must contain a limit cycle. □

## 5.7 Forced oscillations

The investigation of the effect of an external oscillatory disturbance on a nonlinear system is very difficult. If the nonlinearity is small some insight can be gained by the method of Section 5.6. To illustrate the technique and to display the new features that can be present we shall consider van der Pol's



**FIGURE 5.6.3:** A trapping region  $D$  for the system (5.6.8), (5.6.9), together with the vector field (arrows indicating the direction of movement in the phase plane) and two sample trajectories (dashed lines). Note that the direction arrows all point into the trapping region. Poincaré–Bendixson theorem predicts the existence of a stable limit cycle in  $D$  and both trajectories can be seen tending towards this limit cycle. See Section 5.8 for details of how to produce plots of this type in MATLAB.

equation with a forcing term, namely

$$\ddot{x} - \epsilon(1 - x^2)\dot{x} + x = E \sin \Omega t \quad (5.7.1)$$

where  $E$  and  $\Omega$  are constants, and  $\epsilon$  is small.

When  $E = 0$ , previous theory indicates a stable limit cycle  $x = 2 \cos t$  and the system can be expected to go into a **self-excited oscillation**. If  $\epsilon = 0$  but  $E \neq 0$ , there is a forced oscillation with harmonic time variation of argument  $\Omega t$  (Section 2.6). When neither  $\epsilon$  nor  $E$  is zero, both types of oscillation have to be allowed for.

Therefore, try

$$x = A \cos(\omega t + \gamma) + C \sin(\Omega t + \theta) \quad (5.7.2)$$

where the constants  $\gamma$  and  $\theta$  have been incorporated to cover the influence of the self-excited and forced oscillations on each other. It is to be expected that  $\omega$  is near 1 and, if  $\Omega^2$  is not near 1, that  $C$  is not far from  $E/(1 - \Omega^2)$  with  $\theta$  small. No estimate of  $A$  and  $\gamma$  can be made at this stage.

The substitution of (5.7.2) in (5.7.1) leads to

$$\begin{aligned}
 & A(1 - \omega^2) \cos(\omega t + \gamma) + C(1 - \Omega^2) \sin(\Omega t + \theta) - E \sin \Omega t \\
 &= \epsilon [\Omega C(1 - \frac{1}{2}A^2 - \frac{1}{4}C^2) \cos(\Omega t + \theta) \\
 &\quad + (\frac{1}{4}A^2 + \frac{1}{2}C^2 - 1)\omega A \sin(\omega t + \gamma) \\
 &\quad - \frac{1}{4}A^2 C(\Omega + 2\omega) \cos\{(\Omega + 2\omega)t + 2\gamma + \theta\} + \frac{1}{4}A^3 \omega \sin 3(\omega t + \gamma) \\
 &\quad + \frac{1}{4}A^2 C(2\omega - \Omega) \cos\{(\Omega - 2\omega)t - 2\gamma + \theta\} + \frac{1}{4}\Omega C^3 \cos 3(\Omega t + \theta) \\
 &\quad - \frac{1}{4}AC^2(\omega + 2\Omega) \sin\{(\omega + 2\Omega)t + \gamma + 2\theta\} \\
 &\quad + \frac{1}{4}AC^2(2\Omega - \omega) \sin\{(\omega - 2\Omega)t + \gamma - 2\theta\}].
 \end{aligned} \tag{5.7.3}$$

All the sinusoidal terms on the right-hand side of (5.7.3) have different time variations unless  $\Omega = 0, \frac{1}{3}\omega, \omega$  or  $3\omega$ . The case  $\Omega = 0$  has already been disposed of (being the same as  $E = 0$ ) and, for the moment, the exceptional cases in which  $\Omega = \frac{1}{3}\omega, \omega$  or  $3\omega$  will be ignored. On the two sides of (5.7.3) equate the coefficients of  $\cos(\omega t + \gamma)$ ,  $\sin(\omega t + \gamma)$ ,  $\cos \Omega t$  and  $\sin \Omega t$  to obtain

$$A(1 - \omega^2) = 0, \tag{5.7.4}$$

$$(\frac{1}{4}A^2 + \frac{1}{2}C^2 - 1)\omega A = 0, \tag{5.7.5}$$

$$C(1 - \Omega^2) \sin \theta - \epsilon \Omega C(1 - \frac{1}{2}A^2 - \frac{1}{4}C^2) \cos \theta = 0, \tag{5.7.6}$$

$$C(1 - \Omega^2) \cos \theta + \epsilon \Omega C(1 - \frac{1}{2}A^2 - \frac{1}{4}C^2) \sin \theta = E. \tag{5.7.7}$$

It is evident from (5.7.6) and (5.7.7) that

$$C = E \cos \theta / (1 - \Omega^2), \quad \tan \theta = \epsilon \Omega (1 - \frac{1}{2}A^2 - \frac{1}{4}C^2) / (1 - \Omega^2).$$

Thus, so long as  $\Omega^2$  is not near 1,  $\theta$  is small and  $\cos \theta$  may be replaced by unity. In other words, *the nonlinearity has very little effect on the forced oscillation*. On the other hand, (5.7.4) and (5.7.5) imply that

$$\omega = 1, \quad A = (4 - 2C^2)^{1/2}.$$

As  $C^2$  increases from 0,  $A$  decreases from 2 to 0 and finally becomes imaginary for  $C^2 > 2$ . Imaginary values of  $A$  are not permitted and so, when  $C^2 > 2$ , the only possible solution to (5.7.4) and (5.7.5) is  $A = 0$ . The influence of the forcing term, therefore, is to reduce the amplitude of the self-excited oscillation when  $C^2 < 2$  and *to extinguish it completely* when  $C^2 \geq 2$ , i.e.,  $E^2 \geq 2(1 - \Omega^2)^2$ .

Thus, even in the unexceptional case, a strong enough external vibration can obliterate totally the self-excited oscillation.

Turn now to the exceptional case in which  $\Omega = 3\omega$ . Since  $\omega$  is near 1 this can occur only when  $\Omega = 3 + \delta$  and  $\omega = 1 + \frac{1}{3}\delta$  with  $|\delta| \ll 1$ . The terms involving  $\sin 3(\omega t + \gamma)$  and  $\cos\{(\Omega - 2\omega)t - 2\gamma + \theta\}$  in (5.7.3) cannot now be neglected. Equate the coefficients of  $\cos(\omega t + \gamma)$ ,  $\sin(\omega t + \gamma)$ ,  $\sin 3\omega t$  and

$\cos 3\omega t$  on the two sides of (5.7.3); then

$$A(1 - \omega^2) + \frac{1}{4}\epsilon A^2 C \omega \cos(3\gamma - \theta) = 0, \quad (5.7.8)$$

$$\left(\frac{1}{4}A^2 + \frac{1}{2}C^2 - 1\right)\omega A - \frac{1}{4}A^2 C \omega \sin(3\gamma - \theta) = 0, \quad (5.7.9)$$

$$C(1 - 9\omega^2) \cos \theta - E + 3\epsilon \omega C \left(1 - \frac{1}{2}A^2 - \frac{1}{4}C^2\right) \sin \theta = \frac{1}{4}\epsilon A^3 \omega \cos 3\gamma, \quad (5.7.10)$$

$$C(1 - 9\omega^2) \sin \theta - 3\epsilon \omega C \left(1 - \frac{1}{2}A^2 - \frac{1}{4}C^2\right) \cos \theta = \frac{1}{4}\epsilon A^3 \omega \sin 3\gamma. \quad (5.7.11)$$

From (5.7.10) and (5.7.11),  $C = -E/8$  to the first order and  $\tan \theta$  is small so that the forced oscillation is virtually the same as in the absence of the nonlinearity. In (5.7.8) and (5.7.9) put  $\omega = 1 + \frac{1}{3}\delta$ , put  $A \cos(3\gamma - \theta) = \xi$  and put  $A \sin(3\gamma - \theta) = \eta$ . Then, if only dominant terms are retained,

$$\begin{aligned} -\frac{2}{3}\delta + \frac{1}{4}\epsilon C\xi &= 0, \\ \xi^2 + (\eta - \frac{1}{2}C)^2 &= 4 - 7C^2/4. \end{aligned}$$

Consequently,  $\xi$  and  $\eta$  (and thereby  $A$  and  $\gamma$ ) are determined by the intersection of a straight line and a circle. The circle is imaginary if  $C^2 > 16/7$  and the line does not intersect the circle if  $4 - 7C^2/4 < (8\delta/3\epsilon C)^2$ . Therefore  $A$  is nonzero only if  $4 - 7C^2/4 > (8\delta/3\epsilon C)^2$ ; this inequality cannot be satisfied unless  $(\delta/\epsilon)^2 < 9/28$ . Hence, if  $(\delta/\epsilon)^2 < 9/28$  and  $4 - 7C^2/4 > (8\delta/3\epsilon C)^2$ ,

$$A^2 = 4 - \frac{3}{2}C^2 \pm \{4C^2 - \frac{7}{4}C^4 - (8\delta/3\epsilon)^2\}^{1/2}. \quad (5.7.12)$$

Having found  $A$ , we can proceed to determine  $\gamma$ . However,  $\gamma$  occurs only in the form  $3\gamma$  so that if  $\gamma_0$  is a possible value so are  $\gamma_0 + 2\pi/3$  and  $\gamma_0 + 4\pi/3$ . Thus, to a given amplitude of the self-excited oscillation there correspond three possible distinct phases.

According to this approximation when the magnitude of the forcing term is small,  $C^2$  is not large enough for  $A$  to exist. Thus application of a small forcing term of three times the natural frequency extinguishes the self-excited oscillation. As the magnitude of the forcing term and  $C^2$  grow there comes a point where  $A$  is nonzero and self-excited oscillations can occur. Further increase of  $C^2$  will eventually reach a point when  $A$  disappears again so that the self-excited oscillation is absent when the magnitude of the forcing term is large enough. For the upper sign in (5.7.12),  $A$  will increase with  $C^2$  to a maximum before dropping back to zero. The possibility of  $A$  increasing and of surpassing the value 2 are new aspects. For instance,  $A \geq 2$  if  $0 \leq C^2 \leq 1$  when  $\delta = 0$ . In general,  $A \geq 2$  if  $(8\delta/3\epsilon)^2 < 1$  and

$$-\{1 - (8\delta/3\epsilon)^2\}^{1/2} \leq 2C^2 - 1 \leq \{1 - (8\delta/3\epsilon)^2\}^{1/2}.$$

This magnification of the self-excited oscillation by the application of a vibration, which is an integer multiple of the self-excited, is known as **subharmonic resonance**.

Our investigation does not permit us to say which of the many oscillations that have been uncovered can be reached by the system starting from given

initial conditions. A more refined analysis reveals that the negative square root in (5.7.12) corresponds to an unstable state, whereas the upper sign provides a stable state. Therefore, under conditions in which  $A$  is nonzero, the system tends to adopt the larger value of  $A$ ; the phase will depend on the initial conditions.

The exceptional case  $\Omega = \frac{1}{3}\omega$  may be discussed in a similar manner. As regards the special case  $\Omega = \omega$ , there is no necessity to have both terms in (5.7.2) and  $C$  can be placed equal to zero.

---

## 5.8 Numerical solution of systems of equations

The numerical methods for solving initial value problems consisting of systems of differential equations of the form

$$\dot{y}_i = f_i(y_1, \dots, y_n), \quad i = 1, \dots, n \quad (5.8.1)$$

are essentially the same as those for solving initial value problems with a single first-order equation (see Section 1.7). The only difference is that the dependent variable  $\mathbf{y} = (y_1, \dots, y_n)$  is a vector rather than a scalar and, to MATLAB, this makes very little difference. If using the built-in solver `ode45`, the only change from the usage described in Section 1.7 is that the user-defined function `myrhs` now takes a vector  $\mathbf{y}$  as its second input and returns a vector of values for  $\dot{\mathbf{y}}$  as output, and that the initial condition is also a vector.

For example, suppose we wish to solve the general Volterra–Lotka competition model (4.5.12) and (4.5.13). First note that this system may be written in vector form  $\dot{\mathbf{y}} = \mathbf{f}(t, \mathbf{y})$  (see Section 3.5 for vector–matrix notation), where  $\mathbf{f}$  is defined by

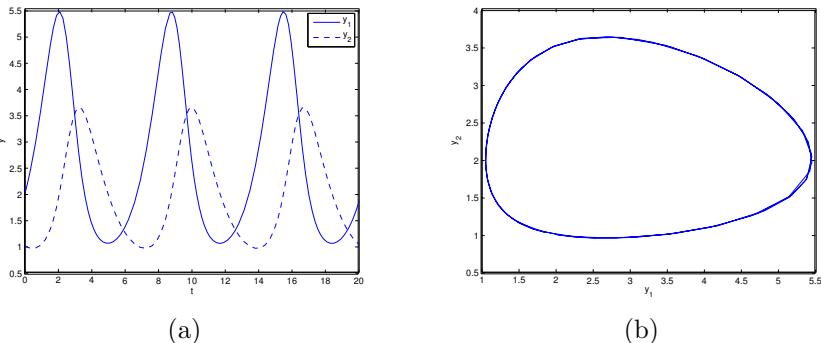
$$\mathbf{f}(t, \mathbf{y}) = \begin{bmatrix} y_1(a_0 - a_1y_1 - a_2y_2) \\ y_2(b_0 - b_1y_1 - b_2y_2) \end{bmatrix}, \quad (5.8.2)$$

and the  $a_i$  and  $b_i$  are constants. Picking some example values for the constants, we can write a MATLAB function to evaluate the vector function  $\mathbf{f}$  as follows:

```
function f = volterra_lotka(t, y)
a0 = 1.2; a1 = 0; a2 = 0.6; b0 = -0.8; b1 = -0.3; b2 = 0;
f = zeros(2, 1); % define f as a 2x1 matrix
f(1) = y(1)*(a0-a1*y(1)-a2*y(2));
f(2) = y(2)*(b0-b1*y(1)-b2*y(2));
```

Note the syntax for defining the column vector  $\mathbf{f}$ : first  $\mathbf{f}$  is defined as a  $2 \times 1$  matrix (i.e., a column vector of length 2); then each component is calculated in turn.

The initial value problem with initial conditions of, for example,  $\mathbf{y}(0) = [2, 1]$  may then be solved and plotted, up to say  $t = 20$ , with the following commands:



**FIGURE 5.8.1:** Numerical solution to the Volterra–Lotka system (5.8.2) with  $a_0 = 1.2$ ,  $a_1 = 0$ ,  $a_2 = 0.6$ ,  $b_0 = -0.8$ ,  $b_1 = -0.3$ ,  $b_2 = 0$ , calculated using `ode45`. (a) Time-series of both species ( $y_1$  and  $y_2$  against  $t$ ). (b) Phase diagram ( $y_2$  against  $y_1$ ).

```
>> t0 = 0; t1 = 20; y0 = [2; 1];
>> [t, Y] = ode45(@volterra_lotka, [t0, t1], y0);
>> plot(t, Y)
```

Note that, since the solution  $\mathbf{y}(t)$  consists of a vector of 2 values,  $\mathbf{Y}$  is a matrix with 2 columns. The  $k^{\text{th}}$  row of  $\mathbf{Y}$  contains the solution vector corresponding to the  $k^{\text{th}}$  value of  $t$ . The above plot command graphs each component of the solution (i.e., the total population of each species) against  $t$  as a separate curve, and is equivalent (for the two-variable case) to the command

```
>> plot(t, Y(:, 1), t, Y(:, 2))
```

The resulting plot is shown in Figure 5.8.1(a), which reveals the cyclical nature of the solution.

It is often useful to plot the solution to a system of differential equations in the phase plane, i.e., plot  $y_2$  against  $y_1$ , instead of plotting each dependent variable against  $t$ . This can be done with the command

```
>> plot(Y(:, 1), Y(:, 2))
```

which produces the graph shown in Figure 5.8.1(b). The key observation here is that the trajectory in the phase plane is a closed curve, which implies that, after some amount of time, the system returns to its previous state. This demonstrates that the solution is periodic.

Often, it is desirable to plot not just a single trajectory corresponding to a particular initial condition in the phase plane, but to produce a plot that is representative of the entire phase portrait. One way to achieve this is to

plot the **vector field** that defines the system of differential equations. The MATLAB command for plotting vector fields is **quiver**. Before using this command, we must define the grid we are working on. This can be done with the following commands:

```
>> y1 = y1_min:h:y1_max; y2 = y2_min:h:y2_max;
>> [Y1, Y2] = meshgrid(y1, y2);
```

Here the **meshgrid** command takes the vectors  $y_1$  and  $y_2$  and creates matrices  $Y_1$  and  $Y_2$  corresponding the  $y_1$  and  $y_2$  coordinates throughout the rectangular region. The minimum and maximum values of  $y_1$  and  $y_2$ , and the plotting resolution  $h$  can be adjusted as required. Now we can define and plot the vector field. For the example given in equation (5.8.2), the following commands produce the graph shown in Figure 5.8.2 (compare this graph to the particular solution shown in Figure 5.8.1(b)):

```
>> F1 = Y1.* (a0-a1*Y1-a2*Y2);
>> F2 = Y2.* (b0-b1*Y1-b2*Y2);
>> quiver(Y1, Y2, F1, F2)
```

Note the use of the **array multiplication operator**  $\cdot*$ . This takes two arrays of the same size and multiplies them component-by-component, as opposed to  $*$ , which will attempt to calculate the matrix product of two arrays. For example,

$$\begin{aligned} [a, b] \cdot* [c, d] &= [a*c, b*d] \\ [a, b] * [c; d] &= a*c+b*d \end{aligned}$$

Similar operations can be achieved with  $./$  for component-by-component division, and  $.^$  for raising components to a power.

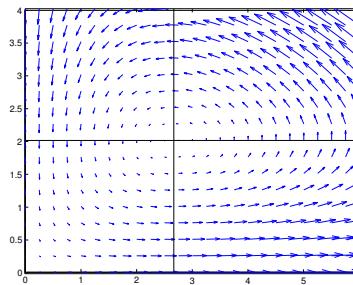
It is also useful to add the **nullclines** of the system to the graph in Figure 5.8.2. These are the curves on which one of the variables is stationary. Hence the  $y_1$ -nullcline is defined by the equation  $y_1 = 0$  and likewise for the  $y_2$ -nullcline. For this example, the nullclines are simply given by

$$y_2 = (a_0 - a_1 y_1)/a_2 \quad \text{and} \quad y_1 = (b_0 - b_2 y_2)/b_1.$$

These may be added to the graph with the following commands:

```
>> hold on
>> nullcline1 = (a0-a1*y1)/a2; nullcline2 = (b0-b2*y2)/b1;
>> plot(y1, nullcline1, nullcline2, y2)
```

(The command **hold on** causes the following **plot** command to be superimposed on the existing Figure.) Notice that the vector field has no  $y_1$  (horizontal) component on the  $y_1$ -nullcline, and no  $y_2$  (vertical) component on the  $y_2$ -nullcline. In this particular example, the nullclines are straight lines, either vertical or horizontal. However, in general the nullclines will be a pair of



**FIGURE 5.8.2:** Vector field plot for the Volterra–Lotka system (5.8.2) with  $a_0 = 1.2$ ,  $a_1 = 0$ ,  $a_2 = 0.6$ ,  $b_0 = -0.8$ ,  $b_1 = -0.3$ ,  $b_2 = 0$ . The solid lines are the nullclines for the system.

curves and are very useful for building a picture of the phase portrait. Points where the nullclines intersect are fixed points (equilibria) of the system.

### Example 5.8.1

Plot solutions to van der Pol's equation (5.6.1) (with  $\epsilon = 0.25$ ) in the phase plane for two initial conditions: (i)  $x(0) = 3$ ,  $\dot{x}(0) = 0$ ; (ii)  $x(0) = 1$ ,  $\dot{x}(0) = 0$ . Draw the corresponding vector field plot, including the nullclines of the system.

First we write (5.6.1) in the form of a first-order system:

$$\dot{y}_1 = y_2, \quad \dot{y}_2 = \epsilon(1 - y_1^2)y_2 - y_1.$$

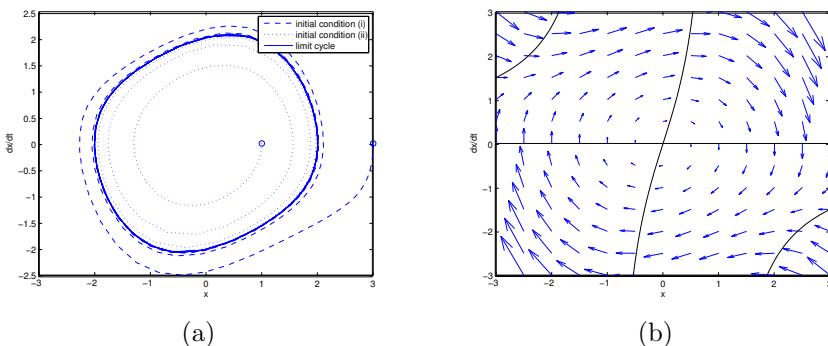
From this, we can write a MATLAB function that calculates the right-hand sides of the differential equations:

```
function f = van_der_pol(t, y)
epsilon = 0.25;
f = [ y(2) ; epsilon*(1-y(1)^2)*y(2)-y(1) ];
```

We can calculate the solution for say  $0 \leq t \leq 25$  and initial condition (i), and plot in the phase plane with the commands:

```
>> [t, Y] = ode45(@van_der_pol, [0, 10], [3; 0]);
>> plot(Y(:, 1), Y(:, 2))
```

Including the other initial condition gives the two trajectories shown in Figure 5.8.3(a). This graph clearly shows the existence of a limit cycle, which both trajectories tend towards. The limit cycle itself can be calculated and added to the graph by plotting the solution after a long period of time has elapsed (say  $500 \leq t \leq 510$ ).



**FIGURE 5.8.3:** van der Pol's equation (5.6.1) with  $\epsilon = 0.25$ : (a) numerical solutions in the phase plane for two different initial conditions (indicated by open circles); (b) vector field plot (solid lines are nullclines).

To plot the vector field, we define a grid  $[Y_1, Y_2]$  as described above, and then draw the graph:

```
>> F1 = Y2; F2 = epsilon*(1-Y1.^2).*Y2-Y1;
>> quiver(Y1, Y2, F1, F2)
```

The nullclines for the system are given by

$$y_2 = 0 \quad \text{and} \quad y_2 = y_1 / (\epsilon(1 - y_1^2)),$$

which can be added to the vector field plot (Figure 5.8.3(b)) as follows:

```
>> hold on
>> nullcline1 = 0*y1;  nullcline2 = y1./(epsilon*(1-y1.^2));
>> plot(y1, nullcline1, y1, nullcline2)
```

Notice that the  $y_2$ -nullcline has vertical asymptotes at  $y_1 = \pm 1$ .

## 5.9 Symbolic computation on first-order systems of equations and higher-order equations

MATLAB's `dsolve` command, introduced in Section 1.8, can handle systems of equations and higher-order equations. This is best illustrated by example. It is worth remembering that exact analytical solutions to nonlinear systems are rarely available, and the following examples are therefore linear.

**Example 5.9.1**

Consider the linear system

$$\dot{y}_1 = -y_1 - 10y_2, \quad (5.9.1)$$

$$\dot{y}_2 = 10y_1 - y_2. \quad (5.9.2)$$

The appropriate MATLAB input/output is

```
>> y_sol = dsolve('Dy1=-y1-10*y2', 'Dy2=10*y1-y2')
y_sol =
y1: [1x1 sym]
y2: [1x1 sym]
```

Here, because there is more than one dependent variable, the output from `dsolve` (`y_sol`) is what is called a **structure**. In this case, there are two dependent variables, so the structure `y_sol` has two **fields** `y1` and `y2`. The syntax for accessing the fields of a structure is `structure.field`. Hence to extract the contents of the two fields to two new variables `y1_sol` and `y2_sol`, the appropriate commands are

```
>> y1_sol = y_sol.y1
y1_sol =
exp(-t)*(C1*cos(10*t)-C2*sin(10*t))
>> y2_sol = y_sol.y2
y2_sol =
exp(-t)*(C1*sin(10*t)+C2*cos(10*t))
```

We can now plot these solutions, with initial conditions of say  $y_1(0) = 1$  and  $y_2(0) = 0$ , following the same procedure as in Example 1.8.1. The time series and phase plane plot are shown in Figure 5.9.1.

**Example 5.9.2**

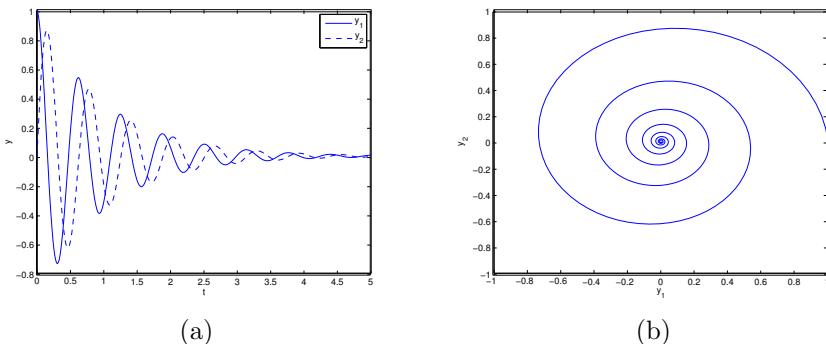
**Bessel's equation** is an equation that frequently arises when solving steady-state temperature distribution problems in cylindrical coordinates ( $t$  here represents location rather than time):

$$t^2 \frac{d^2y}{dt^2} + t \frac{dy}{dt} + (t^2 - v^2)y = 0. \quad (5.9.3)$$

We could write this as a system of first-order equations (see Section 3.2). Alternatively, we can try to solve it directly. The command we need for this is

```
>> y_sol = dsolve('t^2*D2y+t*Dy+(t^2-v^2)*y=0')
```

which returns the solution as



**FIGURE 5.9.1:** Solution to the linear system (5.9.1) and (5.9.2), with initial conditions  $y_1(0) = 1$  and  $y_2(0) = 0$ , calculated using symbolic computation. (a) Time-series of both variables ( $y_1$  and  $y_2$  against  $t$ ). (b) Phase diagram ( $y_2$  against  $y_1$ ).

```
y_sol = C1*besselj(v,t)+C2*bessely(v,t)
```

Here `besselj` and `bessely` are special functions called **Bessel functions**. Note that, because equation (5.9.3) is a second-order differential equation, there are two linearly independent solutions (with arbitrary constants  $C_1$  and  $C_2$ ) in the general solution. MATLAB will happily plot these functions (see Section 1.8), if given a value of  $v$  and suitable initial conditions. For a second-order differential equation, we need two initial conditions. For example, the solution with  $v = 1$  and initial conditions  $y(0) = 0$ ,  $y'(0) = 1$  may be found by

```
>> y_sol = dsolve('t^2*D2y+t*Dy+(t^2-1)*y=0', 'y(0)=0',
                  'Dy(0)=1')
y_sol = 2*besselj(1,t)
```

□

## 5.10 Numerical solution of nonlinear boundary value problems

Although linear boundary value problems were dealt with in Chapter 3, nonlinear boundary value problems are more difficult to solve as they entail

solving a system of  $n + 1$  nonlinear equations. In general, this is a nontrivial task, and root-finding methods for nonlinear systems, such as Newton's method, usually require a reasonable approximation of the solution to use as a starting point.

Newton's method for a system of equations

$$\begin{aligned} f_1(x_1, \dots, x_n) &= 0, \\ &\vdots \\ f_n(x_1, \dots, x_n) &= 0, \end{aligned}$$

is based on the multivariable Taylor expansion of the vector function  $\mathbf{f}(\mathbf{x}) = (f_1, \dots, f_n)$ :

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{J}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + O\left((\mathbf{x} - \mathbf{x}_0)^2\right), \quad (5.10.1)$$

where  $\mathbf{J}$  is the Jacobian matrix of the vector function  $\mathbf{f}$  (see Section 5.3), i.e., the  $(i, j)$  entry of  $\mathbf{J}$  is the partial derivative of  $f_i$  with respect to  $x_j$ :

$$J_{ij} = \frac{\partial f_i}{\partial x_j}.$$

Setting  $\mathbf{x}$  in (5.10.1) to be the root of the vector equation (i.e.,  $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ ) and neglecting terms of order  $(\mathbf{x} - \mathbf{x}_0)^2$  and higher gives

$$\mathbf{J}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) = -\mathbf{f}(\mathbf{x}_0).$$

Given an initial approximation  $\mathbf{x}_0$  to the root, this equation enables an improved estimate for the root  $\mathbf{x}$  to be calculated. Applying this procedure iteratively gives:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \Delta\mathbf{x}, \quad \text{where } \mathbf{J}(\mathbf{x}_n)\Delta\mathbf{x} = -\mathbf{f}(\mathbf{x}_n).$$

The following code is a MATLAB function that attempts to solve a vector equation  $\mathbf{f}(\mathbf{x}) = \mathbf{0}$  by Newton's method.

```
function x = newton(func, x, tol, maxits)
% Function to solve a system of equations f(x)=0 by Newton's
% method
% INPUTS: x - initial approximation to the solution
%          tol - relative tolerance for testing convergence
%          func - handle to a function that calculates f(x)
%          maxits - maximum number of iterations to carry out
% OUTPUTS: x - solution vector
delta_x = 10*tol*x;
count = 0;
while ~norm(delta_x, inf)/norm(x, inf) < tol && count < maxits
```

```
[f, J] = jacobian(func, x);
delta_x = -J\f;
x = x+delta_x;
count = count+1;
end
```

This function takes an initial approximation to the solution, and attempts to improve the accuracy of this approximation by Newton's method, terminating if the relative change in the approximations between successive iterations falls below some specified tolerance `tol`. The function `jacobian` calculates the Jacobian matrix for the user-supplied vector function  $\mathbf{f}(\mathbf{x})$ :

```
function [f0, J] = jacobian(func, x)
% Function to evaluate a vector function and find its Jacobian
% INPUTS: func - handle to the vector function
%          x - argument to vector function
% OUTPUTS: f0 - value of the vector function at x
%          J - Jacobian matrix at x
N = length(x); % length of the vector x
dx = 1e-10; % value of dx to use to calculate the Jacobian
f0 = func(x); % evaluate func at x
J = zeros(N); % initialise J as an NxN matrix
for I = 1:N % loop through columns of J
    x0 = x(I);
    x(I) = x(I)+dx; % perturb x(I) by dx
    f = func(x); % evaluate func at perturbed x
    J(:, I) = (f-f0)../dx; % calculate column I of Jacobian
    x(I) = x0; % restore x(I) to unperturbed value
end
```

The user-supplied function that calculates  $\mathbf{f}(\mathbf{x})$ , which is passed to `newton` in the argument `func`, depends on the specific boundary value problem. This is best illustrated by example.

### **Example 5.10.1**

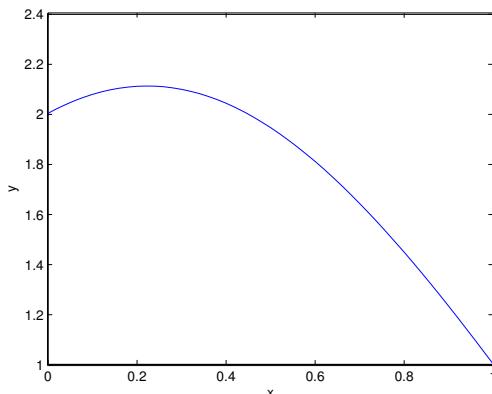
Write MATLAB code to solve the nonlinear boundary value problem

$$\frac{d^2y}{dx^2} + py^2 + q = 0, \quad y(0) = \alpha, \quad y(1) = \beta. \quad (5.10.2)$$

As with a linear problem, the first thing we need to do is to replace derivatives of  $y$  with appropriate finite difference approximations:

$$\frac{y_{i+1} + y_{i-1} - 2y_i}{h^2} + py_i^2 + q = 0.$$

By setting  $i = 1, \dots, n - 1$ , this gives  $n - 1$  equations. The two boundary conditions make a system of  $n + 1$  equations in the  $n + 1$  unknowns  $y_0, \dots, y_n$ .



**FIGURE 5.10.1:** Numerical solution of the boundary value problem (5.10.2) with  $\alpha = 2$ ,  $\beta = 1$ ,  $p = 1$  and  $q = 0$ .

This system can be written in vector form

$$\mathbf{f}(\mathbf{y}) = \mathbf{0}, \quad (5.10.3)$$

where  $\mathbf{y} = (y_0, \dots, y_n)$  and  $\mathbf{f}$  is a vector function of  $\mathbf{y}$ . All we need now is a MATLAB function that calculates  $\mathbf{f}(\mathbf{y})$  for a given vector  $\mathbf{y}$ , and then we can use the Newton's method function to find the solution to the system of equations (5.10.3). At this point, we need to specify values of the boundary conditions  $\alpha$  and  $\beta$  and the coefficients  $p$  and  $q$ . Let's choose  $\alpha = 2$ ,  $\beta = 1$ ,  $p = 1$  and  $q = 0$ . We'll call the function `my_bvp`.

```
function f = my_bvp(y)
global x
% GLOBALS: x - a vector of N+1 equally spaced x values
% INPUTS: y - a vector of N+1 corresponding y values
% OUTPUTS: f - a vector containing the right-hand sides of the
%             system of N+1 equations in N+1 unknowns
alpha = 2; beta = 1; p = 1; q = 0; % define the constants
N = length(x)-1; % number of lattice points
dx = (x(N+1)-x(1))/N; % spacing between lattice points
f = zeros(N+1, 1); % initialise f as a column vector of size N+1
f(1) = y(1)-alpha; % first boundary condition
f(N+1) = y(N+1)-beta; % second boundary condition
% loop through the N-1 equations derived from the DE:
for I = 2:N
    f(I) = (y(I+1)+y(I-1)-2*y(I))/dx^2 + p*y(I)^2 + q;
end
```

Note that the vector of  $x$  values, rather than being passed to this function as an argument, has to be defined as a **global variable**. This means that it is accessible to all MATLAB functions (rather than being local to the function in which it is declared). Hence, we need to define  $\mathbf{x}$  as a global variable in the command window, and set it up as the vector of  $x$  values we are interested in, with a step size of 0.02 :

```
>> global x
>> x = (0:0.02:1)';
```

(The ' transposes  $\mathbf{x}$  to give a column vector rather than a row vector.) Remember, we need an initial approximation to the solution for  $y$ . Let's just try a linear function to see if that works. The linear function that satisfies the boundary conditions  $y(0) = 2$  and  $y(1) = 1$  is  $y = 2 - x$ , which we can apply with the command `>> y = 2-x;`. Now we can call the Newton's method function `newton` to obtain the solution (with a tolerance of  $10^{-4}$ ), and plot  $y$  against  $x$ :

```
>> y = newton(@my_bvp, y, 1e-4, 100);
>> plot(x, y)
```

This produces the graph shown in Figure 5.10.1.

□

The above example is relatively straightforward because the nonlinearity is confined to the dependent variable. If there are nonlinearities in the derivatives of the dependent variables, the problem become more difficult. Nevertheless, the code in the function `my_bvp` is easily adaptable to more complicated equations, although a better initial approximation to the solution may be required.

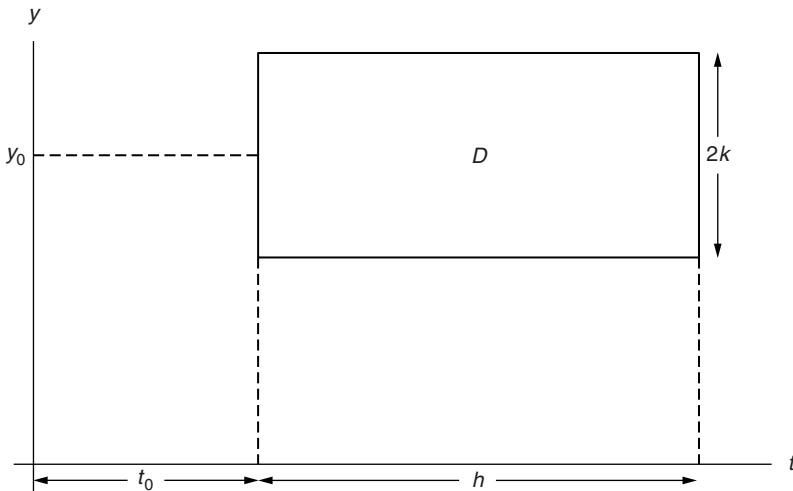
## 5.11 Appendix: existence theory

### 5.11.1 Single first-order equation

The aim of existence theory is to specify conditions under which one can be sure that there is a solution to a differential equation such as

$$\dot{y} = f(t, y). \quad (5.11.1)$$

There is no point in wasting analytical and computational effort on trying to find a solution when there is not one. Basically, there are two ways of demonstrating existence, **nonconstructive** in which no attempt is made to show how one might arrive at a solution, and **constructive** in which a method for building up the solution is described. We shall consider only a constructive



**FIGURE 5.11.1:** The domain of existence.

approach, one that lays the foundation for a numerical attack when that is desired.

The initial value problem for (5.11.1) seeks a solution such that  $y = y_0$  at  $t = t_0$  and this is the problem that we shall discuss in some detail when  $y_0$  is a prescribed constant.

The purpose of the analysis is to show that, under specified conditions, when  $t$  does not stray too far from  $t_0$  there is a solution  $y$  which does not differ from  $y_0$  by more than a certain amount. So we consider what happens as  $t$  ranges from  $t_0$  to  $t_0 + h$  where  $h$  is positive (similar considerations apply when  $h$  is negative). In this range we are prepared to consider deviations of  $y$  from  $y_0$  of magnitude  $k$ , i.e., we expect  $y$  to lie between  $y_0 - k$  and  $y_0 + k$ . The points that originate values of  $f(t, y)$  are then in the domain  $D$  of Figure 5.11.1. Suppose that  $f$  is bounded in  $D$ , say  $|f| < M$ ; then we shall impose the restriction  $h < k/M$ . This can always be arranged by reducing  $h$  if necessary. The constraint is an expression of our expectation that the more  $t$  departs from  $t_0$  the more  $y$  will deviate from  $y_0$ . Further conditions are placed on  $f$  in the following theorem.

### EXISTENCE THEOREM I

Let  $f(t, y)$  be a single-valued continuous function of  $t$  and  $y$  in  $D$  such that

- (a)  $|f(t, y)| < M$  in  $D$ ,
- (b) **(Lipschitz condition)**  

$$|f(t, y) - f(t, y')| < K|y - y'|,$$

$K$  being a finite constant, for any pair of points  $(t, y)$  and  $(t, y')$  in  $D$ . Then,

for  $h < k/M$ , the differential equation (5.11.1) possesses one and only one continuous solution  $y(t)$  in  $t_0 \leq t \leq t_0 + h$  such that  $y(t_0) = y_0$ .

**PROOF** The proof starts by observing that the problem is equivalent to showing that

$$y(t) = y_0 + \int_{t_0}^t f(u, y(u))du \quad (5.11.2)$$

has a solution. For, since  $f$  is bounded, the integral exists and tends to zero as  $t \rightarrow t_0$  with the consequence that  $y(t_0) = y_0$ . Also, a derivative with respect to  $t$  of (5.11.2) returns, because of the assumed continuity of  $f$ , to (5.11.1). So it is sufficient to discuss (5.11.2).

Now solve (5.11.2) by iteration by making a series of approximations. First put  $y(u) = y_0$  in the integral to generate  $y_1(t)$  given by

$$y_1(t) = y_0 + \int_{t_0}^t f(u, y_0)du. \quad (5.11.3)$$

Now produce the sequence  $y_n(t)$  defined by

$$y_n(t) = y_0 + \int_{t_0}^t f(u, y_{n-1}(u))du. \quad (5.11.4)$$

In each approximation the right-hand side can be calculated and a practical mechanism of solution has been erected, provided that the iteration converges to a solution.

Note firstly that (5.11.3) implies that  $y_1(t)$  is a continuous function of  $t$  and, since  $(u, y_0)$  is in  $D$  for  $t < t_0 + h$ ,

$$|y_1(t) - y_0| < M(t - t_0) < Mh < k;$$

thus  $(t, y_1(t))$  is in  $D$  for  $t < t_0 + h$ . The reasoning can now be repeated to show that  $y_2(t)$  is a continuous function of  $t$  such that  $(t, y_2(t))$  is in  $D$  for  $t < t_0 + h$ . It is then clear from (5.11.4) that  $y_n(t)$  is a continuous function of  $t$  such that  $(t, y_n(t))$  is in  $D$  for every  $n$  while  $t < t_0 + h$ .

Suppose now that

$$|y_n(t) - y_{n-1}(t)| < MK^{n-1}(t - t_0)^n/n! \quad (5.11.5)$$

for  $t_0 \leq t \leq t_0 + h$ , a result already proved for  $n = 1$ . Then, from (5.11.4),

$$\begin{aligned} |y_{n+1}(t) - y_n(t)| &\leq \int_{t_0}^t |f(u, y_n(u)) - f(u, y_{n-1}(u))|du \\ &< \int_{t_0}^t K|y_n(u) - y_{n-1}(u)|du \end{aligned}$$

by the Lipschitz condition. Invoking our hypothesis, we have

$$\begin{aligned}|y_{n+1}(t) - y_n(t)| &< MK^n \int_{t_0}^t (u - t_0)^n du / n! \\ &< MK^n (t - t_0)^{n+1} / (n + 1)!\end{aligned}$$

which is the same as (5.11.5) except that  $n$  is replaced by  $n + 1$ . Since it is true for  $n = 1$  it follows by induction that (5.11.5) holds for every  $n$ .

Accordingly

$$\begin{aligned}\left| y_0 + \sum_{r=1}^n \{y_r(t) - y_{r-1}(t)\} \right| &\leq |y_0| + \sum_{r=1}^{\infty} MK^{r-1} (t - t_0)^r / r! \\ &\leq |y_0| + Me^{K(t-t_0)} / K \\ &\leq |y_0| + Me^{Kh} / K\end{aligned}$$

which reveals that the series on the left is absolutely and uniformly convergent (by the Weierstrass  $M$ -test) in  $t_0 \leq t \leq t_0 + h$ . But the sum of a uniformly convergent series of continuous functions is itself continuous and, since

$$y_0 + \sum_{r=1}^{\infty} \{y_r(t) - y_{r-1}(t)\} = y_n(t),$$

it follows that  $\lim_{n \rightarrow \infty} y_n(t)$  exists and is a continuous function  $\hat{y}(t)$  in  $t_0 \leq t \leq t_0 + h$ .

It remains to identify  $\hat{y}$  as a solution of (5.11.2). Now

$$\begin{aligned}\hat{y}(t) &= \lim_{n \rightarrow \infty} y_n(t) = y_0 + \lim_{n \rightarrow \infty} \int_{t_0}^t f(u, y_{n-1}(u)) du \\ &= y_0 + \int_{t_0}^t f(u, \hat{y}(u)) du + \lim_{n \rightarrow \infty} \int_{t_0}^t \{f(u, y_{n-1}(u)) - f(u, \hat{y}(u))\} du.\end{aligned}$$

By the Lipschitz condition, the magnitude of the last integral does not exceed

$$K \int_{t_0}^t |y_{n-1}(u) - \hat{y}(u)| du < K(t - t_0) \max_u |y_{n-1}(u) - \hat{y}(u)|$$

which tends to zero as  $n \rightarrow \infty$ . Equation (5.11.2) has been recovered.

The existence of a continuous solution has now been verified and to complete the theorem it is necessary to show there is no other. Suppose, in fact, there were another solution  $Y(t)$  such that  $Y(t_0) = y_0$ , which is continuous in  $t_0 \leq t \leq t_0 + H$  with  $H \leq h$  and  $|Y(t) - y_0| < k$ . Then

$$Y(t) = y_0 + \int_{t_0}^t f(u, Y(u)) du.$$

Hence, if

$$|Y(t) - y_{n-1}(t)| < K^{n-1}k(t - t_0)^{n-1}/(n - 1)!,$$

$$\begin{aligned} |Y(t) - y_n(t)| &\leq \int_{t_0}^t |f(u, Y(u)) - f(u, y_{n-1}(u))| du \\ &\leq kK^n(t - t_0)^n/n!. \end{aligned}$$

This inequality is valid by induction if it is true for  $n = 1$ . But

$$|Y(t) - y_1(t)| \leq K \int_{t_0}^t |Y(u) - y_0| du \leq Kk(t - t_0)$$

and so the result for  $n = 1$  holds. Letting  $n \rightarrow \infty$  in the inequality we have  $Y(t) = \lim_{n \rightarrow \infty} y_n(t) = \hat{y}(t)$  and uniqueness has been established.  $\square$

### 5.11.2 System of first-order equations

The theory of the preceding section generalises to the system

$$\dot{y}_m = f_m(t, y_1, y_2, \dots, y_n) \quad (m = 1, \dots, n) \quad (5.11.6)$$

under the initial conditions  $y_m = y_{m0}$  at  $t = t_0$ . The region  $D$  is not so simple to depict since it is a rectangular parallelepiped in space of  $n + 1$  dimensions, because each  $y_m$  may change by a different amount from its initial value as  $t$  moves from  $t_0$ . So  $D$  is defined by  $t_0 \leq t \leq t_0 + h$ ,  $|y_m - y_{m0}| \leq k_m$  ( $m = 1, \dots, n$ ). Analogous to the conditions for a single equation the restrictions  $|f_m| < M$  and  $h < k_m/M$  for  $m = 1, 2, \dots$  are imposed.

#### EXISTENCE THEOREM II

Let  $f_m(t, y_1, \dots, y_n)$  be single-valued continuous functions of  $t, y_1, \dots, y_n$  in  $D$  such that for  $m = 1, \dots, n$

(a)  $|f_m(t, y_1, \dots, y_n)| < M$  in  $D$ ,

(b) (**Lipschitz condition**)

$$|f(t, y_1, \dots, y_n) - f(t, y'_1, \dots, y'_n)| < K_1|y_1 - y'_1| + \dots + K_n|y_n - y'_n|,$$

$K_1, \dots, K_n$  being finite constants, for any  $(t, y_1, \dots, y_n)$  and  $(t, y'_1, \dots, y'_n)$  in  $D$ . Then, for  $h < k_m/M$  ( $m = 1, \dots, n$ ) the system (5.11.6) possesses one and only one set of continuous solutions  $y_1(t), \dots, y_n(t)$  in  $t_0 \leq t \leq t_0 + h$  such that  $y_m(t_0) = y_{m0}$  ( $m = 1, \dots, n$ ).

**PROOF** The method of proof runs parallel to that for a single equation. It begins with

$$y_m(t) = y_{m0} + \int_{t_0}^t f_m(u, y_1(u), \dots, y_n(u)) du \quad (m = 1, \dots, n)$$

and an iteration is performed according to

$$y_{mr}(t) = y_{m0} + \int_{t_0}^t f_m(u, y_{1,r-1}(u), \dots, y_{n,r-1}(u)) du.$$

In view of the similarity to the single equation, details will be omitted.  $\square$

### 5.11.3 Differential equation of order n

The above theory can be applied to the order  $n$  differential equation

$$\frac{d^n y}{dt^n} = f\left(t, y, \frac{dy}{dt}, \dots, \frac{d^{n-1}y}{dt^{n-1}}\right) \quad (5.11.7)$$

with the initial conditions

$$y = y_{10}, \quad dy/dt = y_{20}, \quad \dots, \quad d^{n-1}y/dt^{n-1} = y_{n0}$$

at  $t = t_0$ . Make the substitutions

$$y(t) = y_1(t), \quad \frac{dy}{dt} = y_2(t), \quad \dots, \quad \frac{d^{n-1}y}{dt^{n-1}} = y_n(t).$$

The system

$$\begin{aligned} \dot{y}_1 &= y_2, & \dot{y}_2 &= y_3, & \dots, & \dot{y}_{n-1} &= y_n, \\ \dot{y}_n &= f(t, y_1, y_2, \dots, y_n) \end{aligned}$$

is obtained. When this is compared with (5.11.6) we see that

$$f_m(t, y_1, \dots, y_n) = \begin{cases} y_{m+1} & (m = 1, \dots, n-1) \\ f(t, y_1, \dots, y_n) & (m = n). \end{cases}$$

The  $f_m$  for  $m = 1, \dots, n-1$  obviously satisfy the conditions stated in Existence Theorem II. Therefore, if we make  $f_n$  comply with these conditions, that theorem is available for (5.11.7). Accordingly, we have the next theorem.

### **EXISTENCE THEOREM III**

If  $f(t, y_1, \dots, y_n)$  is continuous and

$$|f(t, y_1, \dots, y_n) - f(t, y'_1, \dots, y'_n)| < K_1 |y_1 - y'_1| + \dots + K_n |y_n - y'_n|$$

the differential equation (5.11.7) has one and only one continuous solution  $y(t)$  such that  $dy/dt, \dots, d^{n-1}y/dt^{n-1}$  are continuous for  $t_0 \leq t \leq t_0 + h$  and such that  $y, dy/dt, \dots, d^{n-1}y/dt^{n-1}$  take given values at  $t = t_0$ .

It may be remarked that, if (5.11.7) is linear,

$$f(t, y_1, \dots, y_n) = g(t) - a_0(t)y_1 - a_1(t)y_2 - \dots - a_{n-1}(t)y_n.$$

Since this is a polynomial in  $y_1, \dots, y_n$  the conditions of Existence Theorem III are met except at those values of  $t$  where one or more of  $g, a_0, \dots, a_{n-1}$  are not continuous. Hence, a linear differential equation has a unique continuous solution to the initial value problem, provided that  $t_0$  is not a point where there is lack of continuity on the part of  $g, a_0, \dots, a_{n-1}$ .

---

## Exercises

5.1 At what points are the conditions of Existence Theorem I not satisfied for

- (a)  $\dot{y} = y,$
- (b)  $\dot{y} = |y|^{1/2},$
- (c)  $t^3\dot{y} = y,$
- (d)  $t\dot{y} = y,$
- (e)  $\frac{1}{2}t\dot{y} = t^2 + y?$

Are the points where failure occurs singular points?

5.2 Sketch the trajectories in the phase plane of

- (a)  $\dot{x} = 5x + 2y,$   
 $y = 2x + 2y,$
- (b)  $\dot{x} = 6x + 12y,$   
 $y = 3x + y,$
- (c)  $\dot{x} = 4x + 5y,$   
 $y = -5x - 4y.$
- (d)  $\dot{x} = y - 2x,$   
 $\dot{y} = 4x - 5y,$
- (e)  $\dot{x} = 5x - 5y,$   
 $\dot{y} = 5x - 3y,$
- (f)  $\dot{x} = 5x + 4y,$   
 $\dot{y} = 9x,$
- (g)  $\dot{x} = 4x + 13y,$   
 $\dot{y} = -13x - 6y,$
- (h)  $\dot{x} = 5x - 13y,$   
 $\dot{y} = 13x - 5y.$

In each case, use MATLAB to solve the system numerically for initial condition  $x(0) = 1$ ,  $y(0) = 1$ , and plot the solution as a time series (i.e.,  $x(t)$  against  $t$  and  $y(t)$  against  $t$ ) and in the phase plane (i.e.,  $y(t)$  against  $x(t)$ ).

5.3 Draw the trajectories in the phase plane of

$$\frac{d^2x}{dt^2} + 3\frac{dx}{dt} + 2x = 0.$$

Use MATLAB to solve the system numerically for an initial condition of  $x(0) = 0$ ,  $\dot{x}(0) = 1$ .

5.4 Examine what fixed points arise in the phase plane for

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

where  $a$  and  $b$  are constants ( $a \neq 0$ ).

5.5 Discuss the trajectories of

$$\ddot{x} - 4\dot{x} + 40x = 0$$

in the phase plane by making the substitution  $x = \rho \cos \phi$ ,  $y = \rho \sin \phi$ .

5.6 Examine the possibility of periodic solutions of

$$c\ddot{x} + (2 + 3ax + 4bx^2)x = 0$$

where  $a$ ,  $b$  and  $c$  are constants,  $c$  being positive.

5.7 Sketch the trajectories in the phase plane of

$$\ddot{x} + \sin x = 0.$$

5.8 In the differential equation

$$(1 + a^2x^2)\ddot{x} + (b + a^2\dot{x}^2)x = 0$$

$a$  and  $b$  are constants. Discuss the behaviour of the solution.

5.9 Discuss the trajectories of

$$\ddot{\theta} = (\cos \theta - \mu) \sin \theta$$

for  $-\pi \leq \theta \leq \pi$ , the constant  $\mu$  ( $\neq 1$ ) being positive.

5.10 In the differential equation

$$\ddot{x} + \lambda \dot{x}^2 + x = 0$$

the constant  $\lambda$  is positive. Obtain  $dy/dx$  where  $y = \dot{x}$  and hence derive the differential equation satisfied by  $w = y^2$ . Find the trajectories and determine when they are closed curves.

5.11 Prove that

$$\ddot{x} - \epsilon(1 - x^4)\dot{x} + x = 0$$

has a stable limit cycle and that its amplitude is  $2^{3/4}$  ( $\approx 1.68$ ) when  $0 < \epsilon \ll 1$ .

5.12 Show that

$$\ddot{z} + f(\dot{z}) + z = 0$$

becomes

$$\ddot{x} + f'(x)\dot{x} + x = 0$$

on putting  $\dot{z} = x$ . Hence obtain information about the limit cycle of **Rayleigh's equation**

$$\ddot{z} + \epsilon\left(\frac{1}{3}\dot{z}^2 - \dot{z}\right) + z = 0$$

when  $0 < \epsilon \ll 1$ .

5.13 In **Duffing's equation**

$$\ddot{x} + D\dot{x} + x + \beta x^3 = E \sin \Omega t$$

$\Omega = 1 + \delta$  and  $D, \beta, E$  and  $\delta$  are all small. Show that, if an approximate solution is  $x = A \sin\{(1 + \delta)t + \theta\}$ ,  $A$  satisfies

$$\left(\frac{3}{4}\beta A^3 - 2\delta A\right)^2 + D^2 A^2 = E^2.$$

5.14 After transformation to polar coordinates  $x = r \cos \phi$ ,  $y = r \sin \phi$ , the differential equations of a system are

$$\dot{r} = r - r^3 + \mu r \cos \phi, \quad \dot{\phi} = 1.$$

- (a) In the case  $\mu = 0$ , show that the system has a stable limit cycle. Solve the differential equations directly to verify this result and write down an explicit equation for the limit cycle.
- (b) In the case  $0 < \mu < 1$ , show that the system has a stable limit cycle.

5.15 Show that the polar coordinate system

$$\dot{r} = (r - 1)(a + \sin^2 \phi), \quad \dot{\phi} = 1$$

has stable limit cycles when  $a < -\frac{1}{2}$ . What happens when  $a > -\frac{1}{2}$ ?

5.16 Solve the following nonlinear boundary value problem numerically.

$$y \frac{d^2y}{dx^2} = 10 \cos(x), \quad y(0) = 1, \quad y(1) - 2y'(1) = 0.$$

# Chapter 6

---

## *Mathematics of Heart Physiology*

---

### 6.1 The local model

We begin by recalling the three basic features of the heartbeat cycle upon which a mathematical model is to be developed. These are:

- (a) the model must exhibit an equilibrium state corresponding to diastole;
- (b) it should also contain a threshold for triggering the electrochemical wave emanating from the pacemaker causing the heart to contract into systole;
- (c) it should reflect the rapid return to the equilibrium state.

As in Section 4.2, we suppose the important quantities that model these features to be  $x$ , a typical muscle fibre length, which will necessarily depend on the time  $t$ , and  $b$ , an electrical control variable which governs the electrochemical wave, and which also depends on  $t$ .

In order to make any reasonable progress in modelling the heartbeat cycle, we must assume that the mathematical equations must be drawn from a particular class of equations which have in them the ability to describe at least the main features of heart physiology. The class we choose here is taken from the class of autonomous dynamical systems in two degrees of freedom. In other words we seek a mathematical model of the form

$$\begin{aligned}\frac{dx}{dt} &= f(x, b), \\ \frac{db}{dt} &= g(x, b).\end{aligned}\tag{6.1.1}$$

For this model to satisfy the first important quality (a) above, we ask that it has a unique **stable rest state**. Suppose this occurs at the critical point  $(b_0, x_0)$  of (6.1.1). That is,  $(b_0, x_0)$  satisfy the equations

$$f(x_0, b_0) = g(x_0, b_0) = 0.\tag{6.1.2}$$

If the system (6.1.1) is linearised, as in Chapter 5, about the rest state  $(b_0, x_0)$  we have

$$\begin{aligned}\frac{dx}{dt} &= f(x_0, b_0) + (x - x_0) \frac{\partial f(x_0, b_0)}{\partial x} + (b - b_0) \frac{\partial f(x_0, b_0)}{\partial b} \\ &\quad + \text{higher order terms}, \\ \frac{db}{dt} &= g(x_0, b_0) + (x - x_0) \frac{\partial g(x_0, b_0)}{\partial x} + (b - b_0) \frac{\partial g(x_0, b_0)}{\partial b} \\ &\quad + \text{higher order terms}.\end{aligned}$$

That is,

$$\begin{aligned}\frac{dx}{dt} &= a_{11}(x - x_0) + a_{12}(b - b_0) + \text{higher order terms}, \\ \frac{db}{dt} &= a_{21}(x - x_0) + a_{22}(b - b_0) + \text{higher order terms},\end{aligned}\quad (6.1.3)$$

where

$$a_{11} = \frac{\partial f}{\partial x}, \quad a_{12} = \frac{\partial f}{\partial b}, \quad a_{21} = \frac{\partial g}{\partial x}, \quad a_{22} = \frac{\partial g}{\partial b},$$

each evaluated at  $x = x_0, b = b_0$ .

From Chapter 5 we know that the local stability of the system (6.1.3) in the neighbourhood of  $(b_0, x_0)$  is governed by the roots  $\lambda_1, \lambda_2$  of the quadratic equation

$$\lambda^2 - \lambda(a_{11} + a_{22}) + a_{11}a_{22} - a_{12}a_{21} = 0. \quad (6.1.4)$$

Furthermore we know that our system is stable near  $(b_0, x_0)$  if the real parts of  $\lambda_1, \lambda_2$  are negative. We shall in fact assume a little more and suppose  $\lambda_1, \lambda_2$  to be real and negative, thus eliminating any undesirable oscillatory behaviour in our mathematical model. This requires that

$$\begin{aligned}a_{11} + a_{22} &< 0, \\ a_{11}a_{22} - a_{12}a_{21} &> 0.\end{aligned}\quad (6.1.5)$$

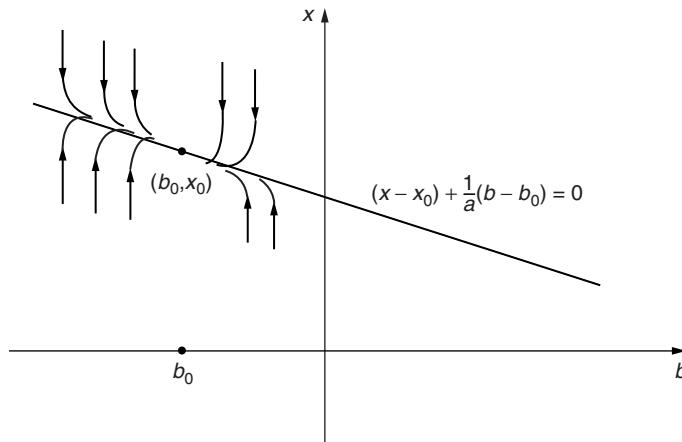
The inequalities (6.1.5) provide an infinite number of ways in which to choose the constants  $a_{ij}, i, j = 1, 2$ . Thus, to make further progress towards a possible model, we make some hypotheses, which, of course, ask for experimental confirmation. The hypotheses we make are the following:

H1: The rate of change of muscle fibre contraction depends, at any particular instant, on the tension of the fibre and on the chemical control.

H2: The chemical control changes at a rate directly proportional to muscle fibre extension.

For the moment we shall not specify under hypothesis H1 precisely how tension should enter into our model but focus on the implication of hypothesis H2. Mathematically, H2 is simply stated through the equation

$$\frac{db}{dt} = x - x_0. \quad (6.1.6)$$



**FIGURE 6.1.1:** Phase plane for the local model (6.1.9).

In other words the function  $g(x, b)$  in (6.1.5) is linear and independent of  $b$  and furthermore  $a_{21} = 1, a_{22} = 0$ . Consequently the inequalities (6.1.5) simplify to

$$a_{11} < 0, \quad a_{12} < 0. \quad (6.1.7)$$

The third of the features required by our mathematical model is that it should reflect the rapid return to the equilibrium state. This quality suggests that  $a_{11}$  be large and negative and since  $\frac{db}{dt}$ , by hypothesis H2, is proportional to  $x - x_0$  we expect  $a_{12}$  to be large and negative as well.

With this information we try

$$a_{11} = -\frac{a}{\epsilon}, \quad a_{12} = -\frac{1}{\epsilon}, \quad (6.1.8)$$

where  $a$  and  $\epsilon$  are positive constants with  $\epsilon$  small.

Putting these remarks and conclusions together, we arrive at the “local linearised model”

$$\begin{aligned} \epsilon \frac{dx}{dt} &= -a(x - x_0) - (b - b_0), \\ \frac{db}{dt} &= x - x_0. \end{aligned} \quad (6.1.9)$$

This model is depicted in the phase plane (Figure 6.1.1) and is constructed as follows. Take  $b_0 < 0, x_0 > 0$ , then along the line

$$(x - x_0) + \frac{1}{a}(b - b_0) = 0, \quad \frac{dx}{dt} = 0.$$

For

$$(x - x_0) + \frac{1}{a}(b - b_0) > 0, \quad \frac{dx}{dt} < 0,$$

that is,  $x$  is decreasing and for  $x > x_0, \frac{db}{dt} > 0$  and so  $b$  is increasing.

Furthermore, away from the line

$$(x - x_0) + \frac{1}{a}(b - b_0) = 0,$$

$|\frac{dx}{dt}|$  is large because of the presence of the small factor  $\epsilon$  and consequently the trajectories are largely vertical. Notice also that as the line

$$(x - x_0) + \frac{1}{a}(b - b_0) = 0$$

is approached from below along a trajectory with  $b < b_0$ , the trajectory crosses the line and is horizontal at the point of crossing. A similar phenomenon occurs when  $b > b_0$  and we approach the line from above. In fact  $(b_0, x_0)$  is a stable node. Figure 6.1.1 summarises these remarks.

---

## 6.2 The threshold effect

In this section we seek to modify the local mathematical model (6.1.9) so as to incorporate the feature (b) of Section 6.1, namely that the model should contain a threshold for triggering the electrochemical wave. Of course we must specify what is meant by the term **threshold**.

To help define these terms and so improve on the model (6.1.9), we recall from Chapter 4 the following facts.

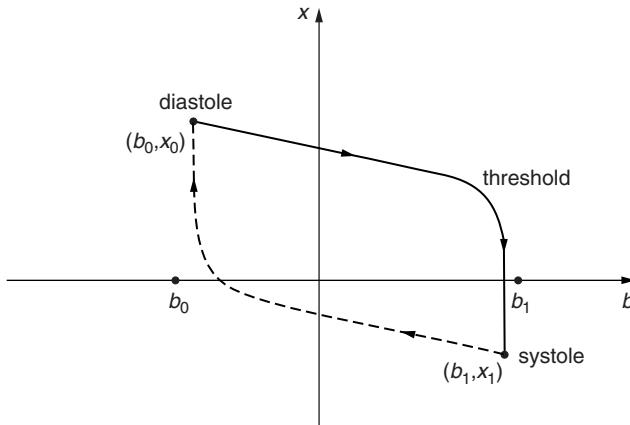
During the heartbeat cycle there are two equilibrium states, namely diastole and systole. The diastolic state is included in our model (6.1.9) whereas the systolic state is not, and so some modifications to include this are required. Furthermore, we observe that the pacemaker triggers off an electrochemical wave, which spreads slowly over the atria causing the muscle fibres to contract fairly slowly. The wave then spreads rapidly causing the whole ventricle to contract into systole.

This discussion suggests that, during the first part of the heartbeat cycle, the muscle fibre  $x$  contracts slowly at first and then at a certain point rapidly contracts further until the systolic equilibrium state is achieved. We shall call the point at which the rapid contraction occurs the **threshold**. While this contraction is going on, the chemical control variable  $b$  will be rising to a value  $b_1$  corresponding to systole.

The remarks we have made so far are summarised in Figure 6.2.1.

Following contraction into systole, the muscle fibres rapidly relax and return the heart to diastole and thus complete the cycle. The return is depicted schematically by the dotted line in Figure 6.2.1.

The problem then is to seek to modify the equations (6.1.9) so as to incorporate the general features shown in Figure 6.2.1. Such a modification is



**FIGURE 6.2.1:** The heartbeat cycle.

largely developed by trial. We propose that the model (6.1.9) is modified to the nonlinear form

$$\begin{aligned}\epsilon \frac{dx}{dt} &= -a(x - x_0) - (b - b_0) - (x - x_0)^3 - 3x_0(x - x_0)^2, \\ \frac{db}{dt} &= x - x_0.\end{aligned}\quad (6.2.1)$$

In order to assist with the development, we wish to write (6.2.1) in the form

$$\begin{aligned}\epsilon \frac{dx}{dt} &= -(x^3 - Tx + b), \quad T > 0, \\ \frac{db}{dt} &= x - x_0.\end{aligned}\quad (6.2.2)$$

If we use Taylor's theorem to expand the right-hand side of the first equation appearing in (6.2.2) about the point  $(b_0, x_0)$ , we can compare the system (6.2.2) with the system (6.2.1). Thus from (6.2.2)

$$\begin{aligned}\epsilon \frac{dx}{dt} &= -[x_0^3 - Tx_0 + b_0 + (3x_0^2 - T)(x - x_0) + 3x_0(x - x_0)^2 \\ &\quad + (x - x_0)^3 + (b - b_0)],\end{aligned}$$

and so we deduce that

$$x_0^3 - Tx_0 + b_0 = 0$$

and

$$3x_0^2 - T = a,$$

from which

$$T = 3x_0^2 - a$$

and the control  $b_0$  is expressed in terms of  $x_0$  and  $a$  via the relation

$$b_0 = 2x_0^3 - ax_0.$$

When tension has been appropriately identified, we shall see in the following section that the system (6.2.2) can be considered as contributing to a “local” model for the heartbeat cycle in that hypotheses H1 and H2 are included in the system and the qualities depicted in Figure 6.2.1 are almost accounted for.

---

### 6.3 The phase plane analysis and the heartbeat model

In order to develop the final form of our model of the heartbeat cycle, we begin by describing the phase plane portrait associated with the system (6.2.2).

To begin with consider the curve

$$x^3 - Tx + b = 0 \quad (6.3.1)$$

shown by the solid line  $ABCD$  in Figure 6.3.1.

On this curve  $dx/dt = 0$  and the flow is parallel to the  $b$  axis. Near the equilibrium  $(b_0, x_0)$  the configuration has that which is depicted in Figure 6.1.1. Furthermore, the direction of flow is determined by the second member of the system (6.2.2).

If we are above the cubic curve (6.3.1), i.e., where

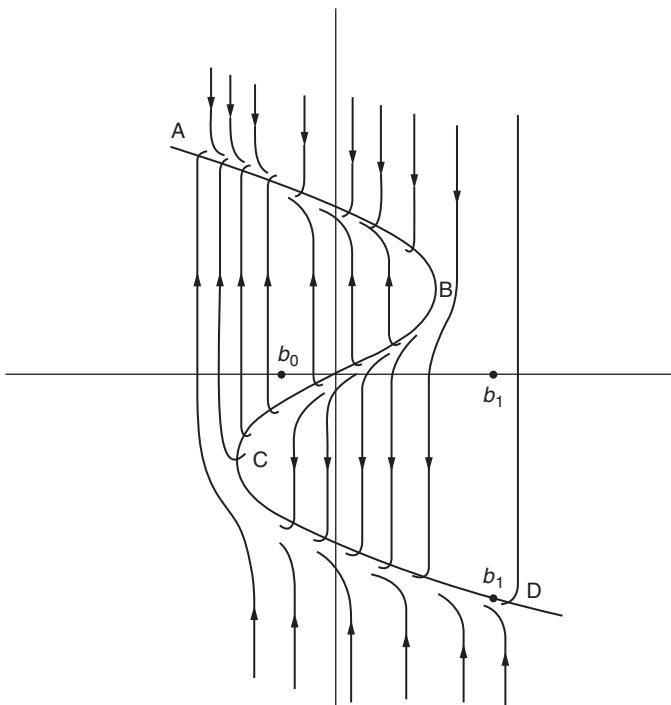
$$x^3 - Tx + b > 0,$$

then we see from (6.2.2) that  $dx/dt$  is large and negative and so the flow is largely vertically downwards, whereas if we are below the curve (6.3.1) the flow is vertically upwards. Thus, in general, the phase portrait consists of vertical trajectories except in the neighbourhood of the cubic curve.

We notice also that the trajectories always flow towards the portions  $AB$  and  $CD$  of (6.3.1) but always away from the portion  $BC$ . It is natural therefore to refer to the segments  $AB, CD$  as attractors for the flow and the segment  $BC$  as a repeller for the flow. The points  $B, C$  are important in that they can be associated with the threshold phenomenon discussed in the previous section.

If we now compare Figures 6.3.1 and 6.2.1, we see that the lower attractor  $CD$  gives rise to trajectories which follow a path back to  $(b_0, x_0)$  similar to the dotted line of Figure 6.2.1. Therefore the systolic state could be represented by a point  $(b_1, x_1)$  on the attractor  $CD$ .

However, the flow on the upper attractor  $AB$  is not in accordance with the general feature depicted by the solid line in Figure 6.2.1 in that there is



**FIGURE 6.3.1:** Phase portrait for the system (6.2.2).

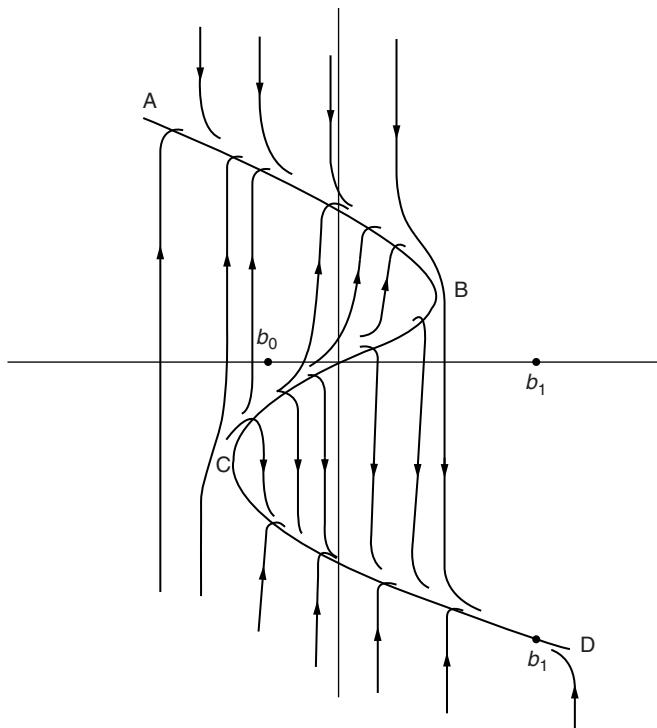
no mechanism for providing a trajectory corresponding to a smooth change of the chemical control from  $b_0$  to  $b_1$ . The mechanism for doing this could be thought of as due to the pacemaker and thus is not present in the model (6.2.2). If we switch the equilibrium state from  $(b_0, x_0)$  to  $(b_1, x_1)$ , then the flow on the upper attractor  $AB$  would provide trajectories similar to the solid line in Figure 6.2.1. This is achieved for the alternative model

$$\begin{aligned} \epsilon \frac{dx}{dt} &= -(x^3 - Tx + b), \\ \frac{db}{dt} &= x - x_1, \end{aligned} \tag{6.3.2}$$

the phase portrait of which is shown in Figure 6.3.2.

In this figure we see that the flow along the upper attractor  $AB$  does conform to the general behaviour of the solid line in Figure 6.2.1 and the point  $B$  can be identified with the threshold. However, the flow along the lower attractor  $CD$  cannot now be identified with the systolic state of the heartbeat cycle.

What is required is a modification of either (6.2.2) or (6.3.2), which incorporates both the desirable features of Figures 6.3.1 and 6.3.2, but excludes



**FIGURE 6.3.2:** Phase portrait for the system (6.3.2).

those features that do not conform to the known physiological behaviour of the heartbeat cycle. A model that does satisfy the above criteria is the following:

$$\begin{aligned} \epsilon \frac{dx}{dt} &= -(x^3 - Tx + b), \\ \frac{db}{dt} &= (x - x_0) + (x_0 - x_1)u, \end{aligned} \quad (6.3.3)$$

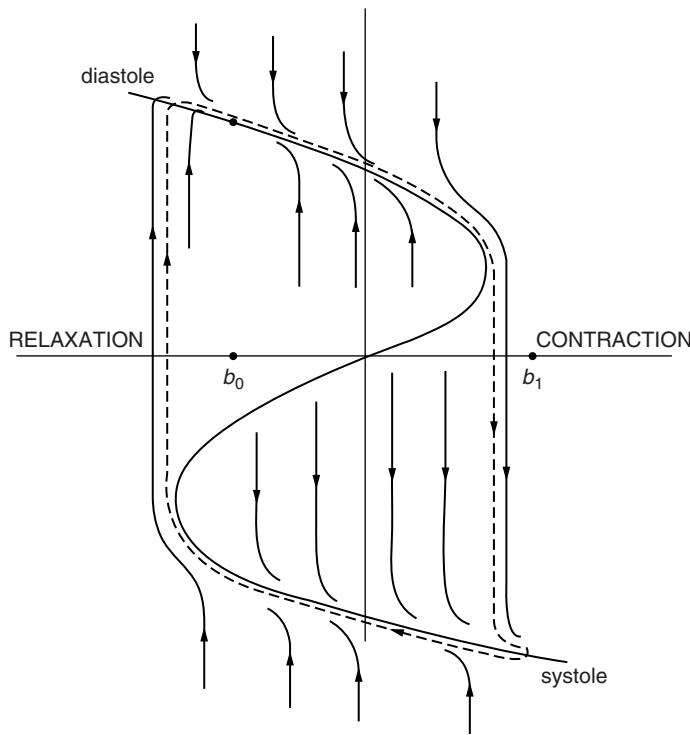
where  $u$  is a control variable associated with the pacemaker and is defined as follows:

(a)  $b_0 \leq b \leq b_1$  and for those values of  $x$  for which  $x^3 - Tx + b > 0$   
 and for (b)  $b > b_1$  and all values of  $x$ .

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

The system (6.3.3) will be called the **heartbeat equations** and the corresponding phase portrait is shown in Figure 6.3.3, where the dashed line indicates the heartbeat cycle.

We should not leave the discussion leading up to the model (6.3.3) without drawing the reader's attention to the mathematical problem of proving that there is a closed trajectory, shown dotted in Figure 6.3.3, corresponding to a



**FIGURE 6.3.3:** Phase portrait for the heartbeat equations.

complete heartbeat cycle. This is a difficult problem and will not be pursued here.

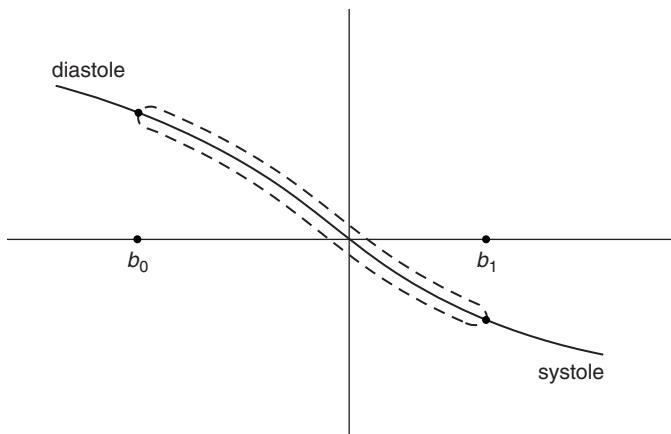
Finally we have to identify the contribution of tension to our model. To help us here, imagine that the muscle fibres are not under tension so that contraction into systole is slow and rather sluggish. In other words, we would not expect the sharp downward trajectories shown in Figure 6.3.3 but rather the slow behaviour shown in Figure 6.3.4.

Referring back to our model (6.3.3) we see that a portrait corresponding to Figure 6.3.4 is obtained if we set  $T = 0$ . We therefore identify  $T$  as tension. In the following section we shall consider this further when we discuss the predictions of the model (6.3.3) compared with known physiological facts.

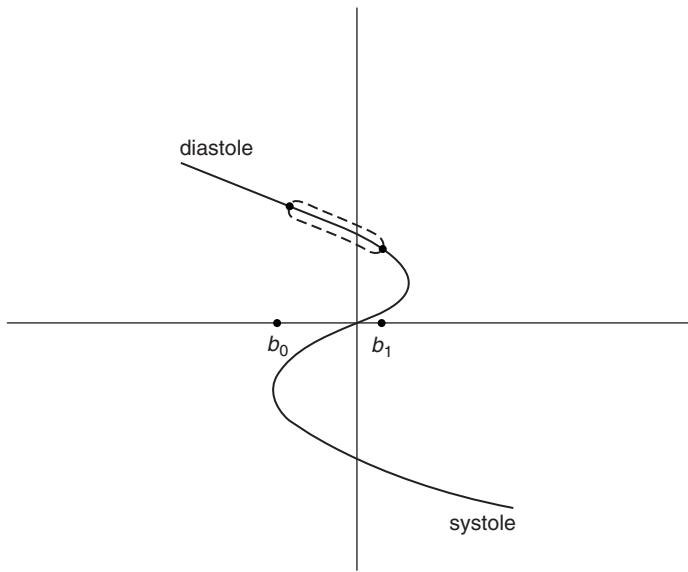
## 6.4 Physiological considerations of the heartbeat cycle

In this section we expand upon some of the physiological aspects of the heartbeat outlined in Chapter 4 and see how they may be interpreted in the context of the heartbeat equations (6.3.3).

Rybäk in 1957 originated the following experiment. If the heart of a frog is



**FIGURE 6.3.4:** Low-tension heartbeat.



**FIGURE 6.4.1:** The overstretched heart leading to cardiac failure.

taken out, then, not surprisingly, it ceases to beat. However, if it is then cut open into a flat membrane and subject to slight tension, it begins to beat once more and continues to do so for some hours. Alternatively, if the pacemaker is removed, then again beating stops. Rybak's experiment is analogous to setting  $T = 0$  in (6.3.3), leading to the sluggish cycle shown in Figure 6.3.4. As  $T$  increases slightly we again obtain a figure similar to Figure 6.3.3, but this time most of the work is done by the pacemaker wave in moving the control  $b$  from

$b_0$  to  $b_1$  and providing the large amount of squeezing necessary to contract the heart into systole. In fact, this low-tension heartbeat corresponds to the small atrial beat described in Chapter 4. Another relevant feature is known as Starling's law. This says that the more the muscle fibres are stretched before beating the more forcible is the beat. Therefore suppose excitement of one form or another causes adrenalin to be injected into the blood stream; the adrenalin then causes the arteries to contract and the pulse rate increases, which in turn causes the blood pressure to rise and the atria to push more blood into the ventricles. Starling's law describes how the stretched ventricles give a larger beat, overcoming the increased arterial back-pressure and circulating the blood faster. Starling's law is present in the model (6.3.3) if  $T$  is large, but not too large.

Finally, if the ventricles are overstretched beyond a certain point, as can happen, for example, when someone with high blood pressure receives a sudden shock, then the heart may fail to beat, or only beat feebly and cardiac failure may result. This particular aspect can be realised in our mathematical model if we increase the tension so much that the threshold extends beyond the systolic equilibrium point  $b_1$ , that is, when

$$T > (27b_1^2/4)^{1/2}. \quad (6.4.1)$$

This condition is such as to prevent the trigger from reaching threshold and so the muscle fibres remain in diastole (see Figure 6.4.1). In other words the heart does not beat and cardiac failure has occurred.

---

## 6.5 A model of the cardiac pacemaker

The heartbeat originates in the sino-atrial node, a region of cells which have the capability of depolarising spontaneously towards the threshold, firing and then recovering. However, the mechanism underlying the pacemaker wave is not fully understood, even though it is the basis of the field of electrocardiography.

In this section we propose a tentative mathematical model of the pacemaker firing mechanism that incorporates the control variable  $u$  introduced in the previous section.

Let us suppose that the pacemaker is characterised by a state  $y$ ,  $0 \leq y \leq 1$ , which satisfies the ordinary differential equation

$$\frac{dy}{dt} = -\gamma y + u, \quad (6.5.1)$$

and that, when  $y = 1$ , the pacemaker fires and jumps back to  $y = 0$ . We regard  $\gamma$  as a small positive number less than  $1/4$  and intrinsic to the pacemaker. Furthermore, it is natural to suppose that the motion of  $y$  is periodic, of

period  $T$  say. Thus if  $\{t_n\}$  denotes the set of firing times we suppose  $u = 1$  for  $t_n < t \leq t_n + T/2$  and  $u = 0$  for  $t_n + T/2 < t \leq t_{n+1}$  where  $t_{n+1} - t_n = T$ . In other words, the times  $\{t_n\}$  correspond to the heart being in diastole while the times  $\{t_n + T/2\}$  correspond to the heart being in systole.

Equation (6.5.1) is a simple differential equation of integrating factor type and is readily solved to give the solution

$$y(t) = \int_{t_n}^t e^{-\gamma(t-t')} u(t') dt', \quad (6.5.2)$$

where  $y(t_n) = 0$  and  $t_n \leq t \leq t_{n+1}$ . The equation for the firing times is therefore

$$1 = \int_{t_n}^{t_{n+1}} e^{-\gamma(t_{n+1}-t')} u(t') dt',$$

which because of the properties of  $u(t)$  can be written in the form

$$1 = \int_{t_n}^{t_n+T/2} e^{-\gamma(t_{n+1}-t')} dt',$$

i.e.,

$$\gamma = e^{-\gamma(t_{n+1}-t_n-T/2)} - e^{-\gamma(t_{n+1}-t_n)}$$

or

$$\gamma = e^{-\gamma T/2} - e^{-\gamma T}. \quad (6.5.3)$$

From (6.5.3) we can compute the period  $T$  as follows: setting  $z = e^{-\gamma T/2}$  we can write (6.5.3) as the quadratic equation

$$z^2 - z + \gamma = 0,$$

the solutions of which are

$$z = \frac{1}{2}(1 \pm \sqrt{(1 - 4\gamma)}),$$

i.e.,

$$2e^{-\gamma T/2} = (1 \pm \sqrt{(1 - 4\gamma)}). \quad (6.5.4)$$

Since we expect  $\gamma T$  to be small, we choose the positive root in (6.5.4), and if we neglect terms of order higher than  $\gamma^2$  we have

$$2e^{-\gamma T/2} \approx 2 - 2\gamma - 2\gamma^2,$$

which further approximates to

$$2(1 - \gamma T/2) \approx 2 - 2\gamma - 2\gamma^2$$

or

$$T \approx 2 + 2\gamma.$$

Thus if  $T$  is known, this equation can be used to estimate the parameter  $\gamma$ , or conversely if  $\gamma$  is known from experiment then  $T$  can be estimated.

## 6.6 Notes

This chapter has been largely motivated by the work of E.C. Zeeman. Section 6.5 was inspired by the paper of B.W. Knight, Dynamics of encoding in a population of neurons, *J. Gen. Physiol.*, **59**, 734–766, 1972.

## Exercises

- 6.1 Verify the phase portrait (Figure 6.1.1) for the “local model” (6.1.9) by solving the system for  $x$  and  $b$  as functions of  $t$ .
- 6.2 Provide a full phase plane analysis for the model (6.2.2).
- 6.3 Provide a full phase plane analysis for the heartbeat equations (6.3.3).
- 6.4 Verify the statements in Section 6.4 regarding Rybak’s experiment and Starling’s law by analysing the behaviour of the phase plane trajectories for the heartbeat equations as the tension  $T$  varies between  $T = 0$  and  $T = (27b_1^2/4)^{1/2}$ .
- 6.5 If the “cubic” term in the first of the heartbeat equations (6.3.3) is replaced by a piecewise linear expression similar to that depicted in Figure 4.3.4, a simplified model is obtained. Specifically, consider the system

$$\begin{aligned}\epsilon \frac{dx}{dt} &= -F(x, b), \\ \frac{db}{dt} &= (x - x_0) + (x_0 - x_1)u,\end{aligned}$$

where

$$\begin{aligned}F(x, b) &= x + b + \sqrt{T}, \quad x < -\frac{1}{2}\sqrt{T}, \\ &= x - b, \quad -\frac{1}{2}\sqrt{T} \leq x \leq \frac{1}{2}\sqrt{T}, \\ &= x + b - \sqrt{T}, \quad x > \frac{1}{2}\sqrt{T}.\end{aligned}$$

Provide a complete analysis of this model and compare the results with those of the system (6.3.3).



# Chapter 7

---

## Mathematics of Nerve Impulse Transmission

---

### 7.1 Excitability and repetitive firing

In this chapter we make a study of some of the principal properties of the simplified model of nerve impulse transmission (4.4.7) due to FitzHugh and Nagumo.

For the space clamped case the model is

$$\begin{aligned}\frac{du}{dt} &= u(1-u)(u-a) - w + I(t), \\ \frac{dw}{dt} &= bu - \gamma w,\end{aligned}\tag{7.1.1}$$

where  $0 < a < 1$ ,  $b > 0$ ,  $\gamma \geq 0$  and  $I(t)$  is the total membrane current, which may be an arbitrary function of time. If the space clamp is removed then the model becomes

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + u(1-u)(u-a) - w, \\ \frac{\partial w}{\partial t} &= bu - \gamma w.\end{aligned}\tag{7.1.2}$$

To begin with we wish to determine whether the model (7.1.1) exhibits the important **threshold** property mentioned in Section 4.3. In mathematical terms this property is the same as asking whether the system (7.1.1) is **excitable**.

Consider the ordinary differential equation,

$$\frac{dy}{dt} = y(1-y)(y-a),\tag{7.1.3}$$

where  $y = 0, a, 1$  are rest states. This equation has stable rest states at  $y = 0, 1$  and an unstable rest state at  $y = a$ . These statements are easily checked by observing that if initially  $y(0) < a$  then  $\frac{dy}{dt} < 0$  and so  $y(t) \rightarrow 0$ , whereas if  $y(0) > a$  then  $\frac{dy}{dt} > 0$  and  $y(t) \rightarrow 1$ . The implication of this is that for “small” (i.e.,  $y(0) < a$ ) initial data the solution is attracted to the rest state

$y = 0$ , whereas if the initial data is “large” (i.e.,  $y(0) > a$ ) then the solution is attracted to  $y = 1$ . We call the parameter  $a$  the **threshold** and call the equation (7.1.3) **excitable**.

Let us now see if a similar property is present in the system (7.1.1). For simplicity we only consider the case where  $I(t) = 0$ . The case  $I(t) \neq 0$  will be discussed later. With  $I(t) = 0$  the system (7.1.1) can be studied using the techniques developed in Chapter 5. First of all it is an important requirement that (7.1.1) has a unique rest state. That is, on setting

$$\frac{du}{dt} = \frac{dw}{dt} = 0,$$

we require the pair (i.e., the nullclines)

$$\begin{aligned} w &= u(1 - u)(u - a), \\ bw &= \gamma w \end{aligned} \tag{7.1.4}$$

to have the unique solution  $(u, w) = (0, 0)$ . That is, the equation

$$u(1 - u)(u - a) = \frac{b}{\gamma}u$$

must have the single solution  $u = 0$ . For this to be the case the quadratic equation,

$$u^2 - (1 + a)u + a + \frac{b}{\gamma} = 0,$$

can only have complex roots. That is, the parameters  $a, b, \gamma$  must be restricted so that

$$(1 - a)^2 < 4\frac{b}{\gamma}, \quad \gamma > 0. \tag{7.1.5}$$

Notice that if  $\gamma = 0$  then (7.1.1) has the unique rest state  $(0, 0)$  without restriction on the parameters  $a$  and  $b$ .

Linearising (7.1.1) about  $(0, 0)$  results in the system

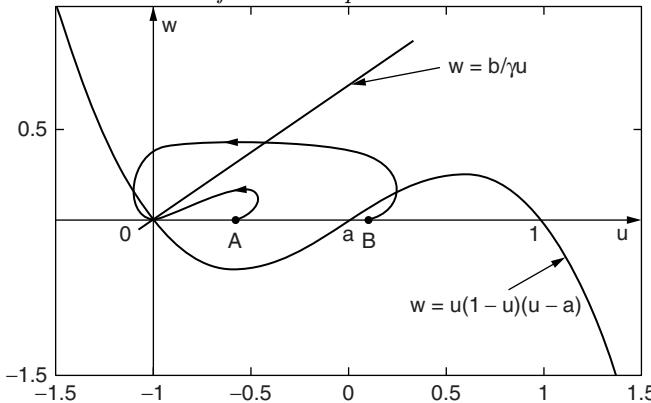
$$\begin{aligned} \frac{du}{dt} &= -au - w, \\ \frac{dw}{dt} &= bu - \gamma w. \end{aligned} \tag{7.1.6}$$

As in Chapter 5 we look for solutions of the form  $u = \alpha \exp \lambda t, w = \beta \exp \lambda t$ . Substituting these into equations (7.1.6) leads to the requirement that

$$\begin{aligned} (a + \lambda)\alpha + \beta &= 0, \\ (\gamma + \lambda)\beta - b\alpha &= 0, \end{aligned}$$

which has nontrivial solutions  $\alpha$  and  $\beta$  only if

$$\lambda^2 + (a + \gamma)\lambda + b + a\gamma = 0. \tag{7.1.7}$$



**FIGURE 7.1.1:** Global phase portrait for the system (7.1.8).

Since  $(b + a\gamma)$  and  $(a + \gamma)$  are positive, it follows that the roots of (7.1.7) have negative real parts and so we conclude that the rest state  $(0, 0)$  is locally stable. Such an analysis does not help to determine global properties of the trajectories. Nevertheless let us try to use the ideas followed in treating the simple problem (7.1.3). We consider the system

$$\begin{aligned}\frac{du}{dt} &= u(1-u)(u-a) - w, \\ \frac{dw}{dt} &= bu - \gamma w,\end{aligned}\tag{7.1.8}$$

subject to super threshold initial data

$$u(0) = u_0, \quad a < u_0 < 1, \quad w(0) = 0.$$

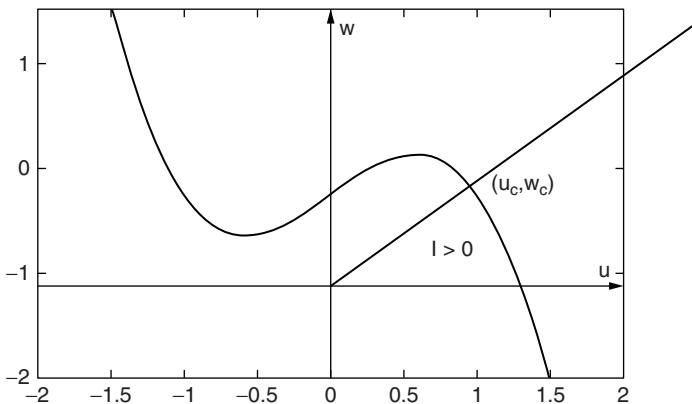
It follows from this that initially  $\frac{du}{dt} > 0$ ,  $\frac{dw}{dt} > 0$  and so  $(u, w)$  moves upwards in the positive quadrant of the phase plane and further away from  $(0, 0)$ . Notice that as  $u$  and  $w$  increase,  $\frac{du}{dt}$  will decrease until it reaches the nullcline  $w = u(1-u)(u-a)$  where  $\frac{du}{dt} = 0$  but  $\frac{dw}{dt}$  is still positive. Beyond this point  $\frac{du}{dt} < 0$  but  $w$  increases until the trajectory meets the nullcline  $\gamma w = bu$ , where  $\frac{dw}{dt} = 0$  and  $\frac{du}{dt} < 0$ . Continuing this argument we see that the trajectory returns to again meet the nullcline  $w = u(1-u)(u-a)$  in the upper-left quadrant where again  $\frac{du}{dt} = 0$  and this time  $\frac{dw}{dt} < 0$ . From here, either the trajectory progresses directly to  $(0, 0)$ , if the solutions of (7.1.7) are real and negative, or spirals towards  $(0, 0)$ , if the solutions of (7.1.7) are complex with negative real parts as shown in Figure 7.1.1.

The same sequence of arguments is used if the initial data are chosen so that

$$u(0) = u_0 > 1, \quad w(0) = 0.$$

Although this is not a conclusive proof, it is enough to convince one that the system (7.1.1) is excitable and that  $a$  is the threshold parameter.

Let us now consider the system in which the current  $I(t)$  is not zero but set at some nonzero value  $I$ , which may be positive or negative. Whether  $I < 0$



**FIGURE 7.1.2(a):** Nullclines for  $I > 0$ .

or  $I > 0$ , the resulting system is still excitable. The only difference being that the unique rest state is a solution to

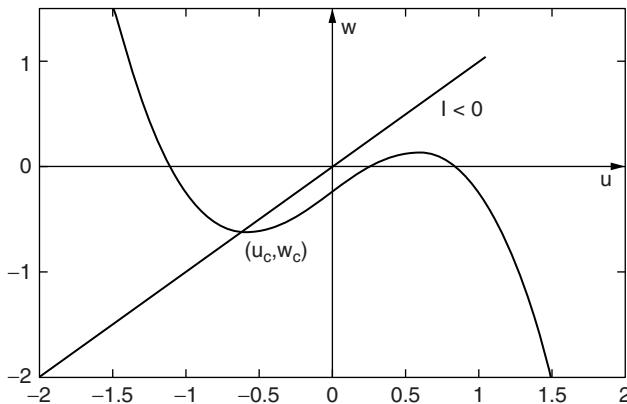
$$\begin{aligned} u(1-u)(u-a) - w + I &= 0, \\ bu &= \gamma w, \end{aligned} \tag{7.1.9}$$

and is no longer at  $(0, 0)$ . That is, for  $I > 0$  the nullclines are as shown in Figure 7.1.2(a) whereas for  $I < 0$  the nullclines are as shown in Figure 7.1.2(b).

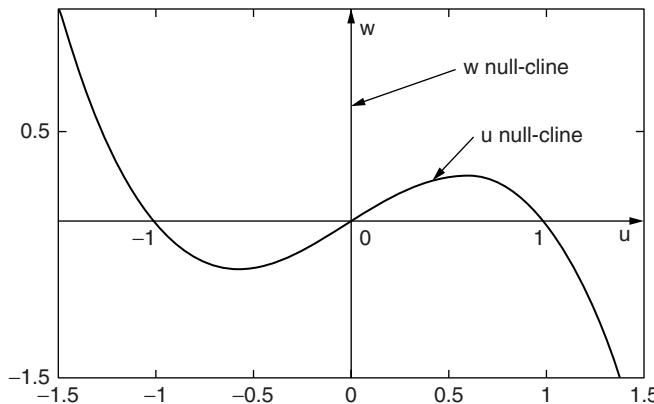
Of particular interest here is the case when  $I > 0$ . In their prize-winning work, Hodgkin and Huxley observed that on applying a constant current to the axon, repetitive firing of the action potential was observed. It is therefore of interest to see whether a similar phenomena is present in the FitzHugh–Nagumo system. Mathematically we ask whether the system (7.1.1) has periodic orbits or limit cycles. A natural tool for exploring this is to use the Poincaré–Bendixson theorem discussed in Chapter 5, Section 5.6. To illustrate the ideas involved we consider the system

$$\begin{aligned} \frac{du}{dt} &= u(1-u^2) - w, \\ \frac{dw}{dt} &= u. \end{aligned} \tag{7.1.10}$$

This system has the unique rest state  $(u, w) = (0, 0)$  and the nullclines are shown in Figure 7.1.3.



**FIGURE 7.1.2(b):** Nullclines for  $I < 0$ .



**FIGURE 7.1.3:** Nullclines of the system (7.1.10).

Linearising (7.1.10) about  $(0, 0)$  we have the system

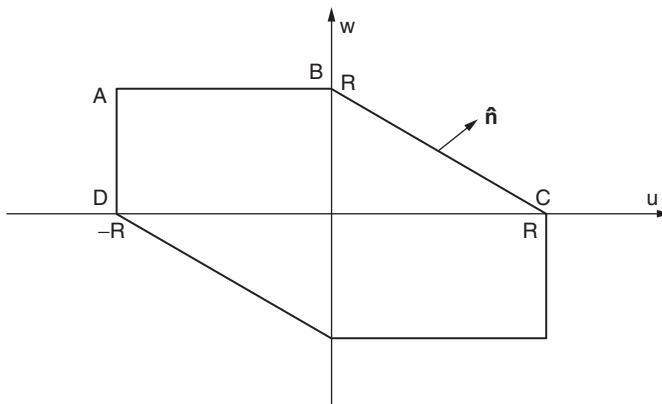
$$\begin{aligned} \frac{du}{dt} &= u - w, \\ \frac{dw}{dt} &= u \end{aligned} \tag{7.1.11}$$

from which the characteristic determinant is

$$\begin{vmatrix} 1 - \lambda & -1 \\ 1 & -\lambda \end{vmatrix} = 0,$$

or

$$\lambda^2 - \lambda + 1 = 0,$$



**FIGURE 7.1.4:** Region  $\Omega$  for the system 7.1.10.

from which we conclude that  $(0, 0)$  is an unstable focus. This means that trajectories starting in a neighbourhood of  $(0, 0)$  move away from  $(0, 0)$ . To be able to make use of the Poincaré-Bendixson theorem we need to show that there is a larger region  $\Omega$  surrounding  $(0, 0)$  within which all trajectories starting in  $\Omega$  remain in  $\Omega$  for all time. Choosing such a region is often an art. A suitable choice for the problem at hand is that shown in Figure 7.1.4.

Notice that since on replacing  $(u, w)$  by  $(-u, -w)$  leaves (7.1.10) unchanged, it is sufficient to consider Figure 7.1.4 with  $u \geq 0$ . On the face  $AB$ ,  $w = R$ ,  $u \leq 0$  and  $\frac{dw}{dt} = u \leq 0$ , with zero only occurring at  $u = 0$ . That is, for any  $R$ , trajectories always enter  $\Omega$  along  $AB$ . Similarly on the face  $AD$ ,  $u = -R$ ,  $0 \leq w \leq R$  and  $\frac{du}{dt} = -R(1 - R^2) - w > 0$  for  $R$  sufficiently large. Thus we conclude that trajectories cross  $AD$  in the positive  $u$  direction for  $R$  large enough.

On the face  $w = R - u$ ,

$$\begin{aligned}\frac{du}{dt} &= u(1 - u^2) - R + u, \\ \frac{dw}{dt} &= u\end{aligned}$$

and so

$$\frac{dw}{du} = \frac{u}{u(1 - u^2) + u - R}.$$

If  $\frac{dw}{du} > -1$  for  $0 \leq u \leq R$ , for  $R$  sufficiently large, then trajectories will cross in  $\Omega$  along  $BC$ . Hence we need

$$\frac{dw}{du} = \frac{u}{u(1 - u^2) + u - R} > -1,$$

for all  $0 \leq u \leq R$ . This means that  $u < -u(1 - u^2) - u + R$ , or

$$H(u) \equiv 3u - u^3 - R < 0,$$

for  $0 \leq u \leq R$ .

Now  $H(u)$  can be rewritten in the form

$$H(u) = -(u+2)(u-1)^2 + 2 - R < 0,$$

if  $R \geq 2$ . So we conclude that if  $R$  is sufficiently large then all trajectories of (7.1.10) remain in  $\Omega$ . It now follows from the Poincaré-Bendixson theorem that all trajectories converge to a limit cycle contained in  $\Omega$ .

An alternative way of considering the flow along the segment  $BC$  is to argue as follows:

Since  $\mathbf{n} = \frac{1}{\sqrt{2}}(1, 1)$  is the outward unit normal vector to  $BC$  and  $(\frac{du}{dt}, \frac{dw}{dt})$  is tangential to all trajectories, then where a trajectory crosses the boundary  $BC$ , the flow will be inwards if

$$\mathbf{n} \cdot \left( \frac{du}{dt}, \frac{dw}{dt} \right) < 0, \quad w = R - u. \quad (7.1.12)$$

That is, we require

$$\begin{aligned} (1, 1) \cdot (u(1 - u^2) - R + u, u) &= 3u - u^3 - R, \\ &= H(u) < 0, \end{aligned}$$

for  $0 \leq u \leq R$  as above.

We now return to the system (7.1.1) with  $I(t) = I > 0$  and constant. The system has the unique rest point  $(u_c, w_c)$  given by the solution to (7.1.8). For excitability we again wish to find an annular region  $\Omega$  in the  $(u, w)$ -plane that does not contain the rest point  $(u_c, w_c)$  and for which the Poincaré-Bendixson theorem can be applied. Suppose we construct a large circle of radius  $R$  in the  $(u, w)$ -plane. We want to show that for  $R$  sufficiently large the flow is always directed inwards. To do this set

$$u = r \cos \theta, \quad w = r \sin \theta. \quad (7.1.13)$$

Then

$$\begin{aligned} \frac{du}{dt} &= \frac{dr}{dt} \cos \theta - r \frac{d\theta}{dt} \sin \theta, \\ &= u(1 - u)(u - a) - w + I, \\ \frac{dw}{dt} &= \frac{dr}{dt} \sin \theta + r \frac{d\theta}{dt} \cos \theta, \\ &= bu - \gamma w, \end{aligned}$$

and so

$$\begin{aligned}\frac{dr}{dt} &= u(1-u)(u-a) \cos \theta - w \cos \theta + I \cos \theta + bu \sin \theta - \gamma w \sin \theta, \\ &= r \cos^2 \theta (1 - r \cos \theta) (r \cos \theta - a) - r \sin \theta \cos \theta + I \cos \theta \\ &\quad + br \sin \theta \cos \theta - \gamma r \sin^2 \theta.\end{aligned}\tag{7.1.14}$$

Now let  $r \rightarrow \infty$ . Due to the presence of the cubic term in  $u$  on the right-hand side of (7.1.14) we have  $\frac{dr}{dt} \leq 0$ , for all  $\theta$ . From the Poincaré-Bendixson theorem we claim that there will be a limit cycle in the phase plane if the rest point  $(u_c, w_c)$  is an unstable node or focus. To investigate this we linearise (7.1.1) about  $(u_c, w_c)$ .

Set

$$\begin{aligned}u &= u_c + \xi, \\ w &= w_c + \eta,\end{aligned}\tag{7.1.15}$$

and expand  $f(u) \equiv u(1-u)(u-a)$  as a Taylor series about  $(u_c, w_c)$  giving

$$f(u) = f(u_c) + f'(u_c)\xi + O(\xi^2).\tag{7.1.16}$$

The linearised system is

$$\begin{pmatrix} \frac{d\xi}{dt} \\ \frac{d\eta}{dt} \end{pmatrix} = \begin{pmatrix} f'(u_c) & -1 \\ b & -\gamma \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}.\tag{7.1.17}$$

The characteristic determinant is therefore

$$\begin{vmatrix} f'(u_c) - \lambda & -1 \\ b & -\gamma - \lambda \end{vmatrix} = 0,$$

or

$$\lambda^2 + (\gamma - f'(u_c))\lambda + b - f'(u_c)\gamma = 0.\tag{7.1.18}$$

For  $(u_c, w_c)$  to be unstable we need

$$\begin{aligned}\gamma - f'(u_c) &< 0 & b - f'(u_c)\gamma &> 0, \\ \gamma &< f'(u_c) & &< b/\gamma.\end{aligned}\tag{7.1.19}$$

An interpretation of the inequality (7.1.19) is that the slope of the  $u$  nullcline at  $(u_c, w_c)$  must be less than the slope of the  $w$  nullcline at that point. Finally we conclude from (7.1.19) and the Poincaré-Bendixson theorem that there is a limit cycle. That is, the FitzHugh–Nagumo system exhibits repetitive firing as originally observed by Hodgkin and Huxley.

## 7.2 Travelling waves

We now investigate travelling wave solutions to the system (7.1.2). Such solutions are of the form

$$u = \phi(x + ct), \quad w = \psi(x + ct), \quad (7.2.1)$$

subject to the requirement that, with  $\xi = x + ct$ ,

$$\lim_{|\xi| \rightarrow \infty} \phi(\xi) = \lim_{|\xi| \rightarrow \infty} \phi''(\xi) = \lim_{|\xi| \rightarrow \infty} \psi(\xi) = 0. \quad (7.2.2)$$

Among the many problems to be investigated in relation to travelling waves are the following:

- (a) provide analytic evidence to support the conjecture that the potential  $\phi(x + ct)$  has a form similar to that depicted in Figure 4.3.2;
- (b) obtain estimates for the wave speed  $c$ .

In the course of investigating these problems, we shall see that the parameters  $a, b$  and  $\gamma$  entering in the system (7.1.2) cannot be chosen arbitrarily but must satisfy some simple constraints.

As a first step to analysing the above problems we substitute the supposed solutions (7.2.1) into (7.1.2) to obtain the coupled system of ordinary differential equations

$$\begin{aligned} \phi'' &= c\phi' - \phi(1 - \phi)(\phi - a) + \psi, \\ c\psi' &= b\phi - \gamma\psi, \end{aligned} \quad (7.2.3)$$

where the primes denote differentiation with respect to  $\xi$ . If we set  $\theta = \phi'$  then (7.2.3) can be written as the system

$$\begin{aligned} \theta' &= c\theta - \phi(1 - \phi)(\phi - a) + \psi, \\ \phi' &= \theta, \\ \psi' &= \frac{b}{c}\phi - \frac{\gamma}{c}\psi. \end{aligned} \quad (7.2.4)$$

The critical points of this system corresponding to “rest states” of the system (7.2.4) are given by

$$\left(0, \phi_i, \frac{b}{\gamma}\phi_i\right), \quad i = 1, 2, 3, \quad (7.2.5)$$

provided  $\gamma > 0$  and  $\phi_i$  is a root of the equation

$$x \left[ (x - a)(1 - x) - \frac{b}{\gamma} \right] = 0. \quad (7.2.6)$$

An important requirement of the system (7.2.4) is that it has a unique “rest state” and this can only be achieved if the cubic equation (7.2.6) has only the real root  $x = 0$ . In other words the quadratic equation

$$x^2 - (1 + a)x + a + \frac{b}{\gamma} = 0$$

must have complex roots. As argued in Section 7.1 we know that this leads to the restriction

$$(1 - a)^2 < 4\frac{b}{\gamma}, \quad \gamma > 0. \quad (7.2.7)$$

Again if  $\gamma = 0$  then (7.2.4) has the unique rest state  $(0, 0)$  without any restrictions on the parameters  $a, b$ .

A complete “phase plane” analysis of the system (7.2.4) is quite difficult due to the fact that it is a third order system rather than a second order one. The methods we employ in this situation will, as shown in the next section, be somewhat different from those discussed in Chapter 5. As a prelude to this consider the special case in which  $b = 0$ . Here the system (7.2.4) partially decouples in that  $\psi$  now satisfies the equation

$$\psi' = -\frac{\gamma}{c}\psi,$$

from which it follows that

$$\psi = A \exp\left(\frac{\gamma \xi}{c}\right).$$

For such a solution to satisfy the conditions (7.2.2) it is clear that  $A = 0$ . In other words  $\psi \equiv 0$  and the system (7.2.4) simplifies to the second order system

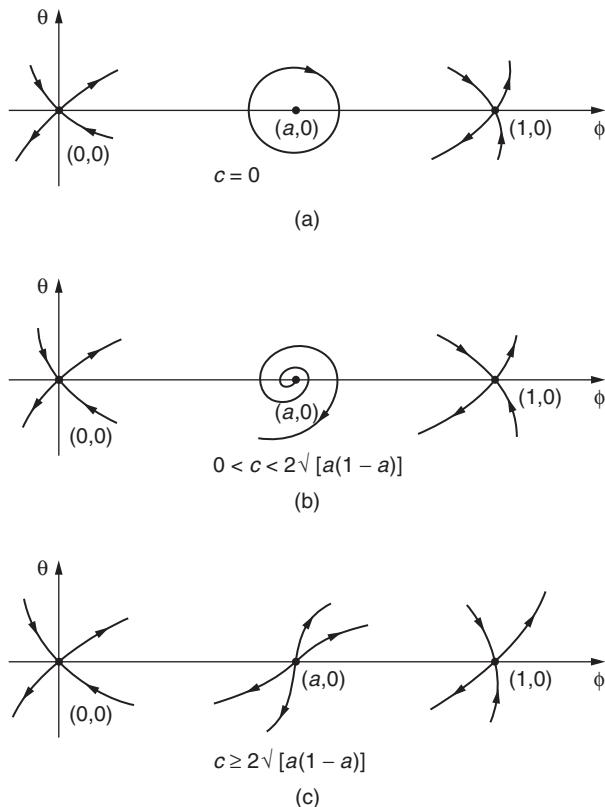
$$\begin{aligned} \theta' &= c\theta - \phi(1 - \phi)(\phi - a), \\ \phi' &= \theta. \end{aligned} \quad (7.2.8)$$

In the  $(\phi, \theta)$  phase plane there are three finite singular points, namely

$$(0, 0), \quad (a, 0), \quad (1, 0). \quad (7.2.9)$$

Following the treatment of Chapter 5, we see that  $(0, 0)$  and  $(1, 0)$  are saddle-points whereas  $(a, 0)$  is a centre if  $c = 0$ , a repulsive spiral for  $0 < c < 2\sqrt{[a(1 - a)]}$  or a repulsive node if  $c \geq 2\sqrt{[a(1 - a)]}$  (see Figure 7.2.1).

To provide a complete, that is global, analysis of the  $(\phi, \theta)$  phase plane is quite a formidable task and will not be pursued here. However, a glance at Figure 7.2.1 leads one to ask whether there is a trajectory which leaves  $(0, 0)$  and enters  $(1, 0)$ . The answer to this is certainly “yes” and one such trajectory



**FIGURE 7.2.1:** Local behaviour of the phase plane.

is provided by the answer to Exercise 4.8. That is, if  $c = \sqrt{2}(\frac{1}{2} - a)$ ,  $0 < a < \frac{1}{2}$ , we have the solution

$$\phi(\xi) = \frac{1}{1 + \exp -\xi/\sqrt{2}}, \quad \theta(\xi) = \frac{\exp -\xi/\sqrt{2}}{\sqrt{2}(1 + \exp -\xi/\sqrt{2})^2}. \quad (7.2.10)$$

In fact this is the only trajectory that connects the points  $(0, 0)$ ,  $(1, 0)$ .

### 7.3 Qualitative behaviour of travelling waves

If we look at Figure 4.3.2 we observe that the graph of the membrane potential crosses the  $\xi$  axis at one point. Evidence that supports the same type of behaviour for  $\phi$  can be demonstrated as follows. Integrate (7.2.3) over

the range  $-\infty < \xi < \infty$  and use the conditions (7.2.2) to obtain

$$\int_{-\infty}^{\infty} \phi(1 - \phi)(\phi - a) d\xi = \int_{-\infty}^{\infty} \psi d\xi$$

and

$$\int_{-\infty}^{\infty} \psi d\xi = \frac{b}{\gamma} \int_{-\infty}^{\infty} \phi d\xi$$

provided  $\gamma \neq 0$ . Eliminating  $\psi$  between these identities shows that

$$\int_{-\infty}^{\infty} \left( \phi(1 - \phi)(\phi - a) - \frac{b}{\gamma} \phi \right) d\xi = 0.$$

This means that the function  $\phi(1 - \phi)(\phi - a) - \frac{b}{\gamma} \phi$  must change sign and, since we require the inequality (7.2.7) to hold, it follows that the equation

$$\phi(1 - \phi)(\phi - a) - \frac{b}{\gamma} \phi = 0$$

has only the root  $\phi = 0$ . In other words  $\phi = 0$  for at least one value of  $\xi$ . This argument of course does not preclude the possibility that  $\phi = 0$  for more than one value of  $\xi$  and so does not completely confirm Figure 4.3.2.

In what follows we assume that  $\phi$  and  $\psi$  are sufficiently well behaved for all integrals that occur to converge.

Multiply the first of equations (7.2.3) by  $\phi$  and the second by  $\psi$ , and then integrate the resulting expressions with respect to  $\xi$  from  $-\infty$  to  $\infty$ . This together with the conditions (7.2.2) gives

$$\begin{aligned} \int_{-\infty}^{\infty} \phi^2(1 - \phi)(\phi - a) d\xi &= \int_{-\infty}^{\infty} (\phi')^2 d\xi + \int_{-\infty}^{\infty} \phi\psi d\xi, \\ \gamma \int_{-\infty}^{\infty} \psi^2 d\xi &= b \int_{-\infty}^{\infty} \phi\psi d\xi. \end{aligned} \quad (7.3.1)$$

Next we repeat the process but this time we multiply by  $\phi'$  and  $\psi'$ , respectively. The result is

$$\begin{aligned} c \int_{-\infty}^{\infty} (\phi')^2 d\xi &= - \int_{-\infty}^{\infty} \phi'\psi d\xi, \\ c \int_{-\infty}^{\infty} (\psi')^2 d\xi &= b \int_{-\infty}^{\infty} \phi\psi' d\xi. \end{aligned} \quad (7.3.2)$$

But

$$\int_{-\infty}^{\infty} \phi'\psi d\xi = - \int_{-\infty}^{\infty} \phi\psi' d\xi$$

and so from (7.3.2) we find that

$$b \int_{-\infty}^{\infty} (\phi')^2 d\xi = \int_{-\infty}^{\infty} (\psi')^2 d\xi. \quad (7.3.3)$$

Finally, multiply the second of equations (7.2.3) by  $\phi$  and integrate, using (7.3.1), to get

$$b^2 \int_{-\infty}^{\infty} (\phi)^2 d\xi = bc \int_{-\infty}^{\infty} \phi \psi' d\xi + \gamma^2 \int_{-\infty}^{\infty} (\psi)^2 d\xi. \quad (7.3.4)$$

If we now eliminate  $\int_{-\infty}^{\infty} (\phi)^2 d\xi$ ,  $\int_{-\infty}^{\infty} (\psi)^2 d\xi$ ,  $\int_{-\infty}^{\infty} \phi \psi' d\xi$  and  $\int_{-\infty}^{\infty} \phi \psi d\xi$  between equations (7.3.1)–(7.3.4) we obtain

$$b \int_{-\infty}^{\infty} \left( \frac{b}{\gamma} \phi^2 - \phi^2 (1 - \phi)(\phi - a) \right) d\xi = \frac{(c^2 - \gamma)}{\gamma} \cdot \int_{-\infty}^{\infty} (\psi')^2 d\xi. \quad (7.3.5)$$

If, instead of eliminating  $\int_{-\infty}^{\infty} \psi^2 d\xi$ , we eliminate  $\int_{-\infty}^{\infty} (\psi')^2 d\xi$  then we obtain the identity

$$b \int_{-\infty}^{\infty} \left( \frac{b}{c^2} \phi^2 - \phi^2 (1 - \phi)(\phi - a) \right) d\xi = \frac{\gamma}{c^2} (\gamma - c^2) \int_{-\infty}^{\infty} \psi^2 d\xi. \quad (7.3.6)$$

Now consider the integrand on the left-hand side of (7.3.5). We have

$$\phi^2 \left( \frac{b}{\gamma} + (a - \phi)(1 - \phi) \right) = \phi^2 \left[ \left( \phi - \frac{(1+a)}{2} \right)^2 + \frac{b}{\gamma} - \frac{(1-a)^2}{4} \right] \geq 0,$$

on noting the inequality (7.2.7). Thus since the left-hand side of (7.3.5) is positive, the same must be true of the right-hand side and this means that

$$c^2 > \gamma, \quad (7.3.7)$$

which provides a lower bound on the wave speed  $c$ . A better lower bound can be achieved as follows. Using the inequality (7.3.7) we deduce from (7.3.6) that

$$\int_{-\infty}^{\infty} \left( \frac{b}{c^2} \phi^2 - \phi^2 (1 - \phi)(\phi - a) \right) d\xi < 0$$

and so the expression  $\frac{b}{c^2} - (1 - \phi)(\phi - a)$  must take negative values. This is possible if and only if

$$(1 + a)^2 > 4 \left( \frac{b}{c^2} + a \right),$$

i.e.,

$$(1 - a)^2 > 4 \frac{b}{c^2}$$

and so

$$c^2 > 4b(1 - a)^2 > \gamma. \quad (7.3.8)$$

In establishing this lower bound on  $c$  we can obtain further insight into the role of  $a$  as a threshold parameter. Since the expression  $\frac{b}{c^2} - (1 - \phi)(\phi - a)$

must take negative values, it follows that the maximum value  $\phi_{max}$  of  $\phi$  must exceed the smallest root of the equation

$$\frac{b}{c^2} - (1 - \phi)(\phi - a) = 0.$$

That is,

$$\phi_{max} > \frac{1+a}{2} - \frac{1}{2}\left((1-a)^2 - \frac{4b}{c^2}\right)^{1/2}.$$

But more than this, it is easy to see that

$$a < \frac{1+a}{2} - \frac{1}{2}\left((1-a)^2 - \frac{4b}{c^2}\right)^{1/2},$$

and so

$$\phi_{max} > a.$$

From this fact we again see that  $a$  plays the role of a threshold parameter.

To conclude our discussion of the qualitative behaviour of the travelling wave  $\phi(\xi)$  we remark that there are a number of problems remaining:

- (a) Obtain an upper bound for the wave speed  $c$ .
- (b) Is the travelling wave  $\phi$  stable with respect to small perturbations?
- (c) Are there other wave-like solutions?

These and other problems have been discussed in the literature but we shall not pursue them further here.

## 7.4 Piecewise linear model

In Chapter 4 it was mentioned that further insight into the behaviour of travelling waves can be gained by studying the more tractable of the system (7.1.2) in which the cubic term is replaced by the piecewise linear form shown in Figure 4.3.4. Here we consider the case where  $\theta = \frac{\pi}{2}$  and study travelling wave solutions of the system

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} - u + H(u - a) - w, \\ \frac{\partial w}{\partial t} &= bu - \gamma w, \end{aligned} \tag{7.4.1}$$

where  $H$  is the Heaviside step function defined as  $H(x) = 1$  for  $x > 0$ ,  $H(x) = 0$  for  $x < 0$ .

We seek a travelling wave solution  $u = U(z)$ ,  $w = W(z)$  where  $u$  is of the form shown in Figure 4.3.2 and  $z = x + ct$ . From (7.4.1) it follows the  $U(z)$  and  $W(z)$  must satisfy the system

$$\begin{aligned}\ddot{U} - c\dot{U} - U + H(U - a) - W &= 0, \\ bU - c\dot{W} - \gamma W &= 0.\end{aligned}\quad (7.4.2)$$

Since this system is autonomous we can specify  $U(0) = a$  and that for some  $z_1 > 0$  we assume  $U(z_1) = a$ . Furthermore at  $z = 0$ ,  $U$  increases through the threshold  $U = a$  and the nonlinear term jumps by 1 and so there must be corresponding jump of  $-1$  in  $\ddot{U}$  at  $z = 0$ ,

$$\left[ \ddot{U} \right]_{0^-}^{0^+} = -1. \quad (7.4.3)$$

Similarly at  $z = z_1$ ,  $U$  is assumed to decrease through  $a$  and so there is a further jump discontinuity in  $\ddot{U}$  at  $z = z_1$ :

$$\left[ \ddot{U} \right]_{z_1^-}^{z_1^+} = 1. \quad (7.4.4)$$

If we differentiate the first of equations (7.4.2) and eliminate  $W$  then  $U$  is seen to satisfy

$$\ddot{U} - (c - \frac{\gamma}{c})\dot{U} - (1 + \gamma)\dot{U} - \frac{(b + \gamma)}{c}U = 0, \quad (7.4.5)$$

for  $z \in (-\infty, \infty) \setminus \{0, z_1\}$ .

The characteristic polynomial associated with (7.4.5) is

$$p(\lambda) = \lambda^3 - (c - \frac{\gamma}{c})\lambda^2 - (1 + \gamma)\lambda - \frac{(b + \gamma)}{c}. \quad (7.4.6)$$

By using the same arguments that led to the inequality (7.3.7) we find that the  $c$  satisfies the same lower bound  $c^2 > \gamma$  in the piecewise linear model. As a consequence we conclude that the characteristic polynomial (7.4.6) has one positive root  $\alpha_1$  say and either two negative real roots  $\alpha_3 < \alpha_2 < 0$  or two complex roots with negative real part. Now  $U(z)$  is given by

$$\begin{aligned}U &= ae^{\alpha_1 z}, \quad z < 0, \\ &= Ae^{\alpha_1 z} + Be^{\alpha_2 z} + Ce^{\alpha_3 z}, \quad z \in (0, z_1), \\ &= De^{\alpha_2 z} + Ee^{\alpha_3 z}, \quad z > z_1,\end{aligned}\quad (7.4.7)$$

where  $B$ ,  $C$ ,  $D$ ,  $E$  and  $F$  are constants to be determined by demanding that  $U$  and  $\dot{U}$  are continuous at  $z = 0$  and  $z = z_1$  together with the jump discontinuities (7.4.3) and (7.4.4). Thus at  $z = 0$  we deduce that

$$\begin{aligned}A + B + C &= a, \\ \alpha_1 A + \alpha_2 B + \alpha_3 C &= a\alpha_1, \\ \alpha_1^2 A + \alpha_2^2 B + \alpha_3^2 C &= a\alpha_1^2 - 1.\end{aligned}\quad (7.4.8)$$

It is convenient to write this system of equations in matrix form:

$$\begin{bmatrix} 1 & 1 & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1^2 & \alpha_2^2 & \alpha_3^2 \end{bmatrix} \begin{bmatrix} A - a \\ B \\ C \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

Now if we define

$$\rho_i = \frac{dp}{d\lambda} |_{\lambda=\alpha_i},$$

then we find that

$$A - a = -\frac{1}{\rho_1}, \quad B = -\frac{1}{\rho_2}, \quad C = -\frac{1}{\rho_3}. \quad (7.4.9)$$

Similarly for the conditions to be satisfied at  $z = z_1$  we have the system:

$$\begin{bmatrix} 1 & 1 & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1^2 & \alpha_2^2 & \alpha_3^2 \end{bmatrix} \begin{bmatrix} Ae^{\alpha_1 z_1} \\ (B - D)e^{\alpha_2 z_1} \\ (C - E)e^{\alpha_3 z_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix},$$

from which we deduce that

$$D = \frac{e^{-\alpha_2 z_1} - 1}{\rho_2}, \quad E = \frac{e^{-\alpha_3 z_1} - 1}{\rho_3}, \quad (7.4.10)$$

together with

$$e^{-\alpha_1 z_1} = (1 - a\rho_1). \quad (7.4.11)$$

Finally, continuity at  $z = 0$  and  $z = z_1$  give the conditions

$$\begin{aligned} A + B + C &= a, \\ Ae^{\alpha_1 z_1} + Be^{\alpha_2 z_1} + Ce^{\alpha_3 z_1} &= a, \\ De^{\alpha_2 z_1} + Ee^{\alpha_3 z_1} &= a. \end{aligned} \quad (7.4.12)$$

Using equations (7.4.11) and (7.4.12) we obtain the relation

$$a = -\frac{1}{\rho_1} - \frac{(1 - a\rho_1)}{\rho_2} e^{-\alpha_2/\alpha_1} - \frac{(1 - a\rho_1)}{\rho_3} e^{-\alpha_3/\alpha_1}. \quad (7.4.13)$$

For  $a, b$  and  $\gamma$  fixed the roots  $\alpha_i$  as well as the derivatives  $\rho_i$  are all functions of the wave speed  $c$ . This means that (7.4.13) is a transcendental equation to be solved for  $c$ , then the above equations can be solved for the unknowns  $A, B, C, D$  and  $E$ .

It turns out that it is better to consider  $c$  as fixed and think of (7.4.13) as an equation for  $a$ . To this end define

$$s = (1 - a\rho_1) = e^{-\alpha_1 z_1},$$

then (7.4.13) can be written as

$$F(s) = 0,$$

where

$$F(s) = 2 - s + \frac{\rho_1}{\rho_2} e^{-\alpha_2/\alpha_1} + \frac{\rho_1}{\rho_3} e^{-\alpha_3/\alpha_1}. \quad (7.4.14)$$

Since  $s = e^{-\alpha_1 z_1} < 1$  it is clear that we seek a solution of  $F(s) = 0$  in  $(0, 1)$ . Using the relations between  $\alpha_i$  and  $\rho_i$  it is easy to show that

$$F(0) - 2 = 0 = F(1) = \frac{dF(1)}{ds},$$

and

$$\frac{d^2F}{ds^2}(1) = \frac{\rho_1}{\alpha_1^2} - 2 = 1 - \frac{2}{\alpha_1} \left(c - \frac{\gamma}{c}\right) - \frac{(1+\gamma)}{\alpha_1^2}.$$

If  $F$  has a maximum at  $s = 1$  then it must have a root in  $(0, 1)$  and this can only happen if

$$\alpha_1^3 - 2\alpha_1^2 \left(c - \frac{\gamma}{c}\right) - \alpha_1(1+\gamma) < 0.$$

Combining this with (7.4.6), i.e.,  $p(\alpha_1) = 0$  we obtain the estimate

$$\alpha_1 > \left(\frac{b+\gamma}{c^2-\gamma}\right)^{1/2}. \quad (7.4.15)$$

Since  $c^2 - \gamma > 0$  this estimate implies that  $p\left(\left(\frac{b+\gamma}{c^2-\gamma}\right)^{1/2}\right) < 0$ , which gives the inequality

$$\left(\frac{b+\gamma}{c^2-\gamma}\right)^{3/2} - \left(\frac{b+\gamma}{c^2-\gamma}\right)^{1/2} (1+\gamma) - 2\frac{(b+\gamma)}{c} < 0.$$

It is tedious to extract an estimate of the wave speed  $c$  from this inequality in general; however in the case where  $\gamma = 0$  the inequality simplifies to

$$\left(\frac{b}{c^2}\right)^{3/2} - \left(\frac{b}{c^2}\right)^{1/2} - 2\frac{b}{c} < 0,$$

leading to

$$c^2 > \frac{b}{(1+2b^{1/2})}. \quad (7.4.16)$$

If with  $b$  fixed, in this case, we allow  $c$  to range over a set of values greater than this and solve  $F = 0$  for  $s \in (0, 1)$ , and so for  $a$ , we can make a plot of  $c$  against  $a$  for which a travelling wave exists. It turns out that for each value of  $a$  there are two possible values of  $c$ . The higher speed corresponds to a stable wave, while the slower wave is unstable.

## 7.5 Notes

A detailed discussion of the phase plane depicted in Figure 7.2.1 together with the global behaviour of the system (7.2.8) is contained in the paper by H.P. McKean, Nagumo's equation, *Adv. Math.*, **4** 209–223, 1970. A detailed account of the piecewise linear model considered in Section 7.4 is to be found in the paper by J. Rinzel and J. B. Keller, Travelling wave solutions of a nerve conduction equation, *Biophysical Journal*, **13** 1313–1337, 1973.

---

## Exercises

- 7.1 Consider the space clamp model (7.1.1) where the induced current  $I$  is assumed constant. Determine the critical points of this system when the current takes the values  $I = \frac{b}{\gamma}$ ,  $I = \frac{ab}{\gamma}$ .
- 7.2 Prove that the system (7.1.1) has a single unstable critical point only when it lies on that part of the  $u$ -nullcline which has a positive gradient.
- 7.3 Verify the local behaviour of the phase plane depicted in Figure 7.2.1 for the system (7.2.8).
- 7.4 Consider the general system:

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + f(u) - w, \\ \frac{\partial w}{\partial t} &= bu - \gamma w,\end{aligned}$$

where  $f(0) = 0$ .

Derive the equations to be satisfied by a travelling wave  $u(x, t) = \phi(x + ct)$ ,  $w(x, t) = \psi(x + ct)$ , and determine conditions on  $f(\phi)$  so that the resulting system has a unique rest state.

- 7.5 Derive the identities (7.3.5) and (7.3.6).
- 7.6 Suppose the function  $f(u)$  in Exercise 7.4 is specialised to

$$\begin{aligned}f(u) &= -u, \quad u \leq a/2, \\ &= u - a, \quad a/2 \leq u \leq (1+a)/2, \\ &= 1 - u, \quad (1+a)/2 \leq u.\end{aligned}$$

Does this system exhibit the same types of behaviour as found in the original FitzHugh–Nagumo system?

- 7.7 Suppose the function  $f(u)$  in Exercise 7.4 is specialised to

$$f(u) = -u + H(u - a).$$

Determine travelling periodic wave trains. In the notation of Section 7.4 construct solutions  $U(z)$  of period  $T$  satisfying the periodic, continuity and jump conditions

$$\begin{aligned} U(0^+) &= U(T^-) = a, \\ U_z(0^+) &= U_z(T^-), \\ U_{zz}(0^+) &= U_{zz}(T^-) - 1, \\ U(z_1^+) &= U(z_1^-) = a, \\ U_z(z_1^+) &= U_z(z_1^-), \\ U_{zz}(z_1^+) &= U_{zz}(z_1^-) + 1. \end{aligned}$$

For fixed  $a$ ,  $b$  and  $\gamma$  these conditions yield eight conditions for nine unknowns. Show that as  $T$  is varied one may solve, in principle, for the wave speed  $c$  and  $z_1$ . Compute the dispersion curve of  $c$  versus  $T$ .

- 7.8 Suppose the function  $f(u)$  is as given in Exercise 7.6. Show that in this case the FitzHugh–Nagumo system has three rest states provided

$$\frac{b}{\gamma} < \frac{1-a}{a+1}.$$

In this case construct travelling wave solutions connecting the rest states  $(0, 0)$  and  $(\frac{\gamma}{\gamma+b}, \frac{b}{\gamma+b})$ .

- 7.9 Under the conditions of Exercise 7.8 construct periodic wave solutions of period  $T$  to the FitzHugh–Nagumo system of the form

$$u(x, t) = U(kx + \omega t), \quad w(x, t) = W(kx + \omega t),$$

where the parameters  $k$  and  $\omega$  satisfy the conditions

$$k^2 = \frac{1-\gamma}{4\pi^2} T^2, \quad \omega^2 = \frac{b-\gamma^2}{4\pi^2} T^2.$$

- 7.10 Use the Poincaré Bendixson theorem to establish the existence of closed orbits to the FitzHugh–Nagumo system of Exercise 7.9.
- 7.11 Write MATLAB programs to compute travelling waves and periodic travelling waves to the FitzHugh–Nagumo system.



# Chapter 8

---

## Chemical Reactions

---

### 8.1 Wavefronts for the Belousov–Zhabotinskii reaction

In this chapter we discuss a simplification of the Belousov–Zhabotinskii reaction model described in Section 4.4. In particular, we give a qualitative analysis of the front of certain travelling concentration waves which have been observed frequently in experiments.

To begin with, we assume that the wave front depends primarily on the concentrations of bromous acid ( $HBrO_2$ ), which we have denoted by  $X$ , and the bromide ion ( $Br^-$ ) denoted by  $Y$  and to a lesser extent on the concentration of the oxidised state  $Ce(IV)$  denoted by  $Z$ . We also assume that the diffusion coefficients  $D_X, D_Y$  are constant and that  $D_X = D_Y = D$ . Furthermore, for simplicity, we consider only one space dimension  $x$ .

Therefore, if we neglect the concentration  $Z$  and take note of the above assumptions, the reaction–diffusion system (4.5.9)–(4.5.11) reduces to

$$\begin{aligned}\frac{\partial X}{\partial t} &= k_1AY - k_2XY + k_3AX - 2k_4X^2 + D\frac{\partial^2 X}{\partial x^2}, \\ \frac{\partial Y}{\partial t} &= -k_1AY - k_2XY + D\frac{\partial^2 Y}{\partial x^2}.\end{aligned}\quad (8.1.1)$$

For later purposes it is convenient to nondimensionalise (8.1.1) by setting

$$\begin{aligned}u &= \frac{2k_4X}{k_3A}, \quad v = \frac{k_2Y}{k_3Ar}, \\ x' &= \left(\frac{k_3A}{D}\right)^{1/2}x, \quad t' = k_3At, \\ L &= \frac{2k_4k_1}{k_2k_3}, \quad M = \frac{k_1}{k_3}, \quad b = \frac{k_2}{2k_4},\end{aligned}\quad (8.1.2)$$

where  $r$  is a suitable parameter. The reason for making these transformations is that the solutions  $u$  and  $v$ , which are of interest, lie in the interval  $0 \leq u, v \leq 1$ .

With the transformations (8.1.2), the system (8.1.1) takes the form

$$\begin{aligned}\frac{\partial u}{\partial t'} &= Lrv + u(1 - u - rv) + \frac{\partial^2 u}{\partial x'^2}, \\ \frac{\partial v}{\partial t'} &= Mv - buv + \frac{\partial^2 v}{\partial x'^2}.\end{aligned}$$

From experimental measurements it is found that  $r$  varies from between 10 to about 50 while

$$L \sim 8.4 \times 10^{-6}, \quad M \sim 2.1 \times 10^{-4} \quad \text{and} \quad b \sim 2.5 \times 10.$$

Since  $u$  and  $v$  are of order 1, the values of  $L$  and  $M$  allow us to neglect the first terms on the right of the above system. With this further simplification and removing the primes from  $x', t'$ , we arrive at the travelling front model:

$$\begin{aligned} \frac{\partial u}{\partial t} &= u(1 - u - rv) + \frac{\partial^2 u}{\partial x^2}, \\ \frac{\partial v}{\partial t} &= -buu + \frac{\partial^2 v}{\partial x^2}, \end{aligned} \quad (8.1.3)$$

where  $x \in (-\infty, \infty)$  and  $0 \leq u, v \leq 1$ .

The solutions we are interested in are those that satisfy the conditions

$$\begin{aligned} \lim_{x \rightarrow -\infty} u(x, t) &= 0 = \lim_{x \rightarrow \infty} v(x, t), \\ \lim_{x \rightarrow \infty} u(x, t) &= 1 = \lim_{x \rightarrow -\infty} v(x, t). \end{aligned} \quad (8.1.4)$$

One class of solutions that satisfy these conditions precisely are the desired travelling fronts. However, before discussing these further we observe that if  $u = 1 - v$  and  $b = 1 - r, r \leq 1$ , the system (8.1.3) reduces to the single equation

$$\frac{\partial u}{\partial t} = bu(1 - u) + \frac{\partial^2 u}{\partial x^2}, \quad (8.1.5)$$

where  $b = 1 - r$ . This equation is called **Fisher's equation** and arises in the study of population genetics. It also occurs in many other problems and we shall come across it again in a different context in Chapter 11. It is therefore sufficiently important to warrant some attention before going on to the general case.

## 8.2 Phase plane analysis of Fisher's equation

We seek a solution  $u$  of (8.1.5) in the form of a travelling wave, viz.

$$u(x, t) = \phi(x + ct), \quad (8.2.1)$$

where  $0 \leq \phi \leq 1$  and  $c (>0)$  is the wave speed.

Bearing in mind the conditions (8.1.4) we see that, as a function of  $\xi = x + ct$ ,  $\phi$  must satisfy

$$\lim_{\xi \rightarrow -\infty} \phi(\xi) = 0, \quad \lim_{\xi \rightarrow \infty} \phi(\xi) = 1. \quad (8.2.2)$$

On substituting (8.2.1) into (8.1.5) we see that  $\phi$  satisfies the nonlinear ordinary differential equation

$$\phi'' - c\phi' + b\phi(1 - \phi) = 0, \quad (8.2.3)$$

where the primes denote differentiation with respect to  $\xi$ .

By writing  $\phi' = \psi$ , (8.2.3) can be rewritten as the system

$$\begin{aligned} \psi' &= c\psi - b\phi(1 - \phi), \\ \phi' &= \psi. \end{aligned} \quad (8.2.4)$$

The singular points of this system in the  $(\phi, \psi)$  plane are  $(0, 0)$  and  $(1, 0)$ , respectively.

If we analyse the character of these singular points we find that  $(0, 0)$  is an unstable node if  $c \geq 2\sqrt{b}$ , a stable focus if  $0 < c < 2\sqrt{b}$  and a centre if  $c = 0$ . The point  $(1, 0)$  is a saddle-point for all  $c \geq 0$ .

Observe that since we are looking for the particular solution of (8.2.4) satisfying the condition  $0 \leq \phi \leq 1$  and (8.2.2), the range  $0 \leq c < 2\sqrt{b}$  is inadmissible, since in this situation  $\phi$  would become negative near the singular point  $(0, 0)$ . Thus for the required solution we must have  $c \geq 2\sqrt{b}$ .

In fact it can be shown that there exists a unique trajectory leaving  $(0, 0)$  and entering  $(1, 0)$ , which remains inside the strip  $(0 \leq \phi \leq 1, \psi \geq 0)$ , but we shall not prove this here.

### 8.3 Qualitative behaviour in the general case

Let us now look for wavefront solutions

$$u(x, t) = \phi(x + ct), \quad v(x, t) = \psi(x + ct), \quad (8.3.1)$$

to the system (8.1.3), (8.1.4). As before if we set  $\xi = x + ct$ , then  $\phi, \psi$  satisfy the equations

$$\begin{aligned} \phi'' - c\phi' + \phi(1 - \phi - r\psi) &= 0, \\ \psi'' - c\psi' - b\phi\psi &= 0, \end{aligned} \quad (8.3.2)$$

where

$$\begin{aligned} \lim_{\xi \rightarrow \infty} \phi(\xi) &= 1 = \lim_{\xi \rightarrow -\infty} \psi(\xi), \\ \lim_{\xi \rightarrow -\infty} \phi(\xi) &= 0 = \lim_{\xi \rightarrow \infty} \psi(\xi). \end{aligned} \quad (8.3.3)$$

It is possible to write (8.3.2) as an equivalent system of first-order equations. However, the resulting system is of fourth order and a phase plane analysis would be extremely difficult to perform. Instead, we shall treat (8.3.2) by

methods similar to those used to discuss the FitzHugh–Nagumo equations of Chapter 7.

Recall that we are looking for wavefronts  $\phi, \psi$  such that  $0 \leq \phi, \psi \leq 1$ . If we suppose  $\phi \geq 0$  then we shall show that  $\phi$  and  $\psi$  have the following properties:

$$(i) \quad \psi \geq 0, \quad (ii) \quad \psi' \leq 0, \quad (iii) \quad \phi' \geq 0. \quad (8.3.4)$$

In other words, we shall show that  $\phi$  is monotonic increasing from 0 to 1 while  $\psi$  is monotonic decreasing from 1 to 0 as  $\xi$  varies from  $-\infty$  to  $+\infty$ .

To prove (i) suppose  $\psi < 0$  for some range of  $\xi$ . Then since  $\psi(-\infty) = 1$  and  $\psi(\infty) = 0$  there must be a negative minimum for which  $\psi' = 0$  and  $\psi'' > 0$ . But from (8.3.2)  $\psi'' = b\phi\psi < 0$  at such a point and this contradicts the above inequalities. Thus we must have  $\psi \geq 0$  for all values of  $\xi$ .

To show that  $\psi' \leq 0$  we argue as follows. Since  $\phi \geq 0$  by assumption and we have just proved that  $\psi \geq 0$ , then from (8.3.2) we have

$$\psi'' - c\psi' = b\phi\psi \geq 0.$$

Now  $\psi$  cannot have a positive maximum since this would mean that for some point  $\xi_0$ ,  $\psi'(\xi_0) = 0$ ,  $\psi'' < 0$ , which is impossible. Thus either  $\psi' \geq 0$  or  $\psi' \leq 0$  for all  $\xi$ . Since  $\psi \geq 0$ ,  $\psi(-\infty) = 1$ ,  $\psi(\infty) = 0$ , we have  $\psi' \leq 0$ .

The statement (iii) that  $\phi' \geq 0$  is a little more involved and has to be proved in steps depending on whether  $0 < r < 1$ ,  $r = 1$  or  $r > 1$ . To begin with we note that  $1 - \phi - r\psi \neq 0$ , since if it were not so (8.3.2) would yield

$$\phi'' - c\phi' = 0$$

giving  $\phi = A + B \exp c\xi$ , which cannot satisfy the conditions (8.3.3) for any choice of  $A$  and  $B$ . Set

$$F = 1 - \phi - r\psi \quad (8.3.5)$$

and combine equations (8.3.2) to give

$$F'' - cF' - \phi F = -br\phi\psi \leq 0, \quad (8.3.6)$$

from which it follows that  $F$  cannot have a negative minimum. From the boundary conditions (8.3.3) we have

$$\lim_{\xi \rightarrow -\infty} F(\xi) = 1 - r, \quad \lim_{\xi \rightarrow \infty} F(\xi) = 0. \quad (8.3.7)$$

There are now three cases,  $0 < r < 1$ ,  $r = 1$ ,  $r > 1$ , to consider.

Take the case  $0 < r < 1$ ; then since  $F(-\infty) > 0$  and the fact that  $F$  cannot have a negative minimum we must have  $F \geq 0$  for all  $\xi$ . Furthermore,  $F$  cannot have a positive minimum followed by a positive maximum. Suppose this could happen; then there would be points  $\xi_{\min}, \xi_{\max}$  where  $\xi_{\min} < \xi_{\max}$  such that

$$F' = 0, \quad F'' > 0$$

at  $\xi = \xi_{\min}$ , i.e.,

$$F'' = \phi F - br\phi\psi > 0$$

or

$$F > br\psi$$

at  $\xi = \xi_{\min}$ . Similarly, at  $\xi = \xi_{\max}$  we have

$$F < br\psi$$

at  $\xi = \xi_{\max}$ .

Since we have supposed that  $F(\xi_{\max}) > F(\xi_{\min})$ , the above inequalities could only hold if  $\psi(\xi_{\max}) > \psi(\xi_{\min})$ , which is impossible because we have already proved that  $\psi' \leq 0$ . This contradiction proves the statement. It is possible, however, for  $F$  to have a positive maximum. In any case, we have proved that  $F \geq 0$  for all  $\xi$  and so from (8.3.2)

$$\phi'' - c\phi' = -\phi F \leq 0.$$

Suppose  $\phi$  has a positive maximum, then in order to satisfy the condition  $\phi \rightarrow 1$  as  $\xi \rightarrow \infty$ ,  $\phi$  would have to approach 1 from above, which would mean that  $F = 1 - \phi - r\psi < 0$  for some range of  $\xi$ . This is a contradiction and so  $\phi' \geq 0$  for all  $\xi$  and  $0 < r < 1$ . The possibility  $\phi' \leq 0$  is excluded by virtue of the boundary conditions (8.3.3).

Similar but more straightforward arguments are used to show that  $\phi' \geq 0$  for all  $\xi$  when  $r \geq 1$ .

Let us now turn to the problem of estimating the wavefront speed  $c$ . Throughout remember that  $0 \leq \phi, \psi \leq 1$ . Integrating the equations (8.3.2) from  $-\infty$  to  $\infty$  gives

$$c = \int_{-\infty}^{\infty} \phi(1 - \phi - r\psi) d\xi, \quad (8.3.8)$$

and

$$c = b \int_{-\infty}^{\infty} \phi\psi d\xi. \quad (8.3.9)$$

Substituting for

$$\int_{-\infty}^{\infty} \phi\psi d\xi$$

then gives

$$c = \frac{b}{b+r} \int_{-\infty}^{\infty} \phi(1 - \phi) d\xi. \quad (8.3.10)$$

Now as  $\xi \rightarrow -\infty$ , the boundary conditions (8.3.3) require  $\phi \rightarrow 0$  and  $\psi \rightarrow 1$  and so (8.3.2) behave near  $\xi = -\infty$  as the linearised equations

$$\phi'' - c\phi' + \phi(1 - r) = 0,$$

$$\psi'' - c\psi' - b\phi = 0,$$

the solutions of which are oscillatory unless

$$c^2 > 4(1 - r), \quad r \leq 1. \quad (8.3.11)$$

Oscillatory solutions are not possible asymptotically when  $r > 1$ . Thus (8.3.11) gives a lower bound on the wave speed.

In order to obtain more useful information about the wave speed  $c$  and in particular on its variation with  $b$ , we proceed as follows. Multiply the first equation in (8.3.2) by  $\phi$  and integrate to get

$$I \equiv \int_{-\infty}^{\infty} \phi'^2 d\xi = -\frac{c}{2} + \int_{-\infty}^{\infty} \phi^2(1 - \phi - r\psi) d\xi. \quad (8.3.12)$$

Similarly, repeating the process but multiplying first by  $\phi'$  gives

$$\begin{aligned} cI &= \int_{-\infty}^{\infty} \phi' \phi d\xi - \int_{-\infty}^{\infty} \phi' \phi^2 d\xi - r \int_{-\infty}^{\infty} \phi' \phi \psi d\xi \\ &= \frac{1}{2} - \frac{1}{3} - r \int_{-\infty}^{\infty} \phi' \phi \psi d\xi, \end{aligned}$$

i.e.,

$$I = \frac{1}{6c} - \frac{r}{c} \int_{-\infty}^{\infty} \phi' \phi \psi d\xi. \quad (8.3.13)$$

We now estimate the integral appearing on the right-hand side of equation (8.3.13) via **Schwarz's inequality** which states that if  $F$  and  $G$  are integrable then

$$\int_a^b FG d\xi \leq \left( \int_a^b F^2 d\xi \right)^{1/2} \left( \int_a^b G^2 d\xi \right)^{1/2}.$$

In our case we take  $a = -\infty, b = \infty, F = \phi', G = \phi \psi$  to obtain

$$\begin{aligned} \int_{-\infty}^{\infty} \phi' \phi \psi d\xi &\leq \left( \int_{-\infty}^{\infty} \phi'^2 d\xi \right)^{1/2} \left( \int_{-\infty}^{\infty} (\phi \psi)^2 d\xi \right)^{1/2} \\ &= I^{1/2} \left( \int_{-\infty}^{\infty} (\phi \psi)^2 d\xi \right)^{1/2} \\ &\leq I^{1/2} \left( \int_{-\infty}^{\infty} (\phi \psi) d\xi \right)^{1/2} \\ &= I^{1/2} \left( \frac{c}{b} \right). \end{aligned} \quad (8.3.14)$$

Substituting this into (8.3.13) results in the inequality

$$I > \frac{1}{6c} - \frac{r}{c} \left( \frac{c}{b} \right) I^{1/2}$$

or

$$I + \frac{r}{(cb)^{1/2}} I^{1/2} - \frac{1}{6c} > 0,$$

i.e.,

$$\begin{aligned} \left( I^{1/2} + \frac{r}{2(cb)^{1/2}} \right)^2 &> \frac{1}{6c} + \frac{r^2}{4cb} \\ &= \frac{1}{2c} \left( \frac{1}{3} + \frac{r^2}{2b} \right) \end{aligned}$$

and so

$$I^{1/2} > -\frac{r}{2(cb)^{1/2}} + \frac{1}{(2c)^{1/2}} \left( \frac{1}{3} + \frac{r^2}{2b} \right)^{1/2},$$

i.e.,

$$I > \frac{1}{4bc} \left[ \left( r^2 + \frac{2}{3}b \right)^{1/2} - r \right]^2. \quad (8.3.15)$$

Next, since  $0 \leq \phi, \psi \leq 1$  we have from (8.3.12)

$$\begin{aligned} I &= -\frac{c}{2} + \int_{-\infty}^{\infty} \phi^2(1-\phi)d\xi - r \int_{-\infty}^{\infty} \phi^2\psi d\xi, \\ &\leq -\frac{c}{2} + \int_{-\infty}^{\infty} \phi(1-\phi)d\xi, \\ &= -\frac{c}{2} + c \left( \frac{b+r}{b} \right), \end{aligned}$$

i.e.,

$$I < \frac{c}{2b}(b+2r). \quad (8.3.16)$$

Combining (8.3.16) and (8.3.15) to eliminate  $I$  we find that

$$\frac{1}{4bc} \left[ \left( r^2 + \frac{2}{3}b \right)^{1/2} - r \right]^2 \leq \frac{c}{2b}(b+2r)$$

from which it follows that

$$c^2 \geq \frac{\left[ \left( r^2 + \frac{2}{3}b \right)^{1/2} - r \right]^2}{2(b+2r)}.$$

Numerical results show that  $c^2 \leq 4$  and so we have the final result

$$\frac{\left[ \left( r^2 + \frac{2}{3}b \right)^{1/2} - r \right]^2}{2(b+2r)} \leq c^2 \leq 4. \quad (8.3.17)$$

Again from numerical results (8.3.11) is a sharper lower bound for  $c^2$  when  $0 < r < 1$ .

## 8.4 Spiral waves and $\lambda - \omega$ systems

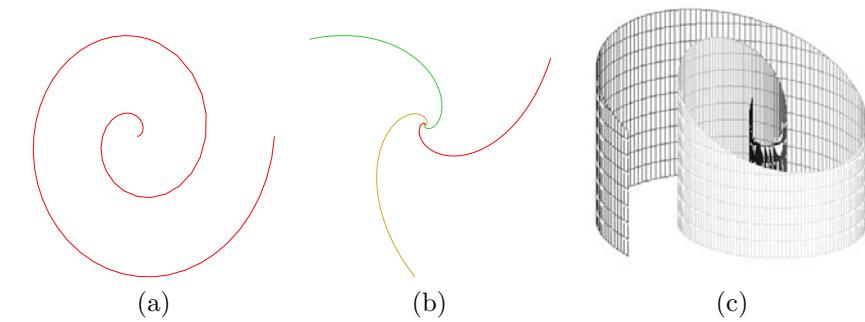
Spiral waves occur in a wide variety of biological, physiological and chemical contexts. In the case of the latter the Belousov–Zhabotinskii mechanism has been shown both experimentally and mathematically to exhibit many fascinating spiral type waves in two dimensions, as well as in three dimensions where they appear as scroll waves. The form of spiral wave observed depends very much on the chemical concentrations used as well as the experimental set up. Numerous examples of spiral waves arising in physiology can be found in the book by Keener and Sneyd (see the notes at the end of this chapter). For example, brain tissue can exhibit electrochemical waves of spreading depression which spread through the cortex. These waves are typified by depolarisation of the neuronal membrane and decreased neural activity. Indeed spiral waves have been seen to rotate about a lesion in brain tissue from the cortex of a rat. In another context it has been suggested that sudden cardiac death could involve three-dimensional spiral type waves of electrical activity, which suddenly become unstable when the heart thickness exceeds some critical value. In biology spiral waves have also been observed in the signalling patterns of the slime mould *Dictyostelium discoideum*.

In mathematical terms a spiral wave is a rotating time periodic solution of an appropriate system of reaction–diffusion equations. At a given time the solution appears as a typical spiral pattern and, as time varies, the whole spiral moves like a rotating bicycle wheel with curved spokes. Suppose the concentration of a reactant is described by its phase,  $\phi$ . Then on using polar coordinates  $r$  and  $\theta$  we can write down a simple spiral wave as a periodic function of  $\phi$  given by

$$\phi = \Omega t \pm m\theta + \psi(r), \quad (8.4.1)$$

where  $\Omega$  is the frequency,  $m$  is the number of arms of the spiral,  $\psi(r)$  is a function of  $r$  which describes the type of spiral and  $\pm$  in the  $m\theta$  term specifies the sense of rotation. Suppose we set  $\phi = 0$  and consider the steady state then we can get a description of spiral forms by specifying the function  $\psi(r)$  in (8.4.1). For example if we choose  $\psi(r) = ar$  and  $m = 1$  with the constant  $a > 0$  we obtain the one-armed Archimedean spiral shown in Figure 8.4.1(a). Similarly if  $\psi(r) = a \ln r$  and  $m = 3$  we obtain the three-armed logarithmic spiral shown in Figure 8.4.1(b). A three-dimensional spiral of scroll type is also shown in Figure 8.4.1(c).

The study of spiral wave solutions to general systems of reaction diffusion equations has been widely pursued and often involves quite complex mathematical analysis. A particular system which exhibits the main features of spiral wave solutions and is more accessible to analytic treatment is the model  $\lambda - \omega$  system.



**FIGURE 8.4.1:** Spiral wave formations: (a) one-armed Archimedean spiral; (b) three-armed logarithmic spiral; (c) three-dimensional scroll-type spiral.

This system for two reactants takes the form:

$$\frac{\partial}{\partial t} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \lambda(a) & -\omega(a) \\ \omega(a) & \lambda(a) \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + D \nabla^2 \begin{pmatrix} u \\ v \end{pmatrix}, \quad (8.4.2)$$

where  $u$  and  $v$  are the concentrations of two reactants,  $a^2 = u^2 + v^2$ ,  $\omega(a)$  and  $\lambda(a)$  are real functions of  $a$ , and  $\nabla^2$  denotes the **Laplacian operator**, defined in two dimensions by

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}. \quad (8.4.3)$$

The first and crucial step in finding spiral wave solutions is to ensure that when we set  $D = 0$  in (8.4.2) the resulting system of ordinary differential equations has a stable limit cycle. This requirement means that  $\lambda(a)$  must have an isolated zero at  $a = a_0 > 0$ ,  $d\lambda/da < 0$  at  $a = 0$  and  $\omega(a_0) \neq 0$ . The limit cycle is given by  $u^2 + v^2 = a_0$  with cycle frequency  $\omega(a_0)$ .

If we introduce the complex variable  $w = u + iv$  then we can write (8.4.2) as the single complex equation

$$\frac{\partial w}{\partial t} = (\lambda + i\omega)w + D \nabla^2 w. \quad (8.4.4)$$

We now look for a solution of the form

$$w = A \exp(i\phi), \quad (8.4.5)$$

where  $A$  is the amplitude of  $w$  and  $\phi$  its phase. Substituting this into (8.4.4) and equating real and imaginary parts we obtain the following pair of equa-

tions for  $A$  and  $\phi$ :

$$\begin{aligned}\frac{\partial A}{\partial t} &= A\lambda(a) - DA |\nabla\phi|^2 + D\nabla^2 A, \\ \frac{\partial\phi}{\partial t} &= \omega(a) + 2A^{-1}D(\nabla A \cdot \nabla\phi) + D\nabla^2\phi.\end{aligned}\quad (8.4.6)$$

Since polar coordinates are appropriate for the study of spiral waves we look for solutions to (8.4.6) of the form

$$A = A(r), \quad \phi = \Omega t + m\theta + \psi(r), \quad (8.4.7)$$

where  $\Omega$  is the unknown frequency and  $m$  is the number of arms of the spiral. On substituting these forms into equation (8.4.6) we obtain the system of ordinary differential equations:

$$\begin{aligned}D\frac{d^2A}{dr^2} + \frac{D}{r}\frac{dA}{dr} + \left(\lambda(a) - D\left(\frac{d\psi}{dr}\right)^2 - \frac{Dm^2}{r^2}\right)A &= 0 \\ D\frac{d^2\psi}{dr^2} + D\left(\frac{1}{r} + \frac{2}{A}\frac{dA}{dr}\right)\frac{d\psi}{dr} &= \Omega - \omega(a).\end{aligned}\quad (8.4.8)$$

Multiplying the second of these equations by  $rA^2$  and integrating gives

$$\frac{d\psi}{dr}(r) = \frac{1}{DrA^2(r)} \int_0^r sA^2(s)[\Omega - \omega(A(s))]ds. \quad (8.4.9)$$

The next step before trying to solve the system (8.4.8) is to decide on appropriate boundary conditions. Clearly we need solutions to be regular at  $r = 0$  and to remain bounded as  $r \rightarrow \infty$ . So from the form of equations (8.4.8) and (8.4.9) we conclude that

$$A(0) = 0, \quad \frac{d\psi}{dr}(0) = 0. \quad (8.4.10)$$

Now suppose  $A \rightarrow A_\infty$  as  $r \rightarrow \infty$ , then from (8.4.9) we have the asymptotic approximation

$$\begin{aligned}\frac{d\psi}{dr} &\sim \frac{1}{DrA_\infty^2} \int_0^r sA_\infty^2[\Omega - \omega(A_\infty)]ds \\ &= \frac{(\Omega - \omega(A_\infty))r}{2D}.\end{aligned}\quad (8.4.11)$$

From this we see that  $d\psi/dr$  is bounded only if

$$\Omega = \omega(A_\infty). \quad (8.4.12)$$

If we let  $r \rightarrow \infty$  in the first of equations (8.4.8) we find that

$$\frac{d\psi}{dr}(\infty) = \left(\frac{\lambda(A_\infty)}{D}\right)^{1/2}. \quad (8.4.13)$$

The above pair of equations is the dispersion relation and shows how the amplitude of the spiral at infinity determines the frequency.

To get some idea of the structure of the spiral near  $r = 0$  we look for a solution of the form

$$A(r) = r^c \sum_{n=0}^{\infty} a_n r^n, \quad a_0 \neq 0. \quad (8.4.14)$$

Substituting (8.4.14) into the first of equations (8.4.8) and equating powers of  $r$  we see that the lowest-order term requires

$$c(c - 1) + c - m^2 = 0 \quad (8.4.15)$$

and so for  $A$  to be regular at  $r = 0$  we must choose the positive root  $c = m$ . Finally from equations (8.4.5) and (8.4.7) the structure of  $u$  and  $v$  near  $r = 0$  is found to be

$$\begin{pmatrix} u \\ v \end{pmatrix} \approx \begin{pmatrix} r^m \cos(\Omega t + m\theta + \psi(0)) \\ r^m \sin(\Omega t + m\theta + \psi(0)) \end{pmatrix} \quad (8.4.16)$$

Analytic solutions to the system (8.4.8) are difficult to obtain and so insight into the structure of spiral waves is gained by numerical analysis or by investigating their behaviour by carrying out an approximate analysis as indicated above near  $r = 0$  or as  $r \rightarrow \infty$  (see Exercise 8.8 ).

## 8.5 Notes

For further information about the Belousov–Zhabotinskii reaction and many related topics see J. D. Murray, *Mathematical Biology*, Springer-Verlag, Heidelberg, 2003.. Many examples of spiral waves arising in physiology are discussed in J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, Heidelberg, 1998. A general geometrical theory approach to the study of spiral waves is given in P. Grindrod, *The Theory and Applications of Reaction-Diffusion Equations, Patterns and Waves*, Oxford University Press, Oxford, 1996. See also A. T. Winfrey, *When Time Breaks Down: The Three-Dimensional Dynamics of Electrochemical Waves and Cardiac Arrhythmias*, Princeton University Press, Princeton, NJ, 1987.

## Exercises

- 8.1 Show that the transformations (8.1.2) reduce equations (8.1.1) to the equations (8.1.3).

- 8.2 Analyse the behaviour of trajectories for the system (8.2.4) in the neighbourhood of the singular points  $(0, 0)$ ,  $(1, 0)$  for all positive values of the wavefront speed  $c$ .
- 8.3 Consider the system (8.3.2), (8.3.3). Use the arguments of Section 8.3 to show that  $\phi' \geq 0$  for  $\xi \in (-\infty, \infty)$  when  $r \geq 1$ .
- 8.4 In the inequality (8.3.17) show that as  $b \rightarrow 0$  or  $r \rightarrow \infty$  it reduces to  $c^2 \geq 0$  and that if  $b \rightarrow \infty$  or  $r \rightarrow 0$  the inequality reduces to  $c^2 \geq 1/3$ . In the latter case, compare the result with that predicted by the inequality (8.3.11).
- 8.5 Travelling bands of microorganisms, chemotactically directed, move into a food source, consuming it as they go. A mathematical model for this is given by

$$\frac{\partial a}{\partial t} = -kb, \quad \frac{\partial b}{\partial t} = \frac{\partial}{\partial x} \left[ D \frac{\partial b}{\partial x} - \frac{b\chi}{a} \frac{\partial a}{\partial x} \right],$$

where  $b(x, t)$  and  $a(x, t)$  are the concentrations of bacteria and nutrient, respectively, and  $D$ ,  $\chi$  and  $k$  are positive constants. Look for travelling wave solutions, as functions of  $z = x - ct$ , where  $c$  is the wave speed, with boundary conditions  $b \rightarrow 0$  as  $|z| \rightarrow \infty$ ,  $a \rightarrow 0$  as  $z \rightarrow -\infty$ ,  $a \rightarrow 1$  as  $z \rightarrow \infty$ . Hence show that  $b(z)$  and  $a(z)$  satisfy the ordinary differential equations

$$\frac{db}{dz} = \frac{b}{cD} \left[ \frac{kb\chi}{a} - c^2 \right], \quad \frac{da}{dz} = \frac{kb}{c}.$$

Find a differential relationship between  $b(z)$  and  $a(z)$ .

In the case where  $\chi = 2D$  show that

$$a(z) = \left[ 1 + Ke^{-cz/D} \right]^{-1}, \quad b(z) = \frac{c^2}{kD} e^{-cz/D} \left[ 1 + Ke^{-cz/D} \right]^{-2},$$

where  $K$  is an arbitrary constant which is equivalent to a linear translation and may be set to 1. Sketch the wave solutions and explain what is happening biologically.

- 8.6 In the absence of diffusion show that equation (8.4.2) has a limit cycle provided  $\lambda(a)$  has an isolated zero at some value  $a = a_0 > 0$ ,  $\lambda'(a_0) < 0$  and  $\omega(a_0) \neq 0$ .
- 8.7 Determine the dispersion relations for (a)  $\lambda(A) = 1 - A^2$ ,  $\omega(A) = -\beta A^2$ ,  $\beta > 0$ , (b)  $\lambda(A) = 1 - A$ ,  $\omega(A) = 1 + \alpha(A - 1)$ ,  $\alpha > 0$ .
- 8.8 To investigate the structure of spiral waves as  $r \rightarrow \infty$  assume that  $A$  and  $\psi'$  are approximated by

$$A(r) = A_0 + \frac{A_1}{r^2}, \quad \psi'(r) = \frac{B_0}{r}.$$

By substituting these expressions into equation (8.4.8), in the case where  $\lambda(A) = 1 - A^2$ ,  $\omega(A) = -\beta A^2$ , show that the unknown coefficients are given by

$$A_0 = 1, \quad A_1 = -\frac{Dm^2 + c^2}{2}, \quad B_0 = c,$$

where

$$c = \frac{1}{D} \int_0^\infty s A^2(s) [\Omega + \beta A^2(s)] ds.$$

- 8.9 The 1-dimensional analogue of a spiral wave is a pulse emitted from the core. For the choices  $\lambda(a) = 1 - A^2$ ,  $\omega(a) = \gamma A^2$  and  $D = 1$  the system (8.4.8) takes the form

$$\begin{aligned} \frac{d^2 A}{dx^2} + A \left( 1 - A^2 - \left( \frac{d\psi}{dx} \right)^2 \right) &= 0, \\ \frac{d^2 \psi}{dx^2} + \frac{2}{A} \frac{dA}{dx} \frac{d\psi}{dx} &= \Omega - \gamma A^2. \end{aligned}$$

By imposing the boundary conditions

$$A(x) \sim A_0 x \text{ as } x \rightarrow \infty, \quad \frac{d\psi}{dx}(0) = 0,$$

$$A(x), \quad \frac{d\psi}{dx} \text{ bounded as } x \rightarrow \infty,$$

show that a solution is

$$\begin{aligned} A(x) &= \left( \frac{\Omega}{\gamma} \right)^{1/2} \tanh(x\sqrt{2}), \\ \frac{d\psi}{dx}(x) &= \left( 1 - \frac{\Omega}{\gamma} \right)^{1/2} \tanh(x\sqrt{2}), \quad \Omega^2 + \frac{9\Omega}{2\gamma} - \frac{9}{2} = 0. \end{aligned}$$



# Chapter 9

---

## Predator and Prey

---

### 9.1 Catching fish

Often one species uses another as food. For example, man uses fish, lions consume gazelles and some wasps eat the caterpillars of moths. The presence of one species can therefore have effects (sometimes irreversible) on another. The study of such dynamic interactions is the topic of this chapter. For simplicity, consideration will be limited to two species, though it must be understood that in nature the situation is frequently much more complicated, with several species involved and interacting in a complex manner. Nevertheless, useful conclusions can be drawn from even relatively simple models and the influence of various actions on the level of available resources ascertained.

Usually, the species regarded as food is called the **prey** and the consuming species the **predator**. The aim is to construct predator–prey models that are relevant, whether it is man killing deer, wasps eating caterpillars or man harvesting timber, though humans often behave differently from other organisms; some general aspects have already been treated in Section 4.5. To fix on a specific illustration for the simplest model, we shall discuss how fishing can affect the population of fish.

Let  $N(t)$  be the number of fish in a designated zone at time  $t$ . It will be assumed that  $N$  varies continuously with  $t$ ; in fact,  $N$  changes only by integers but the error introduced by our approximation should not be appreciable except possibly when  $N$  is small. In the absence of fishing, new fish can be expected to be born at a rate proportional to existing numbers, say  $bN(t)$  where  $b$  is a constant. Similarly, the rate at which fish die will be taken as  $dN(t)$  with  $d$  constant. Then

$$\frac{dN}{dt} = (b - d)N. \quad (9.1.1)$$

The solution of this differential equation has been found in Section 1.1 and is

$$N(t) = N(0)e^{(b-d)t} \quad (9.1.2)$$

where  $N(0)$  is the population at time  $t = 0$ .

The solution (9.1.2) reveals that the fish population eventually disappears if the birth rate is less than the death rate, i.e.,  $b < d$ , is in equilibrium if  $b = d$

and increases exponentially if  $b > d$ , i.e., the reproductive rate is greater than the mortality rate. Exponential growth in which the population doubles in every interval  $0.69/(b - d)$  of time is indeed exhibited by many species under ideal conditions when they have boundless space, food and the resources they need available. However, such growth cannot continue indefinitely in a finite world and there must come a time when shortages of supplies inhibit the exponential growth.

The preparation of models that allow for environmental constraints has followed a variety of routes. One of the most popular is to replace (9.1.1) by

$$\frac{dN}{dt} = a \left(1 - \frac{N}{N_0}\right) N \quad (9.1.3)$$

where  $a$  and  $N_0$  are positive constants. According to Section 1.3, this leads to logistic growth and the population always ends up at  $N_0$  no matter what level it started at. Thus  $N_0$  can be regarded as the maximum population that can be sustained under (9.1.3). Equation (9.1.3) often appears in ecology in the form

$$\frac{dN}{dt} = r \left(1 - \frac{N}{K}\right) N.$$

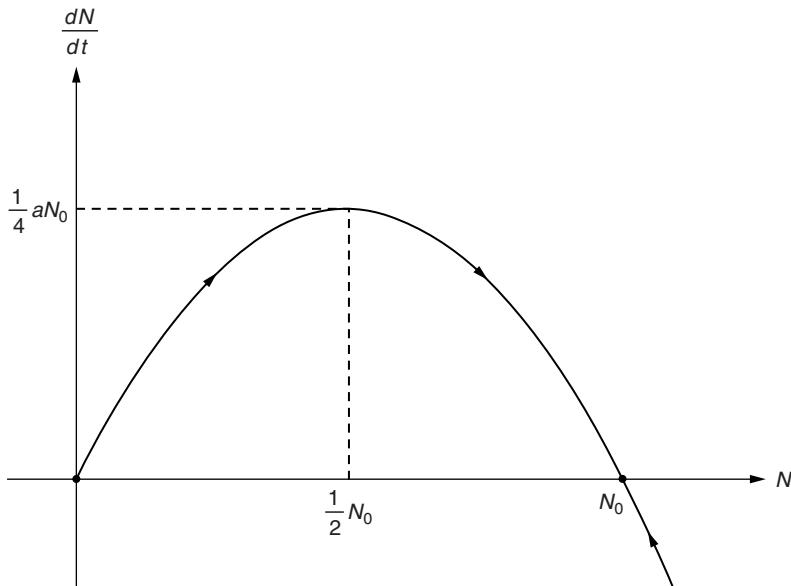
Whether (9.1.3) is applicable to any particular population depends upon either producing a convincing argument for the presence of each term or comparing its predictions with actual observations of the numbers in a population. For instance, yeast grown in laboratory cultures follows the logistic curve admirably. On the other hand, the human population of the United States, while fitting the logistic curve well from 1790 to 1910, grew much more strongly than predicted from 1920 onwards and departed widely from the curve. In general, it seems that the more complex the life history of an organism the less likely its population is to sit on the logistic curve. This notwithstanding, there are often sufficient periods of time for which the logistic curve can be applied to justify retaining it as a model.

Another view is that environmental factors do not instantaneously change the population but take some time to permeate through it. In that case the rate of change depends upon the size of the population at an earlier time and a possible model is

$$\frac{dN(t)}{dt} = a \left(1 - \frac{N(t-T)}{N_0}\right) N(t) \quad (9.1.4)$$

where  $T$  is a constant representing the delay that transpires before effects are felt. Equation (9.1.4) is commonly solved on the computer but its solution bears little resemblance to the logistic curve in many circumstances and can display oscillatory behaviour.

Although (9.1.3) can be solved analytically it is desirable to see what can be deduced by employment of the phase plane since that method may be available when analytical solution is precluded. Note first that there is equilibrium

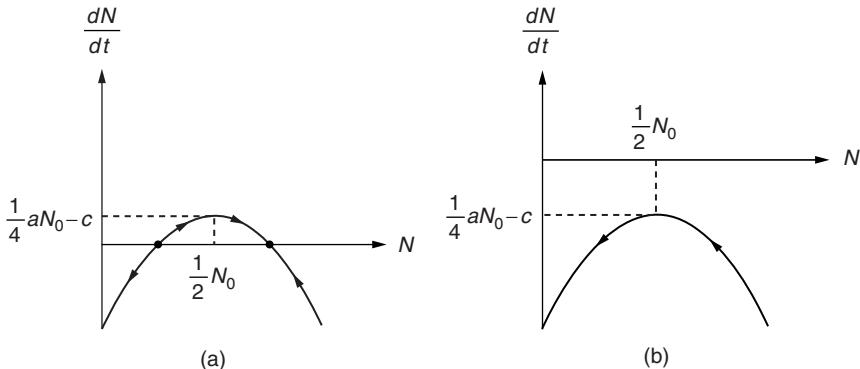


**FIGURE 9.1.1:** The trajectory for logistic growth.

when  $dN/dt = 0$ , i.e., when  $N = 0$  or  $N = N_0$ . Secondly, if  $dN/dt$  is plotted against  $N$  in the phase plane the trajectory obtained is the parabola shown in Figure 9.1.1. The parabola has a maximum of  $\frac{1}{4}aN_0$  at  $N = \frac{1}{2}N_0$ . Also  $dN/dt$  is positive when  $N$  is slightly less than  $N_0$  and negative when  $N$  is a bit larger than  $N_0$ . Consequently the directions on the trajectory are those indicated by the arrows on the diagram. From these it is clear that  $N$  always approaches the equilibrium value  $N_0$  whereas a slight disturbance will cause it to depart from the equilibrium at  $N = 0$ . Thus  $N = N_0$  corresponds to stable equilibrium and  $N = 0$  to unstable equilibrium. Evidently, the main features of the behaviour can be extracted from the phase plane as well as from the explicit solution.

## 9.2 The effect of fishing

A stable environment allows the fish population to adopt a pattern as set out in the previous section, but the disturbance caused by removing some of the fish for food may be profound. Some idea of how the population is affected and whether equilibrium still occurs can be achieved by supposing the fish are caught at the constant rate  $c$  but the population is otherwise governed by (9.1.3). Then



**FIGURE 9.2.1:** The trajectory for constant fishing rate: (a)  $c < \frac{1}{4}aN_0$ ; (b)  $c > \frac{1}{4}aN_0$ .

$$\frac{dN}{dt} = a\left(1 - \frac{N}{N_0}\right)N - c. \quad (9.2.1)$$

The differential equation (9.2.1) can be solved explicitly but, instead of finding the analytical solution, we shall consider it in the phase plane. The corresponding parabolic trajectory is displayed in Figure 9.2.1. If  $c < \frac{1}{4}aN_0$  there are two points of equilibrium of which the lower is unstable and the higher stable. Therefore, if

$$N(0) > \frac{1}{2}N_0\{1 - (1 - 4c/aN_0)^{1/2}\},$$

the population will eventually arrive at the value of  $\frac{1}{2}N_0\{1 + (1 - 4c/aN_0)^{1/2}\}$ . In contrast, if

$$N(0) < \frac{1}{2}N_0\{1 - (1 - 4c/aN_0)^{1/2}\},$$

the fish will steadily decrease in number until they disappear altogether.

When  $c > \frac{1}{4}aN_0$  it is transparent from Figure 9.2.1(b) that there are no points of equilibrium so that the fish population steadily decays and dies out.

For  $c = \frac{1}{4}aN_0$  the picture is similar to Figure 9.2.1(b) but the parabola just touches the  $N$ -axis at  $N = \frac{1}{2}N_0$ . So there is equilibrium at  $N = \frac{1}{2}N_0$  but any perturbation which tends to reduce  $N$  will result in the fish becoming extinct.

We conclude that, if the fish population is not to be decimated, the catching rate must not exceed  $\frac{1}{4}aN_0$  and the initial population must not be too small. If the catching rate is near the maximum the population will be reduced to about half the natural level of  $N_0$  and will be subject to extinction if any other adverse factor enters the scene. It is therefore advisable to have the catching rate significantly below the maximum permitted.

The above simple model has led to a useful recommendation on the catching rate to be adopted if the fish population is to survive, but it must be recognised that the model becomes increasingly unrealistic as the population falls. The scarcer the fish become the more effort has to be expended in catching the same number of fish per unit time. Since there is a limit on the resources for fishing, the catch is likely to become proportional to the fish population when it drops below a certain value. Moreover, at low levels the population is likely to be in a state where exponential growth is pertinent if fishing were to cease. Therefore, when considering survival prospects, the appropriate model is

$$\frac{dN}{dt} = (b - d - f)N \quad (9.2.2)$$

where  $f$  is the factor which accounts for the loss due to fishing. There is equilibrium if  $f = b - d$  so that the population will not vary if the catch is restricted to  $(b - d)N$ . Fishing at a greater rate will lead to extinction. By putting  $f = 0$  in (9.2.2) we can determine how long it will take for the population to attain an acceptable level when there is no fishing.

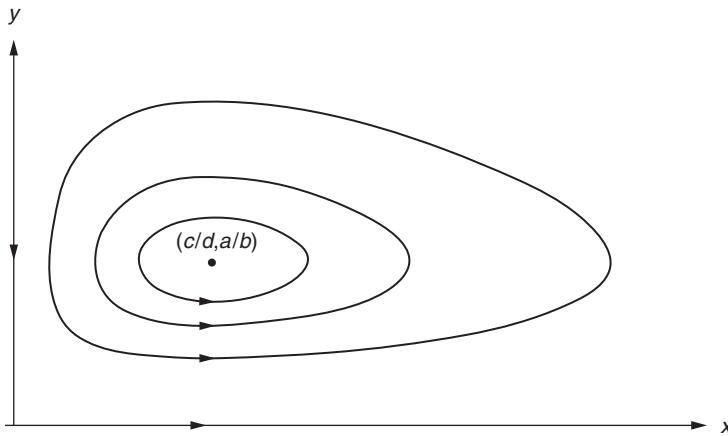
Before any model can be applied in a practical context, it is necessary to be able to measure the parameters that occur. This is a far from trivial task. For example, an estimate of  $b$  may require a consideration of the breeding cycle, the sex ratio at birth and environmental factors. Again, the mortality rate may be different for infants, adults and different sexes. Furthermore, there is a distinct possibility that  $b$  and  $d$  vary with  $N$ . We shall say no more than that adequate estimates are feasible for some species because we now want to examine the complicated matter of interaction between species.

### 9.3 The Volterra–Lotka model

Imagine that there is an island occupied by humans whose sole source of food consists of fish. This is an example of a predator–prey problem and, to illustrate the ideas, we shall consider a very much simplified model in which humans do not evolve new methods of fishing as the fish population changes. Let  $x(t)$  be the population of fish at time  $t$  and  $y(t)$  that of humans. Assume that the fish, when undisturbed, grow exponentially and that, in the absence of fish, the starvation rate of the human population is greater than the birth rate so that it decays exponentially. The number of fish consumed will depend upon the frequency of encounter with humans. If we assume that, when they meet, there is a constant probability that the human will catch a fish and the time taken to eat it is negligible we obtain

$$\dot{x} = ax - bxy, \quad (9.3.1)$$

$$\dot{y} = -cy + dxy, \quad (9.3.2)$$



**FIGURE 9.3.1:** Phase plane for a predator–prey problem.

where  $a, b, c$  and  $d$  are all positive constants. Notice that only solutions in which  $x \geq 0$  and  $y \geq 0$  are of interest.

The system of (9.3.1) and (9.3.2) has critical points where

$$ax - bxy = 0, \quad -cy + dxy = 0,$$

i.e.,  $x = 0$  and  $y = 0$  or  $x = c/d$  and  $y = a/b$ . By linearisation about  $x = 0$  and  $y = 0$  as in Section 5.3, we obtain the system

$$\dot{\xi} = a\xi, \quad \dot{\eta} = -c\eta,$$

which shows that the origin is a saddle-point (Section 5.4). Linearisation near  $(c/d, a/b)$  leads to

$$\dot{\xi} = -bc\eta/d, \quad \dot{\eta} = da\xi/b$$

showing that this critical point is a centre. Also when  $x = c/d$ , (9.3.1) reveals that  $\dot{x}$  is positive and  $x$  increasing when  $y < a/b$ . Therefore the trajectories have the structure of Figure 9.3.1.

The precise equations of the trajectories can be derived by observing that

$$\frac{dy}{dx} = \frac{(dx - c)y}{(a - by)x}$$

which is a separable first-order differential equation. Its solution is

$$a \ln y - by + c \ln x - dx = K \quad (9.3.3)$$

where  $K$  is an arbitrary constant. The left-hand side has a minimum at  $(c/d, a/b)$  and so the contours (9.3.3) must form closed curves in the neighbourhood of this point, in agreement with its being a centre and with Figure 9.3.1.

Since a typical trajectory is a closed curve, the variation of  $x$  and  $y$  with time must be periodic unless either  $x$  or  $y$  is initially zero, a possibility that

will be excluded from further consideration. It is clear from Figure 9.3.1 that a maximum (minimum) in  $y$  occurs about a quarter period after a maximum (minimum) in  $x$ . An interpretation is that when the human population is at a maximum the fish population is declining and that decline induces a drop in the number of humans. The reduction in predators allows the fish to thrive followed by an increase in the humans and the cycle repeats itself. The actual levels attained depend upon which trajectory is traced. A change in the environment may switch the system from one trajectory to another but the periodic fluctuation continues with no tendency to an equilibrium state.

In place of exponential growth for the fish when there are no predators, it can be supposed that the growth is logistic so that

$$\dot{x} = ax - gx^2 - bxy, \quad (9.3.4)$$

$$\dot{y} = -cy + dxy, \quad (9.3.5)$$

which includes (9.3.1) as a special case when  $g = 0$ . The system (9.3.4) and (9.3.5) is said to constitute a **Volterra–Lotka model** for the predator–prey problem. It can be discussed in the phase plane in a similar fashion to that when  $g = 0$ .

Variants to the Volterra–Lotka model have been proposed in order to incorporate other facets of the interaction. In one version it is supposed that  $x_0$  of the fish can find a refuge that makes it impossible for them to be caught. The modification to (9.3.4) and (9.3.5) is then

$$\dot{x} = ax - gx^2 - by(x - x_0), \quad (9.3.6)$$

$$\dot{y} = -cy + dy(x - x_0). \quad (9.3.7)$$

On the other hand, more general assumptions have been discussed. For example, the **Rosenzweig–MacArthur model** has

$$\dot{x} = f(x) - h(x, y), \quad (9.3.8)$$

$$\dot{y} = -cy + kh(x, y), \quad (9.3.9)$$

where  $f$  and  $h$  are appropriate functions. Again logistic growth might be preferred for  $y$  and (9.3.5) replaced by

$$\dot{y} = c'y - g'y^2 + dxy,$$

$c'$  and  $g'$  being constants.

An indication of how information can be extracted from the phase plane can be given by examining

$$\dot{x} = (\lambda(x) - \mu(x, y))x, \quad (9.3.10)$$

$$\dot{y} = (v(y) + \rho(x, y))y. \quad (9.3.11)$$

Since predators consume prey we expect that  $\partial\mu/\partial y > 0$ . Likewise, if the number of prey increases there will be more prey per predator, so  $\partial\rho/\partial x > 0$ .

As the predators increase there will be less prey per predator and so

$$\frac{\partial}{\partial y}(\nu + \rho) < 0.$$

We shall also assume that

$$\frac{\partial}{\partial x}(\lambda - \mu) \leq 0,$$

consistent with (9.3.1) and (9.3.4).

If there is a point of equilibrium  $(x_0, y_0)$  where neither species is extinct, linearisation gives

$$\dot{\xi} = \alpha\xi + \beta\eta, \quad \dot{\eta} = \gamma\xi + \delta\eta$$

where

$$\begin{aligned}\alpha &= x_0 \frac{\partial}{\partial x}(\lambda - \mu), & \beta &= -x_0 \frac{\partial}{\partial y}(\nu + \rho), \\ \gamma &= y_0 \frac{\partial}{\partial x}(\lambda - \mu), & \delta &= y_0 \frac{\partial}{\partial y}(\nu + \rho),\end{aligned}$$

the partial derivatives being evaluated at  $(x_0, y_0)$ . Thus  $\alpha \leq 0, \beta < 0, \gamma > 0$  and  $\delta < 0$ . Consequently, the equilibrium is stable, being either a node or a focus (see Section 5.4). It is a focus (i.e., there are oscillations) if and only if

$$\left( x_0 \frac{\partial}{\partial x}(\lambda - \mu) - y_0 \frac{\partial}{\partial y}(\nu + \rho) \right)^2 - 4x_0y_0 \frac{\partial\mu}{\partial y} \frac{\partial\rho}{\partial x} < 0. \quad (9.3.12)$$

This can be met if the first term is not too large. If  $\alpha = \delta = 0$  were also an allowable case (as for example in (9.3.1), (9.3.2)), the equilibrium point would be a centre and periodic fluctuations would be likely. Another possible case is if

$$\frac{\partial}{\partial x}(\lambda - \mu) > 0.$$

This opens up the possibility of the equilibrium being unstable. In particular, if

$$x_0 \frac{\partial}{\partial x}(\lambda - \mu) + y_0 \frac{\partial}{\partial y}(\nu + \rho) > 0,$$

and condition (9.3.12) holds, then the equilibrium is an unstable centre and it is possible that there is a stable limit cycle. Whether any of these states could be reached from specified initial conditions would depend on global considerations, which have been mentioned in Section 5.6.

The discussion of this section has centred on the phase plane but the reader should be aware that a change of variable can sometimes lead to a simpler situation. For instance, the obvious choice to reduce the second-order differential equation

$$\ddot{x} - 12x\dot{x} + 16x^3 = 0$$

to a system of first order is

$$\dot{x} = y, \quad \dot{y} = 12xy - 16x^3,$$

but one might be better off with the choice

$$\dot{x} = 2(x^2 - y^2), \quad \dot{y} = 4xy.$$

Moreover, the phase plane can be helpful in other ways. The differential equation

$$\frac{dy}{dx} = \frac{x+y}{x-y}$$

is of homogeneous type and can be solved as in Section 1.5. The solution is

$$y = x \tan\{\ln[K'(x^2 + y^2)^{1/2}]\},$$

$$K' = (x_0^2 + y_0^2)^{-1/2} \exp\{\tan^{-1}(y_0/x_0)\},$$

when  $y = y_0$  at  $x = x_0$ . However, in the phase plane we have

$$\dot{x} = x - y, \quad \dot{y} = x + y$$

with solution

$$x = e^t(x_0 \cos t - y_0 \sin t), \quad y = e^t(y_0 \cos t + x_0 \sin t)$$

which is much easier to handle than the previous form.

### Example 9.3.1

Show that the predator-prey system

$$\dot{x} = \left(1 - x - \frac{ay}{x+b}\right)x, \tag{9.3.13}$$

$$\dot{y} = c\left(1 - \frac{y}{x}\right)y, \tag{9.3.14}$$

with  $a = 1$  and  $b = \frac{1}{10}$  has a stable equilibrium if  $c > c^*$  and a stable limit cycle if  $0 < c < c^*$ , where  $c^*$  is some threshold value.

This system corresponds to (9.3.10), (9.3.11) with

$$\lambda(x) = 1 - x, \quad \mu(x, y) = \frac{ay}{x+b}, \quad \nu(y) = 0, \quad \rho(x, y) = c\left(1 - \frac{y}{x}\right).$$

Any equilibrium points  $(x_0, y_0)$  in which both  $x$  and  $y$  are both positive must satisfy

$$y_0 = x_0, \tag{9.3.15}$$

$$0 = 1 - x_0 - \frac{ax_0}{x_0 + b}. \tag{9.3.16}$$

The only positive solution of (9.3.16) is

$$x_0 = \frac{1}{2} \left(1 - a - b + \sqrt{(1 - a - b)^2 + 4b}\right),$$

so there is a unique equilibrium for  $x > 0, y > 0$ . Linearisation gives

$$\begin{aligned}\alpha &= x_0 \left( -1 + \frac{ax_0}{(x_0+b)^2} \right), & \beta &= -\frac{ax_0}{x_0+b}, \\ \gamma &= c, & \delta &= -c.\end{aligned}$$

Note that these are the coefficients in the Jacobian matrix (see Section 5.3) for the system (9.3.13), (9.3.14) at the equilibrium point  $(x_0, y_0)$ . Recall that the equilibrium is stable if and only if the eigenvalues  $\lambda$  of this matrix both have negative real part. The eigenvalues are given by

$$\lambda = \frac{1}{2} \left( \alpha + \delta \pm \sqrt{(\alpha + \delta)^2 - 4(\alpha\delta - \beta\gamma)} \right).$$

We can see that

$$\begin{aligned}\alpha\delta - \beta\gamma &= cx_0 \left( 1 - \frac{ax_0}{(x_0+b)^2} \right) + \frac{acx_0}{x_0+b} \\ &= cx_0 \left( 1 + \frac{ab}{(x_0+b)^2} \right) > 0.\end{aligned}$$

Hence the real parts of both eigenvalues have the same sign, and the equilibrium is stable if and only if  $\alpha + \delta < 0$ . For the values given for  $a$  and  $b$ ,  $x_0 = \frac{-1+\sqrt{41}}{20}$ , and  $\alpha + \delta \approx 0.2625 - c$ . Therefore,  $c^* \approx 0.2625$  and the equilibrium is stable if  $c > c^*$  and unstable if  $c < c^*$ . In the latter case, we can use Poincaré–Bendixson theorem (see Section 5.6) to show that there is a stable limit cycle if we can construct a trapping region in  $x > 0, y > 0$ . We can see directly from (9.3.13), (9.3.14) that if  $x = 0$  then  $\dot{x} = 0$  and if  $y = 0$  then  $\dot{y} = 0$ . Hence no trajectory can ever cross the  $x$ -axis or  $y$ -axis. Furthermore, if  $x = 1$  then  $\dot{x} \leq 0$  and if  $y = 1$  and  $x \leq 1$ , then  $\dot{y} \leq 0$ . Hence no trajectory can ever depart the square region  $0 < x < 1, 0 < y < 1$ , i.e., it is a trapping region. Since this region does not contain a stable equilibrium, it must contain a stable limit cycle.

□

## Exercises

- 9.1 For whales, evidence suggests that  $b$  and  $d$  in (9.2.2) are given by  $b = 0.14$  and  $d = 0.09$  when  $t$  is measured in years. How long would the population take to increase from 800 to 8500 in the absence of fishing? If there were 4000 whales, what catch per year would you recommend? If  $f = 0.75$ , show that the population will take about 14 years to recover from a year's fishing.

9.2 For buffaloes  $b = 0.081$  and  $d = 0.05$  in (9.2.2) when time is calculated in years. If there were 40 million, what is the maximum number to be shot if the herd is to be sustained?

9.3 Discuss the Volterra–Lotka model of (9.3.4) and (9.3.5) in the phase plane when none of the constants is zero.

9.4 Discuss the system

$$\begin{aligned}\dot{x} &= ax - gx^2 - bxy, \\ \dot{y} &= c'y - g'y^2 + dxy\end{aligned}$$

in the phase plane, all the constants being positive. Will the predators die out if there is no prey?

9.5 Show that the system

$$\begin{aligned}\dot{x} &= -ax + b(1 - e^{-ky})x, \\ \dot{y} &= -ay + c(1 - e^{-ky})x,\end{aligned}$$

where  $a, b, c$  and  $k$  are positive constants with  $b > a$ , has equilibrium points at  $(0, 0)$  and  $(\{-b/ck\} \ln\{(b-a)/b\}, -\{1/k\} \ln\{(b-a)/b\})$  of which the second is a saddle-point. Examine the trajectories in the phase plane.

9.6 Compare in the phase plane the systems

$$\begin{aligned}\dot{x} &= ax - bxy, \\ \dot{y} &= c'y - d'y^2/x\end{aligned}$$

and

$$\begin{aligned}\dot{x} &= ax^2 - bx^2y, \\ \dot{y} &= c'xy - d'y^2.\end{aligned}$$

9.7 Discuss the behaviour of the system

$$\begin{aligned}\dot{x} &= -ax + gx^2 - bxy, \\ \dot{y} &= -cy + dxy\end{aligned}$$

where  $a, b, c, d$  and  $g$  are all positive constants.

9.8 Draw the trajectories of (9.3.1) and (9.3.2) when  $a = 10, b = 0.2, c = 10^{-4}$  and  $d = 0.2$ .

- 9.9 A model in which the prey  $x$  and predator  $y$  are removed by an external mechanism as well as interacting is

$$\begin{aligned}\dot{x} &= xf(x, y) - A, \\ \dot{y} &= yg(x, y) - B\end{aligned}$$

where  $A$  and  $B$  are negative constants (if they are negative, restocking is occurring). Assume that  $\partial f/\partial x < 0$ ,  $\partial g/\partial x > 0$ ,  $\partial g/\partial y \leq 0$  for all  $x > 0$  and  $y > 0$ . Show that, when  $A = B = 0$ , the prey can grow from a small initial population when  $f(0, 0) \geq 0$  provided that  $f(0, y)$  is positive for  $y > 0$ ; if  $f(0, y)$  is positive for  $0 < y < y_0$  and  $f(0, y_0) = 0$  show that the prey can still establish itself so long as the initial predator population is less than  $y_0$ .

- 9.10 As a specific model for Exercise 9.9 take

$$\begin{aligned}f(x, y) &= 2 - \frac{x}{30} - \frac{y}{x + 10}, \\ g(x, y) &= \frac{x - 20}{3(x + 10)}.\end{aligned}$$

Show that in the phase plane there can be two critical points, of which one is a saddle-point and the other,  $P$ , is not.

If  $A = 13$  and  $B = 2$ ,  $P$  is stable and there is a trajectory from the saddle-point to  $P$ ; both species can coexist. The same is true when  $A = 3$  and  $B = 3$ .

If  $A = 3$  and  $B = \frac{1}{2}$ ,  $P$  is unstable and there is a trajectory from  $P$  to the saddle-point; one of the species becomes extinct.

If  $A = 3$  and  $B = 2$  or  $\frac{3}{2}$ ,  $P$  is unstable but is surrounded by a stable limit cycle; the populations can exhibit oscillatory behaviour.

- 9.11 Use MATLAB to solve the system (9.3.13), (9.3.14) numerically for two different values of  $c$ , one less than  $c^*$  and one greater than  $c^*$ , and for an appropriate initial condition. Check that your numerical results agree with the analytical results. In the  $c < c^*$  case, plot the vector field (see Section 5.8) to verify that the region  $0 < x < 1$ ,  $0 < y < 1$  is a trapping region.

# Chapter 10

---

## Partial Differential Equations

---

### 10.1 Characteristics for equations of the first order

In this section we set out a method of finding solutions of linear partial differential equations of the first order. We start by considering a particular example, namely

$$2\frac{\partial u}{\partial x} + 3\frac{\partial u}{\partial y} = 0. \quad (10.1.1)$$

Associated with this are certain curves in the  $(x, y)$ -plane specified by obliging them to have the slope which satisfies

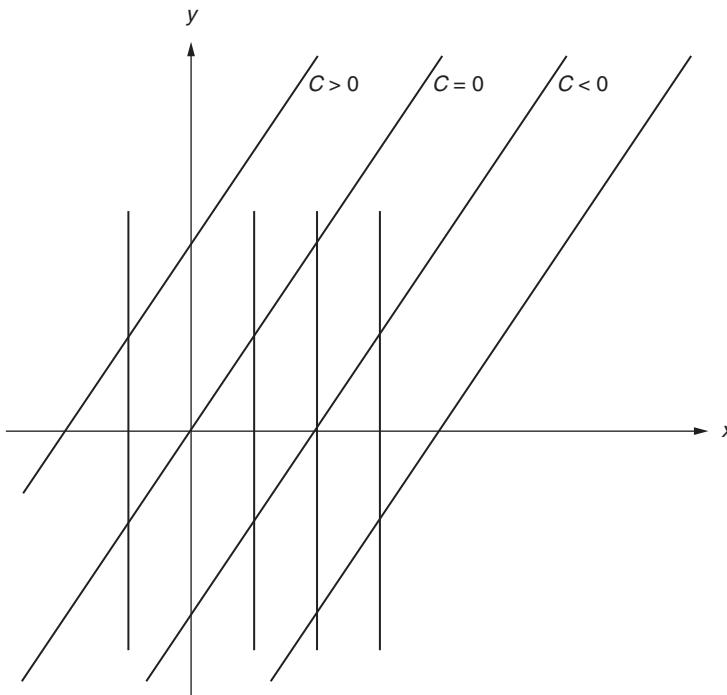
$$\frac{dy}{dx} = \frac{3}{2}, \quad (10.1.2)$$

the right-hand side being the ratio of the coefficients of the two partial derivatives in (10.1.1). The general solution of (10.1.2) is  $y = \frac{3}{2}x + C$  where  $C$  is a constant and the curves are, in fact, straight lines. They are drawn for various values of  $C$  in Figure 10.1.1. These special curves are known as the **characteristics** of the partial differential equation (10.1.1).

The idea in solving (10.1.1) is to introduce a new set of coordinates, called  $\xi$  and  $\eta$ , in which the characteristics can be expressed as  $\xi = \text{constant}$ . The other coordinate  $\eta$  must be selected so that to a given point  $(\xi, \eta)$  there corresponds only one point in the  $x$ - and  $y$ -coordinates. In other words, we want the curve  $\eta = \text{constant}$  to intersect the curve  $\xi = \text{constant}$  in the  $(x, y)$ -plane in one, and only one, point. Now, in Figure 10.1.1, a line parallel to the  $y$ -axis meets a characteristic in one, and only one, point. Therefore make the change of coordinates

$$\xi = y - \frac{3}{2}x, \quad \eta = x. \quad (10.1.3)$$

Notice that (10.1.3) does supply just one  $(x, y)$  for a given  $(\xi, \eta)$ ; in fact  $x = \eta$  and  $y = \xi + \frac{3}{2}\eta$ . Other choices of  $\eta$  are acceptable providing that they meet the criterion that the curve  $\eta = \text{constant}$  intersects  $\xi = \text{constant}$  once and once only; for example,  $\eta = y$  and  $\eta = 3y + 2x$  are other possibilities.



**FIGURE 10.1.1:** The characteristics of (10.1.1).

From (10.1.3), by the chain rule for partial derivatives,

$$\begin{aligned}\frac{\partial u}{\partial x} &= \frac{\partial u}{\partial \xi} \frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial \eta} \frac{\partial \eta}{\partial x} = -\frac{3}{2} \frac{\partial u}{\partial \xi} + \frac{\partial u}{\partial \eta}, \\ \frac{\partial u}{\partial y} &= \frac{\partial u}{\partial \xi} \frac{\partial \xi}{\partial y} + \frac{\partial u}{\partial \eta} \frac{\partial \eta}{\partial y} = \frac{\partial u}{\partial \xi}\end{aligned}$$

since  $\partial \xi / \partial x = -3/2$ ,  $\partial \xi / \partial y = 1$ ,  $\partial \eta / \partial x = 1$ ,  $\partial \eta / \partial y = 0$ . Substitution in (10.1.1) gives

$$2 \frac{\partial u}{\partial \eta} = 0. \quad (10.1.4)$$

Thus the change of coordinates (10.1.3) has simplified the partial differential equation and, in fact, one partial derivative  $\partial u / \partial \xi$  has been removed altogether. It was with the object of causing this partial derivative to disappear that the special choice of  $\xi$  was made.

Equation (10.1.4) tells us that  $u$  is independent of  $\eta$  and so it must be a function of  $\xi$  only, i.e.,

$$u = F(\xi) = F(y - \frac{3}{2}x) \quad (10.1.5)$$

on employing (10.1.3).

The formula (10.1.5) shows that  $u$  is a constant on a characteristic since  $y - \frac{3}{2}x$  is constant on a characteristic. Hence, if  $u$  is known at one point of a characteristic it is known at every point of the characteristic. The constant can be different on different characteristics, but we cannot say how it will change without additional information.

Suppose, indeed, that it is required that  $u = 3y$  on  $x = 0$  for  $y \geq 0$ . Then, putting  $x = 0$  in (10.1.5), we are obliged to have

$$F(y) = 3y$$

for  $y \geq 0$ . Using this expression for  $F$  we have  $F(y - \frac{3}{2}x) = 3(y - \frac{3}{2}x)$  and so, from (10.1.5),

$$u = 3y - 9x/2$$

for  $y \geq 3x/2$ . Thus  $u$  has been found at every point on and above the line  $y = 3x/2$ . Nothing can be said about the behaviour of  $u$  below this line, other than that it is a function of  $y - 3x/2$ , because there is insufficient information. In other words, the solution is restricted to that part of the  $(x, y)$ -plane that is covered by the characteristics that intersect the arc on which the initial data are given.

We might have been asked to find  $u$  so that  $u = 3y$  on the line  $y = 3x/2$  for  $y \geq 0$ . Then (10.1.5) would require  $F(0) = 3y$  for  $y \geq 0$ , which no  $F$  will satisfy because the left-hand side is constant whereas the right-hand side is not. Indeed, the only problem which can be solved when  $u$  is given on  $y = 3x/2$  is for  $u$  to be specified as a constant there. A similar situation arises for the characteristic  $y - 3x/2 = \text{constant}$ . Thus the initial data can be specified arbitrarily for  $u$  only when given on an arc that does not coincide with a characteristic. When initial data are prescribed on a characteristic they must take a special form if inconsistency is to be avoided.

Let us turn now to the more general linear partial differential equation

$$a(x, y)\frac{\partial u}{\partial x} + b(x, y)\frac{\partial u}{\partial y} = c(x, y)u + d(x, y). \quad (10.1.6)$$

We succeeded in solving (10.1.1) because we were able to discover a substitution which eliminated one partial derivative and led to (10.1.4). So we want to try to find a similar substitution for (10.1.6). Now, if  $a \equiv 0$  or  $b \equiv 0$  in the region under consideration, only one partial derivative occurs anyway so it is only when neither  $a$  nor  $b$  vanishes identically that further discussion is necessary. We then consider the curves whose slope satisfies

$$\frac{dy}{dx} = \frac{b(x, y)}{a(x, y)}. \quad (10.1.7)$$

They are known as **characteristics** and have no multiple points provided that  $a$  and  $b$  do not vanish at the same point, a possibility which will be excluded here. The characteristics need not be straight lines as they were for (10.1.1).

Let the equation for the characteristics be written in the form

$$\phi(x, y) = C \quad (10.1.8)$$

where  $C$  is a constant which gives different characteristics for different values. A derivative of (10.1.8) with respect to  $x$  supplies

$$\frac{\partial \phi}{\partial x} + \frac{\partial \phi}{\partial y} \frac{dy}{dx} = 0.$$

But, since the slope is forced to be the same as that of (10.1.7), we must have

$$a(x, y) \frac{\partial \phi}{\partial x} + b(x, y) \frac{\partial \phi}{\partial y} = 0. \quad (10.1.9)$$

Suppose now that there is a set of curves with equation

$$\psi(x, y) = C_1,$$

$C_1$  being constant, with the property that each curve meets each characteristic in one and only one point in the region of interest while  $\psi$  varies continuously along a characteristic. There are usually many possible  $\psi$  which are open to the solver and the one most convenient to the problem in hand should be chosen.

Make the substitution

$$\xi = \phi(x, y), \quad \eta = \psi(x, y). \quad (10.1.10)$$

The conditions placed on  $\phi$  and  $\psi$  ensure that a point may be fixed by either  $(x, y)$  or  $(\xi, \eta)$ . Since

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi} \frac{\partial \phi}{\partial x} + \frac{\partial u}{\partial \eta} \frac{\partial \psi}{\partial x}, \quad \frac{\partial u}{\partial y} = \frac{\partial u}{\partial \xi} \frac{\partial \phi}{\partial y} + \frac{\partial u}{\partial \eta} \frac{\partial \psi}{\partial y}$$

(10.1.6) transforms to

$$\left( a \frac{\partial \psi}{\partial x} + b \frac{\partial \psi}{\partial y} \right) \frac{\partial u}{\partial \eta} = cu + d \quad (10.1.11)$$

on account of (10.1.9). Equation (10.1.11) is often called the **characteristic form** of the partial differential equation.

If values for  $x, y$  in terms of  $\xi, \eta$  are inserted in (10.1.11) from (10.1.10), the coefficients in (10.1.11) become known functions of  $\xi$  and  $\eta$ . For fixed  $\xi$  (10.1.11) is an ordinary differential equation of the first order for  $u$  in terms of  $\eta$ . Therefore  $u$  can be determined but contains an arbitrary constant (which may be different for different values of  $\xi$ ). Thus, if  $u$  is specified at a point of  $\xi = \text{constant}$  it is known at all points of that curve. In other words, if  $u$  is given at one point of a characteristic it can be found at all points of the

characteristic. In particular, when  $u$  is designated to have certain values on an arc that is met at most once by a characteristic it is determined in the region covered by the characteristics intersecting the arc. Again, as for (10.1.1), data cannot be prescribed arbitrarily all along a characteristic; they must comply with (10.1.11) if there is to be a solution.

The existence theorems of Chapter 5 can be employed to check the existence and uniqueness of solutions to (10.1.11).

### **Example 10.1.1**

Find the solution of

$$2x \frac{\partial u}{\partial x} + (x+1) \frac{\partial u}{\partial y} = y$$

in  $x > 0$  such that  $u = 2y$  when  $x = 1$ .

The differential equation for the characteristics is, from (10.1.7),

$$\frac{dy}{dx} = \frac{x+1}{2x} = \frac{1}{2} \left(1 + \frac{1}{x}\right).$$

Therefore the characteristics are

$$y = \frac{1}{2}(x + \ln x) + C. \quad (10.1.12)$$

The slope of the characteristics is positive in  $x > 0$ . Also  $y \rightarrow -\infty$  as  $x \rightarrow 0$  and  $y \rightarrow \infty$  as  $x \rightarrow \infty$ . The shape of the characteristics is, consequently, that shown in Figure 10.1.2, the larger  $C$  the closer the characteristic is to the  $y$ -axis.

A set of curves that intersects characteristics once and once only is furnished by lines parallel to the  $x$ -axis; another set is provided by lines parallel to the  $y$ -axis. Choose those parallel to the  $y$ -axis since the initial values are intimated on  $x = 1$ . Then, according to (10.1.10),

$$\xi = y - \frac{1}{2}(x + \ln x), \quad \eta = x. \quad (10.1.13)$$

Note that it is always worthwhile considering parallel straight lines first for the  $\psi$  curves since they lead to the simplest form for  $\eta$ . Now

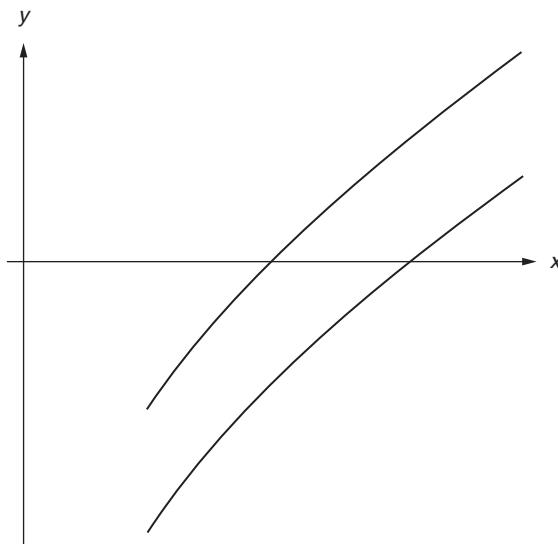
$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi} \left(-\frac{1}{2} - \frac{1}{2x}\right) + \frac{\partial u}{\partial \eta}, \quad \frac{\partial u}{\partial y} = \frac{\partial u}{\partial \xi}$$

and the given partial differential equation is converted to

$$2x \frac{\partial u}{\partial \eta} = y.$$

From (10.1.13),  $x = \eta$  and  $y = \xi + \frac{1}{2}(\eta + \ln \eta)$  so that

$$2\eta \frac{\partial u}{\partial \eta} = \xi + \frac{1}{2}(\eta + \ln \eta)$$



**FIGURE 10.1.2:** The characteristics of (10.1.12).

has to be solved. From

$$\frac{\partial u}{\partial \eta} = \frac{\xi}{2\eta} + \frac{1}{4} \left( 1 + \frac{1}{\eta} \ln \eta \right) \quad (10.1.14)$$

is deduced

$$u = \frac{1}{2}\xi \ln \eta + \frac{1}{4}\eta + \frac{1}{8}(\ln \eta)^2 + F(\xi).$$

The arbitrary function  $F(\xi)$  of  $\xi$  is used rather than a constant because of the partial derivative in (10.1.14).

It is required that  $u = 2y$  when  $x = 1$ , i.e.,  $u = 2\xi + \eta + \ln \eta$  when  $\eta = 1$  or  $u = 2\xi + 1$  when  $\eta = 1$ , and this is to hold for all  $\xi$ . Therefore

$$2\xi + 1 = \frac{1}{4} + F(\xi)$$

whence

$$u = \frac{1}{2}\xi \ln \eta + \frac{1}{4}\eta + \frac{1}{8}(\ln \eta)^2 + 2\xi + \frac{3}{4}.$$

Replacing  $\xi, \eta$  by  $x, y$  via (10.1.13) we obtain

$$u = \left( y - \frac{1}{2}x - \frac{1}{2} \ln x \right) \left( 2 + \frac{1}{2} \ln x \right) + \frac{1}{4}x + \frac{3}{4} + \frac{1}{8}(\ln x)^2. \quad (10.1.15)$$

This solution is valid everywhere in  $x > 0$ , because all the characteristics intersect  $x = 1$  and every point in  $x > 0$  lies on some characteristic.  $\square$

### Example 10.1.2

Find the solution of

$$-x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = 1$$

in  $0 < x < y$  such that  $u = 2x$  on  $y = 3x$ .

The characteristics satisfy

$$\frac{dy}{dx} = -\frac{y}{x}$$

whence their equation is  $\ln y = -\ln x + \text{constant}$ , which may be simplified to

$$xy = \text{constant}.$$

Each characteristic in the first quadrant is therefore the branch of a hyperbola (Figure 10.1.3). The intersecting curves may again be chosen as straight lines and to illustrate the fact that they need not be parallel to the coordinate axes we select them to be parallel to  $y = x$ . (What would be the objection to making them parallel to  $y = -x$ ?) Thus the substitution

$$\xi = xy, \quad \eta = y - x \quad (10.1.16)$$

is made.

With this change of variable

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi} y - \frac{\partial u}{\partial \eta}, \quad \frac{\partial u}{\partial y} = \frac{\partial u}{\partial \xi} x + \frac{\partial u}{\partial \eta}$$

and the partial differential equation is transformed to

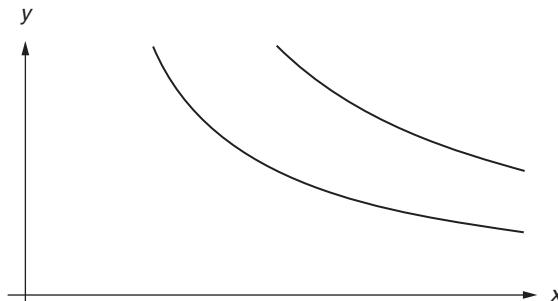
$$(x + y) \frac{\partial u}{\partial \eta} = 1.$$

Now, if  $y$  is eliminated from (10.1.16),

$$x^3 + \eta x - \xi = 0$$

so that  $x = \frac{1}{2}\{-\eta \pm (\eta^2 + 4\xi)^{1/2}\}$ . However,  $x$  and  $\xi$  must both be positive in the first quadrant so that the upper sign must be selected, i.e.,

$$x = \frac{1}{2}\{-\eta + (\eta^2 + 4\xi)^{1/2}\}, \\ y = \frac{1}{2}\{\eta + (\eta^2 + 4\xi)^{1/2}\}.$$



**FIGURE 10.1.3:** The characteristic hyperbolae.

Hence we are led to

$$(\eta^2 + 4\xi)^{1/2} \frac{\partial u}{\partial \eta} = 1.$$

It follows that

$$\begin{aligned} u &= \ln\{\eta + (\eta^2 + 4\xi)^{1/2}\} + F(\xi) \\ &= \ln 2y + F(xy) \end{aligned}$$

on inserting (10.1.16) and the formulae for  $x$  and  $y$  just found.

Since  $u = 2x$  when  $y = 3x$ ,

$$2x = \ln 6x + F(3x^2)$$

which implies that

$$F(z) = \frac{2}{\sqrt{3}}z^{1/2} - \ln(6z^{1/2}/\sqrt{3})$$

when  $z$  is positive. Consequently

$$\begin{aligned} u &= \ln 2y + 2(xy)^{1/2}/\sqrt{3} - \ln\{6(xy)^{1/2}/\sqrt{3}\} \\ &= \ln(y/3x)^{1/2} + 2(xy)^{1/2}/\sqrt{3}. \end{aligned} \tag{10.1.17}$$

Since each characteristic intersects  $y = 3x$  this solution is valid throughout the first quadrant. Nothing can be said about what happens outside the first quadrant because the initial data are given only on characteristics that are confined to the first quadrant.  $\square$

## 10.2 Another view of characteristics

Suppose that  $u$  satisfies

$$a(x, y) \frac{\partial u}{\partial x} + b(x, y) \frac{\partial u}{\partial y} = c(x, y)u + d(x, y) \tag{10.2.1}$$

and we are given that  $u = f(y)$  on  $x = 0$ . It is plausible to try to find a solution near  $x = 0$  by making a series expansion in powers of  $x$ , i.e.,

$$u(x, y) = u(0, y) + x \left[ \frac{\partial}{\partial x} u(x, y) \right]_{x=0} + \frac{1}{2}x^2 \left[ \frac{\partial^2}{\partial x^2} u(x, y) \right]_{x=0} + \dots$$

The first term is known from the information given, namely  $u(0, y) = f(y)$ . Can we find  $\partial u / \partial x$  at  $x = 0$  from the governing partial differential equation?

Clearly,  $\partial u(0, y)/\partial y = f'(y)$  and so putting  $x = 0$  in (10.2.1) furnishes

$$a(0, y) \left[ \frac{\partial}{\partial x} u(x, y) \right]_{x=0} = c(0, y)f(y) + d(0, y) - b(0, y)f'(y). \quad (10.2.2)$$

So long as  $a(0, y) \neq 0$ , this determines  $\partial u/\partial x$  at  $x = 0$  as a known function of  $y$ .

Now take a partial derivative with respect to  $x$  of (10.2.1). Then

$$a(x, y) \frac{\partial^2 u}{\partial x^2} = \text{terms in } u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial^2 u}{\partial x \partial y}.$$

All the terms on the right-hand side are available at  $x = 0$  because

$$\frac{\partial}{\partial y} \left[ \frac{\partial}{\partial x} u(x, y) \right]_{x=0}$$

can be deduced from (10.2.2). Therefore  $\partial^2 u/\partial x^2$  can be found at  $x = 0$  when  $a(0, y) \neq 0$ . Further partial derivatives with respect to  $x$  will supply higher terms in the expansion provided that  $a, b, c$  and  $d$  are sufficiently differentiable. Thus we can hope to find  $u$  off  $x = 0$  when  $a(0, y) \neq 0$ .

If, however,  $a(0, y) = 0$  the process fails immediately because the term involving  $\partial u/\partial x$  disappears from (10.2.2). In this case  $u$  cannot be determined away from  $x = 0$ . In fact,  $f$  cannot be selected arbitrarily because (10.2.2) forces

$$c(0, y)f(y) + d(0, y) = b(0, y)f'(y) \quad (10.2.3)$$

so that  $f$  is indeterminate only to the extent of a constant. To put it another way, the value of  $f$  can be prescribed at a single point of  $x = 0$ ; at all other points its value is obtained from (10.2.3).

These properties coincide with those of a characteristic as discovered in the preceding section. Hence  $a(0, y) = 0$  makes  $x = 0$  a characteristic.

Let us now invoke this result after making a transformation of coordinates. Then we can say that if, after the change of variables  $\xi = \phi(x, y)$ ,  $\eta = \psi(x, y)$ , the coefficient of  $\partial u/\partial \xi$  vanishes when  $\xi = 0$  then  $\xi = 0$  is a characteristic. It follows that, if the coefficient vanishes whatever  $\xi$ , then  $\xi = \text{constant}$  is a characteristic.

This way of looking at characteristics will be useful in discussing partial differential equations of the second order.

### 10.3 Linear partial differential equations of the second order

The linear partial differential equation of the second order that will be discussed in this section is

$$\begin{aligned} & a(x, y) \frac{\partial^2 u}{\partial x^2} + 2b(x, y) \frac{\partial^2 u}{\partial x \partial y} + c(x, y) \frac{\partial^2 u}{\partial y^2} \\ & + e(x, y) \frac{\partial u}{\partial x} + f(x, y) \frac{\partial u}{\partial y} + g(x, y)u = 0 \end{aligned} \quad (10.3.1)$$

where  $a, \dots, g$  are real. As a first step, let us examine the possibility of finding a solution by means of a power series, as in the preceding section, given initial data on  $x = 0$ . Two pieces of data will now be appropriate because the differential equation is of the second order. Suppose, indeed, that

$$u(0, y) = F(y), \quad \left[ \frac{\partial}{\partial x} u(x, y) \right]_{x=0} = G(y).$$

Then

$$\begin{aligned} \frac{\partial}{\partial y} u(0, y) &= F'(y), \quad \frac{\partial^2}{\partial y^2} u(0, y) = F''(y), \\ \frac{\partial}{\partial y} \left[ \frac{\partial}{\partial x} u(x, y) \right]_{x=0} &= G'(y) \end{aligned}$$

so that, when  $x = 0$ , (10.3.1) becomes

$$\begin{aligned} a(0, y) \left[ \frac{\partial^2 u}{\partial x^2} \right]_{x=0} &= -2b(0, y)G'(y) - c(0, y)F''(y) - e(0, y)G(y) \\ &\quad - f(0, y)F'(y) - g(0, y)F(y). \end{aligned} \quad (10.3.2)$$

Therefore, if  $a(0, y) \neq 0$ , we can find the second partial derivative of  $u$  with respect to  $x$  and higher derivatives can be determined from derivatives of (10.3.1) with  $x = 0$ . In this case  $u$  can be continued off  $x = 0$ .

If, however,  $a(0, y) = 0$  the process fails; moreover,  $F$  and  $G$  cannot both be prescribed arbitrarily because (10.3.2) insists that

$$2b(0, y)G'(y) + c(0, y)F''(y) + e(0, y)G(y) + f(0, y)F'(y) + g(0, y)F(y) = 0.$$

By analogy with the first-order equation, we say that  $x = 0$  is a characteristic.

It is now desirable to generalise the approach so that other curves may play the role of characteristics. Change variables by means of the transformation

$$\xi = \phi(x, y), \quad \eta = \psi(x, y).$$

Then

$$\begin{aligned}\frac{\partial u}{\partial x} &= \frac{\partial u}{\partial \xi} \frac{\partial \phi}{\partial x} + \frac{\partial u}{\partial \eta} \frac{\partial \psi}{\partial x}, \quad \frac{\partial u}{\partial y} = \frac{\partial u}{\partial \xi} \frac{\partial \phi}{\partial y} + \frac{\partial u}{\partial \eta} \frac{\partial \psi}{\partial y}, \\ \frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 u}{\partial \xi^2} \left( \frac{\partial \phi}{\partial x} \right)^2 + 2 \frac{\partial^2 u}{\partial \xi \partial \eta} \frac{\partial \phi}{\partial x} \frac{\partial \psi}{\partial x} + \frac{\partial^2 u}{\partial \eta^2} \left( \frac{\partial \psi}{\partial x} \right)^2 + \frac{\partial u}{\partial \xi} \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial u}{\partial \eta} \frac{\partial^2 \psi}{\partial x^2}, \\ \frac{\partial^2 u}{\partial y^2} &= \frac{\partial^2 u}{\partial \xi^2} \left( \frac{\partial \phi}{\partial y} \right)^2 + 2 \frac{\partial^2 u}{\partial \xi \partial \eta} \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y} + \frac{\partial^2 u}{\partial \eta^2} \left( \frac{\partial \psi}{\partial y} \right)^2 + \frac{\partial u}{\partial \xi} \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial u}{\partial \eta} \frac{\partial^2 \psi}{\partial y^2}, \\ \frac{\partial^2 u}{\partial x \partial y} &= \frac{\partial^2 u}{\partial \xi^2} \frac{\partial \phi}{\partial x} \frac{\partial \phi}{\partial y} + \frac{\partial^2 u}{\partial \xi \partial \eta} \left( \frac{\partial \phi}{\partial x} \frac{\partial \psi}{\partial y} + \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial x} \right) + \frac{\partial^2 u}{\partial \eta^2} \frac{\partial \psi}{\partial x} \frac{\partial \psi}{\partial y} \\ &\quad + \frac{\partial u}{\partial \xi} \frac{\partial^2 \phi}{\partial x \partial y} + \frac{\partial u}{\partial \eta} \frac{\partial^2 \psi}{\partial x \partial y}.\end{aligned}$$

Substitution of these formulae into (10.3.1) leads to

$$a_1 \frac{\partial^2 u}{\partial \xi^2} + 2b_1 \frac{\partial^2 u}{\partial \xi \partial \eta} + c_1 \frac{\partial^2 u}{\partial \eta^2} + e_1 \frac{\partial u}{\partial \xi} + f_1 \frac{\partial u}{\partial \eta} + g_1 u = 0 \quad (10.3.3)$$

where

$$a_1 = a \left( \frac{\partial \phi}{\partial x} \right)^2 + 2b \frac{\partial \phi}{\partial x} \frac{\partial \phi}{\partial y} + c \left( \frac{\partial \phi}{\partial y} \right)^2, \quad (10.3.4)$$

$$b_1 = a \frac{\partial \phi}{\partial x} \frac{\partial \psi}{\partial x} + b \left( \frac{\partial \phi}{\partial x} \frac{\partial \psi}{\partial y} + \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial x} \right) + c \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y}, \quad (10.3.5)$$

$$c_1 = a \left( \frac{\partial \psi}{\partial x} \right)^2 + 2b \frac{\partial \psi}{\partial x} \frac{\partial \psi}{\partial y} + c \left( \frac{\partial \psi}{\partial y} \right)^2, \quad (10.3.6)$$

$$e_1 = a \frac{\partial^2 \phi}{\partial x^2} + 2b \frac{\partial^2 \phi}{\partial x \partial y} + c \frac{\partial^2 \phi}{\partial y^2} + e \frac{\partial \phi}{\partial x} + f \frac{\partial \phi}{\partial y}, \quad (10.3.7)$$

$$f_1 = a \frac{\partial^2 \psi}{\partial x^2} + 2b \frac{\partial^2 \psi}{\partial x \partial y} + c \frac{\partial^2 \psi}{\partial y^2} + e \frac{\partial \psi}{\partial x} + f \frac{\partial \psi}{\partial y}, \quad (10.3.8)$$

$$g_1 = g. \quad (10.3.9)$$

From our earlier discussion,  $\xi = \text{constant}$  will be a characteristic if the coefficient of  $\partial^2 u / \partial \xi^2$  vanishes for this value of  $\xi$ . This requires  $a_1 = 0$  and so, from (10.3.4),  $\phi(x, y) = \text{constant}$  is a characteristic when

$$a \left( \frac{\partial \phi}{\partial x} \right)^2 + 2b \frac{\partial \phi}{\partial x} \frac{\partial \phi}{\partial y} + c \left( \frac{\partial \phi}{\partial y} \right)^2 = 0. \quad (10.3.10)$$

We may also seek to make  $\psi = \text{constant}$  a characteristic. Then  $c_1 = 0$  and

$$a \left( \frac{\partial \psi}{\partial x} \right)^2 + 2b \frac{\partial \psi}{\partial x} \frac{\partial \psi}{\partial y} + c \left( \frac{\partial \psi}{\partial y} \right)^2 = 0. \quad (10.3.11)$$

The slope of the curve  $\phi = \text{constant}$  is given by

$$\frac{dy}{dx} = -\frac{\partial\phi/\partial x}{\partial\phi/\partial y}$$

and consequently, from (10.3.10), it satisfies

$$a\left(\frac{dy}{dx}\right)^2 - 2b\frac{dy}{dx} + c = 0. \quad (10.3.12)$$

Comparison of the structures of (10.3.10) and (10.3.11) reveals that the slope of the characteristic  $\psi = \text{constant}$  also complies with (10.3.12). Thus the slopes of both sets of characteristics are provided by the solution of (10.3.12), namely

$$\frac{dy}{dx} = \frac{1}{a}\{b \pm (b^2 - ac)^{1/2}\}. \quad (10.3.13)$$

There are three possible cases to consider.

- (I) If  $ac > b^2$  in some region D there are two families of complex characteristics. The partial differential equation is then said to be **elliptic** in D. A typical example of an elliptic differential equation is Laplace's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

in which  $a = 1$ ,  $b = 0$  and  $c = 1$ .

- (II) If  $ac < b^2$  in some region D there are two families of real characteristics. The partial differential equation is called **hyperbolic** in D. A standard illustration of a hyperbolic differential equation is the equation of one-dimensional wave propagation

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{\alpha^2} \frac{\partial^2 u}{\partial y^2} = 0$$

in which  $a = 1$ ,  $b = 0$  and  $c = -1/\alpha^2$ .

- (III) If  $ac = b^2$  throughout some region D the two slopes coincide and there is one family of real characteristics. The partial differential equation is now **parabolic** in D. The equation of heat conduction

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial u}{\partial y}$$

in which  $a = 1$ ,  $b = 0$  and  $c = 0$  is parabolic.

The behaviour of the solutions of the three types of partial differential equations is quite different and the initial or boundary conditions that they have to obey are usually distinctive. In applications, they normally originate from attempts to model totally different phenomena. One may expect the problems in which the partial differential equation is elliptic in some parts of the  $(x, y)$ -plane and hyperbolic in others are particularly troublesome. We shall now undertake a more detailed study of the various types.

## 10.4 Elliptic partial differential equations

In the elliptic case we shall take the slope of  $\phi = \text{constant}$  as

$$\frac{dy}{dx} = \frac{1}{a} \{b + (b^2 - ac)^{1/2}\} \quad (10.4.1)$$

and that of  $\psi = \text{constant}$  as

$$\frac{dy}{dx} = \frac{1}{a} \{b - (b^2 - ac)^{1/2}\}. \quad (10.4.2)$$

Since the square root is purely imaginary, the two slopes are complex conjugates. It follows that  $\phi$  and  $\psi$  are complex conjugates, i.e.,  $\psi = \phi^*$  and  $\eta = \xi^*$ , the asterisk denoting a complex conjugate.

In terms of  $\phi$  and  $\psi$ , (10.4.1) and (10.4.2) can be expressed as

$$\frac{\partial \phi}{\partial x} = -\frac{1}{a} \{b + (b^2 - ac)^{1/2}\} \frac{\partial \phi}{\partial y}, \quad (10.4.3)$$

$$\frac{\partial \psi}{\partial x} = -\frac{1}{a} \{b + (b^2 - ac)^{1/2}\} \frac{\partial \psi}{\partial y}. \quad (10.4.4)$$

The characteristics have been chosen so that  $a_1 = 0$  and  $c_1 = 0$ , but notice that neither  $a$  nor  $c$  can be zero because of the condition for ellipticity. For  $b_1$  substitute from (10.4.3) and (10.4.4) into (10.3.5) to obtain

$$\begin{aligned} b_1 &= \frac{1}{a} \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y} \{b^2 - (b^2 - ac)\} - \frac{b}{a} \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y} 2b + c \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y} \\ &= \frac{2}{a} \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y} (ac - b^2). \end{aligned} \quad (10.4.5)$$

Since  $\phi$  and  $\psi$  are complex conjugates (10.4.5) may be rewritten as

$$b_1 = \frac{2}{a} \left| \frac{\partial \phi}{\partial y} \right|^2 (ac - b^2)$$

showing that  $b_1$  is real and has the same sign as  $a$ .

Furthermore, (10.3.7) and (10.3.8) indicate that  $f_1 = e_1^*$  so that the partial differential equation has been transformed to

$$2b_1 \frac{\partial^2 u}{\partial \xi \partial \eta} + e_1 \frac{\partial u}{\partial \xi} + e_1^* \frac{\partial u}{\partial \eta} + g_1 u = 0. \quad (10.4.6)$$

In this form, sometimes called the **characteristic form** of an elliptic equation, the partial differential equation involves complex coordinates. To avoid these a further transformation is made. Let

$$\lambda = \frac{1}{2}(\xi + \eta), \quad \mu = \frac{1}{2i}(\xi - \eta).$$

Then  $\lambda$  is the real part of  $\xi$  and  $\mu$  is the imaginary part of  $\xi$  so that both  $\lambda$  and  $\mu$  are real. Moreover

$$\begin{aligned}\frac{\partial u}{\partial \xi} &= \frac{1}{2} \frac{\partial u}{\partial \lambda} + \frac{1}{2i} \frac{\partial u}{\partial \mu}, & \frac{\partial u}{\partial \eta} &= \frac{1}{2} \frac{\partial u}{\partial \lambda} - \frac{1}{2i} \frac{\partial u}{\partial \mu}, \\ \frac{\partial^2 u}{\partial \xi \partial \eta} &= \frac{1}{4} \frac{\partial^2 u}{\partial \lambda^2} + \frac{1}{4} \frac{\partial^2 u}{\partial \mu^2}\end{aligned}$$

so that (10.4.6) becomes

$$\frac{1}{2} b_1 \left( \frac{\partial^2 u}{\partial \lambda^2} + \frac{\partial^2 u}{\partial \mu^2} \right) + \frac{1}{2} \frac{\partial u}{\partial \lambda} (e_1 + e_1^*) + \frac{1}{2i} \frac{\partial u}{\partial \mu} (e_1 - e_1^*) + g_1 u = 0.$$

By division by  $\frac{1}{2} b_1$  we derive

$$\frac{\partial^2 u}{\partial \lambda^2} + \frac{\partial^2 u}{\partial \mu^2} + e_2 \frac{\partial u}{\partial \lambda} + f_2 \frac{\partial u}{\partial \mu} + g_2 u = 0 \quad (10.4.7)$$

where  $e_2, f_2$  and  $g_2$  are all real; in particular,

$$g_2 = 2g_1/b_1 = ga \Big/ (ac - b^2) \left| \frac{\partial \phi}{\partial y} \right|^2 \quad (10.4.8)$$

from (10.3.9) and (10.4.5).

Equation (10.4.7) is known as the **normal form** of an elliptic partial differential equation. The normal form appears in its simplest guise as

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + k^2 u = 0 \quad (10.4.9)$$

where  $k^2$  is a real constant. If  $k^2 = 0$  it reduces to Laplace's equation and occurs in problems in potential theory. If  $k^2$  is positive it arises in the study of two-dimensional harmonic waves such as those produced by a vibrating membrane. The case when  $k^2$  is negative typically results from trying to solve the two-dimensional wave equation by certain methods.

Usually, solutions of elliptic partial differential equations are sought that satisfy prescribed boundary conditions. There are four of common occurrence.

- (1) *Dirichlet problem.* (a) In the *interior problem*  $u$  is given at every point of a closed curve  $C$  and has to be found inside  $C$ . (b) For the *exterior problem*  $u$  is given on a closed curve  $C$  and has to be determined outside  $C$ . A supplementary condition specifying the behaviour at infinity is normally imposed. For example, there might be a requirement that  $\partial u / \partial r \rightarrow 0$  as  $r = (x^2 + y^2)^{1/2} \rightarrow \infty$  in potential theory.
- (2) *Neumann problem.* (a) In the interior problem the derivative of  $u$  along the normal to a closed curve  $C$ , i.e.,  $\partial u / \partial n$ , is prescribed and  $u$  is required inside  $C$ . Often  $\partial u / \partial n$  cannot be specified arbitrarily on  $C$ ; for

instance, if  $k^2 = 0$  in (10.4.9),  $\int_C (\partial u / \partial n) ds = 0$  by the divergence theorem so that any values ascribed to  $\partial u / \partial n$  on  $C$  must be compatible with this relation. (b)  $\partial u / \partial n$  is given on the closed curve  $C$  and  $u$  is to be found outside, usually subject to a condition at infinity as in (1b).

- (3) *Mixed problem.* Here the closed curve  $C$  is split into two portions  $C_1$  and  $C_2$ . On  $C_1$ ,  $u$  is given and on  $C_2$ ,  $\partial u / \partial n$  is specified. Again, both the interior and exterior cases may arise.
- (4) *Impedance or Robin problem.* In this case  $hu + \partial u / \partial n$  is prescribed on the closed curve  $C$ ,  $h$  being a known function. It can be regarded as including the three preceding cases.

Before leaving the subject we give a theorem which guarantees that the solution to the boundary value problem is unique in suitable circumstances.

### **THEOREM 10.4.1**

Let the partial differential equation (10.3.1) be elliptic ( $ac > b^2$ ) in a simply connected domain  $D$  with  $g \leq 0$  and  $a, c$  positive. Then, if  $u = 0$  on a simple closed curve  $C$  in  $D$ ,  $u = 0$  at all points inside  $C$ .

Remark that  $a$  and  $c$  must have the same sign in order to meet the condition for an elliptic equation.

**PROOF** Change the variables so that the equation goes over to its normal form (10.4.7). The transformation maps  $D$  into a domain  $D_1$  and  $C$  into a simple closed curve  $C_1$ . Note that the mapping is one-to-one because the Jacobian

$$\frac{\partial \phi}{\partial x} \frac{\partial \psi}{\partial y} - \frac{\partial \psi}{\partial x} \frac{\partial \phi}{\partial y} = -\frac{2i}{a} (ac - b^2)^{1/2} \left| \frac{\partial \phi}{\partial y} \right|^2.$$

If  $\partial \phi / \partial y$  vanishes so does  $\partial \phi / \partial x$  on account of (10.4.3), contrary to a characteristic having a definite slope. Therefore the Jacobian is nonzero and the mapping is one-to-one.

Also, from (10.4.8) and the hypotheses of the theorem,  $g_2 \leq 0$ .

Suppose that  $u$  is continuous and bounded inside  $C$ . It will have the same properties inside  $C_1$ . As a continuous function  $u$  attains both its upper bound  $M$  (which cannot be negative since  $u = 0$  on  $C_1$ ) and its lower bound  $m$ . Let  $M$  be achieved at  $\lambda = \lambda_0, \mu = \mu_0$ . It follows that

$$\left( \frac{\partial u}{\partial \lambda} \right)_0 = 0, \quad \left( \frac{\partial u}{\partial \mu} \right)_0 = 0$$

where  $(f)_0$  means the value of  $f$  at  $\lambda = \lambda_0, \mu = \mu_0$ . Putting  $\lambda = \lambda_0, \mu = \mu_0$  in (10.4.7) we obtain

$$\left( \frac{\partial^2 u}{\partial \lambda^2} \right)_0 + \left( \frac{\partial^2 u}{\partial \mu^2} \right)_0 + g_2 M = 0. \quad (10.4.10)$$

But, for a maximum

$$\left( \frac{\partial^2 u}{\partial \lambda^2} \right)_0 \leq 0, \quad \left( \frac{\partial^2 u}{\partial \mu^2} \right)_0 \leq 0, \quad (10.4.11)$$

$$\left( \frac{\partial^2 u}{\partial \lambda^2} \right)_0 \left( \frac{\partial^2 u}{\partial \mu^2} \right)_0 > \left( \frac{\partial^2 u}{\partial \lambda \partial \mu} \right)_0^2. \quad (10.4.12)$$

If  $g = 0$  then  $g_2 = 0$  and we could satisfy (10.4.10) and (10.4.11) but not (10.4.12); the same is true if  $M = 0$ . On the other hand, if  $M > 0$  and  $g < 0$ , then  $g_2 M < 0$  and to satisfy (10.4.10) one at least of  $(\partial^2 u / \partial \lambda^2)_0, (\partial^2 u / \partial \mu^2)_0$  must be positive, contradicting (10.4.11). Thus the conditions for a maximum cannot be met inside  $C_1$  and so  $u$  must not be greater than zero inside  $C_1$ .

A similar argument based on the lower bound  $m$  reveals that  $u$  cannot be less than zero inside  $C_1$ .

□ Hence  $u$  is zero inside  $C_1$  and therefore zero inside  $C$ . The theorem is proved.

### **COROLLARY 10.4.1 (Uniqueness Property)**

*Under the conditions of Theorem 10.4.1 there is only one  $u$  which solves (10.3.1) and takes given values on  $C$ .*

Another way of describing this corollary is to say that the solution of the interior Dirichlet problem is unique for an elliptic partial differential equation in which  $a$  and  $g$  have opposite signs.

**PROOF** Suppose there were two possible solutions  $u_1$  and  $u_2$ . Then  $u_1 - u_2$  is zero on  $C$  and also satisfies (10.3.1). Consequently, Theorem 10.4.1 implies that  $u_1 - u_2$  vanishes inside  $C$  and the corollary has been demonstrated. □

It is important to observe that when  $a$  and  $g$  have the same signs Theorem 10.4.1 fails, in general. A counter-example is furnished by

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + (m^2 + n^2)u = 0$$

where  $m$  and  $n$  are integers. The solution

$$u = A \sin mx \sin ny$$

vanishes on the perimeter of the square  $0 \leq x \leq \pi, 0 \leq y \leq \pi$  but is clearly

not zero at all interior points.

---

## 10.5 Parabolic partial differential equations

In the parabolic case  $ac = b^2$ , there is only one family of real characteristics whose slope, according to (10.3.13), is given by

$$\frac{dy}{dx} = \frac{b}{a}.$$

One variable is fixed by  $\xi = \phi(x, y)$  where

$$-\frac{\partial \phi}{\partial x} = \frac{b}{a} \frac{\partial \phi}{\partial y} \quad (10.5.1)$$

and the other variable  $\eta = \psi(x, y)$  is selected in the same way as for first-order equations where only one set of characteristics is available. Thus  $a_1 = 0$  and, assuming  $a \neq 0, c = b^2/a$  so that

$$b_1 = \frac{\partial \psi}{\partial x} \left( a \frac{\partial \phi}{\partial x} + b \frac{\partial \phi}{\partial y} \right) + \frac{\partial \psi}{\partial y} \left( b \frac{\partial \phi}{\partial x} + \frac{b^2}{a} \frac{\partial \phi}{\partial y} \right) = 0$$

from (10.5.1). Hence (10.3.3) reduces to

$$c_1 \frac{\partial^2 u}{\partial \eta^2} + e_1 \frac{\partial u}{\partial \xi} + f_1 \frac{\partial u}{\partial \eta} + g_1 u = 0. \quad (10.5.2)$$

If  $a \equiv 0$  then  $b \equiv 0$  and the original equation (10.3.1) is already in the form (10.5.2) without introducing new variables. Thus (10.5.2) is the characteristic or normal form for a parabolic partial differential equation; here, in contrast to the elliptic case, there is no distinction between the characteristic and normal forms.

It may happen that  $\psi$  can be chosen so that  $f_1 = 0$  and then (10.5.2) has the simpler structure of

$$\frac{\partial^2 u}{\partial \eta^2} + e_2 \frac{\partial u}{\partial \xi} + g_2 u = 0. \quad (10.5.3)$$


---

## 10.6 Hyperbolic partial differential equations

In the hyperbolic case there are two families of real characteristics, since  $b^2 > ac$ . We may take  $\phi$  and  $\psi$  to be defined as in (10.4.1) and (10.4.2) when  $a \neq 0$  but they are now real. Thus  $a_1 = 0, c_1 = 0$  and  $b_1$  is given by (10.4.5),

i.e.,

$$b_1 = 2 \frac{\partial \phi}{\partial y} \frac{\partial \psi}{\partial y} \frac{ac - b^2}{a}.$$

Consequently,  $b_1$  is nonzero. If  $a \equiv 0$ , put  $\phi = x$  and specify the slope of  $\psi =$  constant by  $dy/dx = c/2b$ . Then  $b_1 = b\partial\psi/\partial y$  which is nonzero by virtue of  $b$  not being permitted to be zero in order to retain the hyperbolic character. Hence the characteristic or normal form of a hyperbolic partial differential equation is

$$\frac{\partial^2 u}{\partial \xi \partial \eta} + e_2 \frac{\partial u}{\partial \xi} + f_2 \frac{\partial u}{\partial \eta} + g_2 u = 0. \quad (10.6.1)$$

It is now convenient to collect together the normal forms of the three types:

*Elliptic:*

$$\frac{\partial^2 u}{\partial \lambda^2} + \frac{\partial^2 u}{\partial \mu^2} + e_2 \frac{\partial u}{\partial \lambda} + f_2 \frac{\partial u}{\partial \mu} + g_2 u = 0.$$

*Parabolic:*

$$\frac{\partial^2 u}{\partial \eta^2} + e_2 \frac{\partial u}{\partial \xi} + f_2 \frac{\partial u}{\partial \eta} + g_2 u = 0.$$

*Hyperbolic:*

$$\frac{\partial^2 u}{\partial \xi \partial \eta} + e_2 \frac{\partial u}{\partial \xi} + f_2 \frac{\partial u}{\partial \eta} + g_2 u = 0.$$

Note that, in all cases, the variables are real and the type fixed by the structure of the partial derivatives of the second order solely; the remaining terms in the partial differential equation play no role in determining the type, although their presence or absence will influence the solution.

## 10.7 The wave equation

The one-dimensional wave equation

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{a_0^2} \frac{\partial^2 u}{\partial t^2} = 0 \quad (10.7.1)$$

in which  $a_0$  is a positive constant is hyperbolic. For the characteristics we have

$$\left( \frac{dt}{dx} \right)^2 = \frac{1}{a_0^2}$$

so that

$$\frac{dt}{dx} = \pm \frac{1}{a_0}$$

with solutions  $x \pm a_0 t = \text{constant}$ . Hence the characteristic variables are  $\xi = x - a_0 t$ ,  $\eta = x + a_0 t$ . In terms of them (10.7.1) becomes

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0. \quad (10.7.2)$$

This tells us that  $\partial u / \partial \eta$  is a function of  $\eta$  only and hence that

$$u = f(\xi) + g(\eta) \quad (10.7.3)$$

where  $f$  and  $g$  are arbitrary functions of  $\xi$  and  $\eta$ , respectively. The formula (10.7.3) constitutes the general solution of (10.7.2). Correspondingly,

$$u = f(x - a_0 t) + g(x + a_0 t) \quad (10.7.4)$$

provides the general solution of (10.7.1).

An initial value problem for the wave equation can be solved by means of the general solution. Suppose the initial conditions are

$$u(x, 0) = h(x), \quad \left[ \frac{\partial}{\partial t} u(x, t) \right]_{t=0} = k(x)$$

for  $A \leq x \leq B$ . Then, from (10.7.4),

$$f(x) + g(x) = h(x), \quad (10.7.5)$$

$$a_0 \{-f'(x) + g'(x)\} = k(x), \quad (10.7.6)$$

for  $A \leq x \leq B$ . Integration of (10.7.6) gives

$$-f(x) + g(x) = \frac{1}{a_0} \int_A^x k(v) dv + C \quad (10.7.7)$$

where  $C$  is a constant. Combining (10.7.5) and (10.7.7) we obtain

$$g(x) = \frac{1}{2} \left( h(x) + \frac{1}{a_0} \int_A^x k(v) dv + C \right),$$

$$f(x) = \frac{1}{2} \left( h(x) - \frac{1}{a_0} \int_A^x k(v) dv - C \right)$$

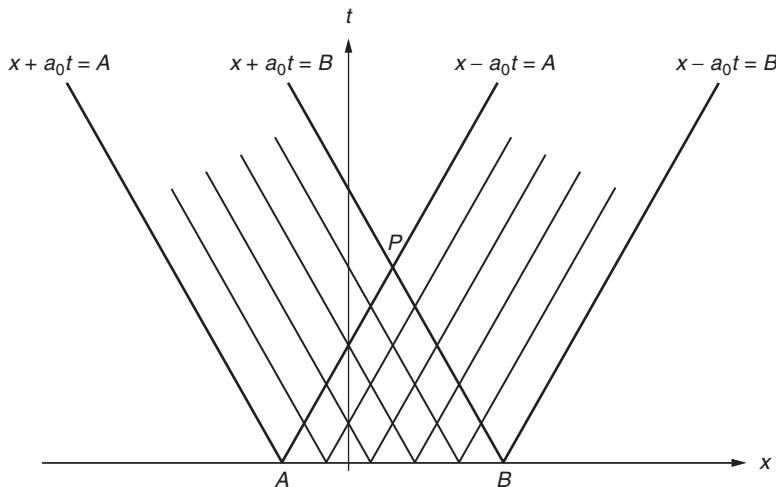
for  $A \leq x \leq B$ . Hence, if  $A \leq x - a_0 t \leq B$  and  $A \leq x + a_0 t \leq B$ ,

$$u = \frac{1}{2} \left( h(x - a_0 t) + \frac{1}{a_0} \int_A^{x - a_0 t} k(v) dv + C \right)$$

$$+ \frac{1}{2} \left( h(x + a_0 t) - \frac{1}{a_0} \int_A^{x + a_0 t} k(v) dv - C \right)$$

from which we deduce that

$$u = \frac{1}{2} \{h(x - a_0 t) + h(x + a_0 t)\} + \frac{1}{2a_0} \int_{x - a_0 t}^{x + a_0 t} k(v) dv. \quad (10.7.8)$$



**FIGURE 10.7.1:** Region where solution of the initial value problem is available.

This constitutes the solution to the initial value problem subject to the restrictions imposed on  $x - a_0 t$  and  $x + a_0 t$ . It is known as **d'Alembert's solution**.

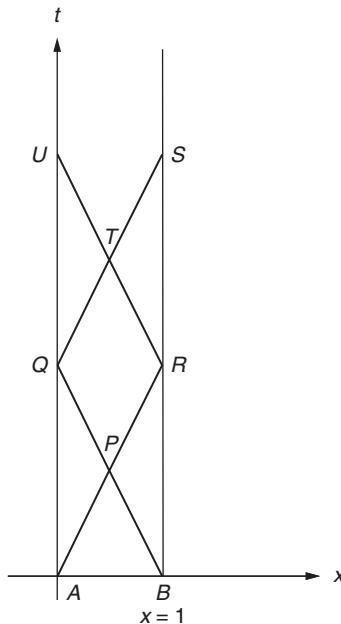
The restrictions on  $x \pm a_0 t$  confine the solution to the triangular region  $APB$  of Figure 10.7.1, which is the part of the  $(x, y)$  plane in  $t > 0$  common to the characteristics which intersect  $AB$ . Nothing can be said about the solution outside this triangle without additional information. It may occur that these initial data are given on the whole  $x$ -axis; in these circumstances  $A$  is  $-\infty$  and  $B$  is  $+\infty$  so that d'Alembert's solution is valid in the entire half-plane  $t > 0$ .

Quite often, the initial value problem will be mixed up with a boundary value problem. Suppose that the initial data are prescribed as before but in the interval  $(0, 1)$  of the  $x$ -axis and that there are also boundary conditions  $u(0, t) = 0, u(1, t) = 0$  for all  $t$ . To be consistent with these conditions it is necessary that  $h(0) = 0, h(1) = 0$ . The form (10.7.4) may still be used and, from the initial conditions,

$$f(x) = \frac{1}{2} \left( h(x) - \frac{1}{a_0} \int_0^x k(v) dv - C \right),$$

$$g(x) = \frac{1}{2} \left( h(x) + \frac{1}{a_0} \int_0^x k(v) dv + C \right)$$

for  $x$  in  $(0, 1)$ . Thus (10.7.8) continues to be the solution in the triangle  $APB$  of Figure 10.7.2. We know  $f$  in  $PRB$  but not  $g$  because  $x + a_0 t - 1$  is positive there. Similarly,  $f$  is needed in  $APQ$ . The boundary conditions are used to



**FIGURE 10.7.2:** Reflection of characteristics by a boundary.

supply the missing links; they imply that

$$f(-a_0t) + g(a_0t) = 0, \quad f(1 - a_0t) + g(1 + a_0t) = 0$$

for  $t \geq 0$  and therefore that

$$f(-y) + g(y) = 0, \quad (10.7.9)$$

$$f(1 - y) + g(1 + y) = 0 \quad (10.7.10)$$

for  $y \geq 0$ . From (10.7.9), it follows that

$$f(y) = -g(-y) = -\frac{1}{2} \left( h(-y) + \frac{1}{a_0} \int_0^{-y} k(v) dv + C \right) \quad (10.7.11)$$

when  $-1 \leq y \leq 0$ . This gives  $f(x - a_0t)$  in the parallelogram  $ARSQ$  where  $-1 \leq x - a_0t \leq 0$ . In particular, for  $APQ$  where  $-1 \leq x - a_0t \leq 0$  and  $0 \leq x + a_0t \leq 1$ ,

$$u = \frac{1}{2} \{h(x + a_0t) - h(a_0t - x)\} + \frac{1}{2a_0} \int_{a_0t-x}^{a_0t+x} k(v) dv.$$

In a similar way, (10.7.10) supplies

$$g(y) = -f(2 - y) = -\frac{1}{2} \left( h(2 - y) - \frac{1}{a_0} \int_0^{2-y} k(v) dv - C \right) \quad (10.7.12)$$

for  $1 \leq y \leq 2$ ;  $g(y)$  for  $2 \leq y \leq 3$  may also be obtained by means of the formula for  $f(y)$  in (10.7.11). Via (10.7.12) we have  $g(x + a_0 t)$  in the parallelogram  $BQUR$ . Thus, in  $BPR$  where  $0 \leq x - a_0 t \leq 1$  and  $1 \leq x + a_0 t \leq 2$ ,

$$u = \frac{1}{2}\{h(x - a_0 t) - h(2 - x - a_0 t)\} + \frac{1}{2a_0} \int_{x-a_0 t}^{2-x-a_0 t} k(v) dv$$

and in  $PQTR$  where both (10.7.11) and (10.7.12) are applicable:

$$u = -\frac{1}{2}\{h(a_0 t - x) + h(2 - a_0 t - x)\} + \frac{1}{2a_0} \int_{a_0 t - x}^{2-x-a_0 t} k(v) dv.$$

To find  $u$  in  $Q TU$  it is necessary to employ (10.7.9) again but it is clear that we can determine  $u(x, t)$  for  $0 \leq x \leq 1$  and any  $t$ . The diagram makes it evident that the characteristics are being reflected at the boundaries and that, at each reflection, a new form of the solution has to be constructed.

A more general form of the wave equation is

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{\{a(x)\}^2} \frac{\partial^2 u}{\partial t^2} \quad (10.7.13)$$

where  $a$  is a function of  $x$  only. The characteristic coordinates are now given by

$$\xi = t - \int^x \frac{dv}{a(v)}, \quad \eta = t + \int^x \frac{dv}{a(v)}.$$

Hence

$$\begin{aligned} \frac{\partial u}{\partial x} &= -\frac{1}{a} \frac{\partial u}{\partial \xi} + \frac{1}{a} \frac{\partial u}{\partial \eta}, & \frac{\partial u}{\partial t} &= \frac{\partial u}{\partial \xi} + \frac{\partial u}{\partial \eta}, \\ \frac{\partial^2 u}{\partial x^2} &= \left( \frac{\partial^2 u}{\partial \xi^2} - 2 \frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \eta^2} \right) \frac{1}{a^2} - \left( \frac{\partial u}{\partial \eta} - \frac{\partial u}{\partial \xi} \right) \frac{a'}{a^2}. \\ \frac{\partial^2 u}{\partial t^2} &= \frac{\partial^2 u}{\partial \xi^2} + 2 \frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \eta^2}. \end{aligned}$$

Accordingly (10.7.13) transforms to

$$4 \frac{\partial^2 u}{\partial \xi \partial \eta} + a' \left( \frac{\partial u}{\partial \eta} - \frac{\partial u}{\partial \xi} \right) = 0. \quad (10.7.14)$$

To express  $a'$  in terms of  $\xi$  and  $\eta$ , notice that

$$\eta - \xi = 2 \int^x \frac{dv}{a(v)}$$

so that  $x$  is a function of  $\eta - \xi$ . Thus  $a'$  will also be some function of  $\eta - \xi$ .

In the particular case when  $a = Ax$ ,  $a' = A$  and the substitution  $\xi = X/A$ ,  $\eta = Y/A$  converts (10.7.14) to

$$4 \frac{\partial^2 u}{\partial X \partial Y} + \frac{\partial u}{\partial Y} - \frac{\partial u}{\partial X} = 0.$$

More generally, if  $a = Ax^\nu$ ,  $a' = \nu Ax^{\nu-1}$  and, when  $\nu \neq 1$ ,

$$\eta - \xi = -2/A(\nu - 1)x^{\nu-1}$$

with the consequence that

$$a' = \frac{-2\nu}{\nu - 1} \frac{1}{\eta - \xi}.$$

In this case (10.7.14) becomes

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = \frac{\nu}{2(\nu - 1)} \frac{1}{\xi - \eta} \left( \frac{\partial u}{\partial \xi} - \frac{\partial u}{\partial \eta} \right). \quad (10.7.15)$$

This is related to the Euler–Darboux equation to be discussed in Section 10.9, where explicit solutions are found for certain values of  $\nu$ .

---

## 10.8 Typical problems for the hyperbolic equation

In order to see what problems could be solved by tracing characteristics it is plausible to assume, in view of the importance of the second derivative in the classification, that, when the conditions are such that

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0 \quad (10.8.1)$$

can be resolved, a similar problem will be reasonably posed for (10.6.1) and thereby for the general hyperbolic partial differential equation. Some examples will now be given where resolution is possible, together with the corresponding interpretation for the general equation.

### 10.8.1 The two-characteristic problem

Suppose that  $u$  is given on the portion of the characteristic  $0 \leq \xi \leq a$ ,  $\eta = 0$ , say  $u = h(\xi)$ , and on the piece of characteristic  $\xi = 0$ ,  $0 \leq \eta \leq b$  by  $u = k(\eta)$  (Figure 10.8.1). From (10.7.3)

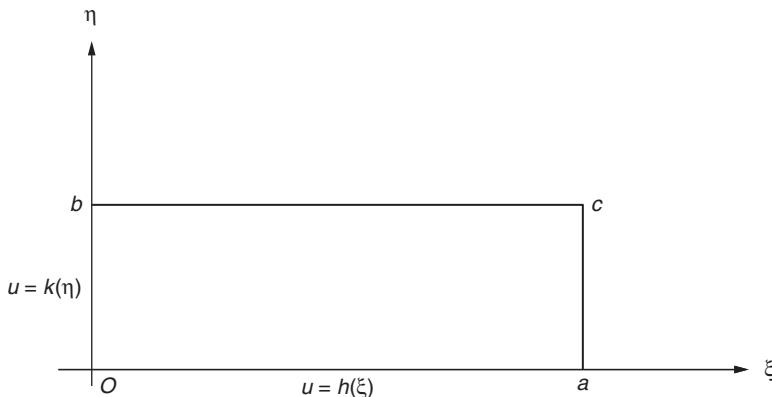
$$u = f(\xi) + g(\eta).$$

By putting  $\eta = 0$  we obtain

$$f(\xi) + g(0) = h(\xi) \quad (10.8.2)$$

for  $0 \leq \xi \leq a$  and by placing  $\xi = 0$

$$f(0) + g(\eta) = k(\eta) \quad (10.8.3)$$



**FIGURE 10.8.1:** The two-characteristic problem.

for  $0 \leq \eta \leq b$ . Hence

$$\begin{aligned} u &= h(\xi) + k(\eta) - f(0) - g(0) \\ &= h(\xi) + k(\eta) - h(0) \end{aligned}$$

on making  $\xi = 0$  in (10.8.2). The same result is obtained from (10.8.3) since  $h(0) = k(0)$  for the data on  $u$  to be consistent at the origin.

This solution will hold everywhere in the rectangle  $Oacb$ , which is the region common to the characteristics that intersect the data lines.

Translating this back to the  $(x, y)$ -plane and general partial differential equation, we see that giving  $u$  on the two characteristics  $O'A, O'B$  (Figure 10.8.2) determines  $u$  in the shaded region  $O'ACB$  where  $AC$  and  $BC$  are the other characteristics through  $A$  and  $B$ , respectively.

### 10.8.2 The mixed problem

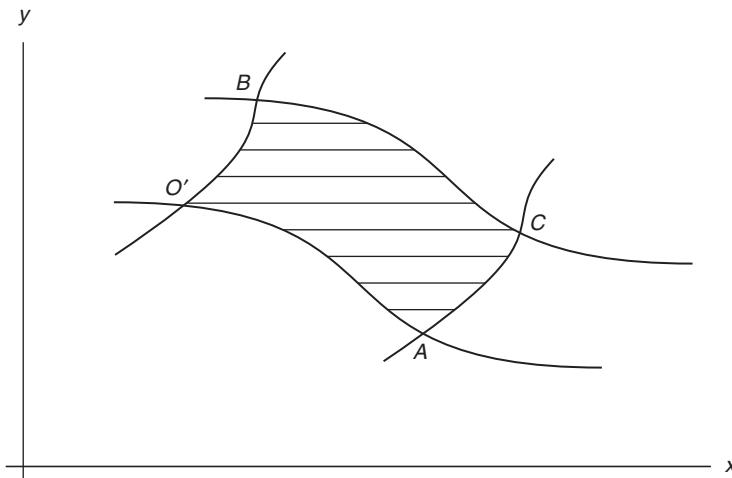
In the mixed problem,  $u = h(\xi)$  on  $0 \leq \xi \leq a, \eta = 0$  as before but either  $u$  or a linear combination of its first partial derivatives is given on an arc  $Oc$  in the first quadrant, the arc being met at most once by any characteristic.

Suppose that the equation of  $Oc$  is  $\eta = \theta(\xi)$ . Take a new variable  $\chi = \theta(\xi)$ . Then  $Oc$  becomes the straight line  $\eta = \chi$  and the form of (10.8.1) is unaltered so that the characteristics remain parallel to the coordinate axes. There is therefore no loss of generality in selecting  $Oc$  as the straight line  $\eta = \xi$  in the first place.

Suppose that  $u = k(\xi)$  on  $Oc$  with  $k(0) = h(0)$ . Then we have

$$f(\xi) + g(0) = h(\xi), \quad (10.8.4)$$

$$f(\xi) + g(\xi) = k(\xi) \quad (10.8.5)$$



**FIGURE 10.8.2:** The two-characteristic problem in the general case.

for  $0 \leq \xi \leq a$ . Therefore

$$u = h(\xi) + k(\eta) - h(\eta)$$

for  $0 \leq \xi \leq a$ ,  $0 \leq \eta \leq a$ . Consequently,  $u$  has been determined everywhere within the square with side  $Oa$  (Figure 10.8.3).

Instead of  $u$  being specified on  $Oc$ , the linear combination

$$\alpha(\xi) \frac{\partial u}{\partial \xi} + \beta(\xi) \frac{\partial u}{\partial \eta} = k(\xi)$$

might be given on  $Oc$ . In that case (10.8.5) will be replaced by

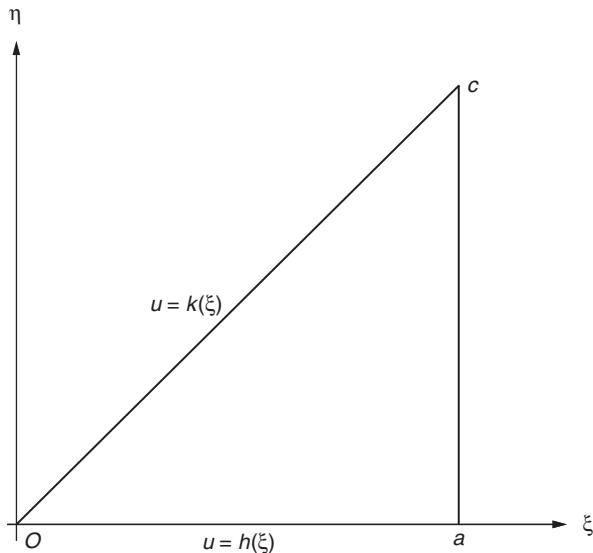
$$\alpha(\xi) f'(\xi) + \beta(\xi) g'(\xi) = k(\xi).$$

Since  $f$  is known from (10.8.4),  $g$  can be calculated by integration and again  $u$  has been discovered in the square.

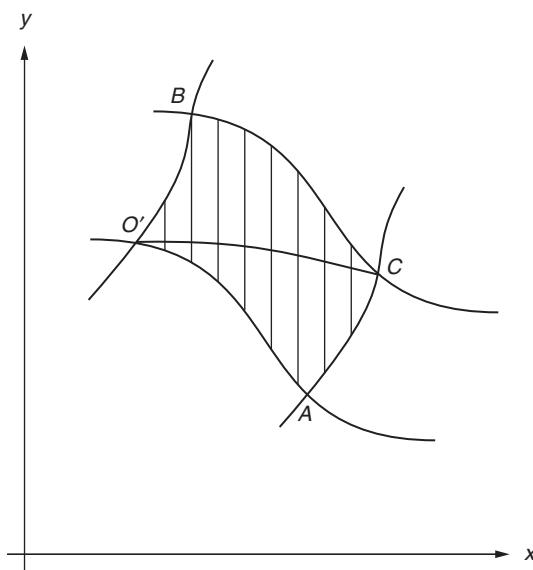
Going back to the  $(x, y)$ -plane we can say that when  $u$  is given on the characteristic arc  $O'A$  (Figure 10.8.4) and either  $u$  or a linear combination of its first partial derivatives is specified on the arc  $O'C$ , which lies between the characteristics  $O'A$ ,  $O'B$  and is met at most once by any characteristic, then  $u$  is determined in the shaded region  $O'ACB$  where  $AC$  and  $BC$  are characteristics.

### 10.8.3 Cauchy's problem

Cauchy's problem is to determine  $u$  when  $u$ ,  $\partial u / \partial \xi$  and  $\partial u / \partial \eta$  are given on  $Oc$ , where  $Oc$  has the same properties as in the previous section, i.e., it is not



**FIGURE 10.8.3:** The mixed problem.



**FIGURE 10.8.4:** The general mixed problem.

a characteristic and is met at most once by a characteristic. As before, there is no loss of generality in taking  $Oc$  to be  $\eta = \xi$ .

Suppose that on  $Oc$

$$u = h(\xi), \quad \frac{\partial u}{\partial \xi} = k(\xi), \quad \frac{\partial u}{\partial \eta} = l(\xi).$$

There is a connection between  $h, k$  and  $l$  because

$$h'(\xi) = \frac{du}{d\xi} = \frac{\partial u}{\partial \xi} \frac{d\xi}{d\xi} + \frac{\partial u}{\partial \eta} \frac{d\eta}{d\xi} = k(\xi) + l(\xi) \quad (10.8.6)$$

since the equation of  $Oc$  is  $\eta = \xi$ .

The imposition of the given conditions supplies

$$\begin{aligned} f(\xi) + g(\xi) &= h(\xi), \\ f'(\xi) &= k(\xi), \quad g'(\xi) = l(\xi). \end{aligned}$$

The three are consistent on account of (10.8.6). With  $f$  found from the second equation and  $g$  from the first,  $u$  is obtained in the same square as in the mixed problem.

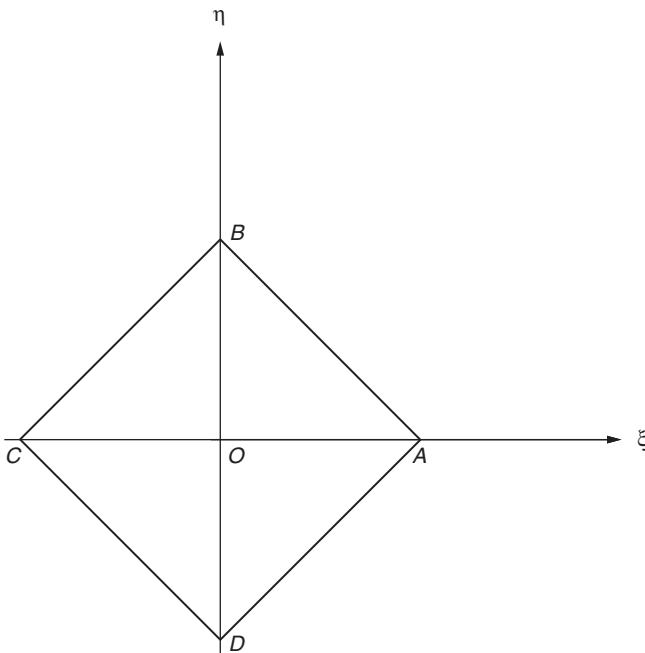
For the general equation, data on  $u$  and its first partial derivative on  $O'C$  in Figure 10.8.4 will fix  $u$  in the characteristic domain  $O'ACB$ .

A return to the reflection problem for the wave equation in Section 10.7 is pertinent here, to examine it from the point of view of this section.  $AB$  of Figure 10.7.2 is not a characteristic and  $u, \partial u / \partial t$  are given on it. Therefore there is a Cauchy problem and, from the above, a solution is available in  $APB$ . From this solution  $u$  is known on  $AP$ , which is a characteristic arc, and  $u = 0$  on  $AQ$ , which is not a characteristic but lies between the two characteristics through  $A$ . Hence there is a mixed problem and  $u$  can be found in  $APQ$ . Similarly, a mixed problem gives  $u$  in  $BPR$ . Now  $u$  is known on  $PQ$  and  $PR$ , which are both characteristic arcs, and so by solving a two-characteristic problem we determine  $u$  in  $PRTQ$ . In this way, a solution to a reflection problem for a general hyperbolic equation can be built up by means of characteristics.

It is possible to find  $u$  when  $u$  is given on two arcs  $Oc, Od$  in the first quadrant when they are not characteristics. If  $Oc$  is in the first quadrant and  $Od$  is in the fourth there is insufficient information for a unique solution. For if we give  $u$  any values on the  $\xi$ -axis we have a mixed problem in the first quadrant and another one in the fourth, so that a solution can be found.

When elliptic partial differential equations were discussed it was pointed out that Dirichlet's problem was an appropriate one. It will now be shown why it is not suitable for hyperbolic equations in general.

Let  $u$  be prescribed on the boundary  $ABCD$  of Figure 10.8.5. The aim is to find  $u$  inside; so assume some values for  $u$  on  $OB$ . Then the form of  $u$  in  $OAB$  is known by solving a mixed problem and similarly in  $OBC$ . From the discovered values of  $u$  on  $OC$ ,  $u$  is determined in  $OCD$  via a mixed



**FIGURE 10.8.5:** The Dirichlet problem for the hyperbolic equation.

problem. Similarly, values in  $OAD$  can be found. In order that the solution be correct, the values of  $u$  on  $OD$  from  $OCD$  and  $OAD$  must agree. This places a condition on the assumed values on  $OB$ . There is no guarantee that this condition can be fulfilled. If it cannot be met there is no solution to the Dirichlet problem. If the condition on  $OB$  can be satisfied there will be at least one solution and there may be several since there may be more than one way of complying with the condition.

We conclude that the Dirichlet problem for the hyperbolic differential equation must be treated with great caution.

## 10.9 The Euler–Darboux equation

This section is devoted to the hyperbolic partial differential equation

$$\frac{\partial^2 u}{\partial x \partial y} = \frac{m}{x-y} \left( \frac{\partial u}{\partial x} - \frac{\partial u}{\partial y} \right) \quad (10.9.1)$$

where  $m$  is a positive integer. This form has already occurred in Section 10.7 and it is called the **Euler–Darboux equation**.

Now

$$\frac{\partial^2}{\partial x \partial y} \{(x-y)v\} = \frac{\partial}{\partial x} \left( (x-y) \frac{\partial v}{\partial y} - v \right) = (x-y) \frac{\partial^2 v}{\partial x \partial y} + \frac{\partial v}{\partial y} - \frac{\partial v}{\partial x}. \quad (10.9.2)$$

Consequently, when  $m = 1$ , (10.9.1) after multiplication by  $x - y$  may be written, on account of (10.9.2), as

$$\frac{\partial^2}{\partial x \partial y} \{(x-y)u\} = 0.$$

It follows that

$$u = \frac{f(x) + g(y)}{x - y},$$

with  $f$  and  $g$  arbitrary, provides the general solution of the Euler–Darboux equation when  $m = 1$ .

Multiply (10.9.1) by  $x - y$  and apply the derivative  $\partial^2/\partial x \partial y$ . Then, by virtue of (10.9.2),

$$(x-y) \frac{\partial^2}{\partial x \partial y} \frac{\partial^2 u}{\partial x \partial y} + \left( \frac{\partial}{\partial y} - \frac{\partial}{\partial x} \right) \frac{\partial^2 u}{\partial x \partial y} = m \left( \frac{\partial}{\partial x} - \frac{\partial}{\partial y} \right) \frac{\partial^2 u}{\partial x \partial y}.$$

Thus, if  $u$  satisfies (10.9.1),  $\partial^2 u/\partial x \partial y$  is a solution of (10.9.1) with  $m$  replaced by  $m + 1$ . The solution already derived for  $m = 1$  now permits the statement that

$$u = \frac{\partial^{2m-2}}{\partial x^{m-1} \partial y^{m-1}} \left( \frac{f(x) + g(y)}{x - y} \right),$$

with  $f$  and  $g$  arbitrary, solves (10.9.1).

## 10.10 Visualisation of solutions

MATLAB provides an effective means of visualising solutions to partial differential equations. Typically, this takes the form of a surface or contour plot for functions of two variables. Alternatively, if one of the variables is time, animations can be used. Below are template MATLAB programs for doing this, followed by some examples for specific problems.

### Surface plotting

This M-file produces a surface plot of a function  $u(x, y)$  against  $x$  and  $y$ . If the command `surf` is replaced by `contour`, a contour plot will be produced instead of a surface. Using `surf c` will produce a combination surface/contour plot.

```

x = x_min:h:x_max; y = y_min:h:y_max;
[X, Y] = meshgrid(x, y);
% Define U here
surf(X, Y, U);
xlabel('x'); ylabel('y'); zlabel('u');

```

In addition to defining the function to be plotted ( $U$ ), the values of the minimum and maximum  $x$  and  $y$  values ( $x_{\min}$ ,  $x_{\max}$ ,  $y_{\min}$ ,  $y_{\max}$ ) and the grid size ( $h$ ) need to be specified. Note that the `meshgrid` command takes the vectors  $x$  and  $y$  and forms matrices  $X$  and  $Y$  corresponding to the  $x$  and  $y$  coordinates in a rectangular grid.

## Animations in one dimension

This M-file produces graphs of a function  $u(x, t)$  against  $x$  for a series of values of  $t$ .

```

x = x_min:h:x_max; t = t_min:dt:t_max;
N_frames = length(t);
wait_time = 0.01;
for I = 1:N_frames
    % Define U here
    plot(x, U)
    xlabel('x'); ylabel('t');
    axis([x_min x_max u_min u_max])
    title(sprintf('t=%4.2e\n', t(I)))
    pause(wait_time);
end

```

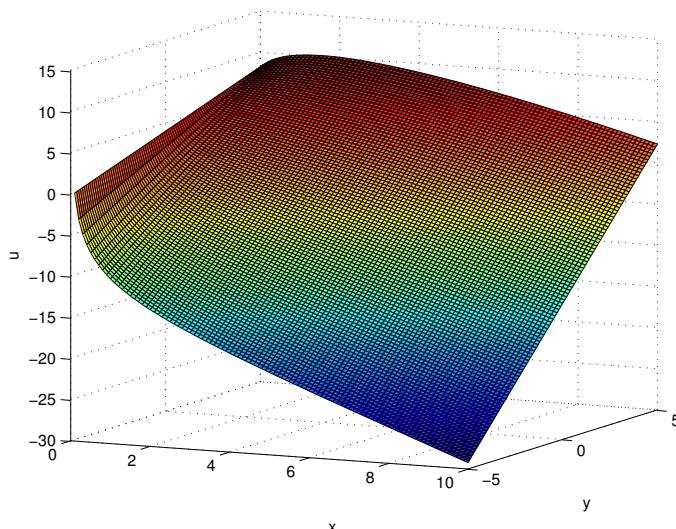
## Animations in two dimensions

This M-file produces surface plots of a function  $u(x, y, t)$  against  $x$  and  $y$  for a series of values of  $t$ .

```

x = x_min:h:x_max; y = y_min:h:y_max; t = t_min:dt:t_max;
[X, Y] = meshgrid(x, y);
N_frames = length(t);
wait_time = 0.01;
for I = 1:N_frames
    % Define U here
    surf(X, Y, U)
    xlabel('x'); ylabel('t'); zlabel('u');
    axis([x_min x_max y_min y_max u_min u_max])
    title(sprintf('t=%4.2e\n', t(I)))
    pause(wait_time);
end

```



**FIGURE 10.10.1:** Surface plot of the solution (10.1.15) for  $u(x, y)$ .

Sometime, as for example in the spiral wave solution (8.4.16) of Chapter 8, a solution to a partial differential equation may be defined in polar coordinates. This situation can be dealt with by computing the solution  $u(r, \theta, t)$  over a polar grid, instead of a rectangular grid. Simply change the commands defining  $x$  and  $y$  in the above by equivalent commands defining appropriate ranges for  $r$  and  $\theta$ :

```
r = 0:dr:r_max; theta = 0:dtheta:2*pi;
[R, Theta] = meshgrid(r, theta);
```

Then define the solution  $U$  in terms of the polar coordinates  $R$  and  $\Theta$ , and replace the **surf** command by

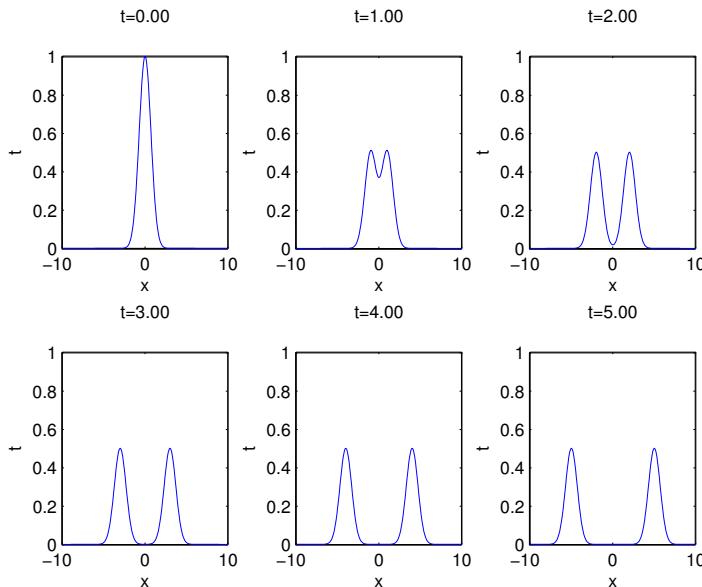
```
surf(R.*cos(Theta), R.*sin(Theta), U)
```

### Example 10.10.1

Plot the solution found in Example 10.1.1.

The solution defined by equation (10.1.15) can be calculated with the following MATLAB command (which should be inserted in place of `% Define U here` in the surface plotting program above).

```
U = (Y-X/2-log(X)/2).*(2+log(X)/2)+X/4+3/4+(log(X)).^2/2;
```



**FIGURE 10.10.2:** Snapshots of d'Alembert's solution (10.7.8) to the wave equation with  $a_0 = 1$ ,  $k(x) \equiv 0$  and  $f(x) = \exp(-x^2)$ .

Note the use of the component-by-component operators `.*` and `.^` (see Section 5.8). Running the program produces the graph shown in Figure 10.10.1. □

### Example 10.10.2

Plot d'Alembert's solution (10.7.8) to the wave equation with  $a_0 = 1$ ,  $k(x) = 0$  and  $h(x) = \exp(-x^2)$ .

According to equation (10.7.8), the solution is

$$u(x, t) = \frac{1}{2} \left( \exp(-(x-t)^2) + \exp(-(x+t)^2) \right).$$

This can be defined in MATLAB by the command:

```
U = (exp(-(x-t(I)).^2)+exp(-(x+t(I)).^2))/2;
```

Figure 10.10.2 shows some snapshots from the animation produced by the program above. □

## Exercises

10.1 If the characteristics of

$$a(x, y) \frac{\partial u}{\partial x} + b(x, y) \frac{\partial u}{\partial y} = 0$$

are  $\phi(x, y) = \text{constant}$ , show that the general solution is  $u = G\{\phi(x, y)\}$  where  $G$  is an arbitrary function.

10.2 Solve

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 2$$

given that  $u = 2x^2$  when  $y = 0$ . Use MATLAB to produce a surface plot of the solution  $u(x, y)$ .

10.3 Solve

$$x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = u$$

given that  $u = 3x$  on  $x + y = 1$ .

10.4 Make the substitution  $x = \rho \cos \theta, y = \rho \sin \theta$  in

$$(x - y) \frac{\partial u}{\partial x} + (x + y) \frac{\partial u}{\partial y} = 4.$$

Hence find the solution in  $0 \leq \theta < 2\pi$  such that  $u = 2r$  when  $\theta = 0$ .

10.5 If  $u = 3y^2$  when  $x = 0$  solve

$$\frac{\partial u}{\partial x} - (x - y - 1) \frac{\partial u}{\partial y} = 3x(y - x).$$

10.6 Given that  $u = y^2$  on  $x = 0$ , find  $u$  in  $y > x > 0$  such that

$$y^2 \frac{\partial u}{\partial x} + xy \frac{\partial u}{\partial y} = x.$$

Use MATLAB to produce a surface plot of the solution  $u(x, y)$ .

10.7 Find the solution in  $x > 0, y > 0$  of

$$x \frac{\partial u}{\partial x} - 2y \frac{\partial u}{\partial y} = x^2 + y^2$$

such that  $u = x^2$  on  $y = 1$ .

10.8 Find the solution in  $x > 0$  of

$$x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = u - 2xy$$

given that  $u = 2y^2 + 2$  on  $x = 2$ . Use MATLAB to produce a surface plot of the solution  $u(x, y)$ .

10.9 If

$$x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = xy$$

and  $u = \frac{1}{2}x^2$  on  $y = x$  determine  $u$ .

10.10 In what regions are the following partial differential equations elliptic, hyperbolic or parabolic?

(a)  $\frac{\partial^2 u}{\partial x^2} + 4 \frac{\partial^2 u}{\partial y^2} = 0$ ,

(b)  $7 \frac{\partial^2 u}{\partial x \partial y} - 3 \frac{\partial u}{\partial y} = 0$ ,

(c)  $x^2 \frac{\partial^2 u}{\partial x^2} + 4y \frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} + 2 \frac{\partial u}{\partial x} = 0$ ,

(d)  $3y \frac{\partial^2 u}{\partial x^2} - x \frac{\partial^2 u}{\partial y^2} = 0$ ,

(e)  $\frac{\partial^2 u}{\partial x^2} + 4 \frac{\partial u}{\partial x} = 0$ ,

(f)  $x^2 y^2 \frac{\partial^2 u}{\partial x^2} + 2xy \frac{\partial^2 u}{\partial x \partial y} + \frac{\partial^2 u}{\partial y^2} = 0$ .

10.11 Make the change of variables  $x = \rho \cos \theta$ ,  $y = \rho \sin \theta$  in

$$(x^2 \cos^2 \alpha - y^2 \sin^2 \alpha) \frac{\partial^2 u}{\partial x^2} + 2xy \frac{\partial^2 u}{\partial x \partial y} + (y^2 \cos^2 \alpha - x^2 \sin^2 \alpha) \frac{\partial^2 u}{\partial y^2} = 0$$

and show that  $\ln \rho \pm \theta \cot \alpha$  are characteristic coordinates. Express the equation in characteristic form.

10.12 Express the partial differential equation

$$\frac{\partial^2 u}{\partial x^2} - e^{-x} \frac{\partial^2 u}{\partial y^2} = 0$$

in characteristic form.

10.13 Find the two families of characteristics of

$$2x^2 \frac{\partial^2 u}{\partial x^2} - 5xy \frac{\partial^2 u}{\partial x \partial y} + 2y^2 \frac{\partial^2 u}{\partial y^2} + 2x \frac{\partial u}{\partial x} + 2y \frac{\partial u}{\partial y} = 0.$$

Convert the equation to characteristic form and hence find  $u$  given that  $u = 2x^2 - 6$  and  $\partial u / \partial y = 3 - x^2$  on  $y = 1$ .

10.14 The characteristic form of

$$\{a(x)\}^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial y^2}$$

is

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = \frac{1}{2(\xi - \eta)} \left( \frac{\partial u}{\partial \xi} - \frac{\partial u}{\partial \eta} \right).$$

Show that  $a(x) = Ae^{Bx}$  where  $A$  and  $B$  are constants.

If the characteristic form is

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = \frac{\nu}{\xi - \eta} \left( \frac{\partial u}{\partial \xi} - \frac{\partial u}{\partial \eta} \right)$$

and  $\nu \neq \frac{1}{2}$ , find a formula for  $a$ .

10.15 The solution of

$$\frac{\partial^2 u}{\partial x^2} + 3 \frac{\partial^2 u}{\partial y^2} + 2x = 0$$

is required in the square  $0 < x < a, 0 < y < a$ . You are asked to advise which of the following sets of boundary conditions on the perimeter of the square are suitable:

- (a)  $u$  given on all four sides;
- (b)  $u$  given on two adjacent sides only;
- (c)  $u$  given on three sides and its normal derivative given on the middle side of the three.

Decide for (a), (b) and (c) whether there is a solution and, if so, whether there is more than one.

10.16 Repeat Exercise 10.15 for

$$\frac{\partial^2 u}{\partial x^2} + 3 \frac{\partial^2 u}{\partial y^2} + u = 0.$$

10.17 Repeat Exercise 10.15 for

$$\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} - 2u = 0.$$

10.18 Repeat Exercise 10.15 for

$$\frac{\partial^2 u}{\partial x \partial y} + 3u = 0.$$

10.19 Use MATLAB to produce a surface plot of the solution (10.1.17) over a suitable range of  $x$  and  $y$ .

- 10.20 Use MATLAB to produce plots or animations of d'Alembert's solution (10.7.8) to the wave equation with  $a_0 = 1$ ,  $k(x) = 0$  and  $h(x) = \cos x$ . What happens if you increase or decrease the value of  $a_0$ ?
- 10.21 Use MATLAB to produce animations for  $0 < r < 4$  of the spiral wave solution to the  $\lambda - \omega$  system (see Section 8.4):

$$u(r, \theta, t) = \cos(\Omega t + m\theta + \psi(r)),$$

where  $\Omega = 1$ ,  $m = 1$  and the function  $\psi(r)$  is:

- (a) an Archimedean spiral  $\psi(r) = r$ ;
- (b) a logarithmic spiral  $\psi(r) = \ln r$ .

Note that the spiral wave is best visualised by plotting an animated filled contour plot, rather than an animated surface plot. This may be achieved by replacing the command `surf` in the program in Section 10.10 by the command `contourf`.

# Chapter 11

---

## Evolutionary Equations

---

### 11.1 The heat equation

The equation

$$\frac{\partial u}{\partial t} = D\nabla^2 u, \quad (11.1.1)$$

where  $D$  is a positive constant and  $\nabla^2$  is the Laplacian (see equation (8.4.3)), arises in the study of problems of heat conduction and for this reason is referred to as the **heat equation**. However, (11.1.1) arises in many other areas in which diffusion processes occur (the subject of Chapter 12) and in this context is referred to as the **classic diffusion equation**.

In this section we shall study (11.1.1) in one space dimension and where  $D = 1$ , that is, we consider the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}. \quad (11.1.2)$$

In addition to the techniques described in Chapter 10, the method of *separation of variables* is a powerful tool with which to solve linear partial differential equations. To begin with, we seek solutions of (11.1.2) in the form

$$u(x, t) = X(x)T(t). \quad (11.1.3)$$

Substitution of this into (11.1.2) leads to the identity

$$X \frac{dT}{dt} = T \frac{d^2 X}{dx^2}$$

or

$$\frac{1}{T} \frac{dT}{dt} = \frac{1}{X} \frac{d^2 X}{dx^2}. \quad (11.1.4)$$

Now  $\frac{1}{T} \frac{dT}{dt}$  is a function of  $t$  only, while  $\frac{1}{X} \frac{d^2 X}{dx^2}$  is a function of  $x$  only. Consequently, both sides of (11.1.4) must be equal to a constant, say  $\lambda$ . Thus  $X$  and  $T$  must satisfy the ordinary differential equations

$$\begin{aligned} \frac{dT}{dt} - \lambda T &= 0, \\ \frac{d^2 X}{dx^2} - \lambda X &= 0. \end{aligned} \quad (11.1.5)$$

These have the general solutions

$$X(x) = \exp \pm \sqrt{\lambda}x, \quad T(t) = \exp \lambda t.$$

Hence (11.1.2) has a particular solution of the form

$$u(x, t) = A \exp(\pm \sqrt{\lambda}x + \lambda t). \quad (11.1.6)$$

Now suppose we are given initial data of the form

$$u(x, 0) = A \exp i\alpha x; \quad (11.1.7)$$

then a solution of (11.1.2) that satisfies (11.1.7) is

$$u(x, t) = A \exp(i\alpha x - \alpha^2 t).$$

It is of interest to see how solutions of this form can be utilised in order to provide a solution to the more general initial value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad -\infty < x < \infty, \\ u(x, 0) &= f(x). \end{aligned} \quad (11.1.8)$$

To carry out this program we shall need the **Fourier Integral Theorem**.

### **THEOREM 11.1.1**

Let  $f(x)$  and  $\frac{df}{dx}$  be continuous and  $\int_{-\infty}^{\infty} |f(x)| dx < \infty$ .

Then

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} f(\xi) e^{-i\alpha\xi} d\xi \right) e^{i\alpha x} d\alpha. \quad (11.1.9)$$

The quantity

$$\hat{f}(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\xi) e^{-i\alpha\xi} d\xi \quad (11.1.10)$$

is called the **Fourier transform** of  $f$  and (11.1.9) provides the **reciprocal relation**

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(\alpha) e^{i\alpha x} d\alpha. \quad (11.1.11)$$

Fourier transforms are another of an important class of techniques available for the solution of partial differential equations. To see this in the present situation, consider a special solution of the heat equation in the form

$$u(x, t; \alpha) = A(\alpha) e^{i\alpha x - \alpha^2 t}, \quad (11.1.12)$$

where  $A(\alpha)$  is an arbitrary function of  $\alpha$ . Since (11.1.2) is a linear partial differential equation, we know that a linear combination of solutions of the

form (11.2.12) is also a solution. This, of course, is the essence of the separation of variables method. With this in mind, we take the reasonable step that the integral with respect to  $\alpha$  of (11.2.12) may also be a solution.

Consider the function

$$u(x, t) = \int_{-\infty}^{\infty} u(x, t; \alpha) d\alpha = \int_{-\infty}^{\infty} A(\alpha) e^{i\alpha x - \alpha^2 t} d\alpha. \quad (11.1.13)$$

What we must do now is first to find a specific form for  $A(\alpha)$  so that  $u(x, t)$  defined by (11.1.13) satisfies the initial condition in (11.1.8) and then to show that (11.1.13) is in fact the (unique) solution to the problem (11.1.8).

Setting  $t = 0$  in (11.1.13) we see that we must have

$$f(x) = \int_{-\infty}^{\infty} A(\alpha) e^{i\alpha x} d\alpha$$

or

$$\frac{f(x)}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} A(\alpha) e^{i\alpha x} d\alpha.$$

In other words,  $A(\alpha)$  must be taken as the Fourier transform of  $f(x)/\sqrt{2\pi}$ . That is,

$$A(\alpha) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\xi) e^{-i\alpha\xi} d\xi$$

and so from (11.1.13) we have

$$u(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(\xi) d\xi \int_{-\infty}^{\infty} e^{i\alpha(x-\xi) - \alpha^2 t} d\alpha. \quad (11.1.14)$$

The result (11.1.14) can be simplified.

Consider

$$\begin{aligned} I &= \int_{-\infty}^{\infty} e^{i\alpha(x-\xi) - \alpha^2 t} d\alpha \\ &= \int_{-\infty}^{\infty} \exp \left[ -t \left( \alpha - \frac{i(x-\xi)}{2t} \right)^2 - \frac{(x-\xi)^2}{4t} \right] d\alpha. \end{aligned}$$

Now set

$$\alpha - \frac{i(x-\xi)}{2t} = \frac{\beta}{\sqrt{t}}$$

to obtain

$$I = \frac{1}{\sqrt{t}} \exp \left( -\frac{(x-\xi)^2}{4t} \right) \int_{-\infty}^{\infty} e^{-\beta^2} d\beta.$$

The integral  $\int_{-\infty}^{\infty} e^{-\beta^2} d\beta$  is known as the *error integral* and takes the value  $\sqrt{\pi}$ . Thus

$$I = \sqrt{\pi/t} \exp \left( -\frac{(x-\xi)^2}{4t} \right)$$

and so on using this in (11.1.14) we obtain the expression

$$u(x, t) = \int_{-\infty}^{\infty} \frac{\exp[-(x - \xi)^2/4t]}{\sqrt{4\pi t}} f(\xi) d\xi. \quad (11.1.15)$$

The function

$$K(x - \xi, t) = \frac{\exp[-(x - \xi)^2/4t]}{\sqrt{4\pi t}}$$

is of special significance in the study of diffusion problems and is called the **fundamental solution** of the heat equation. It has a number of interesting properties that are important if we wish to study the regularity of solutions to heat conduction problems. For example, it is easy to see that  $K \rightarrow 0$  as  $t \rightarrow 0$  except at the point  $\xi = x$ , where  $K \rightarrow \infty$  like the function  $1/\sqrt{t}$ . Furthermore,

$$\int_{-\infty}^{\infty} K(x - \xi, t) d\xi = 1$$

for all  $t \geq 0$

To show that (11.1.15) does indeed solve the problem (11.1.8) can be done by formal substitution. However, it is important to realise that such a procedure requires careful consideration of questions of convergence and the operation of differentiation under the integral. Furthermore, there must be some restriction on the initial data. That this is so is due to the fact that (11.1.15) has been arrived at through the use of the Fourier integral theorem wherein we have assumed that

$$\int_{-\infty}^{\infty} |u(x, 0)| dx = \int_{-\infty}^{\infty} |f(x)| dx < 0.$$

This restriction can be relaxed; in fact (11.1.15) holds if

$$|f(x)| \leq M e^{Nx^2}.$$

Rigorous proofs of the existence and uniqueness of the solution (11.1.15) to the initial value problem (11.1.8) are to be found in most books devoted to the study of partial differential equations of which John (1975) is to be recommended.

## 11.2 Separation of variables

The class of problems treated here are those in which we wish to solve the heat equation in the presence of boundary as well as initial conditions. Such

problems may be attacked by the use of “finite” Fourier transforms or, what is essentially the same thing, by the separation of variables method. Thus, instead of seeking solutions in integral form, as was done in the previous section, we derive solutions expressed in the form of infinite Fourier series.

The following is typical of such problems.

**Example 11.2.1**

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < l, \quad t > 0, \\ u(x, 0) &= f(x), \\ u(0, t) &= u(l, t) = 0. \end{aligned} \tag{11.2.1}$$

As before, we seek solutions in the form

$$u(x, t) = X(x)T(t)$$

in which the functions  $X(x)$  and  $T(t)$  must satisfy the ordinary differential equations

$$\begin{aligned} \frac{dT}{dt} + \lambda T &= 0, \\ \frac{d^2X}{dx^2} + \lambda X &= 0. \end{aligned} \tag{11.2.2}$$

Notice that these equations differ from those in (11.1.5) by the change from  $+\lambda$  to  $-\lambda$ . This is done purely for convenience and, for the present problem, makes the analysis a little easier to handle.

We seek solutions of the second equation in (11.2.2) that satisfy the boundary conditions

$$u(0, t) = u(l, t) = 0. \tag{11.2.3}$$

In other words, we have a typical eigenvalue problem (see Chapter 3, Section 3.6) giving rise to eigenvalues  $\lambda_n = n^2\pi^2/l^2$ ,  $n = 1, 2, \dots$ , and corresponding eigenfunctions  $X_n(x) = \sin(n\pi x/l)$ . Thus for each  $n$  the functions

$$u_n(x, t) = \exp\left(-\frac{n^2\pi^2 t}{l^2}\right) \sin\left(\frac{n\pi x}{l}\right)$$

are solutions of the heat equation that satisfy the boundary conditions (11.2.3). Since the heat equation is linear, a superposition of solutions of the form  $u_n(x, t)$  is also a solution. Thus we look for a solution of the initial boundary value problem in the form of the infinite series

$$u(x, t) = \sum_{n=1}^{\infty} A_n \exp\left(-\frac{n^2\pi^2 t}{l^2}\right) \sin\left(\frac{n\pi x}{l}\right). \tag{11.2.4}$$

If we set  $t = 0$  in (11.2.4), then the initial condition is satisfied if we can choose the unknown coefficients  $A_n, n = 1, 2, \dots$ , so that

$$f(x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{l}\right). \quad (11.2.5)$$

In other words we seek a Fourier sine series expansion of  $f(x)$ .

From Chapter 3 we know that the set of functions  $\sin(n\pi x/l), n = 1, 2, \dots$ , form an orthogonal set on the interval  $0 \leq x \leq l$ , that is,

$$\int_0^l \sin\left(\frac{m\pi x}{l}\right) \sin\left(\frac{n\pi x}{l}\right) dx = \begin{cases} 0, & m \neq n. \\ l/2, & m = n. \end{cases}$$

Thus if we formally multiply (11.2.5) by  $\sin(\frac{m\pi x}{l})$  for some  $m$  and integrate over  $0 \leq x \leq l$  we find that

$$\int_0^l f(x) \sin\left(\frac{m\pi x}{l}\right) dx = \frac{1}{2} A_m$$

or

$$A_m = \frac{2}{l} \int_0^l f(x) \sin\left(\frac{m\pi x}{l}\right) dx.$$

This gives the result

$$u(x, t) = \frac{2}{l} \sum_{n=1}^{\infty} \exp\left(-\frac{n^2\pi^2 t}{l^2}\right) \sin\left(\frac{n\pi x}{l}\right) \int_0^l f(y) \sin\left(\frac{n\pi y}{l}\right) dy. \quad (11.2.6)$$

By construction it is clear that the series (11.2.6) satisfies the boundary and initial conditions of our problem. That it also satisfies the heat equation follows from the fact that the series is convergent for all  $x, 0 \leq x \leq l$ , and  $t \geq 0$  and can be differentiated term by term.

The method of separation of variables is applicable not only to the heat equation but also to problems involving Laplace's equation and the wave equation treated under a variety of initial and boundary conditions. To help to appreciate this we consider the following problem. □

### Example 11.2.2

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - au, \quad 0 < x < \pi, \quad t > 0, \quad (11.2.7)$$

where  $a$  is a positive constant and  $u(x, t)$  is required to satisfy the conditions

$$\begin{aligned} u(0, t) &= \frac{\partial u(\pi, t)}{\partial x} = 0, \\ u(x, 0) &= x(\pi - x). \end{aligned} \quad (11.2.8)$$

Proceeding as before we set

$$u(x, t) = X(x)T(t)$$

to find

$$\frac{1}{T} \frac{dT}{dt} = \frac{1}{X} \frac{d^2X}{dx^2} - a$$

or

$$\frac{dT}{dt} + aT = -\lambda = \frac{d^2X}{dx^2}.$$

In other words, the functions  $X(x), T(t)$  must satisfy the ordinary differential equations

$$\frac{d^2X}{dx^2} + \lambda X = 0, \quad (11.2.9)$$

$$\frac{dT}{dt} + (a + \lambda)T = 0. \quad (11.2.10)$$

We seek solutions of (11.2.9) that satisfy the boundary conditions

$$X(0) = \frac{dX(\pi)}{dx} = 0. \quad (11.2.11)$$

First of all suppose  $\lambda = 0$ ; then (11.2.9) has the general solution  $X(x) = A + Bx$ . On using the boundary conditions (11.2.11), it follows that  $A = B = 0$  and so  $X \equiv 0$  in this case. When  $\lambda \neq 0$  the general solution of (11.2.9) is

$$X(x) = A \sin \sqrt{\lambda}x + B \cos \sqrt{\lambda}x$$

and

$$\frac{dX}{dx} = A\sqrt{\lambda} \cos \sqrt{\lambda}x - B\sqrt{\lambda} \sin \sqrt{\lambda}x.$$

Consequently, the boundary conditions (11.2.11) demand that  $A, B, \sqrt{\lambda}$  be chosen so that

$$\begin{aligned} B &= 0, \\ A\lambda \cos \sqrt{\lambda}\pi - B\lambda \sin \sqrt{\lambda}\pi &= 0. \end{aligned}$$

Thus for  $A \neq 0$  we must have  $\cos \sqrt{\lambda}\pi = 0$ , and so

$$\sqrt{\lambda}\pi = (2n - 1)\pi/2, \quad n = 1, 2, \dots$$

That is the eigenvalues are

$$\lambda_n = \frac{1}{4}(2n - 1)^2,$$

and the corresponding eigenfunctions are

$$X_n(x) = \sin\left(\frac{(2n - 1)x}{2}\right).$$

Therefore we have a solution  $u_n(x, t)$  of (11.2.7) satisfying the boundary conditions in (11.2.8) in the form

$$u_n(x, t) = \exp\{-(2n-1)^2 + 4a)t/4\} \sin[(2n-1)x/2].$$

The complete solution to the problem is then sought in the form

$$u(x, t) = \sum_{n=1}^{\infty} A_n \exp\{-(2n-1)^2 + 4a)t/4\} \sin[(2n-1)x/2]. \quad (11.2.12)$$

By imposing the initial condition in (11.2.8) we determine the unknown coefficients  $A_n$  from the Fourier series

$$x(\pi - x) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{(2n-1)x}{2}\right). \quad (11.2.13)$$

From the previous example we see that

$$A_n = \frac{2}{\pi} \int_0^\pi x(\pi - x) \sin\left(\frac{(2n-1)x}{2}\right) dx.$$

The integral appearing on the right-hand side is best evaluated by “integration by parts.” Consider

$$\begin{aligned} I_n &= \int_0^\pi x(\pi - x) \sin\left(\frac{(2n-1)x}{2}\right) dx \\ &= \left[ -x(\pi - x) \frac{2}{2n-1} \cos\left(\frac{(2n-1)}{2}x\right) \right]_0^\pi \\ &\quad + \frac{2}{2n-1} \int_0^\pi (\pi - 2x) \cos\left(\frac{(2n-1)}{2}x\right) dx \\ &= \left[ (\pi - 2x) \frac{4}{(2n-1)^2} \sin\left(\frac{(2n-1)}{2}x\right) \right]_0^\pi \\ &\quad + \frac{8}{(2n-1)^2} \int_0^\pi \sin\left(\frac{(2n-1)}{2}x\right) dx \\ &= \frac{4\pi}{(2n-1)^2} \cos(n\pi) - \frac{16}{(2n-1)^3} \left[ \cos\left(\frac{(2n-1)}{2}x\right) \right]_0^\pi \\ &= \frac{4\pi}{(2n-1)^2} \cos(n\pi) + \frac{16}{(2n-1)^3}. \end{aligned}$$

Thus

$$A_n = 8[(4/\pi)(2n-1)^{-3} + (-1)^n(2n-1)^{-2}],$$

and we have the solution to the problem in the form

$$\begin{aligned} u(x, t) &= 8 \sum_{n=1}^{\infty} [(4/\pi)(2n-1)^{-3} + (-1)^n(2n-1)^{-2}] \\ &\quad \times \exp\{-(2n-1)^2 + 4a)t/4\} \sin[(2n-1)x/2]. \end{aligned}$$

Consider the problem in two dimensions. □

**Example 11.2.3**

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}, \quad 0 < x < a, \quad 0 < y < b, \quad t > 0,$$

$u(x, y, t) = 0$  on the boundary of the rectangle defined by  $0 \leq x \leq a, 0 \leq y \leq b$  and

$$u(x, y, 0) = f(x, y). \quad (11.2.14)$$

The procedure here is precisely the same as in the previous examples except that we begin by seeking a solution in the form

$$u(x, y, t) = T(t)S(x, y)$$

to find that  $T$  and  $S$  must satisfy the equations

$$\frac{dT}{dt} + \lambda T = 0, \quad (11.2.15)$$

$$\frac{\partial^2 S}{\partial x^2} + \frac{\partial^2 S}{\partial y^2} + \lambda S = 0. \quad (11.2.16)$$

Now (11.2.16) is an elliptic equation of the type considered in Chapter 10 and here we seek a solution  $S(x, y)$  which satisfies the boundary conditions

$$\begin{aligned} S(0, y) &= S(a, y) = 0, \quad 0 \leq y \leq b, \\ S(x, 0) &= S(x, b) = 0, \quad 0 \leq x \leq a. \end{aligned} \quad (11.2.17)$$

This problem is solved by applying the separation of variables technique once more. That is, we set

$$S(x, y) = X(x)Y(y)$$

and find that  $X(x), Y(y)$  must satisfy

$$\frac{1}{X} \frac{d^2 X}{dx^2} + \frac{1}{Y} \frac{d^2 Y}{dy^2} + \lambda = 0,$$

i.e.,

$$\frac{1}{X} \frac{d^2 X}{dx^2} = -\mu = -\left(\frac{1}{Y} \frac{d^2 Y}{dy^2} + \lambda\right),$$

or

$$\frac{d^2 X}{dx^2} + \mu X = 0, \quad (11.2.18)$$

$$\frac{d^2 Y}{dy^2} + (\lambda - \mu)Y = 0. \quad (11.2.19)$$

In addition we require

$$X(0) = X(a) = 0, \quad (11.2.20)$$

$$Y(0) = Y(b) = 0. \quad (11.2.21)$$

The problem (11.2.18), (11.2.20) is of the same form as the problem (11.2.2), (11.2.3) and so we have the eigenvalues

$$\mu_m = \frac{m^2\pi^2}{a^2}, \quad m = 1, 2 \dots$$

and corresponding eigenfunctions

$$X_m(x) = \sin\left(\frac{m\pi x}{a}\right).$$

For a typical eigenvalue  $\mu_m$ , the problem (11.2.19), (11.2.21) becomes

$$\frac{d^2Y}{dy^2} + (\lambda - \mu_m)Y = 0,$$

$$Y(0) = Y(b) = 0.$$

This is another typical eigenvalue problem of the type already considered and leads to the eigenvalues

$$\lambda_n - \mu_m = \frac{n^2\pi^2}{b^2}, \quad n = 1, 2 \dots$$

and corresponding eigenfunctions

$$Y_n(y) = \sin\left(\frac{n\pi y}{b}\right).$$

In other words, a solution of (11.2.16), (11.2.17) is of the form

$$S_{mn}(x, y) = \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right),$$

together with the eigenvalue

$$\lambda_{mn} = \left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)\pi^2.$$

It is important to notice here that since  $S_{mn}$  depends on both  $m$  and  $n$ , the general solution of the problem (11.2.14) must be sought as a “double” series where the summation extends over both  $m$  and  $n$ . Thus we write

$$\begin{aligned} u(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} A_{mn} \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) \\ &\quad \times \exp\left[-\left(\frac{m^2}{a^2} + \frac{n^2}{b^2}\right)\pi^2 t\right]. \end{aligned} \quad (11.2.22)$$

By imposing the initial condition we determine  $A_{mn}$  from the “double” Fourier series

$$f(x, y) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} A_{mn} \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right). \quad (11.2.23)$$

In order to determine the coefficients  $A_{mn}$  we apply the “one-dimensional” arguments used above twice.

Thus if we consider  $y$  as fixed then

$$\frac{a}{2} \sum_{n=1}^{\infty} A_{mn} \sin\left(\frac{n\pi y}{b}\right) = \int_0^a f(x, y) \sin\left(\frac{m\pi x}{a}\right) dx. \quad (11.2.24)$$

Now the right-hand side of (11.2.24) is a function of  $y$  and so (11.2.24) can be considered as a Fourier series in the single variable  $y$ . Thus if we set

$$\int_0^a f(x, y) \sin\left(\frac{m\pi x}{a}\right) dx = g(y),$$

then the coefficients  $A_{mn}$  ( $m$  fixed) in (11.2.24) are given by

$$\frac{ab}{4} A_{mn} = \int_0^b g(y) \sin\left(\frac{n\pi y}{b}\right) dy$$

i.e.,

$$A_{mn} = \frac{4}{ab} \int_0^a \int_0^b f(x, y) \sin\left(\frac{m\pi x}{a}\right) \sin\left(\frac{n\pi y}{b}\right) dx dy. \quad (11.2.25)$$

In conclusion then we have the possible series solution to problem (11.2.14) in the form (11.2.22) where the coefficients  $A_{mn}$  are given by (11.2.25). Again we must verify that (11.2.22) does indeed solve the problem (11.2.14). By construction, the initial and boundary conditions are satisfied and the uniform convergence of the series (11.2.22) justifies differentiation of the series and this in turn allows for verification of the solution by direct substitution into the two-dimensional heat equation. □

### 11.3 Simple evolutionary equations

An evolutionary equation is one of the form

$$\frac{\partial u}{\partial t} = D \nabla^2 u + f(\mathbf{x}, t, u), \quad (11.3.1)$$

where the function  $f$  may be quite a complicated function of its arguments and could even depend on the spatial derivatives of  $u$ . Equations of the form

(11.3.1) have already arisen many times in the context of biological modelling (see Chapter 4) and will be taken up again in Chapter 12.

The expression “simple evolutionary equations” used for this section is somewhat illusory. It does not mean that the problems we shall consider are particularly simple or easy, but refers to equations in which the function  $f$  depends only on the dependent variable  $u$  and is of a simple form such as a polynomial. Thus the equations we shall consider will be of the type

$$\frac{\partial u}{\partial t} = D\nabla^2 u + f(u). \quad (11.3.2)$$

Depending on the context, this type of equation is sometimes also referred to as a reaction–diffusion equation. If  $f(u)$  is linear, say

$$f(u) = Au + B,$$

where  $A$  and  $B$  are constants, then in many instances the equation (11.3.2) can be solved by the separation of variables method. However if, as in many of the applications considered in this book,  $f(u)$  is nonlinear, say

$$f(u) = u(1 - u)(u - a), \quad 0 < u < 1, \quad (11.3.3)$$

then the problem is much more intractable. Indeed, it is not usually possible to obtain “general” analytic solutions and one must solve such problems numerically (see Mitchell and Griffiths, 1980). Despite this, however, many evolutionary equations have “special” or “particular” solutions, which are of fundamental importance to our understanding of biological phenomena modelled by evolutionary equations. In this section we shall explore some of these particular solutions and demonstrate their use in the section to follow.

Suppose that  $f(u)$  is a polynomial in  $u$  and that  $f(u)$  has real roots at  $u = \alpha, \beta, \gamma, \dots$ . Then it is obvious that in this case  $u = \alpha, u = \beta, u = \gamma$ , etc., are all constant solutions of (11.3.2). Such solutions are of importance in the treatment of the pure initial value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= D \frac{\partial^2 u}{\partial x^2} + f(u), \quad -\infty < x < \infty, \\ u(x, 0) &= \phi(x). \end{aligned} \quad (11.3.4)$$

Here, for example, it often happens that the solution  $u(x, t)$  evolves as  $t \rightarrow \infty$  into one of the asymptotic states  $u = \alpha, \beta, \gamma, \dots$ . Which state is reached depends crucially on the form of the initial data  $\phi$  and that of  $f$ . Indeed some of these constant asymptotic states are stable to small perturbations of  $\phi$  while others are not.

As an example of this, consider the problem (11.3.4) with  $f(u)$  given by (11.3.3). Then it can be shown (see Section 11.4) that if

$$\sup_{x \in (-\infty, \infty)} \phi(x) < a,$$

then  $u(x, t) < a$  for all  $x \in (-\infty, \infty)$ ,  $t > 0$ . From this it may be proved that  $u(x, t) \rightarrow 0$  (exponentially) as  $t \rightarrow \infty$ . By a similar argument it can also be shown that if

$$\inf_{x \in (-\infty, \infty)} \phi(x) > a,$$

then  $u(x, t) > a$  for all  $x \in (-\infty, \infty)$ ,  $t > 0$  and  $u(x, t) \rightarrow 1$  as  $t \rightarrow \infty$ . In other words, the problem (11.3.4) has  $u = 0$  or  $u = 1$  as asymptotic states depending on the magnitude of the initial data. The number  $a$  is called a “threshold” parameter.

Another fundamentally important set of solutions to equation (11.3.4) is the class of “travelling wave” solutions. Here we look for solutions of the form

$$u(x, t) = V(x + ct),$$

where  $c$  is a constant, positive or negative, called the “wave speed.” If we make the transformation  $\xi = x + ct$  in (11.3.4) then we see that  $V(\xi)$  satisfies the nonlinear ordinary differential equation

$$V'' + cV' + f(V) = 0, \quad (11.3.5)$$

where the primes represent differentiation with respect to  $\xi$ .

We have encountered travelling waves several times already (Chapters 4, 7 and 8) and have analysed them using the phase plane methods of Chapter 5. They are also of fundamental importance in the study of asymptotic states of the problem (11.3.4). Indeed we may ask the question: When does the solution  $u(x, t)$  “evolve” into the travelling wave  $V(x + ct)$ ? Mathematically we can state the problem as: determine conditions on the initial data  $\phi(x)$  so that

$$\lim_{t \rightarrow \infty} |u(x, t) - V(x + ct)| = 0.$$

In practice, there may be several such travelling waves, some stable and some unstable, and the determination of the appropriate stable travelling wave is often a challenging task.

Travelling waves can be classified as follows:

- (a) wave trains —  $V(\xi)$  periodic;
- (b) wave fronts —  $V(-\infty)$  and  $V(\infty)$  exist and are unequal;
- (c) pulses —  $V(-\infty)$  and  $V(\infty)$  exist and are equal and  $V(\xi)$  is not constant.

Apart from the use of phase plane methods the problem of finding travelling wave solutions is usually solved by numerical methods. However, in some cases, if we are fortunate, travelling waves can be determined analytically.

### **Example 11.3.1**

$$f(u) = u(1 - u)(u - a).$$

Here

$$V(x+ct) = V(\xi) = (1 + e^{-\xi/\sqrt{2}})^{-1}, \quad (11.3.6)$$

where

$$c = \sqrt{2} \left( \frac{1}{2} - a \right).$$

In this example  $V(-\infty) = 0$ ,  $V(\infty) = 1$  and so in this case  $V(\xi)$  is a wave front travelling from right to left with speed  $c = \sqrt{2}(\frac{1}{2} - a)$ .  $\blacksquare$

### Example 11.3.2

$$f(u) = \begin{cases} u, & 0 \leq u < \frac{1}{2}, \\ 1-u, & \frac{1}{2} \leq u. \end{cases} \quad (11.3.7)$$

If we make the substitution  $\xi = x+ct$  then we obtain equation (11.3.5) with  $f$  defined by (11.3.7). Again we look for a nondecreasing (i.e., monotone increasing) solution satisfying the conditions  $0 \leq V(\xi) \leq 1$ ,  $V(-\infty) = 0$ ,  $V(\infty) = 1$ .

As long as  $V < \frac{1}{2}$  we have to solve

$$V'' - cV' + V = 0,$$

which has the solution

$$V = Ae^{\alpha_1 \xi} + Be^{\alpha_2 \xi}, \quad (11.3.8)$$

where  $A$  and  $B$  are constants to be determined and

$$\alpha_1 = \frac{c + \sqrt{(c^2 - 4)}}{2}, \quad \alpha_2 = \frac{c - \sqrt{(c^2 - 4)}}{2},$$

provided  $c^2 \neq 4$ . In order to satisfy the condition  $V(-\infty) = 0$ , it is clear that we must have  $c > 0$ . If  $c < 2$  then (11.3.9) is oscillatory and so we would not have a monotone solution. Consequently we must have  $c \geq 2$ . Suppose  $c > 2$ . Then we have a solution monotone increasing with  $\xi$ . If without loss of generality  $V = 1/2$  when  $\xi = 0$  then we require

$$A + B = \frac{1}{2}. \quad (11.3.9)$$

For  $V \geq \frac{1}{2}$  we must solve the equation

$$V'' - cV' + 1 - V = 0,$$

which has the general solution

$$V = De^{\beta_1 \xi} + Ee^{\beta_2 \xi} + 1, \quad (11.3.10)$$

where  $D, E$  are constants to be determined and

$$\beta_1 = \frac{c + \sqrt{(c^2 + 4)}}{2}, \quad \beta_2 = \frac{c - \sqrt{(c^2 + 4)}}{2}.$$

The requirement  $V(\infty) = 1$  demands that  $D = 0$ . In addition, the condition  $V(0) = \frac{1}{2}$  gives

$$E + 1 = \frac{1}{2},$$

i.e.,

$$E = -\frac{1}{2}.$$

Furthermore, if we require  $V$  to be continuous at  $\xi = 0$  then a differentiation of (11.3.9) and (11.3.11) provides the extra condition

$$\alpha_1 A + \alpha_2 B = \beta_2 E = -\frac{1}{2}\beta_2. \quad (11.3.11)$$

Equations (11.3.9), (11.3.10) can now be solved to give

$$A = \frac{\sqrt{(c^2 + 4)} - 2c + \sqrt{(c^2 - 4)}}{4\sqrt{(c^2 - 4)}},$$

$$B = \frac{\sqrt{(c^2 - 4)} + 2c - \sqrt{(c^2 + 4)}}{4\sqrt{(c^2 - 4)}}.$$

Thus with  $A, B$  and  $E$  determined, (11.3.8), (11.3.10) give the complete solution to the problem for any  $c > 2$ . That is,

$$V(\xi) = \frac{\sqrt{(c^2 + 4)} - 2c + \sqrt{(c^2 - 4)}}{4\sqrt{(c^2 - 4)}} \exp\left[\left(\frac{c + \sqrt{(c^2 - 4)}}{2}\right)\xi\right]$$

$$+ \frac{\sqrt{(c^2 - 4)} + 2c - \sqrt{(c^2 + 4)}}{4\sqrt{(c^2 - 4)}} \exp\left[\left(\frac{c - \sqrt{(c^2 - 4)}}{2}\right)\xi\right], \quad -\infty < \xi < 0$$

$$= -\frac{1}{2} \exp\left[\left(\frac{c - \sqrt{(c^2 + 4)}}{2}\right)\xi\right] + 1, \quad 0 \leq \xi < \infty.$$

Now suppose  $c = 2$ ; then the general solution (11.3.8) must be replaced by one of the form

$$V = Ae^\xi + B\xi e^\xi \quad (11.3.12)$$

and (11.3.10) becomes

$$V = De^{(1+\sqrt{2})\xi} + Ee^{(1-\sqrt{2})\xi} + 1. \quad (11.3.13)$$

By imposing the boundary and continuity conditions we have

$$A = \frac{1}{2} = -E, \quad D = 0$$

and

$$A + B = -\frac{1}{2}(1 - \sqrt{2}),$$

i.e.,

$$B = -\frac{1}{2}(2 - \sqrt{2}).$$

Then, in this case

$$V = \begin{cases} \frac{1}{2}e^\xi - \frac{1}{2}(2 - \sqrt{2})\xi e^\xi, & -\infty < \xi \leq 0, \\ 1 - \frac{1}{2}e^{(1-\sqrt{2})\xi}, & 0 \leq \xi < \infty. \end{cases} \quad (11.3.14)$$

In conclusion then, (11.3.4) with  $f$  given by (11.3.7) has a one parameter family of wavefront solutions for any wave speed  $c \geq 2$ .

Another class of solutions to (11.3.4), which is often useful, are those independent of  $x$ . In this case we have to solve the ordinary differential equation

$$\frac{du}{dt} = f(u). \quad (11.3.15)$$

Suppose we have the initial condition

$$u(0) = u_0 = \text{constant};$$

then (11.3.15) can be solved implicitly as

$$t = \int_{u_0}^u \frac{d\eta}{f(\eta)}, \quad (11.3.16)$$

provided  $u_0$  and  $u$  are such that the integral exists. □

### Example 11.3.3

Suppose  $f(u) = u(1-u)$ . Then (11.3.16) has the solution ( $u_0 \neq 0$ )

$$\begin{aligned} t &= \int_{u_0}^u \frac{d\eta}{\eta(1-\eta)} = \int_{u_0}^u \left( \frac{1}{\eta} + \frac{1}{1-\eta} \right) d\eta, \\ &= \log \left| \frac{u(1-u_0)}{u_0(1-u)} \right|. \end{aligned}$$

We therefore obtain the explicit solution

$$u = \frac{u_0 e^t}{1 - u_0 + u_0 e^t}.$$

This solution shows that if  $0 < u_0 < 1$  then  $u(t) < 1$  and  $u \rightarrow 1$  as  $t \rightarrow \infty$ . Similarly if  $u_0 > 1$  then  $u(t) > 1$  and again  $u \rightarrow 1$  as  $t \rightarrow \infty$ . Note that if  $u_0 = 0$  then  $u(t) \equiv 0$  and if  $u_0 = 1$  then  $u(t) \equiv 1$ .

Finally consider those solutions of (11.3.4) that are independent of  $t$ . In this case we have to solve the ordinary differential equation

$$\frac{d^2u}{dx^2} + f(u) = 0. \quad (11.3.17)$$

In many situations it is possible to solve this equation in the following manner: if  $\frac{du}{dx} \neq 0$  then (11.3.17) can be expressed as

$$\frac{1}{2} \frac{d}{dx} \left( \frac{du}{dx} \right)^2 + f(u) \frac{du}{dx} = 0,$$

i.e.,

$$\left( \frac{du}{dx} \right)^2 + 2 \int^u f(\eta) d\eta = A,$$

for some arbitrary constant  $A$ . Thus if we set

$$\int^u f(\eta) d\eta = F(u)$$

then

$$\frac{du}{dx} = \pm \sqrt{A - 2F(u)}, \quad (11.3.18)$$

and this first-order differential equation can be solved, at least in principle, in the same way as equation (11.3.15).  $\square$

### Example 11.3.4

Consider the positive solutions to the boundary value problem

$$\frac{d^2u}{dx^2} + u(1-u) = 0, \quad (11.3.19)$$

$$u(-l/2) = u(l/2) = 0. \quad (11.3.20)$$

If  $u$  is a solution to this problem with  $u > 0$  for  $-l/2 < x < l/2$  and since  $u = 0$  at the end points it follows that  $u$  must take on its maximum value  $m$  at some intermediate point  $x = a$  where  $-l/2 < a < l/2$ . In other words,  $0 < u(x) \leq u(a) = m$  for  $-l/2 < x < l/2$  and  $u'(a) = 0, u''(a) \leq 0$ . From this observation, equation (11.3.19) shows that

$$u''(a) = -m(1-m) \leq 0$$

and so  $0 < m \leq 1$ . In fact  $0 < m < 1$  for if  $m = 1$  then  $u(x) = 1$  is the unique solution to (11.3.19) with  $u(a) = 1, u'(a) = 0$ . However, such a solution does not satisfy the boundary conditions (11.3.20). With  $0 < m < 1$  we see that  $u''(x) < 0$  for all  $x$  in the interval  $-l/2 < x < l/2$  and so  $u(x) < m$  except at  $x = a$ .

Let us now attempt to construct the solution to the boundary value problem via the method outlined above. Here we can write

$$F(u) = \int_0^u V(1-V)dV = \frac{u^2}{2} - \frac{u^3}{3}$$

and so (11.3.18) can be written as

$$\left(\frac{du}{dx}\right)^2 + 2F(u) = A.$$

When  $x = a$ ,  $\frac{du}{dx} = 0$  and so  $A = 2F(m)$ . Thus

$$\left(\frac{du}{dx}\right)^2 + 2F(u) = 2F(m). \quad (11.3.21)$$

Notice also that  $F$  is a strictly increasing function of  $u$  for  $0 < u < 1$ . Thus if  $x \neq a$  we have from (11.3.21)

$$\frac{du}{dx} = \begin{cases} \sqrt{2}\sqrt{[F(m) - F(u)]}, & -\frac{1}{2}l \leq x < a, \\ -\sqrt{2}\sqrt{[F(m) - F(u)]}, & a < x \leq \frac{1}{2}l. \end{cases}$$

In the first case

$$\int_u^m \frac{d\eta}{\sqrt{[F(m) - F(\eta)]}} = \sqrt{2}(a - x), \quad -\frac{1}{2}l \leq x < a, \quad (11.3.22)$$

and in the second

$$\int_u^m \frac{d\eta}{\sqrt{[F(m) - F(\eta)]}} = \sqrt{2}(x - a), \quad a < x \leq \frac{1}{2}l. \quad (11.3.23)$$

The implicit solution defined by (11.3.22), (11.3.23) is not yet in a desirable form since it involves the two unknown constants  $a$  and  $m$ . On using the boundary conditions we have

$$\begin{aligned} \int_0^m \frac{d\eta}{\sqrt{[F(m) - F(\eta)]}} &= \sqrt{2}(a + l/2), \\ \int_0^m \frac{d\eta}{\sqrt{[F(m) - F(\eta)]}} &= \sqrt{2}(l/2 - a), \end{aligned}$$

and these identities are only compatible if  $a = 0$ . That is,  $u(x)$  takes its maximum value at the midpoint of the interval  $-l/2 \leq x \leq l/2$ . With  $a = 0$  we see that  $l$  and  $m$  are related through

$$l = \sqrt{2} \int_0^m \frac{d\eta}{\sqrt{[F(m) - F(\eta)]}}. \quad (11.3.24)$$

In summary then, the solution to the boundary value problem is given implicitly by the formula

$$\int_0^m \frac{d\eta}{\sqrt{[F(m) - F(\eta)]}} = \sqrt{2}|x|, \quad (11.3.25)$$

where  $m$ , as a function of  $l$ , is given by (11.3.24).

Further important information can be obtained about the problem if we analyse the relationship between  $m$  and  $l$  defined by (11.3.24). In particular it can be shown that

- (a)  $l$  is an increasing function of  $m$  for  $0 \leq m < 1$ ;
- (b)  $l \rightarrow \infty$  as  $m \rightarrow 1$  from below;
- (c)  $l \rightarrow \pi$  as  $m \rightarrow 0$  from above.

These facts imply that there is no positive value of  $m$  satisfying (11.3.24) if  $l < \pi$  while for  $l \geq \pi$ ,  $m$  increases from 0 to 1 as  $l$  increases from  $\pi$  to  $\infty$ . Thus, for  $l \leq \pi$ , (11.3.19) and (11.3.20) have only the trivial solution  $u = 0$ , but for  $l > \pi$  the solution (11.3.25) appears. This type of behaviour is fundamental to the subject of **bifurcation theory** and  $l = \pi$  is called the **bifurcation point**.  $\square$

---

## 11.4 Comparison theorems

Consider the ordinary differential equation

$$\frac{d^2u}{dx^2} = F(x), \quad (11.4.1)$$

defined on the interval  $-\infty \leq a \leq x \leq b \leq \infty$ , where  $F(x) \geq 0$  and  $u$  is a continuous function of  $x$ . If  $F(x) > 0$  then we conclude that  $\frac{d^2u}{dx^2} > 0$  for all  $x$  in the interval  $a \leq x \leq b$ . This means that  $u(x)$  is concave and therefore takes its maximum value  $M$  at either  $x = a$  or  $x = b$  or both. Now suppose that  $F(x) \geq 0$  and that  $u \leq N$  at the end points  $x = a, b$ . Then for any  $\epsilon > 0$  the function  $V = u + \epsilon x^2$  satisfies the equation

$$\frac{d^2V}{dx^2} = F(x) + 2\epsilon > 0$$

and so, from the above observation,  $V$  must attain its maximum at least at one of the endpoints  $x = a, b$ . Thus

$$V(x) \leq \max[u + \epsilon x^2] \leq N + \epsilon \max(|a|^2, |b|^2).$$

Since  $u \leq V$  we have  $u \leq N + \epsilon \max(|a|^2, |b|^2)$  for all  $x$  in the interval  $a \leq x \leq b$  and for any  $\epsilon > 0$ . Letting  $\epsilon \rightarrow 0$ , we conclude that

$$u(x) \leq N, \quad a \leq x \leq b.$$

That is, if  $u$  is a solution to (11.4.1) with  $F \geq 0$ , the value of  $u$  in the interval  $a \leq x \leq b$  cannot exceed its maximum value at the end points  $x = a, b$ .

This relatively simple observation is an example of the **maximum principle**. We shall see that a result of this type also holds for partial differential equations, but, more than this, maximum principles have far-reaching consequences when applied to the study of qualitative properties of the solutions to

evolutionary equations. In this context, we prefer to use the term **comparison principle** or **theorem** for reasons that will become apparent. Before describing these results it is useful to note that the conclusions reached above do not violate the arguments used in Example 11.3.4 where the maximum  $u$  was attained at the midpoint of the interval  $-\frac{1}{2}l \leq x \leq \frac{1}{2}l$ . In that example it was shown that  $0 < u < 1$  and so  $F(x) = -u(x)[1 - u(x)] < 0$ .

Let us now turn to the elliptic partial differential equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = F(x, y) \quad (11.4.2)$$

defined on a bounded domain  $\Omega$  with boundary  $\partial\Omega$ . The arguments we employ as well as the results obtained extend without too much difficulty to higher dimensions, as well as to more general elliptic partial differential operators.

Suppose that  $F > 0$  in  $\Omega$  and that  $u(x, y)$  is continuous on  $\Omega \cup \partial\Omega$ . Then  $u$  attains its maximum value  $M$  somewhere on  $\Omega \cup \partial\Omega$ . If  $u = M$  at some point  $(x_0, y_0)$  of  $\Omega$  then we know that

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial y} = 0$$

and

$$\frac{\partial^2 u}{\partial x^2} \leq 0, \quad \frac{\partial^2 u}{\partial y^2} \leq 0$$

at  $(x_0, y_0)$ . That is,

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \leq 0$$

at  $(x_0, y_0)$ , which contradicts the assumption that  $F > 0$ . Thus the maximum of  $u$  must occur on the boundary  $\partial\Omega$ . If  $F \geq 0$  and  $u \leq N$  on  $\partial\Omega$  then we can apply our previous argument to the function

$$V = u + \epsilon(x^2 + y^2)$$

where  $\epsilon > 0$  is arbitrary. Clearly,  $V$  satisfies the partial differential equation

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = F + 4\epsilon > 0$$

in  $\Omega$ .

Our above conclusions apply and lead to the fact that

$$\begin{aligned} V &\leq \max_{\partial\Omega}[u + \epsilon(x^2 + y^2)] \\ &\leq N + \epsilon R^2 \end{aligned}$$

where  $R$  is the radius of a circle enclosing  $\Omega$ . Furthermore, since  $u \leq V$  we have

$$u \leq N + \epsilon R^2$$

and on letting  $\epsilon \rightarrow 0$  we obtain the result  $u(x, y) \leq N$  in  $\Omega$ . Thus we have the following **maximum principle**.

### THEOREM 11.4.1

*Let  $u$  be a continuous solution of (11.4.2) with  $F \geq 0$ . Then the value of  $u$  in  $\Omega$  cannot exceed its maximum on  $\partial\Omega$ .*

As an application of the maximum principle, consider the following uniqueness problem for the boundary value problem

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = F(x, y)$$

in  $\Omega$  and  $u = f$  on  $\partial\Omega$ .

Suppose the solution to this problem is not unique and that there exist distinct solutions  $u_1$  and  $u_2$  where  $w = u_1 - u_2 \neq 0$ . Since the partial differential equation is linear, it follows that  $w$  is a solution to the problem

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = 0$$

in  $\Omega$  and  $w = 0$  on  $\partial\Omega$ . It follows immediately from the maximum principle that  $w \equiv 0$  and so  $u_1 = u_2$ .

Let us now turn to the heat equation (11.1.2) and consider the function  $u(x, t)$  satisfying the inequality

$$L[u] \equiv \frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} \geq 0$$

defined on the rectangular  $R = \{0 < x < a, 0 < t \leq T\}$ . Suppose that  $M$  is the maximum of the values of  $u$  that occur on the sides  $S = \{x = 0, 0 \leq t \leq T\} \cup \{0 \leq x \leq a, t = 0\} \cup \{x = a, 0 \leq t \leq T\}$ . In addition, suppose that  $u$  takes its overall maximum value  $M_1 > M$  at an interior point  $(x_0, t_0)$  of the rectangle. We shall show that this is impossible.

Define the auxiliary function

$$w(x) = \frac{M_1 - M}{2a^2}(x - x_0)^2.$$

Now since  $u \leq M$  on the three sides  $S$  of the rectangle listed above we have

$$V(x, t) \equiv u(x, t) + w(x) \leq M + \frac{M_1 - M}{2} < M, \quad (11.4.3)$$

on  $S$ . Furthermore

$$V(x_0, t_0) = u(x_0, t_0) = M_1 \quad (11.4.4)$$

and

$$L[V] = L[u] + L[w] = L[u] + \frac{M_1 - M}{a^2} > 0 \quad (11.4.5)$$

throughout  $R$ . Conditions (11.4.3) and (11.4.4) show that  $V$  must assume its maximum either at an interior point of  $R$  or along the open interval

$$\{0 < x < a, t = T\}. \quad (11.4.6)$$

If the maximum occurs at an interior point then

$$\frac{\partial^2 V}{\partial x^2} \leq 0, \quad \frac{\partial V}{\partial t} = 0,$$

and so at such a point  $L[V] \leq 0$ , which contradicts (11.4.5). If the maximum occurs at some point of the interval (11.4.6), then  $\frac{\partial^2 V}{\partial x^2} \leq 0$  and (11.4.5) requires  $\frac{\partial V}{\partial t}$  to be strictly negative. In other words,  $V$  must be larger at an earlier time so that the maximum cannot occur on the interval (11.4.6). Thus the assumption that  $u(x_0, t_0) = M_1 > M$  cannot hold and so the maximum value of  $u$  must be attained somewhere on the sides  $S$ .

This conclusion is an example of the maximum principle for the heat equation. There is a corresponding minimum principle associated with solutions of  $L[u] = 0$ ; we simply replace  $u$  by  $-u$  in the above discussion and repeat the arguments.

A weakness of this type of maximum principle is that it permits the maximum of  $u$  to occur at interior points of a region as well as on the boundary. Thus it is desirable to establish a “strong” form of the maximum principle. In fact, we shall establish a strong maximum principle for a more general differential inequality, which will prove to be of basic importance to the comparison principle, which is our goal.

### **THEOREM 11.4.2 (Strong Maximum Principle)**

Let  $u(x, t)$  satisfy the differential inequality

$$(L + h)[u] \geq 0 \quad (11.4.7)$$

in the rectangular region  $R$  and where the given function  $h(x, t) \leq 0$  in  $R$ . If the maximum  $M$  of  $u$  is attained at an interior point  $(x_1, t_1)$  and if  $M \geq 0$ , then  $u \equiv M$  on all line segments  $t = \text{constant}$ , which lie directly below the horizontal segment of  $R$  containing  $(x_1, t_1)$ .

The proof of this result follows from the following lemmas:

**Lemma 11.4.3** Let  $u(x, t)$  satisfy the differential inequality (11.4.7) in the rectangular region  $R$ . Let  $D$  be a circular disc such that it and its boundary  $\partial D$  are contained in  $R$ . Suppose the maximum of  $u$  in  $R$  is  $M \geq 0$ , that  $u < M$  in the interior of  $D$  and that  $u = M$  at some point  $P(x_1, t_1)$  on the boundary  $\partial D$ . Then the tangent to  $D$  at  $P$  is parallel to the  $x$ -axis. That means  $P$  is either at the top or the bottom of the disc  $D$ .

**PROOF** Let the disc have centre  $(\bar{x}, \bar{t})$  and let  $A$  be the radius of  $D$  (see Figure 11.4.1). The idea is to assume that  $P$  is not at the top or the bottom of the disc and reach a contraction.  $\square$

We can safely assume that  $P$  is the only boundary point where  $u = M$ . For, if otherwise, we can choose a smaller disc interior to  $D$  and tangent to  $D$  at  $P$ . This smaller disc has exactly one point on its boundary where  $u = M$ .

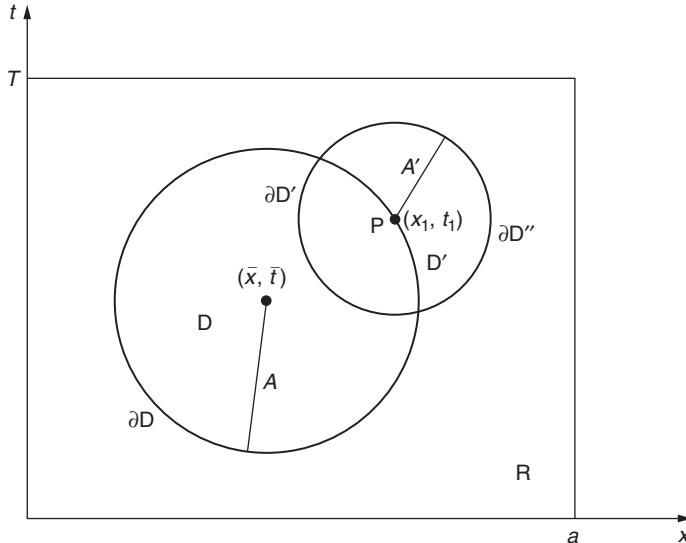
Construct a disc  $D'$  with centre  $(x_1, t_1)$  and of radius  $A'$  where

$$A' < |x_1 - \bar{x}| \quad (11.4.8)$$

and  $D'$  also lies inside  $R$ . Note that since  $P$  is not at the top or the bottom of  $D$ ,  $x_1 \neq \bar{x}$ . Now the boundary of  $D'$  consists of two arcs  $\partial D'$  (which includes its points of intersection with  $\partial D$ ) and  $\partial D''$  (as shown in Figure 11.4.1). Since  $u < M$  on  $\partial D'$  we can find a positive number  $\eta$  so that

$$u \leq M - \eta$$

on  $\partial D'$ .



**FIGURE 11.4.1:** The domain of Lemma 11.4.3.

Furthermore, since  $u \leq M$  throughout the rectangle  $R$ ,

$$u \leq M$$

on  $\partial D''$ . Define the function

$$V(x, t) = \exp\{-\alpha[(x - \bar{x})^2 + (t - \bar{t})^2]\} - e^{-\alpha A^2}$$

where  $\alpha$  is a positive number to be suitably chosen.  $V(x, t)$  has the following properties: (i)  $V(x, t) = 0$  on  $\partial D$ , (ii)  $V(x, t) > 0$  inside  $D$  and (iii)  $V(x, t) < 0$  outside  $D$ . If we substitute  $V(x, t)$  into the differential expression (11.4.7) then

$$(L + h)[V] = 2\alpha \exp\{-[(x - \bar{x})^2 + (t - \bar{t})^2]\} [2\alpha(x - \bar{x})^2 - 1 + (t - \bar{t})] \\ + h(x, t)V(x, t).$$

In the disc  $D'$  and on its boundary we have

$$|x - \bar{x}| > |x_1 - \bar{x}| - A' > 0$$

and so we can choose  $\alpha$  large enough so that

$$(L + h)[V] > 0$$

for all points  $(x, t)$  in  $D' \cup \partial D'$ . Next form the function

$$W(x, t) = u(x, t) + \epsilon V(x, t),$$

where  $\epsilon$  is a positive number to be suitably chosen. Then

$$(L + h)[W] = (L + h)[u] + \epsilon(L + h)[V] > 0 \quad (11.4.9)$$

inside  $D'$ . Since  $u \leq M - \eta$  on  $\partial D'$  we can select  $\epsilon$  so small so that

$$W = u + \epsilon V < M$$

on  $\partial D'$ , and since  $V < 0$  on  $\partial D''$  and  $u \leq M$  we have

$$W = u + \epsilon V < M$$

on  $\partial D''$ . Thus  $W < M$  on the entire boundary of  $D'$ .

In addition, we observe that since  $V = 0$  on  $\partial D$  we have

$$\begin{aligned} W(x_1, t_1) &= u(x_1, t_1) + \epsilon V(x_1, t_1) \\ &= u(x_1, t_1) = M. \end{aligned}$$

Hence the maximum of  $W$  on  $D'$  must occur at an interior point. This means that at such a point

$$\frac{\partial W}{\partial t} = 0, \quad \frac{\partial^2 W}{\partial x^2} \leq 0$$

and  $W = M \geq 0$ , and so since  $h \leq 0$  we have

$$(L + h)[W] \leq 0$$

at the maximum and this contradicts (11.4.9).

Thus the assumption that  $P$  is not at the top or the bottom of  $D$  cannot hold. That is, we have reached a contradiction and the lemma is proved. Note that for  $P$  at the top or the bottom of  $D$  our arguments break down because  $x_1 = \bar{x}$  and so we cannot construct the disc  $D'$  satisfying the condition (11.4.8).

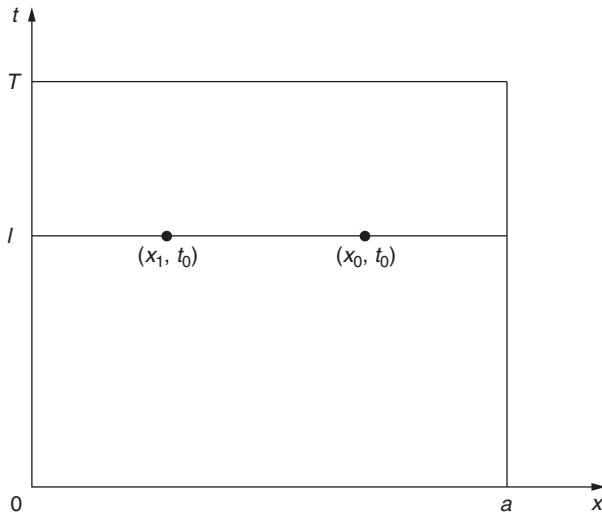


FIGURE 11.4.2: The domain of Lemma 11.4.4.

**Lemma 11.4.4** Suppose that  $u(x, t)$  satisfies (11.4.7) in the rectangular region  $R$ ,  $u < M$  at some interior point  $(x_0, t_0)$  of  $R$  and that  $u \leq M$  throughout  $R$ . If  $l$  is the horizontal line through  $(x_0, t_0)$  (see Figure 11.4.2) then  $u < M$  on  $l$ .

**PROOF** Suppose that  $u = M$  at some interior point  $(x_1, t_0)$  on  $l$  and that  $u < M$  at  $(x_0, t_0)$ . As in the proof of Lemma 11.4.4, we endeavour to reach a contradiction. For convenience we assume that  $x_1 < x_0$  and move  $x_1$  to the right if necessary so that  $u < M$  for  $x_1 < x \leq x_0$ . Let  $d_0$  be either the length  $x_0 - x_1$  or the minimum of the distances from any point of the segment  $x_1 \leq x \leq x_0, t = t_0$  to the boundary of the rectangle  $R$ , whichever is the smaller.

For  $x_1 < x < x_0 + d_0$ , we define  $d(x)$  to be the distance from  $(x, t_0)$  to the nearest point in  $R$  where  $u = M$ . Since  $u(x_1, t_0) = M$ ,  $d(x) \leq x - x_1$ . According to Lemma 11.4.3 this nearest point is directly above or below  $(x, t_0)$ . That is, either  $u(x, t_0 + d(x)) = M$  or  $u(x, t_0 - d(x)) = M$ .

Since the distance from a point  $(x + \delta, t_0)$  to  $(x, t_0 \pm d(x))$  is  $\sqrt{[d(x)^2 + \delta^2]}$  we see that

$$d(x + \delta) \leq \sqrt{[d(x)^2 + \delta^2]} < d(x) + \frac{\delta^2}{2d(x)}. \quad (11.4.10)$$

Similarly by replacing  $x$  by  $x + \delta$  and  $\delta$  by  $-\delta$  we have

$$d(x + \delta) > \sqrt{[d(x)^2 - \delta^2]}. \quad (11.4.11)$$

Suppose then that  $d(x) > 0$  and choose  $0 < \delta < d(x)$ . Now subdivide the interval  $(x, x + \delta)$  into  $n$  equal parts and apply the inequalities (11.4.11) to

get

$$\begin{aligned} d\left(x + \frac{j+1}{n}\delta\right) - d\left(x + \frac{j}{n}\delta\right) &\leq \frac{\delta^2}{2n^2 d(x + j\delta/n)} \\ &\leq \frac{\delta^2}{2n^2 \sqrt{[d(x)^2 - \delta^2]}}, \end{aligned}$$

$j = 0, 1, \dots, n-1$ .

Summing from  $j = 0$  to  $n-1$  gives

$$d(x + \delta) - d(x) \leq \frac{\delta^2}{2n^2 \sqrt{[d(x)^2 - \delta^2]}}$$

for any integer  $n$ . Now if we let  $n \rightarrow \infty$  then it follows that

$$d(x + \delta) \leq d(x), \quad \delta > 0.$$

In other words  $d(x)$  is a nonincreasing function of  $x$ . Since  $d(x) \leq x - x_1$ , which is arbitrarily small for  $x$  sufficiently close to  $x_1$ , we see that  $d(x) \equiv 0$  for  $x_1 < x < x_1 + d_0$ . In other words  $u(x, t_0) \equiv M$  on this interval, which contradicts our assumption that  $u < M$  for  $x_1 < x \leq x_0$ . This establishes the lemma.  $\blacksquare$

**Lemma 11.4.5** Suppose that  $u(x, t)$  satisfies (11.4.7) in the rectangular region  $R$  and that  $u < M$  where  $M \geq 0$  in the horizontal strip contained in  $R$  and defined by  $\{0 \leq x \leq l, t_0 < t < t_1\}$  for some numbers  $t_0$  and  $t_1$ . Then  $u < M$  on the line  $t = t_1$  contained in  $R$ .

**PROOF** Again the proof is by contradiction. Suppose  $P(x_1, t_1)$  is a point on the line  $t = t_1$  where  $u = M$ . Construct a disc  $D$  with centre  $P$  and radius so small that the lower half of  $D$  is entirely in the portion of  $R$  where  $t > t_0$ . (See Figure 11.4.3.)

Introduce the auxiliary function

$$V(x, t) = \exp\{-[(x - x_1)^2 + \alpha(t - t_1)]\} - 1.$$

Substituting this into the differential expression in (11.4.7), we have

$$(L + h)[V] = \exp\{-[(x - x_1)^2 + \alpha(t - t_1)]\}[4(x - x_1)^2 - 2 + \alpha] + h(x, t)V(x, t).$$

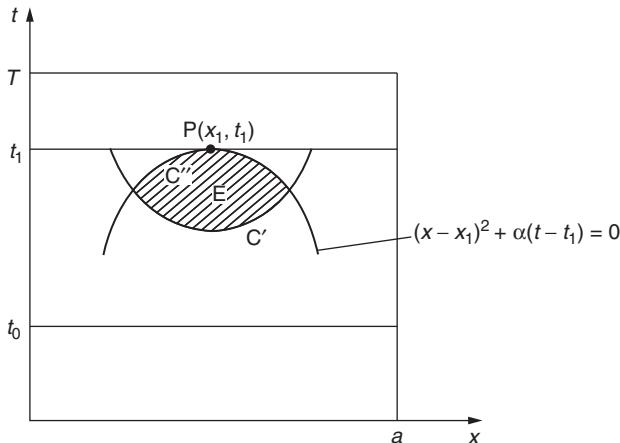
Now choose  $\alpha$  positive and sufficiently large so that

$$(L + h)[V] > 0$$

in  $D$  for  $t \leq t_1$ .

The parabola

$$(x - x_1)^2 + \alpha(t - t_1) = 0$$



**FIGURE 11.4.3:** The domain of Lemma 11.4.5.

is tangent to the line  $t = t_1$  at  $P$ . We denote by  $C'$  the portion of  $\partial D$  that is below the parabola (including the endpoints) and  $C''$  the portion of the parabola lying inside the disc  $D$ .

By hypothesis,  $u < M$  on  $C'$  and so we can find a number  $\eta > 0$  such that

$$u \leq M - \eta$$

on  $C'$ .

Now construct the function

$$W(x, t) = u(x, t) + \epsilon V(x, t)$$

where  $\epsilon$  is a positive number yet to be chosen. Observe that  $V = 0$  on  $C''$  and so if  $\epsilon$  is chosen sufficiently small we see that  $W$  enjoys the following properties:

- (a)  $(L+h)[W] = (L+h)[u] + \epsilon(L+h)[V] > 0$  in the shaded region  $E$  shown in Figure 11.4.3.
- (b)  $W = u + \epsilon V < M$  on  $C'$ .
- (c)  $W = u + \epsilon V \leq M$  on  $C''$ .

Condition (a) shows that  $W$  cannot attain its maximum in  $E$ , and so the maximum of  $W$  is  $M$  and it occurs at  $P$ . We conclude that

$$\frac{\partial W}{\partial t} \geq 0 \quad (11.4.12)$$

at  $P$ .

Also it is easy to verify that

$$\frac{\partial V}{\partial t} = -\alpha < 0 \quad (11.4.13)$$

at  $P$ .

Thus

$$\frac{\partial u}{\partial t} = \frac{\partial W}{\partial t} - \epsilon \frac{\partial V}{\partial t} > 0 \quad (11.4.14)$$

at  $P$ . However since the maximum of  $u$  on  $t = t_1$  occurs at  $P$  we have

$$\frac{\partial u}{\partial x} = 0, \quad \frac{\partial^2 u}{\partial x^2} \leq 0,$$

and  $u = M \geq 0$ , and these inequalities violate the differential inequality (11.4.7). Thus we have again arrived at a contradiction and the lemma is proved.  $\square$

Armed with Lemmas 11.4.3 and 11.4.5 we can now give the proof of Theorem 11.4.2.

**PROOF OF THEOREM 11.4.2** Suppose there is a value of  $t$ , say  $t_0 < t_1$ , such that  $u(x_1, t_1) < M$ . Let  $\tau$  be the least upper bound of values of  $t < t_1$  for which  $u(x_1, t) < M$ . By continuity  $u(x_1, \tau) = M$  while  $u(x_1, t) < M$  for some interval  $\tau_1 < t < \tau$ . Then from Lemma 11.4.3 it follows that  $u < M$  in the strip  $\{0 \leq x \leq l, \tau_1 < t < \tau\}$ , and Lemma 11.4.5 shows that  $u(x_1, t) < M$  on  $t = \tau$ , which is a contradiction and so  $u(x_1, t_0) \equiv M$ . This proves the theorem.  $\square$

Having established the strong maximum principle, we now use it to prove a comparison theorem, which is an invaluable tool with which to study qualitative properties of evolutionary equations of the type (11.3.2).

#### **THEOREM 11.4.6 (The Comparison Principle)**

Let  $u(x, t), v(x, t), A \leq u(x, t), v(x, t) \leq B$  be bounded and continuous functions satisfying the inequalities

$$\begin{aligned} u_t - u_{xx} - f(u) &\geq v_t - v_{xx} - f(v), \quad \text{in } (a, b) \times [0, T], \\ A \leq v(x, 0) &\leq u(x, 0) \leq B, \quad \text{in } (a, b) \end{aligned}$$

where  $-\infty \leq a < b \leq \infty$  and  $0 < T \leq \infty$  and  $f(u)$  is a continuously differentiable function of  $u$ .

Furthermore if  $a > -\infty$  assume that

$$A \leq v(a, t) \leq u(a, t) \leq B \quad \text{on } [0, T]$$

and if  $b < \infty$  assume that

$$A \leq v(a, t) \leq u(a, t) \leq B \quad \text{on } [0, T].$$

Then either  $u \equiv v$  in  $[a, b] \times [0, T]$  or  $u > v$  in  $(a, b) \times [0, T]$ .

**PROOF** From the mean value theorem we can write

$$(u - v)_t - (u - v)_{xx} \geq f(u) - f(v) = f'(v + \theta(u - v))(u - v)$$

for some  $\theta$  where  $0 < \theta < 1$ . Now take

$$\alpha = \max_{[A, B]} f'(u)$$

and define the function

$$W(x, t) = (u - v)e^{-\alpha t}.$$

It then follows that  $W(x, t)$  satisfies the inequality

$$(W)_t - (W)_{xx} \geq \{f'(v + \theta(u - v)) - \alpha\}W.$$

Since the coefficient of  $W$  is nonpositive, the conclusions of the theorem follow from an application of the strong maximum principle theorem.  $\square$

### Example 11.4.1

Let  $u(x, t)$  satisfy the following problem:

$$\begin{aligned} u_t &= u_{xx} + u(1 - u), \\ u &= 0 \quad \text{for } |x| = \frac{1}{2}l, \\ u(x, 0) &= u_0(x) > 0. \end{aligned}$$

For this problem, several applications of the comparison theorem with  $f(u) = u(1 - u)$  can be used to show that  $u(x, t) \geq 0$  and that if  $l > \pi$  then

$$\lim_{t \rightarrow \infty} u(x, t) = V(x) \tag{11.4.15}$$

where  $V(x)$  is the unique positive solution to Example 11.3.4. The proof of these results proceeds in several steps, some of which will be given here and the remaining given as problems at the end of this chapter.

To show that  $u(x, t) \geq 0$  take  $v(x, t) = 0$  in the comparison theorem. In order to establish the result (11.4.15) we proceed to construct appropriate **upper** and **lower** functions  $\bar{u}(x, t)$  and  $\underline{u}(x, t)$ , respectively, so that

$$\underline{u}(x, t) \leq u(x, t) \leq \bar{u}(x, t)$$

and such that

$$\lim_{t \rightarrow \infty} \underline{u}(x, t) = \lim_{t \rightarrow \infty} \bar{u}(x, t) = V(x). \quad (11.4.16)$$

In order to construct  $\bar{u}(x, t)$  we proceed as follows: let  $M = \max(1, M^*)$  where  $M^* = \max u_0(x)$  and  $\bar{u}(x, t)$  denotes the solution to

$$\begin{aligned} \bar{u}_t &= \bar{u}_{xx} + \bar{u}(1 - \bar{u}), \quad |x| < l/2, \quad t > 0, \\ \bar{u}(x, 0) &= M, \quad \text{if } |x| < l/2, \\ \bar{u} &= 0, \quad \text{for } |x| = l/2, \quad t > 0. \end{aligned}$$

On setting  $u = \bar{u}$  and  $v = u(x, t)$  in the comparison theorem it is easy to see that

$$u(x, t) \leq \bar{u} \quad \text{for } |x| \leq l/2, \quad t > 0.$$

It is easy to see that  $\bar{u} \geq 0$ . Next we show that  $\bar{u}(x, t) \leq M$ . This is done by taking  $u = M$  and  $v = \bar{u}$  in the comparison theorem. If, for arbitrary  $h > 0$ , we set  $u = \bar{u}$  and  $v = \bar{u}(x, t + h)$ , then the comparison theorem shows that  $\bar{u}(x, t) \geq \bar{u}(x, t + h)$ . In other words,  $\bar{u}(x, t)$  is a nonincreasing function of  $t$  for each  $x$ . Furthermore,  $\bar{u}(x, t)$  is bounded below and so

$$\lim_{t \rightarrow \infty} \bar{u}(x, t) = W(x)$$

exists and  $0 \leq W(x) \leq M$  for  $|x| \leq l/2$ .

It can now be shown that  $W(x)$  is a nonnegative solution of Example 11.3.4 (see Exercise 11.16).

Now set  $u = \bar{u}(x, t)$  and  $v = V(x)$ . Then the comparison theorem shows that  $W(x) \geq V(x)$ . However,  $V(x)$  is the unique positive solution to Example 11.3.4 and so it must be the case that  $W(x) = V(x)$ . These arguments establish one part of the limit relation (11.4.16). The other is proved in a similar manner (see Exercises 11.15 and 17).  $\square$

### Example 11.4.2

Let  $u(x, t)$  satisfy the problem

$$\begin{aligned} u_t &= u_{xx} + u(1 - u)(u - a), \quad -\infty < x < \infty, \quad t > 0, \\ u(x, 0) &= \phi(x). \end{aligned}$$

Suppose

$$\sup_{-\infty < x < \infty} \phi(x) < a$$

and let  $v(x, t) = a$ . Then the comparison principle applied to the functions  $u(x, t), v(x, t)$  shows that  $u(x, t) < a$  for all  $x$  and  $t$ .

Similarly, if

$$\inf_{-\infty < x < \infty} \phi(x) > a$$

then again the comparison principle says that  $u(x, t) > a$  for all  $x$  and  $t$ .  $\square$

## 11.5 Notes

For further reading on partial differential equations see F. John, *Partial Differential Equations*, Springer-Verlag, Berlin, 1975; and H.F. Weinberger, *Partial Differential Equations*, Blaisdell, New York, 1965.

A comprehensive treatment of maximum and comparison principles is to be found in M.H. Protter and H.F. Weinberger, *Maximum Principles in Differential Equations*, Springer-Verlag, Berlin, 1984 or A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

For an introduction to the numerical solution of partial differential equations see A.R. Mitchell and D.F. Griffiths, *The Finite Difference Method in Partial Differential Equations*, John Wiley & Sons, New York, 1980.

## Exercises

11.1 Show that the bounded solution of  $u_t = u_{xx}$  with initial data

$$u(x, 0) = \begin{cases} 1, & x > 0, \\ 0, & x < 0 \end{cases}$$

is given by

$$u(x, t) = \frac{1}{2}[1 + \phi(x/2\sqrt{t})]$$

where  $\phi(x)$  is the error function defined by

$$\phi(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Use MATLAB to produce plots or animations of  $u(x, t)$ . The error function  $\phi(x)$  can be computed in MATLAB with the command `erff(x)`.

11.2 Find the solution  $u(x, t)$  of the heat equation

$$u_t = u_{xx}, \quad x > 0, \quad t > 0$$

with boundary, initial data

$$\begin{aligned} u(x, 0) &= 0 && \text{for } x > 0, \\ u(0, t) &= 1 && \text{for } t > 0. \end{aligned}$$

11.3 Find the solution of the initial value problem

$$\begin{aligned} u_t &= u_{xx}, \quad 0 < x < 1, \quad t > 0, \\ u(x, 0) &= f(x), \quad 0 < x < 1, \\ u(0, t) &= u(1, t) = 0, \quad t > 0. \end{aligned}$$

## 11.4 Solve

$$\begin{aligned} u_t &= u_{xx}, \quad 0 < x < \pi, \quad t > 0, \\ u(x, 0) &= x(\pi - x), \quad 0 < x < \pi, \\ u(0, t) &= u(\pi, t) = 0, \quad t > 0. \end{aligned}$$

## 11.5 Solve

$$\begin{aligned} u_t &= u_{xx}, \quad 0 < x < \pi, \quad t > 0, \\ u(x, 0) &= \sin x, \quad 0 < x < \pi, \\ \frac{\partial u}{\partial x}(0, t) &= \frac{\partial u}{\partial x}(\pi, t) = 0, \quad t > 0. \end{aligned}$$

## 11.6 Solve

$$\begin{aligned} u_t &= u_{xx} + u_{yy}, \quad x, y \in (0, 1) \times (0, 1), \quad t > 0, \\ u(x, 0, t) &= u(x, 1, t) = 0, \\ u(0, y, t) &= u(1, y, t) = 0, \\ u(x, y, 0) &= xy. \end{aligned}$$

11.7 For the  $n$ -dimensional evolutionary equation

$$u_t = \nabla^2 u + f(u),$$

show that the wavefront  $u(\mathbf{x} \cdot \nu - ct)$ , where

$$\mathbf{x} \cdot \nu = \sum_{i=1}^n x_i \nu_i$$

and  $\nu$  is an arbitrary unit vector, is a solution provided  $u(\xi)$  satisfies the ordinary differential equation

$$u'' + cu' + f(u) = 0.$$

## 11.8 By following the method illustrated in Example 11.3.2, find a wavefront solution to the equation

$$u'' - cu' + f(u) = 0,$$

satisfying the conditions  $c > 0$ ,  $0 \leq u \leq 1$ ,  $u(-\infty) = 0$ ,  $u(\infty) = 1$  where

$$f(u) = \begin{cases} 0, & 0 \leq u < 1/2, \\ u - 1/2, & 1/2 \leq u \leq 3/4, \\ 1 - u, & u > 3/4. \end{cases}$$

11.9 Establish the results (a), (b), (c) following (11.3.25).

11.10 Show that if

$$\frac{d^2u}{dx^2} + A(x) \frac{du}{dx} = 0, \quad 0 < x < 1,$$

where  $A$  is continuous, then  $u$  attains its maximum in the interval  $0 \leq x \leq 1$  either at  $x = 0$  or at  $x = 1$ .

11.11 Show that if  $u$  is a solution of

$$\nabla^2 u + D(x, y) \frac{\partial u}{\partial x} + E(x, y) \frac{\partial u}{\partial y} = -F(x, y)$$

in some bounded two-dimensional region  $\Omega$  with  $F < 0$ , then  $u$  cannot attain its maximum at any point of  $\Omega$ .

11.12 Show that if  $u$  is continuous in a bounded two-dimensional region  $\Omega$  with boundary  $\partial\Omega$  and

$$\frac{\partial^2 u}{\partial x^2} + e^x \frac{\partial^2 u}{\partial y^2} - e^{x+y} u = 0 \quad \text{in } \Omega,$$

$$u \leq 0 \quad \text{on } \partial\Omega,$$

then  $u \leq 0$  in  $\Omega$ .

11.13 Show that if

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \quad 0 < x < 1, \\ \frac{\partial u}{\partial x}(0, t) &= 0, \end{aligned}$$

the maximum of  $u$  for  $0 \leq x \leq 1$ ,  $0 \leq t \leq T$  must occur at  $t = 0$  or at  $x = 1$ .

11.14 The function  $u(x, t) = -(x^2 + 2xt)$  is a solution of the equation

$$x \frac{\partial^2 u}{\partial x^2} - \frac{\partial u}{\partial t} = 0.$$

Is the maximum principle value in the region  $-1 < x < 1$ ,  $0 \leq t \leq 1/2$ ?

11.15 In Example 11.4.1, let  $v_\lambda(x)$  denote the positive solution of the problem

$$\begin{aligned} v'' + v(1 - v) &= 0, \quad |x| < \lambda/2, \quad \pi < \lambda < l, \\ v &= 0, \quad |x| = \lambda/2, \end{aligned}$$

and let  $\underline{u}(x, t)$  denote the solution of the problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + \underline{u}(1 - u), \quad |x| < l/2$$

$$\underline{u} = \begin{cases} v_\lambda(x), & |x| \leq \lambda/2, \\ 0, & |x| > \lambda/2, \end{cases}$$

$$\underline{u}(x, t) = 0 \quad |x| = l/2, \quad t \geq 0.$$

Use the comparison theorem to show that

$$\underline{u} \leq u(x, t),$$

for  $|x| \leq l/2$ ,  $t \geq 0$ .

- 11.16 Use the comparison theorem to show that the “upper” solution  $\bar{u}(x, t)$  constructed in Example 11.4.1 satisfies

$$\bar{u}(x, t) \geq v(x),$$

where  $v(x)$  is the unique positive solution to Example 11.3.4. Hence prove that  $v(x) = w(x)$ , where  $w(x)$  is the limiting function in Example 11.4.1.

- 11.17 By using the comparison theorem, show that  $\underline{u}(x, t)$  and  $v_\lambda(x)$  defined in Exercise 11.15 satisfy the inequality

$$\underline{u}(x, t) \geq v_\lambda(x) = \underline{u}(x, 0), \quad |x| \leq \lambda/2.$$

Hence show that  $\underline{u}(x, t)$  increases to  $v(x)$  as  $t \rightarrow \infty$ .

- 11.18 Use MATLAB to produce plots or animations of the travelling wave solution (11.3.6) with  $a = \frac{1}{4}$  and with  $a = \frac{3}{4}$ . What do you think will happen as  $a \rightarrow \frac{1}{2}$ ?

- 11.19 Use MATLAB to produce plots or animations of the travelling wave solution (11.3.14) with a wavespeed of  $c = 1$ .

# Chapter 12

---

## Problems of Diffusion

---

### 12.1 Diffusion through membranes

The equations we shall study in this chapter are all particular cases of the general system

$$\begin{bmatrix} \frac{\partial u_1}{\partial t} \\ \frac{\partial u_2}{\partial t} \\ \vdots \\ \frac{\partial u_m}{\partial t} \end{bmatrix} = \begin{bmatrix} D_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & D_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \cdots & D_{m-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & D_m \end{bmatrix} \begin{bmatrix} \nabla^2 u_1 \\ \nabla^2 u_2 \\ \vdots \\ \nabla^2 u_m \end{bmatrix} + \begin{bmatrix} f_1(u_1, u_2, \dots, u_m) \\ f_2(u_1, u_2, \dots, u_m) \\ \vdots \\ f_m(u_1, u_2, \dots, u_m) \end{bmatrix}.$$

This system can be conveniently expressed in the matrix form

$$\frac{\partial \mathbf{u}}{\partial t} = \mathbf{D} \nabla^2 \mathbf{u} + \mathbf{f}(\mathbf{u}), \quad (12.1.1)$$

where  $\mathbf{u} = \mathbf{u}(t, x_1, \dots, x_n)$  is a vector-valued function of time  $t$  and the  $n$  variables  $x_i, i = 1, \dots, n$ , and has  $m$  components  $u_i(t, x_1, \dots, x_n), i = 1, \dots, m$ .  $\mathbf{D}$  is an  $m \times m$  diagonal matrix and  $\mathbf{f}$  is an  $m \times 1$  column vector.  $\nabla^2$  is the  $n$ -dimensional Laplacian

$$\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \dots + \frac{\partial^2}{\partial x_n^2}.$$

Equations such as (12.1.1) are called **nonlinear diffusion equations** and the matrix  $\mathbf{D}$  is called the **diffusion matrix**. Our treatment will be confined to  $n = 1$  (one space dimension) and  $m = 1$  or 2.

Most of the biological models we have described so far are formulated as nonlinear diffusion equations and the reader may care to verify this by referring back to Chapter 4. Although we have used the term “diffusion” before, we have done so without motivation. Thus, before describing some of the salient features of diffusion equations, we shall consider the precise meaning of diffusion and the diffusion process.

Consider a fluid called the **solvent** in which some matter has been dissolved. This matter is called a **solute** and the combined fluid is called a **solution**. The solution may be conveniently characterised by its mass concentration, which we denote by  $C$ . Thus  $C$  is the **mass of dissolved matter per unit volume** of liquid.

One of the basic methods by which the solute molecules disperse and are transported through the solvent is by **diffusion**. This mechanism is usually a consequence of thermal motion of the individual solute molecules.

Let us now look at a solution in which simple molecular diffusion is taking place, the fluid being otherwise at rest. The chief mechanism of transport of the solute is governed by concentration differences. Thus we ask, what is the flux of solute molecules going through a unit area in a unit of time? The material flux per unit area  $j$  is according to **Fick's law** related to the rate of change  $\frac{\partial C}{\partial x}$  of  $C$  by

$$j = -D \frac{\partial C}{\partial x}, \quad (12.1.2)$$

where we assume that  $C$  depends only on  $t$  and the space variable  $x$ . The relation (12.1.2) is referred to as **Fick's first law of diffusion** and  $D$  is called the **diffusion coefficient** and is a characteristic of the solute in the fluid. The minus sign in (12.1.2) is important in that it reflects the fact that the molecular flow is from a high concentration to a low concentration region.

Suppose now we consider the solution to occupy some region of three-dimensional space and so the concentration  $C = C(t, x, y, z)$  will be a function of the time  $t$  and the spatial coordinates  $x, y$  and  $z$ . Clearly the flux will be a vector  $\mathbf{j}$  with components  $j_x, j_y, j_z$  and, according to Fick's first law of diffusion,

$$j_x = -D \frac{\partial C}{\partial x}, \quad j_y = -D \frac{\partial C}{\partial y}, \quad j_z = -D \frac{\partial C}{\partial z}. \quad (12.1.3)$$

Here of course we are assuming the solvent to be *homogeneous and isotropic* so that  $D$  is independent of position and is the same in all directions. In other words  $D$  is assumed constant.

Suppose we have a fixed but arbitrary volume of fluid  $V$  with surface  $S$  (see Figure 12.1.1) and we consider the flow of solute through this volume. By the law of conservation of mass we know that the rate at which solute accumulates (or disappears) within  $V$  must be balanced by the net influx (or efflux) across the bounding surface  $S$ . The total amount of solute at time  $t$  in  $V$  is  $\int_V C(t, x, y, z) dV$ , where  $dV$  is a typical volume element, and the total flux across the boundary  $S$  is  $\int_S \mathbf{j} \cdot \mathbf{n} dS$ . Consequently, by the law of conservation of mass,

$$\frac{\partial}{\partial t} \int_V C(t, x, y, z) dV = - \int_S \mathbf{j} \cdot \mathbf{n} dS,$$

i.e.,

$$\int_V \frac{\partial}{\partial t} C(t, x, y, z) dV + \int_S \mathbf{j} \cdot \mathbf{n} dS = 0.$$

By making use of Gauss' divergence theorem we can write this last result as

$$\int_V \left( \frac{\partial}{\partial t} C(t, x, y, z) + \nabla \cdot \mathbf{j} \right) dV = 0,$$

where  $\nabla \cdot \mathbf{j}$  is the divergence of the vector  $\mathbf{j}$ . Written in terms of Cartesian coordinates it is defined as

$$\nabla \cdot \mathbf{j} = \frac{\partial j_x}{\partial x} + \frac{\partial j_y}{\partial y} + \frac{\partial j_z}{\partial z}.$$

But  $V$  is arbitrary and so this equation must hold for all volumes, and this leads to

$$\frac{\partial}{\partial t} C + \nabla \cdot \mathbf{j} = 0. \quad (12.1.4)$$

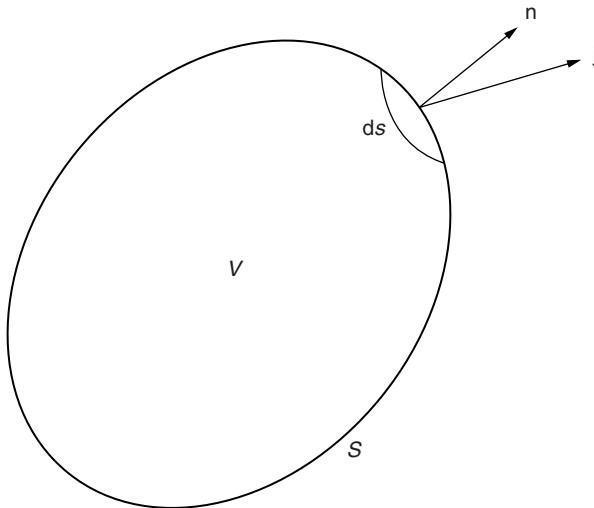
Finally, substituting for  $\mathbf{j}$  from (12.1.3) we have

$$\frac{\partial}{\partial t} C - D \nabla^2 C = 0. \quad (12.1.5)$$

This is the classic diffusion equation. In our context it is an expression of **Fick's second law** and in physics it is the basic equation of heat conduction (cf. Chapter 11). In this case  $C$  is interpreted as temperature and  $D$  heat conductivity.

Let us now consider the application of Fick's laws to the problem of diffusion of a solute into a cell. The diffusion here depends critically on the transport of solution through the membrane surrounding the protoplasm of the cell, namely the **plasma membrane**. This membrane consists of a double layer of lipid molecules called the **bilipid layer**. This layer contains a structure of globular proteins on both its surfaces, some of which penetrate across the entire width of the bilipid layer. However, the process by which these proteins control the diffusion of the solute is not fully understood.

The simplest theory of transport into or out of a cell by diffusion is derived on the assumption that the intracellular diffusion is so rapid that the solute concentration there is uniform in space. In most experiments of diffusion into cells, the volume of extracellular space is made so large that the solute concentration there is not affected to any significant extent by any loss in cells and may be assumed constant. Consider then an idealised cell membrane model to consist of a homogeneous lipid layer separating two aqueous phases, the cell interior and the cell exterior (see Figure 12.1.2a). At each of the two lipid interfaces, a discontinuity in solute concentration exists at equilibrium. This discontinuity is a consequence of the molecular barrier that exists for a solute molecule entering the lipid from the aqueous phase. Suppose  $C_o$  represents



**FIGURE 12.1.1:** The conservation of mass.

the concentration on the outside of the cell and  $\bar{C}_o$  the concentration inside the lipid layer and adjacent to the outer interface. The discontinuity between  $C_o$  and  $\bar{C}_o$  is expressed in terms of the **partition coefficient**  $\Gamma$ , i.e.,

$$\bar{C}_o = \Gamma C_o,$$

where  $\Gamma < 1$  unless the solute is more readily soluble in lipid than in water, in which case  $\Gamma > 1$ .

In a similar manner we have

$$\bar{C}_i = \Gamma C_i$$

where  $C_i$  represents the concentration in the cell interior and  $\bar{C}_i$  the concentration inside the lipid layer and adjacent to the inner surface.

Assume the thickness  $\delta$  of the cell membrane to be small so that, to first order, the concentration gradient in the lipid layer can be expressed as

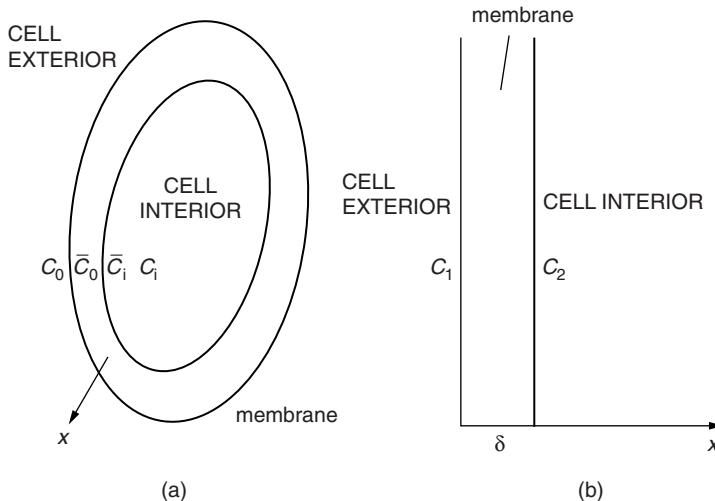
$$\frac{dC}{dx} = \frac{(\bar{C}_o - \bar{C}_i)}{\delta}.$$

On using Fick's first law of diffusion we obtain

$$j = \frac{D}{\delta} (\bar{C}_i - \bar{C}_o), \quad (12.1.6)$$

where  $D$  is the diffusion constant of the solute in the lipid layer. In experiments, measurements are made of  $C_i$  and  $C_o$  and  $\bar{C}_i, \bar{C}_o$  are in general inaccessible to observation and so we prefer to write

$$j = \frac{\Gamma D}{\delta} (C_i - C_o). \quad (12.1.7)$$



**FIGURE 12.1.2:** Idealised cell membrane model.

This is a good approximation to the flux  $j$  provided that  $C_i, C_o$  vary slowly.

Let  $m$  be the mass of solute contained in the cell; then at time  $t$  the law of conservation of mass gives

$$\frac{dm}{dt} = -jA, \quad (12.1.8)$$

the minus sign indicating that when  $j$  is positive, solute flowing outwards,  $m$  decreases. In (12.1.8)  $A$  denotes the surface area of the cell. Since  $C_i$  is assumed uniform and varies only with time, the mass of solute contained in the cell at time  $t$  is

$$m = C_i V \quad (12.1.9)$$

where  $V$  is the volume of the cell. Using this result in (12.1.8) we find

$$V \frac{dC_i}{dt} = -jA,$$

i.e.,

$$\frac{dC_i}{dt} = k(C_o - C_i) \quad (12.1.10)$$

where

$$k = \frac{\Gamma D A}{\delta V}.$$

Suppose at time  $t = 0$ ,  $C_i = C_{io}$ ; then (12.1.10) is readily solved to give

$$C_i(t) = C_o + (C_{io} - C_o)e^{-kt}.$$

Thus as  $t \rightarrow \infty$ ,  $C_i(t) \rightarrow C_o$  and the cell contains solute of concentration  $C_o$  entirely.

This simple result is derived on the assumption that the thickness  $\delta$  of the lipid layer is small. If this is not the case then a more precise study must be made. To this end we consider diffusion through a lipid layer in the form of a slab as shown in Figure 12.1.2(b). For simplicity we shall assume that there is no solute in the membrane initially. Thus we have to solve the diffusion equation (12.1.5) in one spatial dimension  $x$  subject to the boundary conditions

$$\begin{aligned} C(0, t) &= C_1, \\ C(\delta, t) &= C_2, \end{aligned} \quad (12.1.11)$$

and the initial condition

$$C(x, 0) = 0, \quad 0 < x < \delta. \quad (12.1.12)$$

The concentrations  $C_1, C_2$  at the two outer walls of the membrane are assumed to be constant.

The problem defined by (12.1.5), (12.1.11) and (12.1.12) is typical of the heat conduction problems that we have considered in Chapter 11. Thus, if we apply the method of separation of variables, the most general solution of (12.1.5) satisfying the boundary conditions (12.1.11) can be written in the form

$$C(x, t) = C_1 + (C_2 - C_1) \frac{x}{\delta} + \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{\delta}\right) \exp\left(-\frac{n^2\pi^2 D t}{\delta}\right), \quad (12.1.13)$$

where the constants  $A_n, n = 1, 2, \dots$ , are to be determined from the initial condition (12.1.12).

On setting  $t = 0$  in (12.1.12) a simple Fourier analysis (see Chapter 11) shows that

$$A_n = \frac{2}{n\pi} (C_2 \cos n\pi - C_1), \quad n = 1, 2, \dots,$$

and so the complete solution to the problem is given by

$$\begin{aligned} C(x, t) &= C_1 + (C_2 - C_1) \frac{x}{\delta} \\ &+ \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{(C_2 \cos n\pi - C_1)}{n} \sin\left(\frac{n\pi x}{\delta}\right) \exp\left(-\frac{n^2\pi^2 D t}{\delta}\right). \end{aligned} \quad (12.1.14)$$

From this expression we can calculate the rate  $j$  at which the diffusing substance emerges at the interface  $x = 0$  per unit area per unit time. That is,

$$j = -D \left( \frac{\partial C}{\partial x} \right)_{x=0},$$

i.e.,

$$j = -(C_2 - C_1) \frac{D}{\delta} - \frac{2}{\delta} \sum_{n=1}^{\infty} (C_2 \cos n\pi - C_1) \exp\left(-\frac{n^2\pi^2 D t}{\delta}\right).$$

The first term on the right represents the steady-state flux, while the remaining terms represent a transient flux that is significant only for short times.

---

## 12.2 Energy and energy estimates

In most processes that are governed by diffusion equations of the type introduced in equation (12.1.1) one is interested in how the concentration  $\mathbf{u}$  behaves for large times. It is well known that for the linear case wherein  $\mathbf{f}(\mathbf{u}) \equiv \mathbf{0}$  any initial concentration  $\mathbf{u}_0$  is smoothed out and, in fact,  $\mathbf{u}(x, t)$  often tends exponentially to zero as  $t$  tends to infinity. In our case  $\mathbf{f}(\mathbf{u}) \neq \mathbf{0}$  and the behaviour of  $\mathbf{u}(x, t)$  depends crucially on the form of  $\mathbf{f}(\mathbf{u})$ . Furthermore, in many cases, rather specially designed methods have to be developed to examine asymptotic behaviour.

A method that is widely applicable in many cases is the so-called “energy method.” Here we shall illustrate its use in a few fairly simple examples, and examine the method in more detail in the next few sections when we take up the subjects of nerve impulse transmissions and chemical reactions once more.

Consider the reaction-diffusion problem defined by

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2} + f(u), \quad (12.2.1)$$

$$u(x, 0) = u_0(x),$$

$$\frac{\partial}{\partial x} u(0, t) = \frac{\partial}{\partial x} u(1, t) = 0, \quad (12.2.2)$$

and  $f(u)$  is assumed to be such that this problem has solutions that are bounded for all time  $t \geq 0$ .

Define an “energy function”  $E(t)$  by

$$E(t) = \frac{1}{2} \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 dx. \quad (12.2.3)$$

Differentiating this with respect to  $t$  gives

$$\frac{dE}{dt} = \int_0^1 \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial x \partial t} dx, \quad (12.2.4)$$

which, on using the differential equation, can be written in the form

$$\begin{aligned} \frac{dE}{dt} &= \int_0^1 \frac{\partial u}{\partial x} \frac{\partial}{\partial x} \left( D \frac{\partial^2 u}{\partial x^2} + f(u) \right) dx, \\ &= \int_0^1 D \frac{\partial u}{\partial x} \frac{\partial^3 u}{\partial x^3} dx + \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 \frac{\partial f}{\partial u} dx. \end{aligned} \quad (12.2.5)$$

Integrating the first integral appearing in (12.2.5) by parts we find

$$\begin{aligned}\frac{dE}{dt} &= \left[ D \frac{\partial u}{\partial x} \frac{\partial^2 u}{\partial x^2} \right]_0^1 - D \int_0^1 \left( \frac{\partial^2 u}{\partial x^2} \right)^2 dx + \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 \frac{\partial f}{\partial u} dx \\ &= -D \int_0^1 \left( \frac{\partial^2 u}{\partial x^2} \right)^2 dx + \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 \frac{\partial f}{\partial u} dx.\end{aligned}\quad (12.2.6)$$

Let  $m = \max |\partial f / \partial u|$  over all solution values (assumed bounded) of  $u$ ; then (12.2.6) can be replaced by the inequality

$$\frac{dE}{dt} \leq -D \int_0^1 \left( \frac{\partial^2 u}{\partial x^2} \right)^2 dx + m \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 dx,$$

i.e.,

$$\frac{dE}{dt} \leq -D \int_0^1 \left( \frac{\partial^2 u}{\partial x^2} \right)^2 dx + 2mE(t). \quad (12.2.7)$$

In order to proceed further, we would like to express the first integral appearing in (12.2.7) in a more recognisable form. For example, suppose we could prove that for some constant  $\lambda$ ,

$$\int_0^1 \left( \frac{\partial^2 u}{\partial x^2} \right)^2 dx \geq \lambda \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 dx. \quad (12.2.8)$$

Then we could replace (12.2.7) by the inequality

$$\frac{dE}{dt} \leq 2(m - \lambda D)E(t), \quad (12.2.9)$$

from which we conclude that, since  $E(t) \geq 0$ ,

$$\frac{d}{dt}(e^{-2(m-\lambda D)t} E(t)) \leq 0.$$

An integration then reveals that

$$E(t) \leq E(0)e^{2(m-\lambda D)t},$$

where  $E(0) \geq 0$ . From this we conclude that if the diffusion coefficient  $D$  is sufficiently large, i.e.,  $D > m/\lambda$ , then  $E(t)$  decays exponentially to zero as  $t \rightarrow \infty$  and hence  $\partial u / \partial x$  tends to zero for all  $x \in [0, 1]$  as  $t \rightarrow \infty$ . This implies that  $u$  becomes spatially homogeneous regardless of the initial concentration  $u_0(x)$ .

To arrive at the above result we have relied on the assumption that an inequality of the form (12.2.8) holds. Such an inequality is called the **Poincaré inequality**; it occurs widely in problems where estimation depends on establishing certain inequalities and is proved as follows.

Consider the ordinary differential equation

$$\frac{d^2y}{dx^2} + \lambda y = 0 \quad (12.2.10)$$

subject to the boundary conditions

$$\frac{dy(0)}{dx} = \frac{dy(1)}{dx} = 0. \quad (12.2.11)$$

This constitutes a typical “eigenvalue problem” and has eigensolutions

$$\phi_n(x) = \cos n\pi x, \quad n = 0, 1, 2, \dots,$$

provided  $\lambda = \lambda_n = n^2\pi^2$ . Such eigenfunctions and eigenvalues have already been encountered in Chapter 11 and in Section 12.1 above. Indeed we know that for any function  $w(x)$  that satisfies the conditions (12.2.11) and is twice differentiable,  $\frac{d^2w}{dx^2}$  can be expanded as

$$\frac{d^2w}{dx^2} = \sum_{n=0}^{\infty} a_n \cos n\pi x \quad (12.2.12)$$

where

$$a_0 = \int_0^1 \frac{d^2w}{dx^2} dx = \left[ \frac{dw}{dx} \right]_0^1 = 0$$

and

$$a_n = 2 \int_0^1 \frac{d^2w}{dx^2} \cos n\pi x dx, \quad n \geq 1.$$

If we integrate (12.2.12) twice and use the boundary conditions (12.2.11) we find

$$w(x) = \sum_{n=0}^{\infty} -\frac{a_n}{n^2\pi^2} \cos n\pi x + C. \quad (12.2.13)$$

Thus

$$\begin{aligned} \int_0^1 \left( \frac{dw}{dx} \right)^2 dx &= \left[ w \frac{dw}{dx} \right]_0^1 - \int_0^1 w \frac{d^2w}{dx^2} dx \\ &= \int_0^1 \left( \sum_{n=0}^{\infty} \frac{a_n}{n^2\pi^2} \cos n\pi x \right) \left( \sum_{n=0}^{\infty} a_n \cos n\pi x \right) dx \\ &\quad - C \int_0^1 \left( \sum_{n=0}^{\infty} a_n \cos n\pi x \right) dx \\ &= \frac{1}{2} \sum_{n=0}^{\infty} \frac{a_n^2}{n^2\pi^2} \\ &\leq \frac{1}{2\pi^2} \sum_{n=0}^{\infty} a_n^2. \end{aligned} \quad (12.2.14)$$

Now if we multiply (12.2.12) by  $d^2w/dx^2$  and integrate we find

$$\int_0^1 \left( \frac{d^2w}{dx^2} \right)^2 dx = \sum_{n=0}^{\infty} a_n \int_0^1 \frac{d^2w}{dx^2} \cos n\pi x dx = \frac{1}{2} \sum_{n=0}^{\infty} a_n^2. \quad (12.2.15)$$

This is another useful result in the application of Fourier series and is called the **Parseval equality**. Using (12.2.15) in the estimate (12.2.14) leads us to the final result,

$$\int_0^1 \left( \frac{dw}{dx} \right)^2 dx \leq \frac{1}{\pi^2} \int_0^1 \left( \frac{d^2w}{dx^2} \right)^2 dx. \quad (12.2.16)$$

Comparing this result with the estimate (12.2.8) we see that  $\lambda = \pi^2$  and so if the diffusion coefficient  $D > m/\pi^2$  then  $E(t)$  decays at least as fast as  $\exp 2(m - \pi^2 D)t$ .

One should remark that this decay estimate depends on how good the bound  $m$  on  $|\partial f/\partial u|$  is; indeed  $m$  may be only a crude estimate and so  $D$  may have a lower bound smaller than  $m/\pi^2$ . Throughout we have only considered a problem defined in one space dimension; this is no real restriction because the Poincaré inequality (12.2.16) holds in any space dimension with, of course, a different constant  $\lambda$ .

Suppose now that instead of the boundary conditions (12.2.2) we impose the conditions

$$u(0, t) = u(1, t) = 0; \quad (12.2.17)$$

then the energy function (12.2.3) is not really appropriate and is replaced by

$$E(t) = \frac{1}{2} \int_0^1 u^2 dx. \quad (12.2.18)$$

Proceeding in the same way as before we have

$$\begin{aligned} \frac{dE}{dt} &= \int_0^1 u \frac{\partial u}{\partial t} dx \\ &= \int_0^1 u \left( D \frac{\partial^2 u}{\partial x^2} + f(u) \right) dx \\ &= D \int_0^1 u \frac{\partial^2 u}{\partial x^2} dx + \int_0^1 u f(u) dx \\ &= D \left[ u \frac{\partial u}{\partial x} \right]_0^1 - D \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 dx + \int_0^1 u f(u) dx \\ &= -D \int_0^1 \left( \frac{\partial u}{\partial x} \right)^2 dx + \int_0^1 u f(u) dx. \end{aligned} \quad (12.2.19)$$

Unlike the problem treated above, the inequality (12.2.8) does not help us to estimate usefully the first integral appearing in (12.2.9). However, if we look

at (12.2.16) and observe that the substitution  $V(x) = dw/dx$  satisfies the conditions

$$V(0) = V(1) = 0,$$

then we have an inequality for  $V$  of the form

$$\int_0^1 \left( \frac{\partial V}{\partial x} \right)^2 dx \geq \pi^2 \int_0^1 V^2 dx. \quad (12.2.20)$$

Using this in (12.2.19) gives

$$\frac{dE}{dt} \leq -2D\pi^2 E(t) + \int_0^1 u f(u) dx.$$

Now suppose  $|uf(u)| < mu^2$  for all solution values of  $u$ . Then we have

$$\begin{aligned} \frac{dE}{dt} &\leq -2D\pi^2 E(t) + m \int_0^1 u^2 dx \\ &= 2(m - D\pi^2)E(t), \end{aligned}$$

and once again we see that if  $D > m/\pi^2$ ,  $E(t) \rightarrow 0$  as  $t \rightarrow \infty$ , which implies  $u(x, t) \rightarrow 0$  as  $t \rightarrow \infty$ , and this means that the concentration  $u$  decays exponentially to zero regardless of the form of the initial conditions.

In a similar way we may deduce a corresponding behaviour if  $uf(u) \leq -du^2$  for all  $x \in [0, 1]$  and  $t \geq 0$ . Here we do not need the inequality (12.2.20) but simply ignore the first term on the right-hand side of (12.2.19) to arrive at the estimate

$$\begin{aligned} \frac{dE}{dt} &\leq \int_0^1 u f(u) dx \\ &\leq -d \int_0^1 u^2 dx = -2dE(t), \end{aligned}$$

and so

$$E(t) \leq E(0)e^{-2dt}.$$

The conclusions above hold also in this case. It is useful to note that unlike the problem in which there are “zero flux” conditions at  $x = 0$ ,  $x = 1$ , we have obtained our asymptotic results by imposing certain constraints on the form of the reaction term  $f(u)$  other than requiring it to be bounded for all solution values.

We shall look at this again in the following section.

### 12.3 Global behaviour of nerve impulse transmissions

In the study of the various models proposed to govern nerve impulse transmissions, an important feature that requires investigation is the long-term behaviour of the axon potential. This is a formidable task if we consider the Hodgkin–Huxley model. However, the FitzHugh–Nagumo system is a little more tractable and allows for the application of the energy arguments that we used in the previous section.

The problem to discuss is defined as

$$\begin{aligned}\frac{\partial^2 u}{\partial x^2} &= \frac{\partial u}{\partial t} - u(1-u)(u-a) + w, \\ \frac{\partial w}{\partial t} &= bu - \gamma w, \quad x \geq 0, t \geq 0,\end{aligned}\tag{12.3.1}$$

$$\begin{aligned}u(0, t) &= P(t), \quad u(x, 0) = 0, \\ w(0, t) &= b \int_0^t P(s) e^{-\gamma s} ds, \quad w(x, 0) = 0.\end{aligned}\tag{12.3.2}$$

Recall that in the conditions (12.3.2),  $P(t)$  represents the stimulus emanating at  $x = 0$ . In practice, this stimulus acts over a finite interval of time and so we make the additional assumption that

$$P(t) = 0 \quad \text{for } t \geq T$$

for some time  $T$ . Furthermore we assume that  $u(x, t)$ ,  $w(x, t)$  tend sufficiently rapidly to zero as  $x$  tends to infinity so that the integrals  $\int_0^\infty u^2(x, t) dx$ ,  $\int_0^\infty w^2(x, t) dx$  exist and are finite for all  $t$ .

If we multiply the first of equations (12.3.1) by  $bu$  and the second by  $w$  and subtract the resulting equations, we obtain

$$bu \frac{\partial^2 u}{\partial x^2} - w \frac{\partial w}{\partial t} = bu \frac{\partial u}{\partial t} - bu^2(1-u)(u-a) + \gamma w^2,$$

i.e.,

$$bu \frac{\partial u}{\partial t} + w \frac{\partial w}{\partial t} = bu \frac{\partial^2 u}{\partial x^2} + bu^2(1-u)(u-a) - \gamma w^2$$

or

$$\frac{1}{2} \frac{\partial}{\partial t} (bu^2 + w^2) = b \frac{\partial}{\partial x} \left( u \frac{\partial u}{\partial x} \right) - b \left( \frac{\partial u}{\partial x} \right)^2 + bu^2(1-u)(u-a) - \gamma w^2. \tag{12.3.3}$$

Integrating this expression with respect to  $x$  and using the initial conditions (12.3.2) we find, for  $t \geq T$ , that

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \int_0^\infty (bu^2 + w^2) dx &= \int_0^\infty b \frac{\partial}{\partial x} \left( u \frac{\partial u}{\partial x} \right) dx - b \int_0^\infty \left( \frac{\partial u}{\partial x} \right)^2 dx \\ &\quad + \int_0^\infty [bu^2(1-u)(u-a) - \gamma w^2] dx \\ &= -b \int_0^\infty \left( \frac{\partial u}{\partial x} \right)^2 dx + \int_0^\infty [bu^2(1-u)(u-a) - \gamma w^2] dx \\ &\leq \int_0^\infty [bu^2(1-u)(u-a) - \gamma w^2] dx. \end{aligned} \quad (12.3.4)$$

Suppose now that  $u(x, t) < a$  for all  $x, t > 0$ ; then we can find a constant  $\delta$  which depends on  $a$  so that

$$u^2(1-u)(u-a) \leq -\delta u^2.$$

This estimate used in (12.3.4) gives the inequality

$$\frac{1}{2} \frac{d}{dt} \int_0^\infty (bu^2 + w^2) dx \leq - \int_0^\infty (b\delta u^2 + \gamma w^2) dx.$$

If  $C = \min\{\delta, \gamma\}$  then

$$\frac{d}{dt} \int_0^\infty (bu^2 + w^2) dx \leq -2C \int_0^\infty (bu^2 + w^2) dx.$$

From this we conclude that

$$\int_0^\infty (bu^2 + w^2) dx \leq K e^{-2Ct}. \quad (12.3.5)$$

In summarising this result we can say that if  $u(x, t), w(x, t)$  are sufficiently smooth and if the action potential always remains below the potential value  $a$ , then the mean square of  $u(x, t)$  decays exponentially in time. In other words,  $a$  represents a threshold effect. Recall that we have already identified  $a$  as a threshold parameter in Chapter 7 where we considered the dynamics of the space-clamped FitzHugh–Nagumo model. Thus it is reasonable to conclude that, unless the potential  $u(x, t)$  exceeds  $a$ , a pulse will not propagate. Whether the axon reaches threshold or not depends on the stimulus  $P(t)$ .

From numerical and biological evidence there is support for the conjecture that a strong stimulus of short duration or a weak stimulus of long duration is sub-threshold. Mathematically, this conjecture can be shown to be correct as follows.

Suppose  $a > b/\gamma$  and that the stimulus  $P(t)$  is bounded, continuous and satisfies the conditions

- (i)  $P(t) = P(0) = 0$  for all  $t \geq T > 0$ ,
- (ii)  $\sup_{t \geq 0} |P(t)| \leq M$ .

Then

$$\sup_{x \geq 0} |u(x, t)| \leq k \int_0^T |P(t)| dt$$

for all  $t \geq T$ . Furthermore there are constants  $C, k$  and  $\lambda$  such that if  $\int_0^T |P| dt < \lambda$  then

$$\sup_{x \geq 0} |u(x, t)| \leq k \exp(-Ct), \quad t \geq 0.$$

This result, the proof of which is beyond the scope of this book, says that if  $a$  is not too small but the mean absolute value of the stimulus  $P$  is sufficiently small then the action potential  $u$  will decay exponentially. One can prove a further result, namely that if  $M$  in (ii) is sufficiently small then

$$\sup_{x \geq 0} |u(x, t)| \leq k \exp(-Ct), \quad t \geq 0.$$

Throughout our discussion of nerve axons we have always taken the axon to be infinite in extent. A real nerve is of finite extent and the correct problem to be studied is a mixed initial-boundary problem.

Suppose the axon is of length  $l$ . At  $x = 0$  we retain the conditions (12.3.2) while at  $x = l$  we impose the condition

$$u(l, t) = 0, \quad t \geq 0,$$

or

$$\frac{\partial u(l, t)}{\partial x} - \alpha u(l, t) = 0, \quad t \geq 0, \quad \alpha \leq 0. \quad (12.3.6)$$

What the precise conditions prevail in practice at  $x = l$  is not known.

In order to analyse the problem governed by (12.3.1), (12.3.2) and (12.3.6) we again suppose that the stimulus  $P(t)$  is nonzero over a finite time interval, i.e.,  $P(t) = 0, t \geq T$ .

Multiply the first of equations (12.3.1) by  $u$  and the second by  $b^{-1}w$  and subtract the results to get

$$\frac{1}{2} \frac{\partial}{\partial t} \left( u^2 + \frac{w^2}{b} \right) = u \frac{\partial^2 u}{\partial x^2} + u^2(1-u)(u-a) - \frac{\gamma}{b} w^2.$$

Integrating this over the interval  $[0, l]$  we find

$$\frac{1}{2} \int_0^l \frac{\partial}{\partial t} \left( u^2 + \frac{w^2}{b} \right) dx = \int_0^l \left( u \frac{\partial^2 u}{\partial x^2} + u^2(1-u)(u-a) - \frac{\gamma}{b} w^2 \right) dx. \quad (12.3.7)$$

Now

$$\int_0^l u \frac{\partial^2 u}{\partial x^2} dx = \left[ u \frac{\partial u}{\partial x} \right]_0^l - \int_0^l \left( \frac{\partial u}{\partial x} \right)^2 dx.$$

If  $u(0, t) = 0 = u(l, t)$ ,  $t \geq T$ , then

$$\int_0^l u \frac{\partial^2 u}{\partial x^2} dx = - \int_0^l \left( \frac{\partial u}{\partial x} \right)^2 dx.$$

Similarly if

$$u(0, t) = 0 = \frac{\partial}{\partial x} u(l, t) - \alpha u(l, t), \quad t \geq T,$$

then

$$\int_0^l u \frac{\partial^2 u}{\partial x^2} dx = \alpha u^2(l, t) - \int_0^l \left( \frac{\partial u}{\partial x} \right)^2 dx.$$

Thus, in either case, (12.3.7) gives rise to the inequality

$$\frac{1}{2} \frac{\partial}{\partial t} \int_0^l \left( u^2 + \frac{w^2}{b} \right) dx \leq \int_0^l \left[ u^2(1-u)(u-a) - \frac{\gamma}{b} w^2 dx - \left( \frac{\partial u}{\partial x} \right)^2 \right] dx. \quad (12.3.8)$$

In order to analyse (12.3.8) we first observe the quadrature term  $(1-u)(u-a)$  has a maximum at  $u = (1+a)/2$  given by  $(1-a)^2/4$  and so we can write

$$\frac{1}{2} \frac{\partial}{\partial t} \int_0^l \left( u^2 + \frac{w^2}{b} \right) dx \leq \int_0^l \left[ \frac{(1-a)^2}{4} u^2 - \frac{\gamma}{b} w^2 - \left( \frac{\partial u}{\partial x} \right)^2 \right] dx. \quad (12.3.9)$$

Next, we estimate the last term in (12.3.9) using the Poincaré inequality discussed in the previous section. For example, if we use the boundary conditions

$$u(0, t) = 0 = u(l, t), \quad t \geq T,$$

then

$$\int_0^l \left( \frac{\partial u}{\partial x} \right)^2 dx \geq \left( \frac{\pi}{l} \right)^2 \int_0^l u^2 dx. \quad (12.3.10)$$

A similar inequality is obtained if we use the boundary condition

$$\frac{\partial}{\partial x} u(l, t) - \alpha u(l, t) = 0.$$

Thus from (12.3.10) we have

$$\frac{1}{2} \frac{\partial}{\partial t} \int_0^l \left( u^2 + \frac{w^2}{b} \right) dx \leq \int_0^l \left[ \left( \frac{(1-a)^2}{4} - \frac{\pi^2}{l^2} \right) u^2 - \frac{\gamma}{b} w^2 \right] dx.$$

Now if  $(1-a)^2/4 - \pi^2/l^2 < 0$ , i.e.,  $l < 2\pi/(1-a)$ , then with

$$C = \max \left\{ \frac{\pi^2}{l^2} - \frac{(1-a)^2}{4}, \gamma \right\}$$

we have

$$\frac{\partial}{\partial t} \int_0^l \left( u^2 + \frac{w^2}{b} \right) dx \leq -2C \int_0^l \left( u^2 + \frac{w^2}{b} \right) dx.$$

From this we can argue, as we did above, to conclude that

$$\int_0^l \left( u^2 + \frac{w^2}{b} \right) dx$$

decays exponentially in time for  $t \geq T$ . In other words, unless the axon is long enough, all stimuli decay exponentially with time. The reader should bear in mind, from our previous discussion, that the converse is not necessarily true. That is, even an axon of infinite length cannot propagate a stimulus unless that stimulus satisfies certain criteria. The precise nature of  $P(t)$  necessary for the propagation of an action potential is not fully understood.

---

## 12.4 Global behaviour in chemical reactions

In this section we take up once more a study of the simplified model of the Belousov–Zhabotinskii reaction, which we dealt with in Chapter 8. The objective here is to enlarge on the discussion of travelling wavefronts.

The model to be considered is defined through the reaction–diffusion equations

$$\begin{aligned} \frac{\partial u}{\partial t} &= u(1-u-rv) + \frac{\partial^2 u}{\partial x^2}, \\ \frac{\partial v}{\partial t} &= -buv + \frac{\partial^2 v}{\partial x^2}, \quad -\infty < x < \infty, \quad t > 0, \end{aligned} \quad (12.4.1)$$

where  $u$  represents the nondimensionalised concentration of bromous acid ( $HBrO_2$ ) and  $v$  represents the nondimensionalised concentration of bromide ion ( $Br^-$ ). The constants  $r$  and  $b$  are positive and are of the order 10 and  $2.5 \times 10$ , respectively. Furthermore, the concentrations  $u$  and  $v$  satisfy the inequalities

$$0 \leq u \leq 1, \quad 0 \leq v \leq 1.$$

The type of wavefront solutions we seek are those which satisfy the boundary conditions

$$\begin{aligned} u(-\infty, t) &= 0, \quad u(\infty, t) = 1, \\ v(-\infty, t) &= 1, \quad v(\infty, t) = 0. \end{aligned} \quad (12.4.2)$$

There is one special case that can be treated fully; this is the situation in which

$$v = 1 - u. \quad (12.4.3)$$

For now we need only consider the equations

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + (1-r)u(1-u), \\ \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2} + bu(1-u),\end{aligned}\quad (12.4.4)$$

which are the same if  $b = 1 - r$  with  $r < 1$ .

We have shown in Chapter 8 that the equation

$$\frac{\partial \phi}{\partial t} = \frac{\partial^2 \phi}{\partial x^2} + k\phi(1-\phi)$$

has travelling wave solutions  $\phi(x + ct)$  satisfying the conditions  $\phi(-\infty) = 0$ ,  $\phi(\infty) = 1$ , provided  $c \geq 2\sqrt{k}$ . Consequently the system (12.4.4) has travelling wave solutions of speed

$$c \geq 2\sqrt{(1-r)} = 2\sqrt{b}, \quad \text{if } r \leq 1.$$

If, however, we impose initial data of the form

$$u(x, 0) = h(x), \quad (12.4.5)$$

where  $h(x)$  is zero for  $x < 0$  and equals one for  $x > 0$ , then Kolmogorov, Petrovskii and Piskounov, as long ago as 1937, proved that  $u(x, t)$  satisfying (12.4.4), (12.4.5) evolves into a travelling wave  $u(x + ct)$  as  $t \rightarrow \infty$ , where  $c = 2\sqrt{(1-r)} = 2\sqrt{b}$  if  $r \leq 1$ .

This result can be established by following comparison type arguments similar to those used in Chapter 11, and depends very much on the nonlinear term  $u(1-u)$ . We shall not go further into the matter here.

We now demonstrate that any wave solution, in which  $u \geq 0$ ,  $v \geq 0$ , must have a wave speed  $c(r, b) \leq 2$  for all  $b \geq 0$  and all  $r \geq 0$ . To do this, we shall introduce the function  $w(x, t)$  defined as the solution to the problem

$$\begin{aligned}\frac{\partial w}{\partial t} &= \frac{\partial^2 w}{\partial x^2} + w(1-w), \\ w(x, 0) &= h(x), \\ w(-\infty, t) &= 0, \quad w(\infty, t) = 1,\end{aligned}\quad (12.4.6)$$

where  $h(x)$  is of the same form as described above.

We now know that  $w(x, t)$  evolves into a travelling wave speed 2. We write

$$\bar{w}(x, t) = u(x, t) - w(x, t), \quad (12.4.7)$$

where  $u(x, t)$  is the unique solution of (12.4.1) which has the same initial and boundary conditions as  $w(x, t)$ . Subtracting the equation in (12.4.6) from the first of equations (12.4.1), we find

$$\frac{\partial^2 \bar{w}}{\partial x^2} - \frac{\partial \bar{w}}{\partial t} + [1 - (w + u)]\bar{w} = ruv \quad (12.4.8)$$

with the restriction  $w \geq 0$ ,  $u \geq 0$  and with the class of initial conditions we have  $u \leq 1$ ,  $w \leq 1$ . This last statement is a consequence of the comparison theorem in Chapter 11. Thus  $1 - (w + u) \leq 1$ . What we would like to do now is to employ the maximum principle of Chapter 11 in order to estimate  $\bar{w}$ . However, since the coefficient of  $\bar{w}$  (12.4.8) is less than one, some modification must first be carried out. If we set

$$W = \bar{w}e^{-Kt}$$

for some positive constant  $K$  then (12.4.8) becomes

$$\frac{\partial^2 W}{\partial x^2} - \frac{\partial W}{\partial t} + [1 - (w + u) - K]W = ruve^{-Kt} \geq 0.$$

Choose  $K > 1$  so that  $1 - (w + u) - K < 0$ . On noting that  $W$  satisfies the same boundary and initial conditions as  $\bar{w}$ , the maximum principle can now be invoked to conclude that  $W$  has its maximum at  $t = 0$  or at  $|x| = \infty$ .

Thus  $\bar{w}$  has its maximum at either of these possible points. However, the maximum value of  $\bar{w}$  occurs at the maximum value of  $u(x, t) - w(x, t)$ , which is zero at  $t = 0$  and  $x = \pm\infty$ . From this we conclude that

$$u(x, t) \leq w(x, t) \quad (12.4.9)$$

for all  $x$  and  $t > 0$ . But this implies that the solution  $u(x, t)$  is at all points less than or equal to  $w(x, t)$ ; in particular this is true for the travelling wave solution  $w(x, t) = f(x + ct)$  of (12.4.6). Thus, if  $u(x, t)$  does evolve into a travelling wave solution with speed  $c$  then  $c \leq 2$ ; for if the wave speed were greater than two the inequality (12.4.9) would be violated.

To complete our discussion we consider the behaviour of the wave speed  $c(r, b)$  for limiting values of  $r$  and  $b$ . If  $b = 0$  then the second equation of (12.4.1) reduces to the heat equation

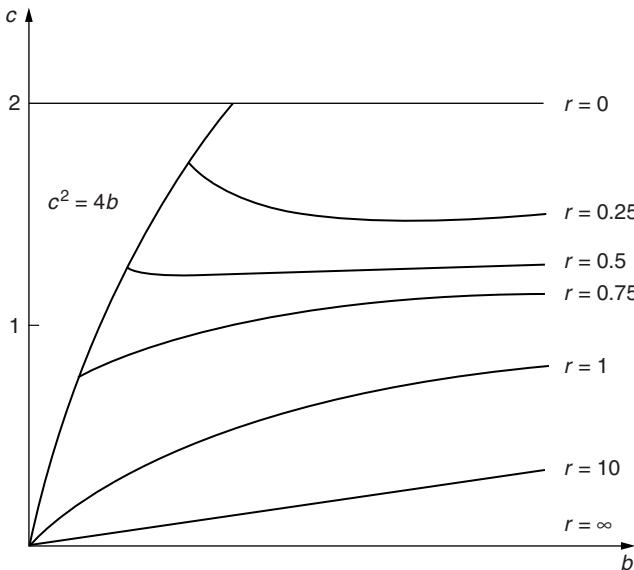
$$\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2},$$

which does not have travelling wave solutions and so neither can the equation for  $u(x, t)$ . Thus we conclude that  $c(r, b) \rightarrow 0$  as  $b \rightarrow 0$  with  $r > 0$ .

If  $b \rightarrow \infty$  then (12.4.1) implies  $v = 0$  (the trivial solution  $u = 0$  is excluded). In this case  $c(b, r) \rightarrow 2$  as  $b \rightarrow \infty$  with  $r \geq 0$ . Next, if  $r = 0$  then (12.4.1) decouples and, in particular, we have to discuss the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + u(1 - u).$$

This equation is simply the equation (12.4.6) for which travelling wave solutions exist with speed  $c = 2$ . Thus with  $r = 0$  both  $u$  and  $v$  evolve into travelling waves with speed  $c = 2$ , i.e.,  $c(r, b) \rightarrow 2$  as  $r \rightarrow 0, b > 0$ . If  $r \rightarrow \infty$



**FIGURE 12.4.1:** Wave speed  $c$  as a function of  $b$  for various values of  $r$ .

then  $u \equiv 0$  or  $v \equiv 0$  and in either case there can be no wave solution and so  $c(r, b) \rightarrow 0$  as  $r \rightarrow \infty$ . These results are summarised in Figure 12.4.1

There are a number of important problems that remain in the discussion of the Belousov–Zhabotinskii reaction and diffusion problems in general. For example, there is the question of existence of periodic solutions, as well as a thorough study of those classes of initial data  $h(x)$  for which  $u$  and  $v$  evolve into travelling waves. That is to say, how do solutions behave under small disturbances of initial and/or boundary data? Such problems are largely beyond the scope of this book, but the interested reader is encouraged to consult the literature cited in the notes at the end of this chapter.

## 12.5 Turing diffusion driven instability and pattern formation

Here we discuss the role of pattern in developmental biology and the mechanisms proposed to analyse it from a mathematical point of view. To begin with we introduce the following biological concepts:

**Embryology** : embryology is that part of biology that is concerned with the formation and development of the embryo from fertilization until birth. Embryological development is a sequential process and follows a ground plan laid down early in gestation.

**Morphogenesis** : morphogenesis is that part of embryology that is concerned with the development of pattern and form. How the development of the ground plan is established is not really understood.

**Pattern and form** : Whatever pattern is observed in the animal world it is almost certain that the process that produced it is unknown. The mechanism must be genetically controlled but the genes themselves cannot create the pattern.

This promotes the question: how is genetic information translated into patterns?

There have been proposed two major concepts that provide mechanisms for pattern formation: (i) positional information due to Lewis Wolpert (1969) and (ii) diffusion driven instability due to Alan Turing (1952).

Wolpert's idea of positional information suggests that cells are pre-programmed to read a chemical (or morphogen) concentration and differentiate accordingly into different kinds of cells leading to the development of, for example, cartilage, bone, tissue, hair, etc. This point of view separates the developmental process into the following steps.

- (i) The creation of a morphogen concentration of spatial pattern (a morphogen is the chemical associated with creating morphogenesis).
- (ii) The establishment of positional information that depends on a chemical pre-specification so that the cell can read out its position in the coordinates of chemical concentration and differentiate, undergo appropriate cell shape change or migrate.
- (iii) Once the pre-pattern is established, morphogenesis is a slave process.

To date it is very much an open problem as to how this idea is to be implemented in mathematical modelling.

In 1952 Alan Turing suggested that under certain conditions chemicals can react and diffuse in such a way as to produce steady state heterogeneous spatial patterns of chemical concentration. It is this counterintuitive idea that we pursue here.

To begin with let us return to the discussion leading to the derivation of Fick's second law. This time, however, instead of considering the flow of solute through an arbitrary volume of fluid, we first suppose the volume  $V$  to contain a mixture of chemicals of concentration  $c_i, i = 1, \dots, n$  and further that there is a chemical source  $\mathbf{f}$  that is vector valued and depends only on the chemical concentrations  $c_i$ . Using the law of conservation of mass we argue as we did before. That is, the rate of change of a chemical concentration in  $V$  must equal the rate of change of flow of chemical across the surface  $S$  together with the chemical concentration created in  $V$ . Consequently for each chemical  $c_i$  we have

$$\frac{\partial}{\partial t} \int_V c_i(\mathbf{x}, t) dV = - \int_S \mathbf{J}_i \cdot \mathbf{n} dS + \int_V f_i dV, \quad (12.5.1)$$

where  $\mathbf{J}_i = -D_i \nabla c_i$  with  $D_i$  being the diffusion constant for  $c_i$ . Again, by noting that  $V$  is arbitrary we arrive at the reaction-diffusion equation

$$\frac{\partial c_i}{\partial t} = D_i \nabla^2 c_i + f_i(c_1, \dots, c_n), \quad i = 1, 2, \dots, n. \quad (12.5.2)$$

In matrix notation

$$\frac{\partial \mathbf{c}}{\partial t} = D \nabla^2 \mathbf{c} + \mathbf{f}(\mathbf{c}), \quad (12.5.3)$$

where  $D$  is a diagonal diffusivity matrix with entries  $D_i, i = 1, \dots, n$ ,  $\mathbf{c} = (c_1, \dots, c_n)$ ,  $\mathbf{f}(\mathbf{c}) = (f_1(c_1, \dots, c_n), \dots, f_n(c_1, \dots, c_n))$ .

Throughout we shall consider the case of chemicals  $c_1, c_2, n = 2$ . That is, we consider the system

$$\begin{aligned} \frac{\partial c_1}{\partial t} &= D_1 \nabla^2 c_1 + f_1(c_1, c_2), \\ \frac{\partial c_2}{\partial t} &= D_2 \nabla^2 c_2 + f_2(c_1, c_2). \end{aligned} \quad (12.5.4)$$

Suppose in the first instance that the spatial domain is one-dimensional and infinite in extent. The situation in which the domain is finite will be considered later. In this case we have the system

$$\begin{aligned} \frac{\partial c_1}{\partial t} &= D_1 \frac{\partial^2 c_1}{\partial x^2} + f_1(c_1, c_2), \\ \frac{\partial c_2}{\partial t} &= D_2 \frac{\partial^2 c_2}{\partial x^2} + f_2(c_1, c_2). \end{aligned} \quad (12.5.5)$$

Suppose that there exists a steady state to this system in which the chemicals are well mixed and uniform. Turing's profound idea is simple. He suggested that if in the absence of diffusion (i.e.,  $D_1 = D_2 = 0$  in (12.5.5)),  $c_1$  and  $c_2$  approach a linearly stable uniform steady state as  $t \rightarrow \infty$ ; then under certain conditions spatially inhomogeneous patterns can evolve by "diffusion driven instability" for  $D_1 \neq D_2 \neq 0$ .

It is this fundamental idea that we shall explore.

Suppose there are solutions

$$c_1 = c_{1,0}, \quad c_2 = c_{2,0} \quad (12.5.6)$$

of (12.5.5) which are independent of time  $t$  and space  $x$ . Notice that if  $c_{1,0}$  and  $c_{2,0}$  are steady states then, from (12.5.5), they must satisfy

$$f_1(c_1, c_2) = f_2(c_1, c_2) = 0. \quad (12.5.7)$$

Consequently there may be several steady states, each determined as a solution of (12.5.7).

We now ask a fundamental question:

Can spatially inhomogeneous patterns spontaneously develop from arbitrarily small perturbations from a uniform **linearly stable** steady state?

To investigate this question let

$$\begin{aligned} c_1(x, t) &= c_{1,0} + d_1, \\ c_2(x, t) &= c_{2,0} + d_2, \end{aligned} \quad (12.5.8)$$

where  $|d_i| \ll 1, i = 1, 2$ .

We now “linearise” (12.5.5) by expanding  $f_i(c_1, c_2)$  as a Taylor series about  $(c_{1,0}, c_{2,0})$ . That is,

$$\begin{aligned} f_i(c_1, c_2) &= f_i(c_{1,0}, c_{2,0}) + d_1 \frac{\partial f_i}{\partial c_1}(c_{1,0}, c_{2,0}) + d_2 \frac{\partial f_i}{\partial c_2}(c_{1,0}, c_{2,0}) \\ &\quad + \text{higher order terms in } d_1, d_2, \quad i = 1, 2. \end{aligned}$$

To a first order approximation in which we neglect the higher order terms and noting that  $f_i(c_{1,0}, c_{2,0}) = 0$  we arrive at the linear system

$$\frac{\partial d_i}{\partial t} = D_i \frac{\partial^2 d_i}{\partial x^2} + d_1 \frac{\partial f_i}{\partial c_1}(c_{1,0}, c_{2,0}) + d_2 \frac{\partial f_i}{\partial c_2}(c_{1,0}, c_{2,0}), \quad i = 1, 2. \quad (12.5.9)$$

For notational convenience and ease of presentation we define the quantities

$$a_{ij} \equiv \frac{\partial f_i}{\partial c_j}(c_{1,0}, c_{2,0}), \quad i, j = 1, 2. \quad (12.5.10)$$

Then (12.5.9) can be written as

$$\begin{aligned} \frac{\partial d_1}{\partial t} &= D_1 \frac{\partial^2 d_1}{\partial x^2} + a_{11}d_1 + a_{12}d_2, \\ \frac{\partial d_2}{\partial t} &= D_2 \frac{\partial^2 d_2}{\partial x^2} + a_{21}d_1 + a_{22}d_2, \end{aligned} \quad (12.5.11)$$

or, in matrix form,

$$\frac{\partial \mathbf{d}}{\partial t} = D \frac{\partial^2 \mathbf{d}}{\partial x^2} + A \mathbf{d}, \quad (12.5.12)$$

where  $\mathbf{d} = (d_1, d_2)^T$ ,  $D = \text{diag}(D_1, D_2)$  and  $A$  is the  $2 \times 2$  matrix with entries  $a_{i,j}$ .

The system (12.5.12) can be solved by the method of separation of variables as described in Chapter 11.

We shall look for solutions of the form

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} e^{\sigma t} \cos kx, \quad (12.5.13)$$

in which  $\alpha_i (i = 1, 2)$ ,  $\sigma$  and  $k$  are constants. This form of solution is not the most general solution to (12.5.12). Indeed we could choose a solution of the form

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} e^{\sigma t} \sin kx, \quad (12.5.14)$$

or even a linear combination of such solutions. Whichever choice is made leads to the same analysis and so there is no loss in generality. The importance of the form (12.5.13) will become apparent when we consider diffusion driven instability for systems defined on finite spatial domains in which boundary conditions have to be taken into account.

On substituting (12.5.13) into (12.5.12) and dividing the resulting equations throughout by  $e^{\sigma t} \cos kx$  we arrive at the pair of algebraic equations

$$\begin{aligned}\alpha_1\sigma &= a_{11}\alpha_1 + a_{12}\alpha_2 - D_1k^2\alpha_1, \\ \alpha_2\sigma &= a_{21}\alpha_1 + a_{22}\alpha_2 - D_2k^2\alpha_2,\end{aligned}$$

i.e.,

$$\begin{aligned}\alpha_1(\sigma - a_{11} + D_1k^2) - \alpha_2a_{12} &= 0, \\ -\alpha_1a_{21} + \alpha_2(\sigma - a_{22} + D_2k^2) &= 0.\end{aligned}\quad (12.5.15)$$

If we think of this pair of equations as simultaneous equations for the determination of  $(\alpha_1, \alpha_2)$  then for a nontrivial solution to exist we must have

$$\begin{vmatrix} \sigma - a_{11} + D_1k^2 & -a_{12} \\ -a_{21} & \sigma - a_{22} + D_2k^2 \end{vmatrix} = 0,$$

i.e.,

$$\sigma^2 + \sigma((D_1 + D_2)k^2 - (a_{11} + a_{22})) + ((a_{11} - D_1k^2)(a_{22} - D_2k^2) - a_{12}a_{21}) = 0. \quad (12.5.16)$$

This is the characteristic equation for the determination of  $\sigma$ . Of fundamental importance is the fact that  $\sigma$  determines the growth or decay in time of the solution (12.5.13). That is, if  $\text{Re}(\sigma) > 0$  then the solution (12.5.13) grows exponentially in time and so is **unstable**. This means that the homogeneous steady state  $(c_{1,0}, c_{2,0})$  is unstable to small spatially inhomogeneous perturbations. Likewise if  $\text{Re}(\sigma) < 0$  then the steady state  $(c_{1,0}, c_{2,0})$  is stable to small spatially inhomogeneous perturbations.

Recall that in the absence of diffusion we require the uniform steady state  $(c_{1,0}, c_{2,0})$  to be linearly stable. So on setting  $D_1 = D_2 = 0$  in (12.5.16) we require the roots  $\sigma_1, \sigma_2$ , say, of

$$\sigma^2 - \sigma(a_{11} + a_{22}) + (a_{11}a_{22} - a_{12}a_{21}) = 0, \quad (12.5.17)$$

to satisfy  $\text{Re}(\sigma_i) < 0, i = 1, 2$ . Now

$$\begin{aligned}\sigma_1 + \sigma_2 &= a_{11} + a_{22}, \\ \sigma_1\sigma_2 &= (a_{11}a_{22} - a_{12}a_{21}),\end{aligned}$$

from which it is clear that whether  $\sigma_1, \sigma_2$  are real or complex,  $\text{Re}(\sigma_i) < 0$  if and only if

$$\begin{aligned}a_{11} + a_{22} &< 0, \\ a_{11}a_{22} - a_{12}a_{21} &> 0.\end{aligned}\quad (12.5.18)$$

That these inequalities hold is crucial as to whether spatially inhomogeneous patterns can evolve by diffusion driven instability.

When  $D_1, D_2 \neq 0$  we require the roots  $\sigma'_1, \sigma'_2$  of (12.5.16) to be such that  $\text{Re}(\sigma'_1)$  and/or  $\text{Re}(\sigma'_2)$  are nonnegative.

Now

$$\sigma'_1 + \sigma'_2 = (a_{11} + a_{22}) - (D_1 + D_2)k^2 \quad (12.5.19)$$

and

$$\sigma'_1 \sigma'_2 \equiv H(k^2) = (a_{11} - D_1 k^2)(a_{22} - D_2 k^2) - a_{12}a_{21}. \quad (12.5.20)$$

Since  $D_1, D_2$  are positive and noting (12.5.18) it is clear that

$$\sigma'_1 + \sigma'_2 < 0. \quad (12.5.21)$$

If  $\sigma'_1, \sigma'_2$  are complex then (12.5.21) implies that  $\text{Re}(\sigma'_i) < 0$ , ( $i = 1, 2$ ) and  $\sigma'_1 \sigma'_2 = |\sigma'_1|^2 > 0$  and so the perturbation (12.5.12) decays exponentially and patterns do not form. So the only possibility is that  $\sigma'_1, \sigma'_2$  are real and  $\sigma'_1 \sigma'_2 < 0$ , i.e.,  $H(k^2) < 0$ .

Let us examine  $H(k^2)$  more closely. As a quadratic equation in  $k^2$  we have

$$H(k^2) = D_1 D_2 k^4 - (D_1 a_{22} + D_2 a_{11})k^2 + a_{11}a_{22} - a_{12}a_{21}. \quad (12.5.22)$$

When  $k = 0, H(k^2) > 0$  and if  $H(k^2) < 0$  for some values of  $k^2$  then (12.5.22) must have real positive roots and so

$$D_1 a_{22} + D_2 a_{11} > 0 \quad (12.5.23)$$

and

$$(D_1 a_{22} + D_2 a_{11})^2 \geq 4D_1 D_2 (a_{11}a_{22} - a_{12}a_{21}). \quad (12.5.24)$$

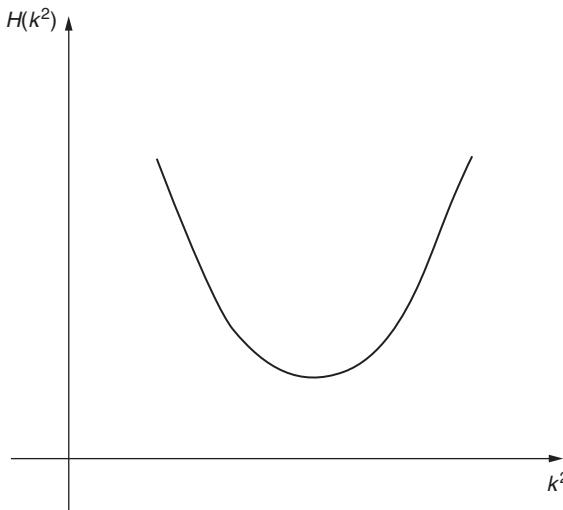
In general the graph of  $H(k^2)$  against  $k^2$  will take one of the forms illustrated in Figures 12.5.1a,b,c.

From (12.5.23), (12.5.24) and Figures 12.5.1a,b,c we arrive at some very important conclusions about the type of reaction diffusion systems that lead to diffusion driven instability. From (12.5.18) and (12.5.23) we immediately conclude that  $D_1 \neq D_2$ . In other words one morphogen must diffuse faster than the other. Again from these inequalities we see that  $a_{11}$  and  $a_{22}$  must be of opposite sign.

Suppose  $a_{11} > 0$  and  $a_{22} < 0$ . Then from (12.5.10) we have

$$a_{11} = \frac{\partial f_1}{\partial c_1}(c_{1,0}, c_{2,0}) > 0, \quad a_{22} = \frac{\partial f_2}{\partial c_2}(c_{1,0}, c_{2,0}) > 0,$$

which means that  $c_1$  is **activated** by its own rate of production whereas  $c_2$  is **inhibited** by its rate of production. In this case we naturally refer to  $c_1$  as the **activator** and  $c_2$  as the **inhibitor**.



**FIGURE 12.5.1a:** Steady state stable; no pattern formation.  $H(k^2) > 0$  and the steady state is stable to small perturbations.

Furthermore we also see that

$$\frac{D_2}{D_1} > -\frac{a_{22}}{a_{11}} > 1,$$

or

$$D_2 > D_1. \quad (12.5.25)$$

In other words the inhibitor  $c_2$  diffuses faster than the activator  $c_1$ .

With the choice  $a_{11} > 0, a_{22} < 0$ , the fact that (see (12.5.18))  $a_{11}a_{22} - a_{12}a_{21} > 0$  means that  $a_{12}a_{21} < 0$ . If  $a_{12} < 0, a_{21} > 0$  then we have an **activator-inhibitor** system whereas if  $a_{12} > 0, a_{21} < 0$  we have a **positive feedback** system.

Turning now to Figures 12.5.1a,b,c, we see that in a) there are no values of  $k^2$  for which  $H(k^2) < 0$  and so small perturbations are stable and no patterns form. Figure 12.5.1b shows the critical case in which there is precisely one value  $k_c^2$  of  $k^2$  occurring when  $H(k^2) = 0$ . This value is easily calculated to be

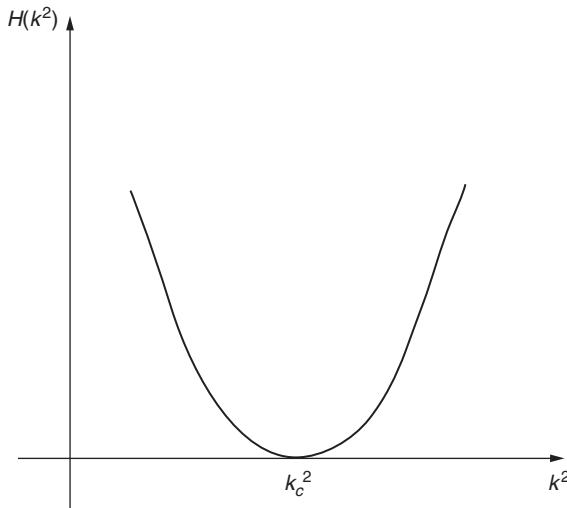
$$k_c^2 = \frac{1}{2} \left( \frac{a_{11}}{D_1} + \frac{a_{22}}{D_2} \right), \quad (12.5.26)$$

where  $D_1$  and  $D_2$  are critical diffusion constants that satisfy

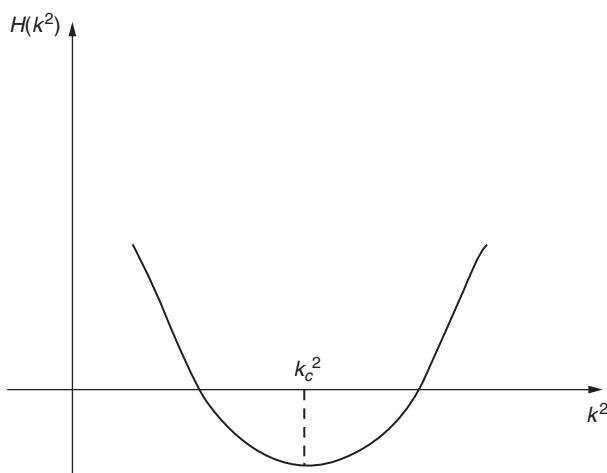
$$(D_1 a_{22} + D_2 a_{11})^2 = 4 D_1 D_2 (a_{11} a_{22} - a_{12} a_{21}).$$

In other words the critical wave number and critical diffusion constants are related by

$$k_c^2 = \left[ \frac{(a_{11} a_{22} - a_{12} a_{21})}{D_1 D_2} \right]^{1/2}. \quad (12.5.27)$$



**FIGURE 12.5.1b:** Critical case; onset of instability.  $H(k^2)$  has a single critical root  $k_c^2$  indicating the onset of instability of the steady state to small perturbations.



**FIGURE 12.5.1c:** Steady state unstable; patterns form.  $H(k^2) < 0$  for a range of  $k^2$  for which the steady state is unstable to small perturbations.

In Figure 12.5.1c,  $H(k^2) < 0$  and so there are a continuum of spatial patterns that can evolve by diffusion driven instability for a continuum of  $k^2$  satisfying

$$k_1^2 < k^2 < k_2^2, \quad (12.5.28)$$

where  $H(k_i^2) = 0$ ,  $i = 1, 2$ .

To illustrate the general analysis we consider the specific example:

$$\begin{aligned} \frac{\partial c_1}{\partial t} &= \frac{c_1^2}{c_2} - bc_1 + D_1 \frac{\partial^2 c_1}{\partial x^2}, \\ \frac{\partial c_2}{\partial t} &= c_1^2 - c_2 + D_2 \frac{\partial^2 c_2}{\partial x^2}, \quad (x, t) \in (-\infty, \infty) \times (0, \infty), \end{aligned} \quad (12.5.29)$$

where  $b$  is a positive constant.

The steady states of the system (12.5.29) are determined by

$$\frac{c_1^2}{c_2} = bc_1, \quad c_1^2 = c_2. \quad (12.5.30)$$

The nontrivial steady state is therefore

$$c_{1,0} = \frac{1}{b}, \quad c_{2,0} = \frac{1}{b^2}. \quad (12.5.31)$$

As above we linearise the system about this steady state by writing

$$c_1 = \frac{1}{b} + d_1, \quad c_2 = \frac{1}{b^2} + d_2. \quad (12.5.32)$$

Furthermore, from (12.5.10), we can write down the coefficients  $a_{ij}$ ,  $i, j = 1, 2$  of the linearised system (12.5.11) as

$$\begin{aligned} a_{11} &= b, & a_{12} &= -b^2, \\ a_{21} &= \frac{2}{b}, & a_{22} &= -1. \end{aligned} \quad (12.5.33)$$

Note that  $a_{11}$  and  $a_{22}$  are of opposite sign and furthermore the inequalities (12.5.18) demand that

$$0 < b < 1. \quad (12.5.34)$$

For the evolution of pattern we require  $H(k^2)$  given by (12.5.22) to be non-positive, i.e.,

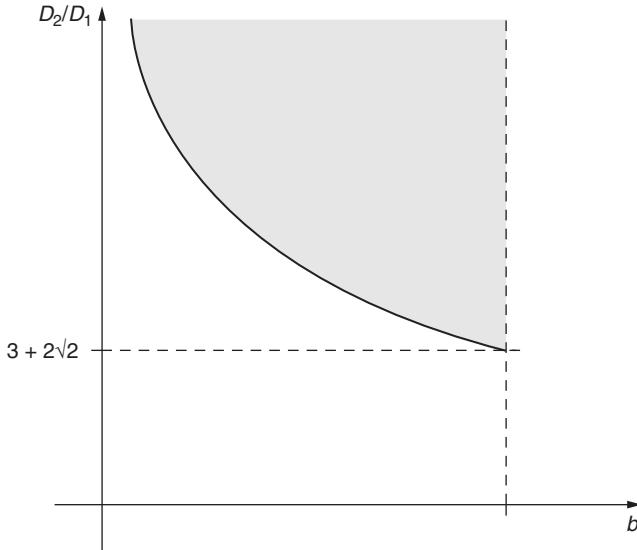
$$D_1 D_2 k^4 - (b D_2 - D_1) k^2 + b \leq 0, \quad (12.5.35)$$

together with the inequalities (see (12.5.23) and (12.5.24))

$$b D_2 - D_1 > 0 \quad (12.5.36)$$

and

$$(b D_2 - D_1)^2 \geq 4 D_1 D_2 b. \quad (12.5.37)$$



**FIGURE 12.5.2:** Turing parameter space for the system (12.5.29).

Consequently

$$\frac{D_2}{D_1} > \frac{1}{b} > 1,$$

showing that the inhibitor  $c_2$  diffuses faster than the activator  $c_1$ .

Also from (12.5.37) we deduce that

$$\frac{D_2}{D_1} b \geq 3 + 2\sqrt{2}. \quad (12.5.38)$$

This inequality defines the “Turing parameter space”, shown in Figure 12.5.2, in which diffusion driven instability occurs.

Next, the critical wave number  $k_c$  is given by

$$k_c^2 = \frac{1}{2} \left( \frac{b}{D_1} - \frac{1}{D_2} \right) \quad (12.5.39)$$

and a continuum of spatial patterns can evolve for all values of  $k^2$  satisfying

$$k_1^2 < k^2 < k_2^2$$

where  $H(k_i^2) = 0$ ,  $i = 1, 2$ . In other words

$$2D_1 D_2 k_1^2 = (bD_2 - D_1) - \sqrt{(bD_2 - D_1)^2 - 4D_1 D_2 b}$$

and

$$2D_1 D_2 k_2^2 = (bD_2 - D_1) + \sqrt{(bD_2 - D_1)^2 - 4D_1 D_2 b}.$$

## 12.6 Finite pattern forming domains

Let us now consider the general system (12.5.4) defined on a bounded domain. In the case of one-space dimension we consider the domain to be the interval  $x \in [0, l]$ . To complete the formulation of the problem we require boundary as well as initial conditions. That is, in addition to (12.5.4) we assume  $c_1(x, t)$ ,  $c_2(x, t)$  satisfy

$$c_1(x, 0) = f_1(x), \quad c_2(x, 0) = f_2(x), \quad (12.6.1)$$

where  $f_1(x)$  and  $f_2(x)$  are given concentrations. Also we impose the “**no flux**” boundary conditions

$$\frac{\partial c_i}{\partial x}(x, t) = 0, \quad x = 0, l, \quad i = 1, 2. \quad (12.6.2)$$

The reason for choosing zero flux boundary conditions is that we are primarily interested in self-organisation of patterns. No flux boundary conditions mean that there is to be no external input of morphogens.

To develop an analysis of Turing driven instabilities in this case we proceed in precisely the same way as we did above for an infinite domain. The crucial point is that the perturbations  $d_1$  and  $d_2$ , which solve the system (12.5.11), must now satisfy the conditions

$$\frac{\partial d_i}{\partial x}(x, t) = 0, \quad x = 0, l, \quad i = 1, 2. \quad (12.6.3)$$

A little thought shows that  $d_1$ ,  $d_2$  must be of the form (12.5.13) where

$$\cos kx = 0 \quad (12.6.4)$$

for  $x = 0, l$ . Consequently, rather than being arbitrary, the parameter  $k$  must take the discrete values  $k_n$ ,  $n = 1, \dots$ , where

$$\sin kl = 0,$$

i.e.,

$$k_n = \frac{n\pi}{l}, \quad n = 1, 2, \dots \quad (12.6.5)$$

This observation is crucial; it is fundamental to deciding which patterns are selected. Proceeding precisely as before we find that the critical wave number  $k_c$  is given by

$$\frac{n^2\pi^2}{l^2} = \frac{1}{2} \left( \frac{a_{11}}{D_1} + \frac{a_{22}}{D_2} \right), \quad (12.6.6)$$

which may or may not be satisfied for a given  $l$ . However if one is allowed to vary  $l$  then it is possible to choose a mode characterised by  $n$  so that (12.6.6)

is satisfied. In other words a critical pattern of a certain form is dependent on the size of the domain.

More generally we see that spatial patterns evolve if we can select integer values of  $n$  so that  $H(k_n^2) < 0$ . This can only happen if we can choose  $n$  so that

$$k_1^2 < \frac{n^2\pi^2}{l^2} < k_2^2 \quad (12.6.7)$$

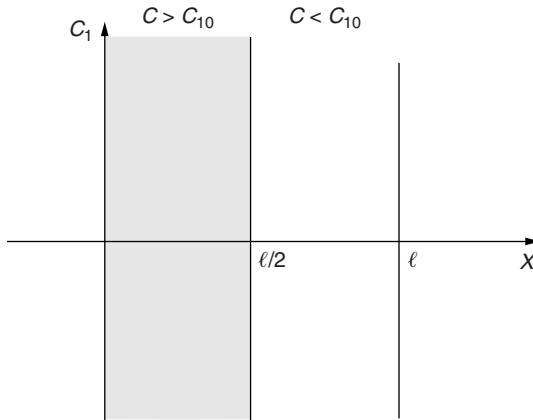
where  $k_i^2$ ,  $i = 1, 2, \dots$ , satisfy  $H(k_i^2) = 0$ . That is,

$$\begin{aligned} \frac{(D_1 a_{22} + D_2 a_{11}) - \sqrt{(D_1 a_{22} + D_2 a_{11})^2 - 4D_1 D_2 A}}{2D_1 D_2} &< \frac{n^2\pi^2}{l^2} \\ &< \frac{(D_1 a_{22} + D_2 a_{11}) + \sqrt{(D_1 a_{22} + D_2 a_{11})^2 - 4D_1 D_2 A}}{2D_1 D_2}, \end{aligned} \quad (12.6.8)$$

where  $A = (a_{11}a_{22} - a_{12}a_{21}) > 0$ . The number of integers  $n$  for which (12.6.8) is satisfied determines the modes of pattern selection. To see this suppose the domain size  $l$  is such that (12.6.8) is satisfied only for  $n = 1$ . The only unstable mode is  $\cos \frac{\pi x}{l}$  and morphogen concentration

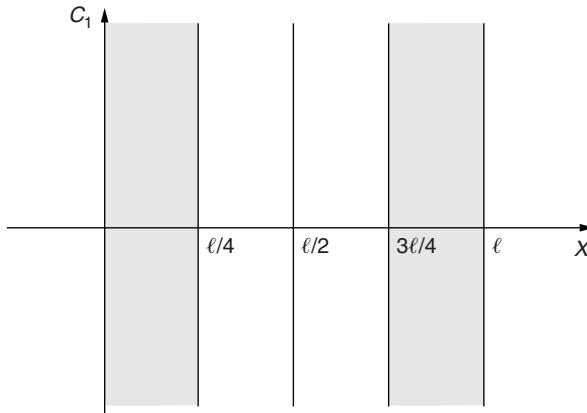
$$c_1(x, t) \sim c_{1,0} + \alpha \exp\left[\lambda\left(\frac{\pi^2}{l^2}\right)t\right] \cos \frac{\pi x}{l}, \quad (12.6.9)$$

where  $\lambda(\frac{\pi^2}{l^2})$  is the positive root of (12.5.6). This unstable mode is the dominant one which emerges as  $t$  increases. If we say that **black** corresponds to a concentration above the steady state  $c_{1,0}$  and **white** corresponds to a concentration below  $c_{1,0}$  then we have the pattern shown in Figure 12.6.1.



**FIGURE 12.6.1:** Morphogen pattern for  $n = 1$ .

Similarly if  $n = 2$  is the only value of  $n$  for which the inequality (12.6.8)



**FIGURE 12.6.2:** Morphogen pattern for  $n = 2$ .

holds then (12.6.9) becomes

$$c_1(x, t) \sim c_{1,0} + \alpha \exp\left[\lambda\left(\frac{4\pi^2}{l^2}\right)t\right] \cos\frac{2\pi x}{l}, \quad (12.6.10)$$

leading to the pattern shown in Figure 12.6.2.

Which mode or combination of modes and hence patterns are selected depends on initial conditions (12.6.1).

Now consider the two-dimensional domain  $\Omega$  defined by  $0 \leq x \leq l$ ,  $0 \leq y \leq h$ , with rectangular boundary  $\partial\Omega$  on which no-flux boundary conditions are imposed.

Once again the theory developed above in the one-dimensional case is followed with only minor modifications. Most importantly we seek solutions of the linearised problem (12.5.11) of the form

$$\begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} e^{\sigma t} \cos k_1 x \cos k_2 y, \quad (12.6.11)$$

where the wave numbers  $k_1$  and  $k_2$  are chosen so that  $d_1$  and  $d_2$  satisfy the boundary conditions:

$$\begin{aligned} \frac{\partial d_i}{\partial x} &= 0, & x = 0, l, \\ \frac{\partial d_i}{\partial y} &= 0, & y = 0, h, \end{aligned} \quad (12.6.12)$$

$i = 1, 2$ .

By the method of separation of variables, or otherwise, we find that

$$k_1 = \frac{m\pi}{l}, \quad k_2 = \frac{n\pi}{h}, \quad (12.6.13)$$

for integers  $m, n = 1, 2, \dots$ .

Now proceed precisely as before to see that if we define

$$K^2 = k_1^2 + k_2^2, \quad (12.6.14)$$

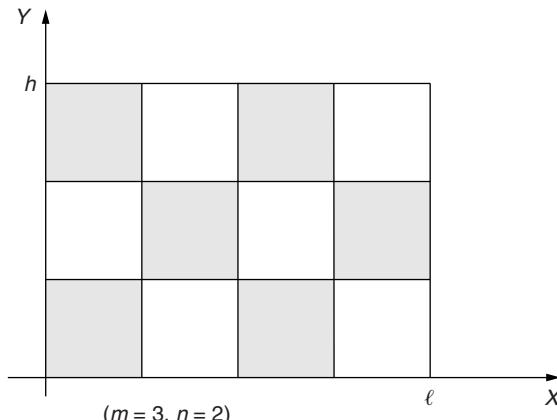
then we again arrive at equations (12.5.16) and (12.5.22) with  $k^2$  simply replaced by  $K^2$ . The critical wave number  $K_c^2$  is given by

$$K_c^2 = \frac{m^2\pi^2}{l^2} + \frac{n^2\pi^2}{h^2} = \frac{1}{2} \left( \frac{a_{11}}{D_1} + \frac{a_{22}}{D_2} \right) \quad (12.6.15)$$

and modes characterised by  $m$  and  $n$  exist if they satisfy the inequalities

$$\begin{aligned} \frac{(D_1 a_{22} + D_2 a_{11}) - \sqrt{(D_1 a_{22} + D_2 a_{11})^2 - 4D_1 D_2 A}}{2D_1 D_2} &< \frac{m^2\pi^2}{l^2} + \frac{n^2\pi^2}{h^2} \\ &< \frac{(D_1 a_{22} + D_2 a_{11}) + \sqrt{(D_1 a_{22} + D_2 a_{11})^2 - 4D_1 D_2 A}}{2D_1 D_2}. \end{aligned} \quad (12.6.16)$$

To illustrate possible modes, suppose the domain size is sufficiently large so that (12.6.16) holds for  $m = 3, n = 2$ . Then the pattern shown in Figure 12.6.3 is possible where the shaded areas indicate regions in which the morphogen concentration is above the steady state.



**FIGURE 12.6.3:** Morphogen pattern for  $m = 3, n = 2$ .

The fundamental assumption of pattern formation via Turing diffusion driven instabilities is that the linearly unstable modes that grow exponentially in time will eventually be bounded by the nonlinear kinetic terms in (12.5.4). This is indeed the case. To prove mathematically that this happens one has to show that in the positive quadrant of morphogen space there is a

compact region about the uniform steady state  $(c_{1,0}, c_{2,0})$  to which  $c_1$  and  $c_2$  are always confined. Such a region is called a confined set or a contracting set. The determination of such sets is often difficult and is beyond the scope of this book. The interested reader is encouraged to consult the literature cited in the notes at the end of this chapter.

---

## 12.7 Notes

### Diffusion through membranes

Most of the material here is based on the results of S.I. Rubinow and we recommend his book *Introduction to Mathematical Biology*, John Wiley & Sons, New York, 1975 for further developments and a host of biological applications. See also J.D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1993.

### Global behaviour of nerve impulse transmissions

For a fuller background to the material here, we refer to the survey article of S.P. Hastings, Some mathematical problems from neurobiology, *Am. Math. Monthly*, **82**, 881–895, 1975 and the detailed results in J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, Berlin, 1983.

### Global behaviour in chemical reactions

An excellent treatment of the Belousov–Zhabotinskii reaction and many other diffusion problems in biology are to be found in J.D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1993.

### Diffusion driven instability and pattern formation; Finite pattern forming domains

The interested reader is recommended to read the fundamental papers: A.M. Turing, The chemical basis of morphogenesis, *Phil. Trans. Roy. Soc. Lond.*, **B237**, 37–72, 1952; and L. Wolpert, Positional information and the spatial pattern of cellular differentiation, *J. Theor. Biol.*, **25**, 1–47, 1969. See also the treatment in J.D. Murray, *Mathematical Biology*, Springer-Verlag, Berlin, 1993.

## Exercises

- 12.1 A cell of solute concentration  $c$  and characteristic length  $\delta$  is placed in a large bathing solution of solute with fixed concentration  $c_0$ . If the region  $0 \leq x \leq \delta$  represents the cell while the regions  $x < 0$ ,  $x > \delta$  represent the cell exterior, show that the solution concentration of the cell is given by

$$c(x, t) = c_0 \left[ 1 - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin\left(\frac{(2n-1)\pi x}{\delta}\right) \exp[-(2n-1)^2 \pi^2 D t] / \delta^2 \right],$$

where  $D$  is the diffusion constant.

- 12.2 In one dimension, let there be a slab of solute of uniform concentration  $c_0$  and of thickness  $2a$ . At  $t = 0$

$$c(x, 0) = \begin{cases} c_0, & -a < x < a, \\ 0, & |x| > a. \end{cases}$$

Find the concentration  $c(x, t)$  for all  $x$  and  $t$ .

- 12.3 A stationary spherical cell of radius  $a$  is metabolising a nutrient that is at uniform concentration  $c_0$  in the surrounding medium initially. Assume that the cell instantaneously metabolises any nutrient molecules that enter it, so that the nutrient concentration at the cell wall is zero at all times. Solve the diffusion equation in spherical polar co-ordinates and find the nutrient concentration  $c = c(r, t)$  in the surrounding medium as a function of the radial position  $r$  and the time  $t$ .

- 12.4 By considering the energy function

$$E(t) = \int_t^{t+T} \left( \frac{du}{dx} \right)^2 dx,$$

show that the ordinary differential equation

$$\frac{du}{dt} = F(u),$$

where  $F(u) = \frac{df(u)}{du}$ , cannot have a periodic solution of period  $T$ .

- 12.5 Consider the reaction-diffusion system

$$\begin{aligned} \frac{\partial u}{\partial t} &= D_1 \frac{\partial^2 u}{\partial x^2} + f(u, v), \\ \frac{\partial v}{\partial t} &= D_2 \frac{\partial^2 v}{\partial x^2} + g(u, v), \end{aligned}$$

defined for  $x, t \in (0, 1) \times (0, \infty)$ , and where  $u, v$  satisfy the following initial and boundary conditions

$$\begin{aligned}\frac{\partial u}{\partial x} &= \frac{\partial v}{\partial x} = 0 \quad \text{for } x = 0, 1, t \geq 0, \\ u(x, 0) &= u_0(x), \quad v(x, 0) = v_0(x) \quad \text{for } 0 \leq x \leq 1, \\ \frac{du_0}{dx} &= \frac{dv_0}{dx} = 0 \quad \text{for } x = 0, 1.\end{aligned}$$

Define the energy function

$$E(t) = \frac{1}{2} \int_0^1 \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial v}{\partial x} \right)^2 \right] dx$$

and show that

$$\begin{aligned}\frac{dE}{dt} &\leq -d \int_0^1 \left[ \left( \frac{\partial^2 u}{\partial x^2} \right)^2 + \left( \frac{\partial^2 v}{\partial x^2} \right)^2 \right] dx + \frac{m}{2} \int_0^1 \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial v}{\partial x} \right)^2 \right] dx \\ &\leq (m - 2\pi^2 d)E,\end{aligned}$$

where  $d = \min(d_1, d_2)$  and

$$m = \max_{u,v} \left[ \left| \frac{\partial f}{\partial u} \right| + \left| \frac{\partial f}{\partial v} \right| + \left| \frac{\partial g}{\partial u} \right| + \left| \frac{\partial g}{\partial v} \right| \right].$$

Here,  $\max_{u,v}$  means the maximum over the solution values of  $u$  and  $v$ . Deduce that if  $m < 2\pi^2 d$  then  $E(t) \rightarrow 0$  as  $t \rightarrow \infty$ . Interpret this result.

- 12.6 From the general theory of diffusion driven instability and using the notation of Section 12.5 derive the inequalities

$$D_2 > D_1 \quad \text{and} \quad D_1 \tau_1 < D_2 \tau_2,$$

where

$$\tau_i = |a_{ii}|^{-1}, \quad i = 1, 2.$$

- 12.7 Consider the reaction-diffusion system

$$\begin{aligned}\frac{\partial u}{\partial t} &= \alpha uv - \beta u^2 + \delta^2 \nabla^2 u, \\ \frac{\partial v}{\partial t} &= v - uv + v^2 + \nabla^2 v,\end{aligned}$$

where  $u$  and  $v$  represent concentrations of morphogens and  $\alpha, \beta$  and  $\delta$  are positive constants. Find the nonzero equilibrium point and determine conditions on  $\alpha$  and  $\beta$  for a Turing instability to occur.

Calculate the values of  $\delta$  for which the Turing instability can take place.

12.8 A model capable of diffusion driven instability is

$$\begin{aligned}\frac{\partial u}{\partial t} &= \gamma(a - u + u^2v) + \frac{\partial^2 u}{\partial x^2}, \\ \frac{\partial v}{\partial t} &= \gamma(b - u^2v) + d\frac{\partial^2 v}{\partial x^2}, \quad (x, t) \in (-\infty, \infty) \times (0, \infty).\end{aligned}$$

Determine the homogeneous steady state and show that it is stable provided

$$(b - a) - (a + b)^3 < 0.$$

By linearising the system about the homogeneous steady state show that diffusion instability can occur if

$$H(k^2) \equiv dk^4 = k^2 \left[ \gamma(a + b)^2 - d\gamma \frac{(b - a)}{(a + b)} \right] + \gamma^2(a + b)^2 < 0.$$

12.9 Derive diffusion driven instability and pattern forming properties of the two dimensional reaction-diffusion system

$$\frac{\partial u}{\partial t} = a - bu - \frac{u^2}{v} + \nabla^2 u,$$

$$\frac{\partial v}{\partial t} = u^2 - v + d\nabla^2 v,$$

where  $a, b$  are positive parameters and  $d$  is a positive diffusion coefficient. The system is defined on the rectangular region  $0 \leq x \leq A, 0 \leq y \leq B$  and the morphogens  $u$  and  $v$  satisfy the boundary conditions

$$\begin{aligned}\frac{\partial u}{\partial x} &= 0, \quad \frac{\partial v}{\partial x} \quad \text{on } x = 0, A \\ \text{and} \quad u &= 0, \quad v = 0 \quad \text{on } y = 0, B.\end{aligned}$$

# Chapter 13

---

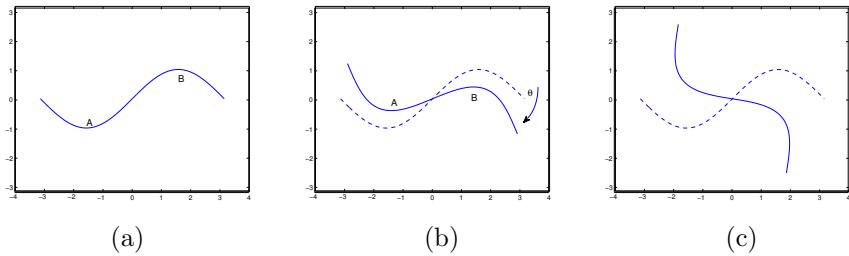
## Bifurcation and Chaos

---

### 13.1 Bifurcation

In Chapter 5, we saw that equilibrium points, or fixed points, are important features of many systems of differential equations. In addition, the local stability of any equilibrium points is significant because it is this property that determines whether solutions of the differential equations that start near to equilibrium will tend towards the equilibrium point, or away from it. Frequently, systems of differential equations contain one or more **parameters**, that are usually assumed to be constants. A natural question to ask, however, is how do the position and stability of an equilibrium change as one of the parameters of the system is varied. Such a parameter will be called a **control parameter** to distinguish it from any parameters which remain fixed.

As a physical example, consider a ball that is free to roll along the frictionless track shown in Figure 13.1.1(a). The motion of the ball can be described by a system of differential equations, but without worrying about the mathematical details, we can immediately say that the system has two equilibrium states, which occur when the ball is stationary at the positions marked A and B in the diagram. Furthermore, our intuition tells us that A is locally stable, whilst B is unstable – any slight perturbation to the ball’s position will cause it to roll away from B. Now suppose that we rotate the track through an angle  $\theta$  (Figure 13.1.1(b)). Here,  $\theta$  is the control parameter ( $\theta = 0$  in Figure 13.1.1(a)), and we are interested in what happens to the two equilibria as  $\theta$  changes. Again, our intuition tells us that, to begin with, the two equilibria move slightly – A is always at the track’s local minimum, B at the local maximum, and their respective stability properties do not change. Eventually, however, if we rotate the track enough, it will reach a point where there is no longer a local minimum or maximum, and hence the equilibria no longer exist (Figure 13.1.1(c)). This is called a **bifurcation**: a sudden, qualitative change in the system’s behaviour, such as a change in the stability or (as in this example) the number of equilibria. The value of the control parameter  $\theta$  at which the behaviour changes is called the **bifurcation point**. This chapter is concerned with identifying different classes of bifurcation and where they occur.



**FIGURE 13.1.1:** A bifurcation caused by tilting a track through angle  $\theta$ : (a) At  $\theta = 0$ , there is a stable equilibrium at A and an unstable equilibrium at B; (b) As  $\theta$  increases a small amount, the equilibria move but their stability is unchanged; (c) For sufficiently large  $\theta$ , both equilibria are destroyed.

### Example 13.1.1

Consider the nonlinear conservative system in which  $\mu \geq 0$  and

$$\dot{x} = y, \quad \dot{y} = \mu \sin x - x.$$

The equilibrium points are given by  $y = 0$  and solutions of

$$\mu \sin x - x = 0. \quad (13.1.1)$$

One solution of (13.1.1) is  $x = 0$ . To see if there are any other solutions we examine the derivative with respect to  $x$  of the left-hand side of (13.1.1), namely  $\mu \cos x - 1$ . The derivative is always negative if  $0 \leq \mu < 1$ . Therefore, for this range of  $\mu$ , the left-hand side of (13.1.1) decreases steadily as  $x$  increases from 0. Since it is zero when  $x = 0$  it cannot vanish subsequently and there is no other solution of (13.1.1) for  $0 \leq \mu < 1$ .

If  $\mu > 1$ , the derivative is positive at  $x = 0$ . As  $x$  increases the derivative decreases, passes through zero and then stays negative up to  $x = \pi$ ; for simplicity we limit  $x$  to  $(-\pi, \pi)$ . Therefore, as  $x$  increases, the left-hand side of (13.1.1) first increases from 0 to some positive maximum where  $\mu \cos x = 1$  and then decreases steadily to  $x = \pi$ . At  $x = \pi$  it is negative and so there is a zero  $x = x_1 \neq 0$  of (13.1.1) when  $\mu > 1$ . By the preceding argument  $\mu \cos x_1 < 1$ . Since (13.1.1) is unaffected by changing the sign of  $x$  there is also a solution  $x = -x_1$ .

Thus, we have found one equilibrium point for  $0 \leq \mu < 1$  and three for  $\mu > 1$ .

To examine the nature of the equilibrium we proceed as in Section 5.3. Near the equilibrium point  $(0,0)$  put  $x = \xi$ ,  $y = \eta$  with  $\xi$  and  $\eta$  small. Then

$$\dot{\xi} = \eta, \quad \dot{\eta} = (\mu - 1)\xi.$$

According to Section 5.4, the equation for the eigenvalues of the Jacobian matrix at the equilibrium is

$$\lambda^2 = \mu - 1.$$

If  $\mu < 1$ , the two values of  $\lambda$  are imaginary complex conjugates and the equilibrium point is a centre. When  $\mu > 1$ , the two values of  $\lambda$  are real but of opposite sign; the equilibrium point is a saddle-point.

When  $\mu > 1$  the equilibrium point  $(x_1, 0)$  can occur. Near it put  $x = x_1 + \xi$ ,  $y = \eta$  to obtain

$$\dot{\xi} = \eta, \quad \dot{\eta} = (\mu \cos x_1 - 1)\xi.$$

In this case

$$\lambda^2 = \mu \cos x_1 - 1.$$

Since  $\mu \cos x_1 - 1$  is negative, as shown above, the equilibrium point is a centre. The same is true for the equilibrium point  $(-x_1, 0)$ .  $\square$

Several features of this example should be noted. The position of the equilibrium point  $x = 0$  does not change as the control parameter  $\mu$  is altered but its character switches as  $\mu$  passes through 1. It could be regarded as going from a stable regime to an unstable one—termed *interchange of stabilities* by Poincaré. At the switch, extra equilibrium points are born; their position varies with the control parameter but not their character. The changing of the character of an equilibrium point and/or the creation of extra ones by alteration of a control parameter is known as **bifurcation**. The value of  $\mu$  where bifurcation occurs is called a **bifurcation point**.

An aid to keeping track of what is going on as the control parameter varies is the **bifurcation diagram**. This is a plot of the positions of the equilibrium points against the control parameter (see Figure 13.1.2). The letters *c* and *s* signify which curves represent centres and saddle-points, respectively. Adding information about  $\lambda$  would render the diagram too complicated; so sometimes there is a separate plot of  $\lambda$  against  $\mu$ . It can be somewhat awkward when  $\lambda$  is complex. To avoid a three-dimensional picture  $\lambda$  can be plotted on the complex  $\lambda$ -plane with values of  $\mu$  attached to the points.

On account of its appearance a bifurcation like that of Figure 13.1.2 is sometimes known as a **pitchfork bifurcation**.

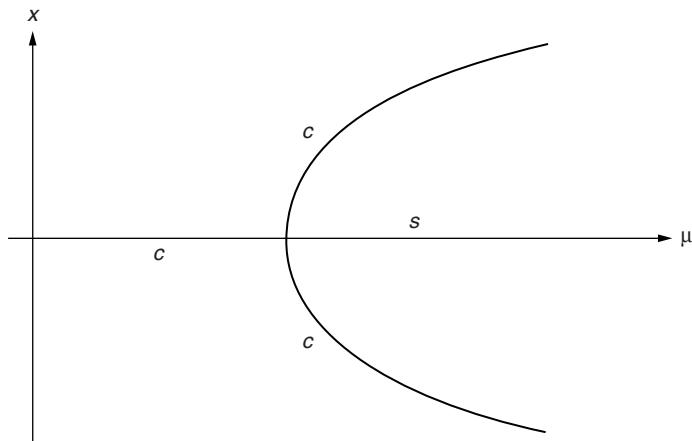
### **Example 13.1.2**

If the equation for  $\dot{y}$  in Example 13.1.1 is changed to

$$\dot{y} = \mu - x^2 - \mu^{1/4}y, \tag{13.1.2}$$

there are no equilibrium points for  $\mu < 0$ . For  $\mu > 0$  there are two equilibria:  $(\sqrt{\mu}, 0)$  and  $(-\sqrt{\mu}, 0)$ . Again, following the method of Section 5.4, the eigenvalues  $\lambda$  for the first of these equilibria satisfy

$$\lambda^2 + \mu^{1/4}\lambda + 2\mu^{1/2} = 0.$$



**FIGURE 13.1.2:** Bifurcation diagram for Example 13.1.1, showing a pitch-fork bifurcation. For values of  $\mu$  below the bifurcation point, there is a single equilibrium that is a centre (c); for values of  $\mu$  above the bifurcation point, there is a saddle-point (s) and two centres (c).

The roots of this equation are  $\lambda = (\mu^{1/4}/2)(-1 \pm \sqrt{7}i)$ , which are a complex conjugate pair with negative real part. Hence this equilibrium is a stable focus. The eigenvalues for the second equilibrium  $(x, y) = (-\sqrt{\mu}, 0)$  satisfy

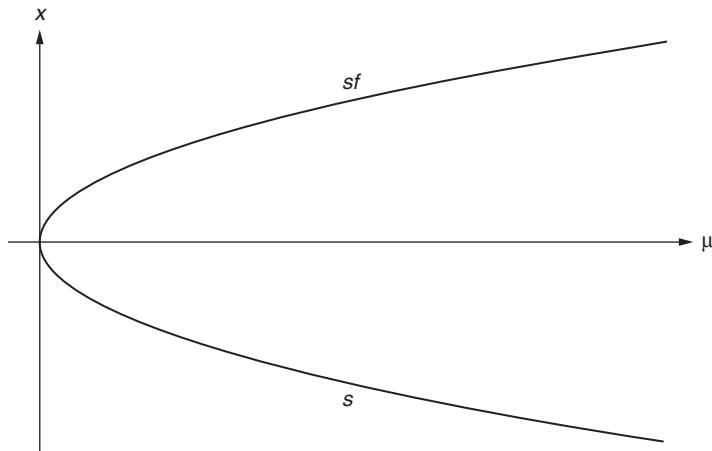
$$\lambda^2 + \mu^{1/4}\lambda - 2\mu^{1/2} = 0.$$

The roots of this equation are  $\lambda = (\mu^{1/4}/2)(-1 \pm 3)$  and hence this equilibrium is a saddle-point. The resulting bifurcation diagram is shown in Figure 13.1.3.  $\square$

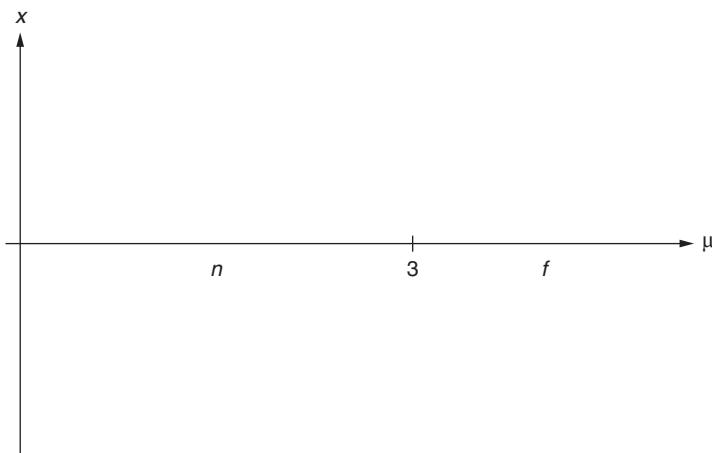
This class of bifurcation, where a stable and an unstable equilibrium collide and destroy one another at the bifurcation point, is called a **fold bifurcation** or **saddle-node bifurcation**. In fact, we have met this type of bifurcation previously, in the the tilting track example in Figure 13.1.1. For small  $\theta$ , there is a stable and an unstable equilibrium. As  $\theta$  increases, and the track is tilted, the equilibria move steadily closer together, until  $\theta$  reaches the bifurcation point, where the equilibria coalesce into a single equilibrium point. As  $\theta$  increases above the bifurcation point, this equilibrium disappears altogether.

Bifurcation is not restricted to the conservative systems that have been studied so far. For example, the system

$$\begin{aligned}\dot{x} &= 5x + (3 - \mu)y + 5x^3, \\ \dot{y} &= x + 5y + x^3\end{aligned}$$



**FIGURE 13.1.3:** Bifurcation diagram for Example 13.1.2, showing a fold bifurcation. For values of  $\mu$  below the bifurcation point, there are no equilibria; for values of  $\mu$  above the bifurcation point, there is a stable focus (sf) and a saddle (s).



**FIGURE 13.1.4:** Bifurcation from node ( $n$ ) to focus ( $f$ ).

with  $\mu \geq 0$  has an equilibrium point at  $(0, 0)$ . It is a node for  $\mu < 3$  and a focus for  $\mu > 3$ . The bifurcation diagram is displayed in Figure 13.1.4, the letters  $n$  and  $f$  indicating node and focus, respectively.

The reader should be warned that often the term bifurcation is employed much more strictly than has been used here. It is kept to denote points where the stability of the system changes. A diagram like Figure 13.1.4 would not be regarded as bifurcation in the strict sense because the stability is unaltered

although the behaviour of the system changes radically from nonoscillatory to oscillatory.

It should be stressed that the identification of the type of equilibrium point has been carried out by the approximation of Section 5.3. Thus, the equations are valid only near the equilibrium point. They represent what is occurring locally. How the local scene fits into the global picture can be very difficult to resolve.

Of course, there may be more than one control parameter as in

$$\dot{x} = y, \quad \dot{y} = \mu_1 x + \mu_2 x^2 + \mu_3 x^3 + x^5,$$

which has three control parameters  $\mu_1$ ,  $\mu_2$  and  $\mu_3$ . Discussion of bifurcation is generally extremely tricky even for the equilibrium point in which  $x = 0$ , let alone the determination of other equilibrium points. Pictorial representation is far from straightforward since three dimensions are occupied already by the control parameters. Accordingly, we shall concentrate on cases in which there is a single control parameter.

Often what we have called equilibrium points are known as *fixed points* in bifurcation theory, for reasons that will become apparent later. Sometimes it will be convenient to use one term in the rest of this chapter and sometimes the other.

---

## 13.2 Bifurcation of a limit cycle

Bifurcation is not confined to the types of equilibrium points in Section 5.4 although they are common in systems of two differential equations. Limit cycles can be involved even with two differential equations. As the number of differential equations increases to three and upwards the possible types of common bifurcations proliferate rapidly. Also discussion of them becomes increasingly complicated, going beyond the scope of this book. So we shall stay mostly with two differential equations.

An example where limit cycles are involved in bifurcation is provided by the differential equations

$$\dot{x} = -y + x(\mu - x^2 - y^2), \quad (13.2.1)$$

$$\dot{y} = x + y(\mu - x^2 - y^2). \quad (13.2.2)$$

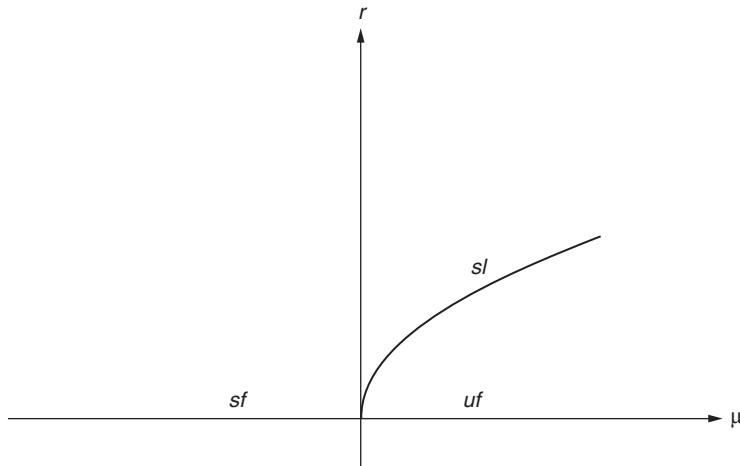
Make the transformation to polar coordinates by putting  $x = r \cos \phi$ ,  $y = r \sin \phi$ . Using the polar coordinate transformation formulae (5.6.12) gives

$$\dot{r} = r(\mu - r^2), \quad (13.2.3)$$

$$\dot{\phi} = 1. \quad (13.2.4)$$

Equation (13.2.4) can be solved immediately to give

$$\phi = \phi_0 + t \quad (13.2.5)$$



**FIGURE 13.2.1:** The Hopf bifurcation, from a stable focus (sf) to an unstable focus (uf) and stable limit cycle (sl).

where  $\phi_0$  is the value of  $\phi$  at  $t = 0$ . Thus, the radius vector (see Figure 5.6.2) sweeps steadily counterclockwise as  $t$  increases. The end of the radius vector describes the trajectory, which starts from  $(r_0, \phi_0)$  where  $r_0$  is the initial value of  $r$ .

When  $\mu < 0$ , (13.2.3) shows that  $\dot{r}$  is negative for any positive  $r$ . Therefore,  $r$  decreases steadily and the trajectory spirals round the origin in a counterclockwise sense and moves inwards as time progresses. Eventually, the trajectory ends up at the equilibrium point  $r = 0$ . The origin is a stable focus.

When  $\mu > 0$ , the right-hand side of (13.2.3) is positive for  $r^2 < \mu$ . Hence  $r$  increases and the spiral round the origin moves outwards. Now the origin is an unstable focus. The outward movement of  $r$  does not continue unabated because  $r = \sqrt{\mu}$  is another equilibrium point ( $r = -\sqrt{\mu}$  is not a possibility because  $r$  is not allowed to be negative). The trajectory with  $r_0 = \sqrt{\mu}$  is the circle  $r = \sqrt{\mu}$ , i.e., a limit cycle. Consequently, a trajectory started with  $r_0^2 < \mu$  spirals counterclockwise and outwards, steadily approaching the limit cycle. If  $r^2 > \mu$ ,  $\dot{r} < 0$  and a trajectory spirals inwards to the limit cycle. Accordingly, the limit cycle is stable. The bifurcation diagram is shown in Figure 13.2.1.

The creation of a limit cycle in this way by variation of a control parameter is known as **Hopf bifurcation**.

Although we have not needed it the solution of the differential equation (13.2.3) is available. It is

$$r^2 = \frac{\mu r_0^2}{r_0^2 + (\mu - r_0^2)e^{-2\mu t}} \quad (13.2.6)$$

where the positive square root of the right-hand side gives  $r$ . You should check that it leads to the same conclusions as have been obtained above.

Notice that the linear approximation of Section 5.3 leads to

$$\dot{r} = \mu r$$

instead of (13.2.3). This does predict correctly that a stable focus switches to an unstable focus as  $\mu$  passes through 0. However, it does not tell us what happens to the trajectories as they spiral away from the unstable focus. As far as the local behaviour is concerned they could be going off to infinity. It is only with the full equation (13.2.3) that we discover that they end up on a limit cycle eventually. This illustrates yet again the care that is necessary in extrapolating local behaviour to the global picture in nonlinear systems.

---

### 13.3 Discrete bifurcation and period-doubling

Often observations on a system have to be taken at intervals that are too wide apart for it to be sensible to use a model based on differential equations. For example, if the crop of grain on a farm is measured just at the end of the year, the frequency would be insufficient for a differential equation on our normal time scale. Nevertheless, phenomena similar to those described above may occur.

As a simple model, suppose that, at the end of the first year, the crop is  $C_1$  and the farmer sets aside a proportion of this as seed for the next year. Assume perfect germination and no loss of seed; the crop at the end of the second year  $C_2$  will be proportional to  $C_1$ , say

$$C_2 = \mu C_1.$$

If the farmer carries out the same process at the end of each year,  $C_3 = \mu C_2$  and generally  $C_{n+1} = \mu C_n$ . Obviously,  $C_{n+1} = \mu^n C_1$ , which shows that the crop increases steadily from year to year when  $\mu > 1$ . In contrast, when  $\mu < 1$ , the crop decays towards zero. There is a bifurcation as  $\mu$  passes through 1 from a stable (but undesirable) state to an unstable state.

Of course, the assumption that  $\mu$  does not vary from year to year is not likely to be valid in a practical situation. This might be overcome by using an average value but there is a more serious objection to the model. It takes no account of losses due to bad weather or parasites. Moreover, there could be a limit to the size of crop which the nutrients in the soil could support. The simplest way of allowing for such effects is to change the relation between  $C_n$  and  $C_{n+1}$  to

$$C_{n+1} = \mu C_n - \nu C_n^2 \quad (13.3.1)$$

with  $\mu$  and  $\nu$  positive constants. Making  $\nu$  positive ensures that the extra term in (13.3.1) represents losses. Furthermore, because it contains  $C_n^2$ , it will inhibit the growth if the crop becomes too large.

Now, make the substitution

$$C_n = \mu x_n / \nu$$

in (13.3.1). This gives

$$x_{n+1} = \mu x_n (1 - x_n). \quad (13.3.2)$$

Since negative crops are not permitted  $x_n$  must be restricted to lie between 0 and 1. Note that  $\mu$  cannot exceed 4. If  $\mu$  is greater than 4 it is possible to generate values of  $x_n$  outside the permitted range. For example,  $x_n = 1/2$  would make  $x_{n+1}$  too large.

Equation (13.3.2) constitutes an **iteration scheme** or **map** in which each value is determined from its predecessor once  $x_1$  has been fixed. Equation (13.3.2) in particular is called the logistic map. If  $x_n$  tends to a limit  $x_0$  as  $n \rightarrow \infty$  then  $x_0$  satisfies

$$x_0 = \mu x_0 (1 - x_0). \quad (13.3.3)$$

The **If** is in bold because we do not yet know whether or not  $x_n$  tends to a limit.

The solutions of (13.3.3) are known as the **fixed points** of the iteration scheme. They are

$$x_0 = 0, \quad x_0 = 1 - 1/\mu. \quad (13.3.4)$$

Since  $x_0$  is forced to lie between 0 and 1 the second possibility does not exist unless  $\mu > 1$ . Consequently, there are two fixed points when  $\mu > 1$  but only one when  $\mu < 1$ . Note that, when  $x_n \rightarrow 1 - 1/\mu$ ,  $C_n \rightarrow (\mu - 1)/\nu$ .

Observe that, if  $x_1 = 1$ ,  $x_2 = 0$  and all subsequent  $x_n$  are zero. So the iteration goes straight to the fixed point  $x_0 = 0$  whenever  $x_1 = 1$ .

To find out what happens near a fixed point, there is a simple test for the stability of fixed points. In general, iteration schemes such as (13.3.2) can be written in the form

$$x_{n+1} = f(x_n), \quad (13.3.5)$$

for some function  $f$ . We now take equation (13.3.5) and carry out a Taylor expansion of  $f(x_n)$  about the fixed point  $x_0$ :

$$x_{n+1} = f(x_0) + f'(x_0)(x_n - x_0) + O((x_n - x_0)^2),$$

where  $f'(x_0)$  is the derivative of  $f(x)$  evaluated at  $x = x_0$ . Noting that, by definition, a fixed point  $x_0$  must satisfy  $x_0 = f(x_0)$ , and writing  $x_n = x_0 + \epsilon_n$ , we get

$$\epsilon_{n+1} = f'(x_0)\epsilon_n + O(\epsilon_n^2). \quad (13.3.6)$$

Here,  $f'(x_0)$  is a constant and  $\epsilon_n$  is the deviation of  $x_n$  from the fixed point  $x_0$ . Hence if

$$|f'(x_0)| < 1, \quad (13.3.7)$$

**TABLE 13.3.1:**  
Iteration for  $\mu = 3.1$ .

$n$	$x_n$
1	0.250
2	0.581
3	0.755
4	0.574
5	0.758
6	0.569
7	0.760
8	0.565
9	0.762
10	0.562
11	0.763
12	0.561
13	0.764
14	0.560
15	0.764
16	0.559
17	0.764
18	0.559
19	0.764
20	0.558

then, provided  $\epsilon_n$  is small so that the terms of order  $\epsilon_n^2$  and higher can be neglected, (13.3.6) implies that  $|\epsilon_{n+1}| < |\epsilon_n|$ , so the deviation decreases steadily and tends to zero as  $n \rightarrow \infty$ . In this case, the fixed point can be regarded as stable. In contrast, if  $|f'(x_0)| > 1$ , then the deviation increases with each iteration, and the fixed point is unstable.

Now let us apply this test to the iteration scheme (13.3.2). Here the function  $f$  is given by  $f(x) = \mu x(1 - x)$ , and differentiating gives  $f'(x) = \mu(1 - 2x)$ . We now simply substitute in the fixed points  $x_0$  found in (13.3.4) to get

$$f'(0) = \mu, \quad f'(1 - 1/\mu) = 2 - \mu.$$

Hence the stability of the fixed points is determined by the value of the parameter  $\mu$ . If  $\mu < 1$ , then  $x_0 = 0$  is stable and  $x_0 = 1 - 1/\mu$  is unstable. If  $1 < \mu < 3$  then  $x_0 = 0$  is unstable and  $x_0 = 1 - 1/\mu$  is stable. Finally, if  $\mu > 3$ , then both fixed points are unstable. There are therefore bifurcations at  $\mu = 1$  and  $\mu = 3$ .

That leaves us with the puzzle of what happens when  $\mu > 3$  and neither fixed point is stable. The possibility of  $x_n$  going off to infinity is prevented by  $x_n$  being confined to  $(0, 1)$ , so something new must be taking place. A clue is offered by a numerical calculation. Table 13.3.1 shows the first 20 values of  $x_n$  when  $x_1 = 0.250$  and  $\mu = 3.1$ . It is clear that alternative entries are pretty

much the same from  $x_{10}$  onwards. In other words, the iteration is oscillating between two values without settling down to one or the other. The bifurcation at  $\mu = 3$  is said to be a **period-doubling bifurcation**, because the period of the iteration has doubled, in this case from period 1 (i.e., a fixed point) to period 2.

What it means, in terms of our original model, is that the crop is high one year, low the next, and so on.

Let us see if analysis supports the numerical evidence. Applying (13.3.5) to  $x_{n+2}$  gives

$$x_{n+2} = f(x_{n+1}) = f(f(x_n)). \quad (13.3.8)$$

If the system arrives at a state, as suggested by Table 13.3.1, where the same value  $x_0$  repeats every second iteration, then (13.3.8) implies that

$$x_0 = f(f(x_0)). \quad (13.3.9)$$

After insertion of  $f$  from (13.3.2), (13.3.9) can be rearranged as

$$x_0 \left( 1 - x_0 - \frac{1}{\mu} \right) \left\{ (1 - x_0)^2 + \left( \frac{1}{\mu} - 1 \right) (1 - x_0) + \frac{1}{\mu^2} \right\} = 0. \quad (13.3.10)$$

That two of the solutions of (13.3.10) are the fixed points found already in (13.3.4) is not surprising because, for them, the value is repeated every iteration (rather than merely every second iteration), so  $x_{n+1} = x_n$ , and certainly  $x_{n+2} = x_n$ . The remaining two solutions of (13.3.10) are given by

$$2(1 - x_0) = 1 - \frac{1}{\mu} \pm \left\{ \left( 1 - \frac{3}{\mu} \right) \left( 1 + \frac{1}{\mu} \right) \right\}^{1/2}. \quad (13.3.11)$$

For  $\mu = 3.1$ , the two values of  $x_0$  in (13.3.11) are 0.5580 and 0.7646 to four significant figures. These values agree well with those at the end of Table 13.3.1.

To check whether the period-2 solution is stable (i.e., whether the iteration will converge to this pair of alternating values), a linear analysis of (13.3.8) is performed, analogous to the linear analysis of (13.3.5) to test for stability of a fixed point. In fact, we note from (13.3.9) that the period-2 solution is nothing more than a fixed point of the function  $g$ , defined by  $g(x) = f(f(x))$ . Hence, we can simply apply the fixed point stability test (13.3.7) to  $g$  by calculating  $g'(x_0)$ . By chain rule,

$$g'(x_0) = f'(x_0)f'(f(x_0)). \quad (13.3.12)$$

Note that, since the two roots in (13.3.11) form a period-2 cycle, one root gives  $x_0$  and the other root gives  $f(x_0)$  (this is easily checked). Substituting these values for  $x_0$  and  $f(x_0)$  into (13.3.12) gives

$$g'(x_0) = 4 + 2\mu - \mu^2.$$

**TABLE 13.3.2:** Bifurcation points for period-doubling.

$\mu_1$	3	period-2
$\mu_2$	3.4495	period-4
$\mu_3$	3.5441	period-8
$\mu_4$	3.5644	period-16
$\mu_5$	3.5687	period-32
$\vdots$	$\vdots$	$\vdots$
$\mu_\infty$	3.569946	

The right-hand side is equal to 1 when  $\mu = 3$  and decreases as  $\mu$  increases until it reaches  $-1$  at  $\mu = 1 + \sqrt{6} = 3.4495$ . Hence the fixed point of  $g$ , which corresponds to the period-2 solution of  $f$ , is stable for  $3 < \mu < 1 + \sqrt{6}$  but not elsewhere. Accordingly, period-2 behaviour occurs in the range  $3 < \mu < 1 + \sqrt{6}$  but something else must happen for  $\mu > 1 + \sqrt{6}$ . There is another bifurcation to be considered.

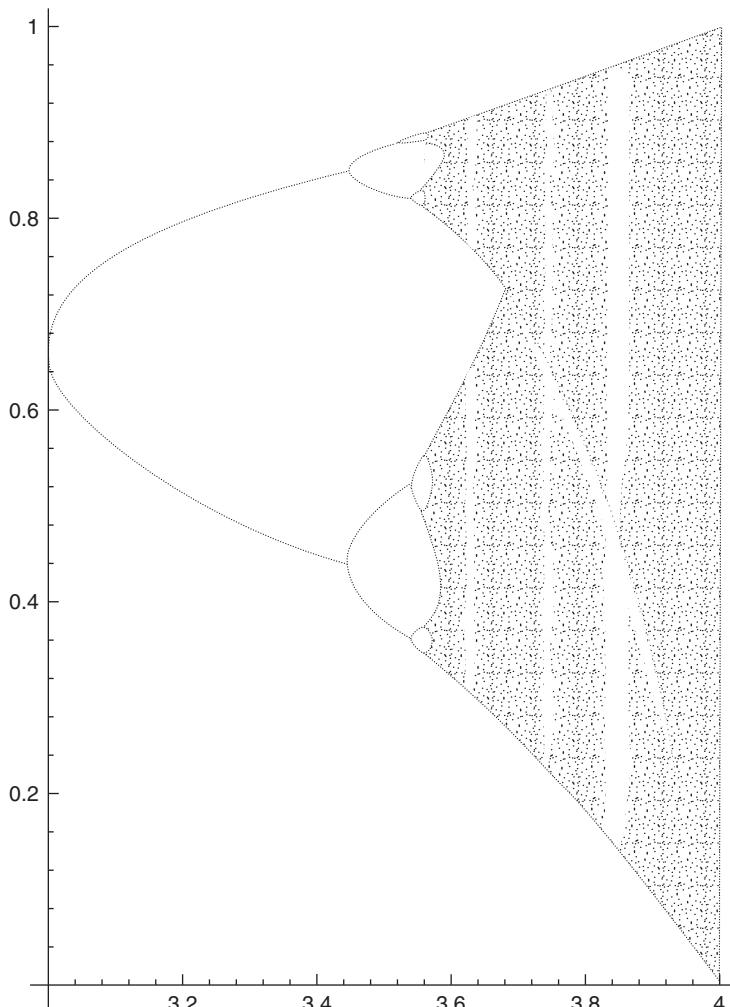
At the bifurcation both points of the period-2 solution split into two. Now, after enough iterations have been carried out, the iteration oscillates between four values. For example, when  $\mu = 3.5$ , the four values are (0.8750, 0.3828, 0.8269, 0.5009); in other words there is period-4 behaviour. Hence the bifurcation at  $\mu = 1 + \sqrt{6}$  is another period-doubling bifurcation, this time with the period doubling from 2 to 4.

However, period-4 behaviour lasts only up to  $\mu = 3.5441$ . Then there is another period-doubling bifurcation resulting in period-8 behaviour. The process continues with the intervals between bifurcation points getting shorter and shorter (see Table 13.3.2).

Finally, the process comes to a halt when  $\mu$  reaches the value  $\mu_\infty \approx 3.5699$ . An increase of  $\mu$  above  $\mu_\infty$  produces an entirely new regime which is best understood from the bifurcation diagram as  $\mu$  goes from 3, where period-doubling commences, to 4, its maximum permitted value (Figure 13.3.1). It can be seen that the iteration seemingly wanders all over the place without converging. Here we have an example of **chaos**.

It should be noted that the bifurcation diagram in Figure 13.3.1 is slightly different from those in earlier sections in that all stable fixed and periodic points are shown (rather than just fixed points), but no unstable points appear. The reason is that this diagram has been generated by a computer starting from some arbitrary point and finding the points of convergence (see Section 14.4 for examples of how to do this in MATLAB). Unless the initial point happens to coincide with an unstable fixed point its presence will never be detected and it will be absent from the bifurcation diagram.

While this simple model of crops shows extremely complex behaviour it is not likely that real crops exhibit the same phenomena because many other factors than those considered may be relevant. Nevertheless, the important



**FIGURE 13.3.1:** Bifurcation diagram for  $\mu$  between 3 and 4.

point is that a simple mathematical relation, involving only a quadratic, can display a rich set of features as the control parameter  $\mu$  varies. For some values of  $\mu$  the chaotic behaviour would entail wild variations in the crop from year to year, which might be interpreted as due to the vagaries of the weather or some other factor when they are, in fact, part and parcel of the model.

One of the difficulties in verifying experimentally whether a system conforms to the model is the closeness of successive bifurcation points. Their difference is so small that it could be hidden easily in experimental error. Notwithstanding, a set of experiments on the flour beetle indicates that populations can exhibit period-doubling and chaotic behaviour with variation of

a control parameter. In this case, the control parameter is the rate of cannibalism of pupae by adults.

The methods following (13.3.5) can be used to discuss other iteration schemes such as

$$x_{n+1} = \mu \sin(\pi x_n),$$

which also displays period-doubling and chaotic behaviour. For these and more complicated iterations, such as that for the flour beetle, it will not be possible to progress very far without extensive computation (see Section 14.4).

NOTE: The logistic growth equation

$$\frac{dp}{dt} = N_0 p - ap^2 \quad (13.3.13)$$

was discussed in Section 1.3. Its solution tends to  $N_0/a$  as  $t \rightarrow \infty$  so long as  $p$  is not initially zero. In solving a differential equation numerically the derivative is often approximated so that an estimate of the solution is obtained at intervals of time, say  $t = 0, \tau, 2\tau, \dots$ . For example, the fact that

$$\frac{dp}{dt} = \lim_{h \rightarrow 0} \{p(t+h) - p(t)\}/h$$

suggests the approximation

$$\frac{dp}{dt} = \{p((n+1)\tau) - p(n\tau)\}/\tau.$$

Writing  $p(nr) = p_n$ , we have the approximation

$$p_{n+1} = (N_0\tau + 1)p_n - a\tau p_n^2$$

for (13.3.13). This has the same form as (13.3.1) with  $\mu = N_0\tau + 1$ . Since  $C_n \rightarrow (\mu - 1)/\nu$  as  $n \rightarrow \infty$  when  $1 < \mu < 3$  it follows that  $p_n \rightarrow N_0/a$  provided that  $0 < N_0\tau < 2$ . Once  $N_0\tau$  exceeds 2 the iteration does not tend to the solution of (13.3.13) but can exhibit period doubling and chaos, an illustration of the importance of choosing the time interval correctly in the numerical approximation. Of course,  $p_n$  tending to the right limit when  $0 < N_0\tau < 2$  does not mean that intermediate values of the iteration are good approximations to the solution of (13.3.13) without further investigation.

## 13.4 Chaos

The chaotic behaviour in Figure 13.3.1 looks as though it is the result of a random process. Nevertheless, the wild wandering is entirely deterministic; each point is determined directly from its predecessor via the iteration scheme.

**TABLE 13.4.1:** Iteration for  $\mu = 3.99$ .

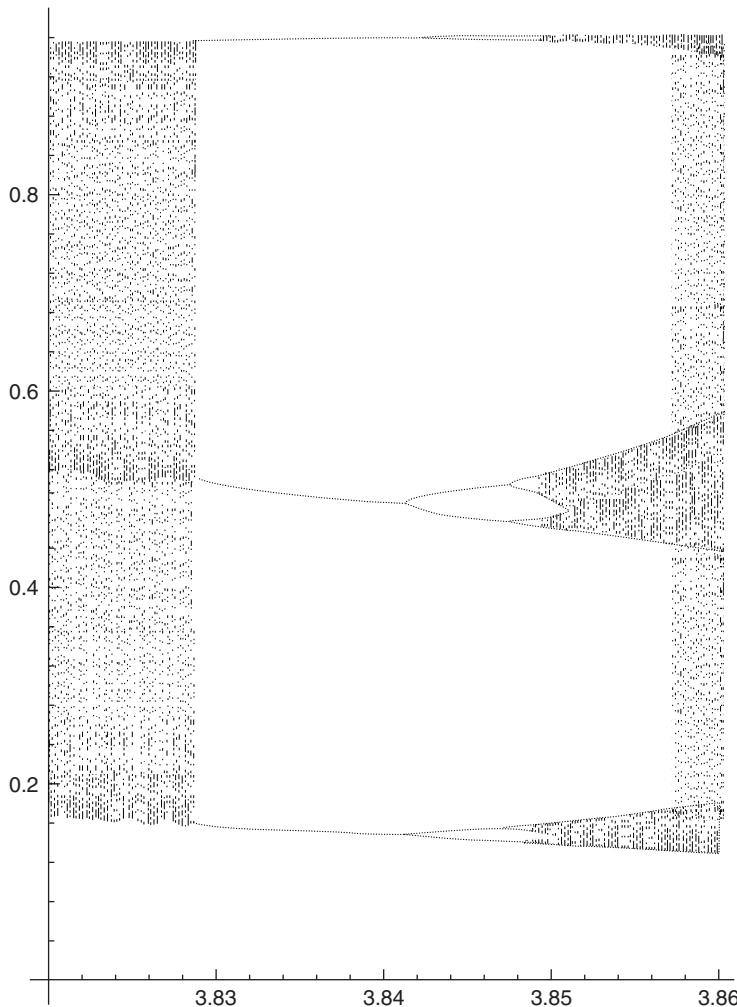
$n$	$x_n$	$x_n$	$x_n$
1	0.5000	0.5005	0.5010
2	0.9975	0.9975	0.9975
3	0.0100	0.0100	0.0100
4	0.0393	0.0393	0.0394
5	0.1507	0.1507	0.1509
6	0.5106	0.5107	0.5112
7	0.9971	0.9970	0.9970
8	0.0117	0.0118	0.0119
9	0.0462	0.0464	0.0471
10	0.1759	0.1767	0.1790
11	0.5784	0.5804	0.5864
12	0.9730	0.9717	0.9677
13	0.1050	0.1097	0.1246
14	0.3750	0.3900	0.4352
15	0.9351	0.9488	0.9807
16	0.2421	0.1937	0.0753
17	0.7321	0.6232	0.2780
18	0.7825	0.9369	0.8008
19	0.6791	0.2358	0.6364
20	0.8694	0.7189	0.9232

No considerations of probability are involved. One property that is not evident from the figure can be seen in Table 13.4.1, which shows the iteration for three different starting values when  $\mu = 3.99$ .

The way that successive values swing about is obvious but there is another feature. The initial values are close together and the first few iterates tally. But soon the iterates in the three columns begin to deviate and after about 16 iterations they bear little or no resemblance to one another. A *small change in the initial value can produce a very large difference in the iterates*. This property, often referred to as **sensitivity to initial conditions**, is one of the hallmarks of chaos.

It must not be supposed that the onset of chaos precludes the possibility of fixed points and bifurcation. An expanded version of the bifurcation diagram with  $\mu$  lying between 3.82 and 3.86 is displayed in Figure 13.4.1. For example, when  $\mu = 3.835$ , there is a stable period-3 solution, consisting of the three points (0.95863, 0.15207, 0.49451). Not only is there a gap between two regions of chaos with period-3 behaviour, but there is also bifurcation with period-doubling (with the period following the sequence 3, 6, 12, ...). Similar ‘periodic windows’ in an otherwise chaotic picture have been found to occur with other iteration schemes.

Another important feature concerns the values of  $\mu$  at bifurcations before



**FIGURE 13.4.1:** Bifurcation diagram for  $\mu$  between 3.82 and 3.86.

the onset of chaos (see Table 13.3.2). Define

$$\delta_n = \frac{\mu_{n+1} - \mu_n}{\mu_{n+2} - \mu_{n+1}}. \quad (13.4.1)$$

A numerical investigation reveals that, as  $n \rightarrow \infty$ ,  $\delta_n \rightarrow \delta = 4.6692016$ ;  $\delta$  is known as *Feigenbaum's number*.

Feigenbaum's number can be useful in two ways. If  $\mu_1$ ,  $\mu_2$  and  $\mu_3$  are known, (13.4.1) suggests that

$$\mu_4 = \mu_3 + (\mu_3 - \mu_2)/\delta. \quad (13.4.2)$$

There is, naturally, no guarantee that  $\mu_4$  exists for an arbitrary iteration scheme but, if it does, (13.4.2) offers a clue as to where to search for the bifurcation point.

More generally, if  $\delta_n$  can be replaced by  $\delta$  in (13.4.1) for  $n \geq N$ ,

$$\mu_{n+2} - \mu_{n+1}(1 + 1/\delta) + \mu_n/\delta = 0. \quad (13.4.3)$$

Try  $\mu_n = a^n$  and then

$$a^2 - a(1 + 1/\delta) + 1/\delta = 0,$$

whence  $a = 1$  or  $1/\delta$ . Hence the general solution of (13.4.3) is

$$\mu_n = A + B/\delta^n.$$

The constants  $A$  and  $B$  can be fixed by putting  $n = N$  and  $n = N + 1$ . The result is that

$$\mu_n = \mu_{N+1} + \frac{(\mu_{N+1} - \mu_N)(1 - \delta^{N+1-n})}{\delta - 1} \quad (13.4.4)$$

for  $n \geq N$ .

As a test of (13.4.4) take  $N = 1$  with  $\mu_1$  and  $\mu_2$  as in Table 13.3.2. Then (13.4.4) predicts that  $\mu_3$  will be 3.5458,  $\mu_4 = 3.5664$  and  $\mu_5 = 3.5708$ . While these differ a little from the values in Table 13.3.2 they are sufficiently close to be a good guide in a search for bifurcation points where period-doubling occurs.

Since  $\delta > 1$  the prediction of (13.4.4) for the onset of chaos is

$$\mu_\infty = \mu_{N+1} + \frac{\mu_{N+1} - \mu_N}{\delta - 1}. \quad (13.4.5)$$

When  $N = 1$ , (13.4.5) predicts  $\mu_\infty = 3.5720$  which is surprisingly close to the correct value.

It is remarkable that such good predictions are obtained from  $N = 1$  and one would expect to do much better with larger values of  $N$ . For other iteration schemes exhibiting period-doubling leading to chaos, one would hope that (13.4.4) and (13.4.5) would prove equally valuable. Unfortunately, it is true only for functions with a single maximum near where a good approximation is a quadratic like (13.3.2). For functions of the type

$$f(x) = 1 - a|x|^m \quad (13.4.6)$$

there is period-doubling and  $\delta$  exists for each  $m$  but the value of  $\delta$  changes with  $m$ . Some values are

$m =$	2	4	6	8
$\delta =$	5.12	9.32	13.37	17.40

where  $\delta$  has been rounded to two decimal places. The more general iteration

$$\begin{aligned}x_{n+1} &= f(x_n, y_n), \\y_{n+1} &= g(x_n, y_n)\end{aligned}$$

has been found to possess bifurcations with period-doubling when

$$\frac{\partial f}{\partial x} \frac{\partial g}{\partial y} - \frac{\partial f}{\partial y} \frac{\partial g}{\partial x} = 1$$

but now  $\delta = 0.72$ . Period-doubling between chaotic regions as in Figure 13.4.1 is governed by yet other values of  $\delta$ .

To sum up, the basic features that have been observed in the iteration are:

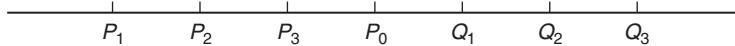
- (i) There are successive bifurcations at which period-doubling occurs, the values of  $\mu$  becoming ever more compressed.
- (ii) Chaotic regions appear for  $\mu > \mu_\infty$ .
- (iii) In a chaotic region the iteration is highly sensitive to changes in initial value.
- (iv) Period-3, period-5, ... and multiples thereof can appear between chaotic regions.
- (v) Positions of bifurcation points can be estimated if an appropriate Feigenbaum number  $\delta$  is available.

As already explained, chaos is not a random process. What distinguishes a chaotic solution to a deterministic equation is that it is highly sensitive to the initial conditions and wanders all over the place in a manner that seems to be random. Solutions that alter sharply after small changes in initial conditions but do not exhibit random behaviour or simply tend to infinity are not classed as chaotic.

## 13.5 Stability of limit cycles

In the discussion of the Hopf bifurcation in Section 13.2 it was shown that the limit cycle was stable. There is another method of demonstrating this property; it uses the **Poincaré section** of a limit cycle.

The Poincaré section is formed by the following procedure. Draw a line segment in the phase plane that intersects the limit cycle. Any convenient line can be chosen so long as it intersects the limit cycle. Let  $P_0$  be a point of intersection with the limit cycle. Start a trajectory from a point near to, but not on, the limit cycle. As we follow the trajectory it will cross the line of the



**FIGURE 13.5.1:** The Poincaré section.

Poincaré section at a point, say  $P_1$  (Figure 13.5.1). The next time around, the crossing will be at  $P_2$ , say, and the next one at  $P_3$ . If the sequence of points  $P_1, P_2, P_3, \dots$  approaches  $P_0$  as in Figure 13.5.1 the trajectory is tending to the limit cycle. If the sequence leaves  $P_0$  as for  $Q_1, Q_2, Q_3, \dots$  in Figure 13.5.1 the trajectory is going away from the limit cycle. Thus, a stable limit cycle can be identified by the sequences on both sides of  $P_0$  tending to  $P_0$ . A sequence always stays on one side of  $P_0$  because distinct trajectories in the phase plane cannot intersect in finite time. Two trajectories are said to be distinct when the starting point of one does not lie on the other.

### Example 13.5.1

The discussion of the Hopf bifurcation in Section 13.2 was based on the polar coordinates  $r$  and  $\phi$ . Explicit formulae for them were obtained, namely

$$\begin{aligned} r^2 &= \frac{\mu r_0^2}{r_0^2 + (\mu - r_0^2)e^{-2\mu t}}, \\ \phi &= \phi_0 + t \end{aligned}$$

(see (13.2.5) and (13.2.6)) for a trajectory starting at  $(r_0, \phi_0)$ .

Take the line  $\phi = \phi_0$  for the Poincaré section. Then  $r$  will give the distance of the point of intersection from the origin. The first return of the trajectory to the Poincaré line occurs when  $t = 2\pi$  and so

$$r_1^2 = \frac{\mu r_0^2}{r_0^2 + (\mu - r_0^2)e^{-4\pi\mu}}.$$

At the  $n$ th return  $t = 2n\pi$  and

$$r_n^2 = \frac{\mu r_0^2}{r_0^2 + (\mu - r_0^2)e^{-4\pi\mu n}}.$$

We deduce that

$$r_{n+1}^2 = \frac{\mu r_n^2 e^{4\pi\mu}}{\mu + r_n^2(e^{4\pi\mu} - 1)}, \quad (13.5.1)$$

which is an iteration scheme to determine  $r_n$  from its initial value. It is easy to check that, when  $\mu > 0$ , the iteration possesses fixed points at  $r = 0$  and  $r = \sqrt{\mu}$ , the second corresponding to the limit cycle, in harmony with what is known already.

Equation(13.5.1) can be rewritten as

$$r_{n+1}^2 = \mu + \frac{\mu(r_n^2 - \mu)}{\mu + r_n^2(e^{4\pi\mu} - 1)} \quad (13.5.2)$$

$$= r_n^2 + \frac{(\mu - r_n^2)r_n^2(e^{4\pi\mu} - 1)}{\mu + r_n^2(e^{4\pi\mu} - 1)}. \quad (13.5.3)$$

These equations are helpful in noting certain features. From (13.5.2) can be inferred  $r_{n+1}^2 \geq \mu$  according as  $r_n^2 \geq \mu$ . Thus the sequence of points on the Poincaré section does stay on one side of  $P_0(r^2 = \mu)$  in conformity with the general theory. Also (13.5.3) shows that  $r_{n+1}^2 \geq r_n^2$  according as  $\mu \geq r_n^2$ . Hence the sequence is steadily increasing (or decreasing) and is bounded above (or below). In either case, the sequence converges and the only possibility is convergence to  $P_0$ . Consequently, the limit cycle is stable, in agreement with the conclusion of Section 13.2.  $\square$

The Poincaré section has two characteristics. Firstly, it transforms a two-dimensional problem into one on a line, thereby saving a dimension. Secondly, instead of working through differential equations with continuous variations in time, it employs values at discrete time intervals, each interval being roughly the time to go round the limit cycle once.

Since a trajectory that starts at  $P_n$  is completely specified, so is  $P_{n+1}$ , i.e.,  $P_{n+1}$  can be inferred from  $P_n$ . To put this on a mathematical footing select any convenient reference point  $P_r$  on the Poincaré line and identify a point on the line by its distance  $s$  from  $P_r$ , with sign depending on which side of  $P_r$  it lies. If  $P_n$  is at  $s_n$  the relation between  $P_{n+1}$  and  $P_n$  can be expressed as

$$s_{n+1} = f(s_n). \quad (13.5.4)$$

The point  $P_0$  is a fixed point of (13.5.4) because the trajectory that starts at  $P_0$  is a limit cycle and, therefore, must cross the Poincaré line exactly at  $P_0$  on every circuit. Accordingly, if  $P_0$  is designated by  $s_0$ ,

$$s_0 = f(s_0). \quad (13.5.5)$$

From (13.5.4) and (13.5.5)

$$s_{n+1} - s_0 = f(s_n) - f(s_0).$$

For  $s_n$  close to  $s_0$ , a Taylor expansion to the first order gives

$$s_{n+1} - s_0 = (s_n - s_0) \left[ \frac{df(s)}{ds} \right]_{s=s_0}.$$

Writing

$$M = \left[ \frac{df(s)}{ds} \right]_{s=s_0} \quad (13.5.6)$$

we have

$$s_{n+1} - s_0 = M^n(s_1 - s_0).$$

It follows that, if  $|M| < 1$ ,  $s_{n+1} \rightarrow s_0$ , i.e., the points on the Poincaré line tend to  $P_0$  and the limit cycle is stable. In contrast, when  $|M| > 1$ ,  $s_{n+1}$  separates further and further from  $s_0$  and the limit cycle is unstable. The case  $|M| = 1$  needs further consideration. Actually,  $M$  cannot be negative. If  $M$  were negative  $s_{n+1}$  and  $s_n$  would be on opposite sides of  $P_0$ . For that to occur the trajectory would have to cross the limit cycle and that is not allowed.

For the scheme of (13.5.3) you can verify that  $M = e^{-4\pi\mu} < 1$ , another confirmation that the limit cycle involved is stable.

The Poincaré section offers another method of determining the character of a limit cycle by using an iteration scheme. The main difficulty in its application is finding the function  $f$  in (13.5.4). This is equivalent to solving the original set of differential equations, a task which may be extremely difficult or even impossible in practice. Without explicit knowledge of  $f$  or some adequate approximation to it the applicability of the Poincaré section is restricted.

We have seen in Section 13.3 that iteration schemes can display chaotic behaviour. That raises the question: Could an iteration scheme for the Poincaré line generate chaos and, if so, would the associated trajectory in the phase plane be chaotic? The question cannot be answered without a decision on what constitutes a chaotic trajectory, which is a solution of differential equations. Confining our attention to systems in which the trajectories stay in a bounded region (so that there is no possibility of the solution becoming infinite) we expect the conditions for chaos to be

- (a) distinct trajectories do not intersect,
- (b) the trajectories are bounded,
- (c) trajectories, which are initially close, diverge widely and rapidly.

Not all of these conditions can be met by trajectories in the phase plane, showing that *chaos cannot arise for the phase plane or Poincaré line*.

The situation is quite different in *phase space* with three or more dimensions. Here the trajectories can intertwine without intersection and stay in a bounded region so that chaos is possible. A chaotic trajectory cannot return to its initial point; if it did, it would correspond to a periodic solution of the differential system. It might, however, return to near the initial point but then, on account of (c), its subsequent path would bear little or no resemblance to the path from the initial point. That chaos can occur in phase space is illustrated in the next section.

### 13.6 The Poincaré plane

Rather than attempt a treatment of general phase space we shall consider only three dimensions. Then a first order system would take the form

$$\begin{aligned}\dot{x} &= f(x, y, z), \\ \dot{y} &= g(x, y, z), \\ \dot{z} &= h(x, y, z),\end{aligned}$$

where  $f$ ,  $g$  and  $h$  do not involve the time derivatives of  $x$ ,  $y$  and  $z$  or the time itself. Control parameters may be present also. Such a system is said to be **autonomous**.

When the functions involve the time as well, as in forced oscillations, the system is called **nonautonomous**. There is a trick for turning a nonautonomous system into the autonomous type in one higher dimension. Suppose

$$\begin{aligned}\dot{x} &= f(x, y, t), \\ \dot{y} &= g(x, y, t).\end{aligned}\tag{13.6.1}$$

Put  $z = t$  so that  $\dot{z} = 1$ . The system becomes

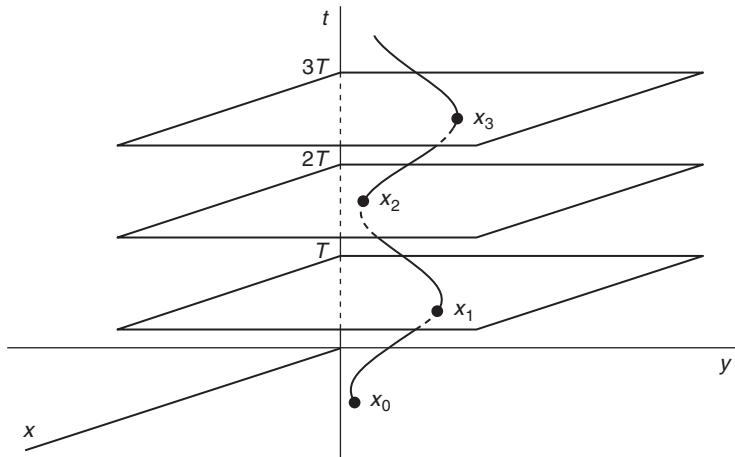
$$\begin{aligned}\dot{x} &= f(x, y, z), \\ \dot{y} &= g(x, y, z), \\ \dot{z} &= 1,\end{aligned}\tag{13.6.2}$$

which is autonomous but in three dimensions instead of the original two. This device enables the treatment of nonautonomous systems on the same footing as autonomous systems. The cost is the necessity to work with an extra dimension in the phase space. Furthermore, the new system has no fixed points because the right-hand side of the equation for  $z$  never vanishes. However, that difficulty can be overcome by generalising the Poincaré line of two dimensions to higher dimensions.

The simplest generalisation is when the functions in (13.6.1) are periodic, i.e., there is a constant  $T$  such that

$$f(x, y, t + T) = f(x, y, t), \quad g(x, y, t + T) = g(x, y, t)$$

for all  $x$ ,  $y$  and  $t$ . If, now, a trajectory of (13.6.2) is drawn in the three-dimensional space from  $t = 0$  it will intersect the planes  $t = T, 2T, \dots$ . Denote  $(x, y)$  by  $\mathbf{x}$  and let the points of intersection starting from  $\mathbf{x}_0$  be  $\mathbf{x}_1, \mathbf{x}_2, \dots$  (see Figure 13.6.1). Plot the points  $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots$  on the  $(x, y)$ -plane (Figure 13.6.2). The resulting diagram is known as the **Poincaré plane**. To put it another way, the Poincaré plane is constructed by plotting the points  $\mathbf{x}(0), \mathbf{x}(T), \mathbf{x}(2T), \dots$  of the trajectory on the  $(x, y)$ -plane.



**FIGURE 13.6.1:** A three-dimensional trajectory.

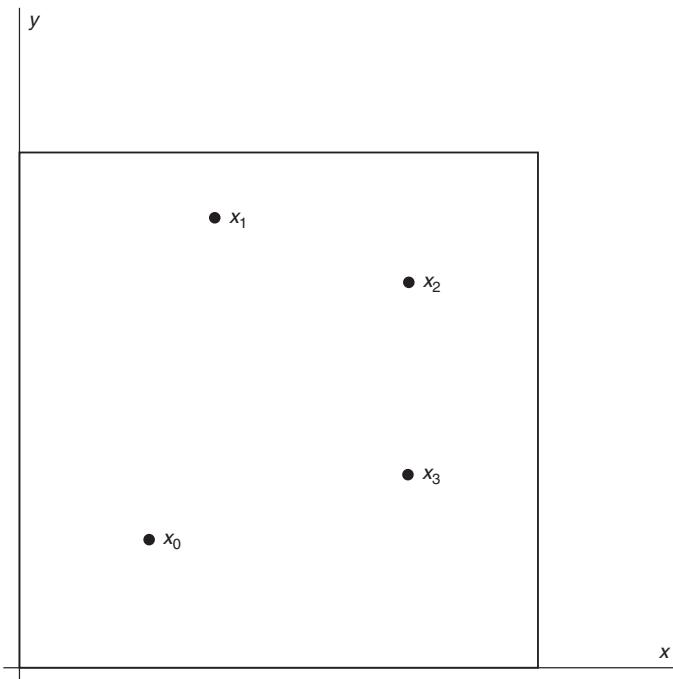
Of course, construction of the Poincaré plane is equivalent to solving the system of differential equations. So it does not simplify the determination of a solution. Fortunately, there are cases where a reliable approximation can be devised as will be seen in the next section. It is evident from Figures 13.6.1 and 13.6.2 that a complete trajectory will produce a scatter of points on the Poincaré plane. As a result it can be awkward to understand what is going on. Therefore, it is generally wise to delay plotting until any transients have died away so that the long-term behaviour is more easily visible. Once the transients are unimportant the picture on the Poincaré plane will consist of a single point if the system tends to periodic motion with period  $T$ . On the other hand, if the period is  $2T$ , the picture will show two points and, generally, for period  $nT$ , there will be  $n$  points. When  $T$  corresponds to the period of a periodic forcing term, two or more points on the Poincaré plane indicate that, in the long-term, the system will oscillate in a subharmonic of the periodic force.

On a trajectory abbreviate  $x(nT)$  and  $y(nT)$  to  $x_n$  and  $y_n$ , respectively. Since the point  $(x_n, y_n)$  determines  $(x_{n+1}, y_{n+1})$  there is a relation

$$\begin{aligned} x_{n+1} &= F(x_n, y_n), \\ y_{n+1} &= G(x_n, y_n), \end{aligned} \tag{13.6.3}$$

i.e., the points on the Poincaré plane satisfy an iteration scheme. The long-term behaviour is dictated by the fixed points of this scheme; they are the solutions of

$$\begin{aligned} x_0 &= F(x_0, y_0), \\ y_0 &= G(x_0, y_0). \end{aligned} \tag{13.6.4}$$



**FIGURE 13.6.2:** The Poincaré plane.

Every solution of (13.6.4) corresponds to a limit cycle in the three-dimensional phase space.

If  $(x_n, y_n)$  is close to  $(x_0, y_0)$  the subtraction of (13.6.4) from (13.6.3) followed by a Taylor expansion leads to

$$\begin{aligned} x_{n+1} - x_0 &= A(x_n - x_0) + B(y_n - y_0), \\ y_{n+1} - y_0 &= C(x_n - x_0) + D(y_n - y_0) \end{aligned} \quad (13.6.5)$$

where

$$A = \frac{\partial F(x, y)}{\partial x}, \quad B = \frac{\partial}{\partial y} F(x, y), \quad C = \frac{\partial}{\partial x} G(x, y), \quad D = \frac{\partial}{\partial y} G(x, y), \quad (13.6.6)$$

all the derivatives being evaluated at  $(x_0, y_0)$ .

To solve (13.6.5) try both  $x_n - x_0$  and  $y_n - y_0$  being proportional to  $\lambda^n$ . Then  $\lambda$  must satisfy

$$\lambda^2 - \lambda(A + D) + AD - BC = 0. \quad (13.6.7)$$

If the solutions of (13.6.7) are  $\lambda_1, \lambda_2$  there are constants  $A_1, A_2, B_1$  and  $B_2$  such that

$$x_n - x_0 = A_1 \lambda_1^n + A_2 \lambda_2^n, \quad y_n - y_0 = B_1 \lambda_1^n + B_2 \lambda_2^n$$

unless  $\lambda_1 = \lambda_2$  when  $\lambda_2^n$  is replaced by  $n\lambda_1^n$ . In any case, for  $(x_n, y_n)$  to tend to  $(x_0, y_0)$ , whatever the initial conditions it is necessary that  $|\lambda_1| < 1$  and  $|\lambda_2| < 1$ . If  $|\lambda_1| > 1$  and  $|\lambda_2| > 1$ ,  $(x_n, y_n)$  will tend to separate from  $(x_0, y_0)$ . When  $|\lambda_1| < 1$  and  $|\lambda_2| > 1$ ,  $(x_n, y_n)$  will approach or deviate from  $(x_0, y_0)$  according as  $A_2$  or  $B_2$  is or is not zero; this will depend on the initial conditions. Limit cycles that are approached from some nearby points and separated from others are often known as *saddle limit cycles*. In summary we have

- |   |                       |
|---|-----------------------|
| $ \lambda_1  < 1$ and $ \lambda_2  < 1$ | stable limit cycle,   |
| $ \lambda_1  > 1$ and $ \lambda_2  > 1$ | unstable limit cycle, |
| $ \lambda_1  < 1$ and $ \lambda_2  > 1$ | saddle limit cycle.   |

Obviously, if varying a control parameter causes  $|\lambda_1|$  or  $|\lambda_2|$  to pass through unity, a bifurcation occurs.

To illustrate these points we consider the motion of a particle that is given a constant velocity for an interval of time, the constant depending upon the position of the particle at the beginning of the interval. The resulting behaviour is amazingly complex.

### Example 13.6.1

For every positive integer  $n$ ,  $(x, y)$  satisfies

$$\begin{aligned} \text{in } n < t < n + 1/3, \quad \dot{x} &= 0, & \dot{y} &= 3(1 - ax^2), \\ \text{in } n + 1/3 < t < n + 2/3, \quad \dot{x} &= 3(b - 1)x(n + 1/3), & \dot{y} &= 0, \\ \text{in } n + 2/3 < t < n + 1, \quad \dot{x} &= 3y(n + 2/3) - 3x(n + 2/3), & \dot{y} &= -\dot{x}. \end{aligned}$$

The quantities  $a$  and  $b$  are nonnegative constants. Integration of the differential equations gives

$$\begin{aligned} x(n + 1/3) &= x(n), & y(n + 1/3) &= y(n) + 1 - ax(n)^2, \\ x(n + 2/3) &= bx(n + 1/3), & y(n + 2/3) &= y(n + 1/3), \\ x(n + 1) &= y(n + 2/3), & y(n + 1) &= x(n + 2/3). \end{aligned}$$

Combining these equations and writing  $x_n$  for  $x(n)$  we have the iteration scheme

$$\begin{aligned} x_{n+1} &= y_n + 1 - ax_n^2, \\ y_{n+1} &= bx_n. \end{aligned} \tag{13.6.8}$$

In the discussion of the iteration it will be assumed that  $0 < b < 1$  and  $a \geq 0$ . Also  $b$  will be kept fixed while  $a$  is regarded as the control parameter.

The values of  $x$  at fixed points satisfy

$$ax^2 + (1 - b)x - 1 = 0$$

with solutions

$$2ax_0^{(1)} = b - 1 + \{4a + (1 - b)^2\}^{1/2}, \quad (13.6.9)$$

$$2ax_0^{(2)} = b - 1 - \{4a + (1 - b)^2\}^{1/2}. \quad (13.6.10)$$

At  $x_0^{(1)}$ , (13.6.6) gives

$$A = -2ax_0^{(1)}, \quad B = 1, \quad C = b, \quad D = 0$$

so that (13.6.7) becomes

$$\lambda^2 + 2ax_0^{(1)}\lambda - b = 0. \quad (13.6.11)$$

Evidently  $\lambda_1$  and  $\lambda_2$  have opposite signs; take  $\lambda_1 > 0$  and  $\lambda_2 < 0$ . When  $a = 0$ ,  $ax_0^{(1)} = 0$  so that  $\lambda_1 = \sqrt{b}$ ,  $\lambda_2 = -\sqrt{b}$ . Since  $b < 1$  there is stability at  $a = 0$ . In

$$\lambda_1 - b/\lambda_1 = -2ax_0^{(1)} \quad (13.6.12)$$

the right-hand side decreases as  $a$  increases. Therefore  $\lambda_1$  decreases and hence so does  $\lambda_2$ . Consequently, stability continues until  $\lambda_2 = -1$  and  $\lambda_1 = b$ . This occurs, according to (13.6.12) and (13.6.9), when

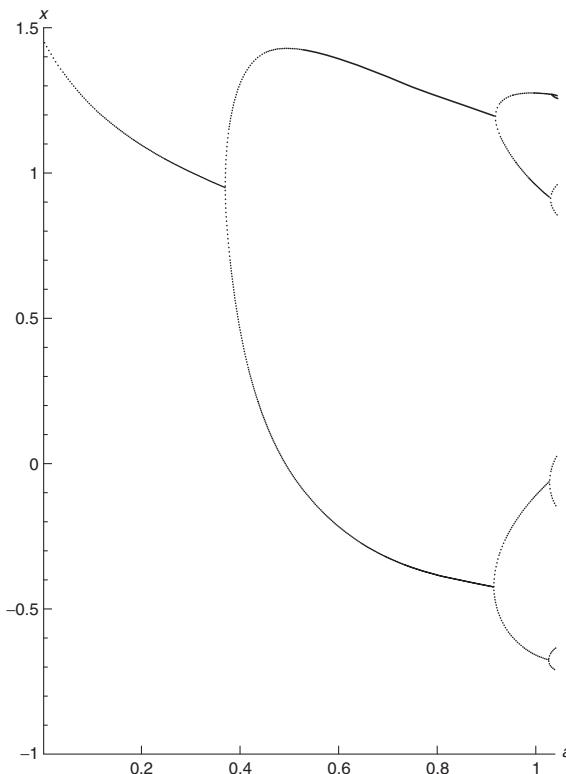
$$4a = 3(1 - b)^2. \quad (13.6.13)$$

Thus (13.6.13) specifies when bifurcation will take place.

For the fixed point  $x_0^{(2)}$  replace  $x_0^{(1)}$  by  $x_0^{(2)}$  in (13.6.11) and (13.6.12). At  $a = 0$ ,  $2ax_0^{(2)} = 2(b - 1)$  and then the left-hand side of (13.6.11) is negative when  $\lambda = 1$  so that  $\lambda_1 > 1$ . Since  $-ax_0^{(2)}$  increases with  $a$  so does  $\lambda_1$  on account of (13.6.12). Hence  $x_0^{(2)}$  is never stable; since  $|\lambda_2| < b$  it corresponds to a saddle limit cycle.

Hence for  $a$  less than the value in (13.6.13), the iteration has two fixed points, of which  $x_0^{(1)}$  is stable and  $x_0^{(2)}$  is not. For  $a > 3(1 - b)^2/4$ , period-doubling occurs. A bifurcation diagram for  $x$  when  $b = 0.3$  is shown in Figure 13.6.3 (fixed points that are not stable do not appear on such diagrams as explained in Section 13.3). Here  $a$  goes from 0 to 1.04 and it is clear that there are further bifurcations with accompanying period-doubling. Moreover, the distance between successive bifurcations grows steadily shorter just as in Section 13.3.

As  $a$  increases beyond 1.04 matters become much more complicated. Figure 13.6.4 shows the continuation of the bifurcation diagram as  $a$  goes from 1.04 to 1.06. More period-doubling takes place and regions of chaotic behaviour make their appearance. Beyond 1.06 the situation is even more complex but discussion of the details would take us too far afield.  $\square$

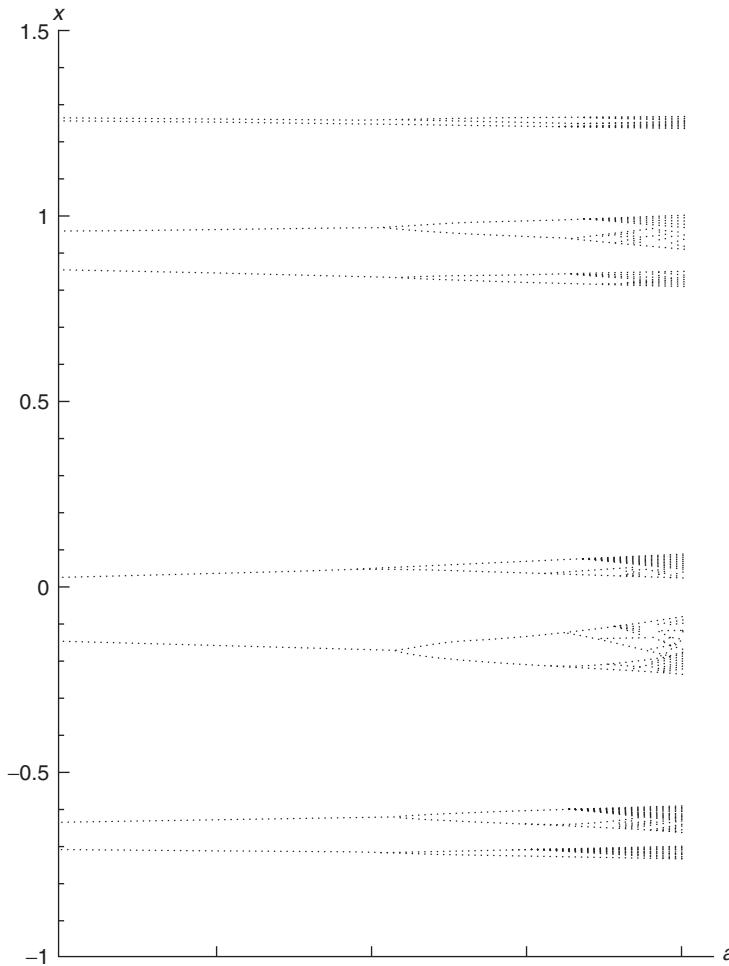


**FIGURE 13.6.3:** Bifurcation diagram for  $x$  when  $b = 0.3$  and  $a$  ranges from 0 to 1.04.

The example has demonstrated that period-doubling and chaos can be present when three variables are involved. It shows also that even a particle moving at constant velocity at intervals can exhibit behaviour that is far from simple. In this case we were able to find the exact formula connecting points on the Poincaré plane but in many circumstances this is not feasible. So we turn now to a method of approximation that allows progress when conditions are suitable.

## 13.7 Averaging

The method of averaging is designed to relate a nonautonomous system to an autonomous one by means of an approximation. It is appropriate for problems where a small perturbation is made to a system that can be solved



**FIGURE 13.6.4:** Bifurcation diagram for  $x$  when  $b = 0.3$  and  $a$  ranges from 1.04 to 1.06.

exactly. Although the perturbation may be small that does not mean that the solutions of the perturbed system stay close to those of the unperturbed. This anomaly is seen in the problem:

$$\ddot{x} + 2\epsilon\dot{x} + x = 0, \quad (13.7.1)$$

subject to the initial conditions

$$x(0) = a, \dot{x}(0) = 0. \quad (13.7.2)$$

The exact solution of this problem is

$$x(t) = ae^{-\epsilon t} \{ \cos(\sqrt{1-\epsilon^2}t) + \frac{\epsilon}{\sqrt{1-\epsilon^2}} \sin(\sqrt{1-\epsilon^2}t) \}, \quad (13.7.3)$$

corresponding to an oscillation with decaying amplitude. Now look for a solution to (13.7.1) in the form of a regular perturbation expansion of the form

$$x(t, \epsilon) = x_0(t) + \epsilon x_1(t) + \epsilon^2 x_2(t) + \dots, \quad (13.7.4)$$

where the unknown functions  $x_0(t)$ ,  $x_1(t)$ ,  $x_2(t)$ ,  $\dots$  are to be determined from (13.7.1) and the initial conditions (13.7.2). Substituting (13.7.4) into (13.7.1) and equating powers of  $\epsilon$  to zero, it is found that

$$\begin{aligned} \ddot{x}_0 + x_0 &= 0, \\ \ddot{x}_1 + 2\dot{x}_0 + x_1 &= 0, \\ \ddot{x}_2 + 2\dot{x}_1 + x_2 &= 0. \end{aligned} \quad (13.7.5)$$

The initial conditions (13.7.2) require that

$$\begin{aligned} x_0(0) &= a, \quad \dot{x}_0(0) = 0, \\ x_1(0) &= 0, \quad \dot{x}_1(0) = 0, \\ x_2(0) &= 0, \quad \dot{x}_2(0) = 0. \end{aligned} \quad (13.7.6)$$

Consequently

$$\begin{aligned} x_0(t) &= a \cos t, \\ x_1(t) &= a \sin t - at \cos t, \end{aligned}$$

and the expansion (13.7.4), to first order in  $\epsilon$ , is

$$x(t, \epsilon) = a \cos t + \epsilon a (\sin t - t \cos t) + \mathcal{O}(\epsilon^2). \quad (13.7.7)$$

This expansion corresponds to an oscillation with increasing amplitude due to the presence of the secular term  $t \cos t$ . In fact (13.7.7) is the beginning of a convergent series expansion of the exact solution provided  $\epsilon t \ll 1$ . However in practice one is interested in the behaviour of the solution for fixed small  $\epsilon$  and not fixed  $t$ . Therefore, just taking the unperturbed system as the first approximation may lead quickly to erroneous results; an approximation with a behaviour similar to that of the perturbed system is needed.

The type of system to be considered is, in vector form,

$$\dot{\mathbf{x}} = \epsilon \mathbf{f}(\mathbf{x}, t, \epsilon) \quad (13.7.8)$$

where the scalar  $\epsilon$  is very small compared with unity and  $\epsilon \geq 0$ . As regards  $\mathbf{f}$  we require it to be periodic in  $t$  so that

$$\mathbf{f}(\mathbf{x}, t + T, \epsilon) = \mathbf{f}(\mathbf{x}, t, \epsilon) \quad (13.7.9)$$

for all  $t$  under consideration. Moreover,  $\mathbf{f}$  has to possess at least two continuous derivatives in its variables and to be bounded for bounded values of its arguments.

The related autonomous system is obtained by averaging  $\mathbf{f}$  over a period. To be specific define  $\mathbf{w}$  to be the solution of

$$\dot{\mathbf{w}} = \frac{\epsilon}{T} \int_0^T \mathbf{f}(\mathbf{w}, t, 0) dt. \quad (13.7.10)$$

In the integration on the right-hand side of (13.7.10)  $\mathbf{w}$  is treated as a constant.

Under the above assumptions it is possible to prove the following:

- (a) If  $\mathbf{x}(t)$  is a solution of (13.7.8) such that  $\mathbf{x}(0) = \mathbf{x}_0$  and  $\mathbf{w}(t)$  is a solution of (13.7.10) such that  $\mathbf{w}(0) = \mathbf{w}_0$ , then, if  $|\mathbf{x}_0 - \mathbf{w}_0| = O(\epsilon)$ ,  $|\mathbf{x}(t) - \mathbf{w}(t)| = O(\epsilon)$  for values of  $t$  up to  $O(1/\epsilon)$ .
- (b) If  $\mathbf{w} \equiv (u, v)$  and (13.7.10) has a fixed point  $\mathbf{w}_p$ , which corresponds to a stable, unstable, or saddle limit cycle, then the Poincaré plane of (13.7.8) has a fixed point at  $\mathbf{w}_p + O(\epsilon)$  with the same stability.

Property (a) shows that solutions of (13.7.8) and (13.7.10), which start close together, stay close for a long time since  $\epsilon$  is small. Thus (13.7.10) does provide a good approximation to the trajectories of (13.7.8). In addition, property (b) indicates that a good idea of the limit cycles of (13.7.8) can be obtained from (13.7.10).

### **Example 13.7.1**

Find an approximate solution of

$$\dot{x} = -\epsilon x \sin^2 t \quad (13.7.11)$$

by the method of averaging.

Actually, the exact solution of (13.7.11) can be found by separation of variables. It gives, if  $x(0) = x_0$ ,

$$x(t) = x_0 \exp \left\{ -\epsilon \left( \frac{1}{2}t - \frac{1}{4} \sin 2t \right) \right\}. \quad (13.7.12)$$

The averaged equation corresponding to (13.7.11) is, from (13.7.10),

$$\dot{w} = -\epsilon w \frac{1}{2\pi} \int_0^{2\pi} \sin^2 t dt = -\frac{1}{2}\epsilon w \quad (13.7.13)$$

since  $w$  is regarded as a constant in the integration. The solution of (13.7.13) such that  $w = w_0$  at  $t = 0$  is

$$w(t) = w_0 \exp(-\frac{1}{2}\epsilon t). \quad (13.7.14)$$

The difference between (13.7.11) and (13.7.14) is

$$\dot{x}(t) - w(t) = \exp(-\frac{1}{2}\epsilon t) [x_0 - w_0 + x_0 \left\{ \exp(\frac{1}{4}\epsilon \sin 2t) - 1 \right\}]. \quad (13.7.15)$$

This shows immediately that, if  $x_0 - w_0 = O(\epsilon)$ , then  $x(t) - w(t) = O(\epsilon)$  for  $t \geq 0$ . Thus, property (a) is confirmed and for a larger range of  $t$  than specified. However, if (13.7.14) were replaced by

$$\dot{x} = \epsilon x \sin^2 t,$$

the only modification to (13.7.15) would be a change in the sign of  $\epsilon$ . Then, as  $t$  increased beyond  $1/\epsilon$ , the first factor of the right-hand side of (13.7.15) could become large enough to offset the  $O(\epsilon)$  of the second factor. Hence the upper limit on  $t$  in (a) cannot be removed.

The only fixed point of (13.7.13) is  $w = 0$ . According to property (b) there is a nearby fixed point associated with (13.7.11) and obviously it is  $x = 0$ , as may be checked from (13.7.12).  $\square$

### Example 13.7.2

Consider the perturbed oscillator

$$\ddot{x} + x = E \cos \omega t - D\dot{x} - Bx^3$$

where  $E$ ,  $D$  and  $B$  are to be regarded as small, positive constants.

Converting the differential equation into a first order system by the substitution  $y = \dot{x}$  we have

$$\dot{x} = y, \tag{13.7.16}$$

$$\dot{y} = -x - Dy - Bx^3 + E \cos \omega t. \tag{13.7.17}$$

The system does not have the structure of (13.7.8) so that the method of averaging cannot be applied directly. To transform it to the requisite form introduce polar coordinates. Write

$$x(t) = r(t) \cos\{\omega t - \theta(t)\}, \tag{13.7.18}$$

$$y(t) = -\omega r(t) \sin\{\omega t - \theta(t)\}. \tag{13.7.19}$$

Substitution in (13.7.16) and (13.7.17) leads to

$$\begin{aligned} \dot{r} \cos(\omega t - \theta) + r\dot{\theta} \sin(\omega t - \theta) &= 0, \\ -\omega\dot{r} \sin(\omega t - \theta) + \omega r\dot{\theta} \cos(\omega t - \theta) &= F \end{aligned}$$

where

$$F = D\omega r \sin(\omega t - \theta) + (\omega^2 - 1)r \cos(\omega t - \theta) - Br^3 \cos^3(\omega t - \theta) + E \cos \omega t.$$

Solving these equations for  $\dot{r}$  and  $\dot{\theta}$  we have

$$\omega\dot{r} = -F \sin(\omega t - \theta), \tag{13.7.20}$$

$$\omega r\dot{\theta} = F \cos(\omega t - \theta). \tag{13.7.21}$$

If, now,  $F$  were replaced by  $\epsilon(F/\epsilon)$  this system would be of the desired form with the various small coefficients in  $F$  divided by  $\epsilon$ . However, it is slightly more convenient to proceed with the original form without explicitly writing in the multiplication and division by  $\epsilon$ .

The averaged equations are obtained from formulae like

$$\frac{\omega}{2\pi} \int_0^{2\pi/\omega} \cos^2(\omega t - \theta) dt = \frac{1}{2},$$

$$\frac{\omega}{2\pi} \int_0^{2\pi/\omega} \cos(\omega t - \theta) \cos \omega t dt = \frac{1}{2} \cos \theta.$$

The result is

$$\omega \dot{r} = -\frac{1}{2} D \omega r + \frac{1}{2} E \sin \theta, \quad (13.7.22)$$

$$\omega r \dot{\theta} = \frac{1}{2} r (\omega^2 - 1) - \frac{3}{8} B r^3 + \frac{1}{2} E \cos \theta. \quad (13.7.23)$$

A fixed point  $(r_0, \theta_0)$  of (13.7.22) and (13.7.23) satisfies

$$E \sin \theta_0 = D \omega r_0, \quad (13.7.24)$$

$$E \cos \theta_0 = \frac{3}{4} B r_0^3 + r_0 (1 - \omega^2). \quad (13.7.25)$$

According to (13.7.24),  $\sin \theta_0$  can never be negative so that  $\theta_0$  can be restricted to the interval  $(0, \pi)$ . With that convention  $\theta_0$  can be determined from (13.7.25) once  $r_0$  is known. The elimination of  $\theta_0$  gives

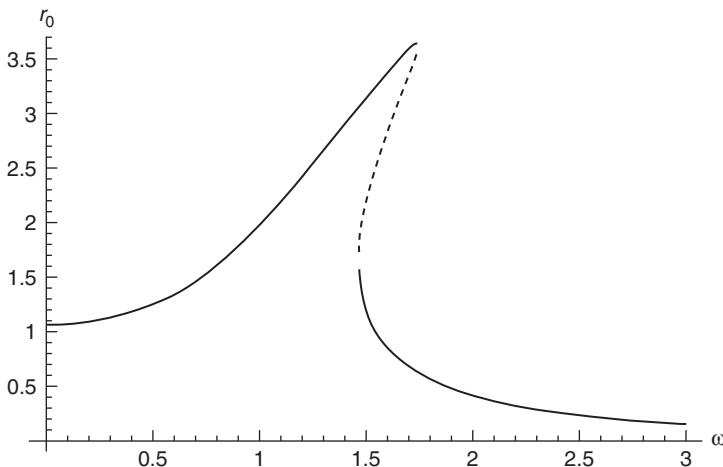
$$r_0^2 \left( \frac{3}{4} B r_0^2 + 1 - \omega^2 \right)^2 + D^2 \omega^2 r_0^2 = E^2 \quad (13.7.26)$$

or, with  $R = r_0^2$ ,

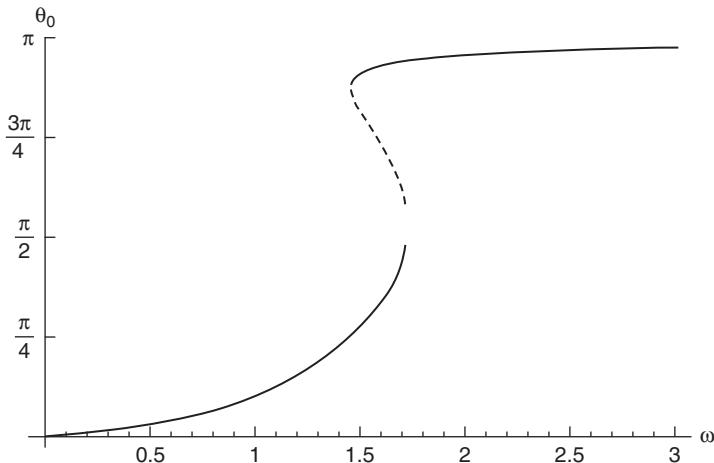
$$9B^2 R^3 + 24BR^2(1 - \omega^2) + 16R\{D^2 \omega^2 + (1 - \omega^2)^2\} - 16E^2 = 0, \quad (13.7.27)$$

which is a cubic equation to fix  $r_0$ . It is transparent from (13.7.26) that no root of the cubic can be negative. Since the left-hand side of (13.7.27) is negative when  $R = 0$  there is at least one positive root. When  $\omega^2 \leq 1$  there is at most one positive root by Descartes's rule of signs. Hence there is precisely one positive root when  $\omega^2 \leq 1$ . For  $\omega^2 > 1$  there may be three positive roots or one.

Keep  $B$ ,  $D$  and  $E$  fixed and treat  $\omega$  as a control parameter. Typical values of  $r_0$  obtained from (13.7.27) are shown in Figure 13.7.1. The graph was drawn for  $B = 0.2$ ,  $D = 0.2$ ,  $E = 1.25$ , values chosen to make what happens clearly visible rather than ensuring that they are small enough to comply with the requirements of the theory of averaging. For most of the range of  $\omega$  the cubic has a single root but for  $\omega$  in the neighbourhood of 1.5 there are three roots. The corresponding values of  $\theta_0$  are displayed in Figure 13.7.2; over much of the range  $\theta_0$  is not far from 0 or  $\pi$ .



**FIGURE 13.7.1:** Behaviour of  $r_0$  as  $\omega$  varies.



**FIGURE 13.7.2:** Variation of phase with  $\omega$ .

Since  $r$  is the length of the radius vector the value of  $r_0$  specifies the radius of a limit cycle. Therefore, we expect the original system to have a limit cycle whose distance from the origin is approximately  $r_0$ . In an interval of  $\omega$  there are three possible limit cycles; which of these does the system adopt?

To answer the question consider a nearby trajectory where  $r = r_0 + \rho$ ,  $\theta = \theta_0 + \eta$ . Then, for a first approximation, (13.7.22) and (13.7.23) become

$$\begin{aligned}\omega\dot{\rho} &= -\frac{1}{2}D\omega\rho + \frac{1}{2}E\eta\cos\theta_0, \\ \omega r_0\dot{\eta} &= \frac{1}{2}(\omega^2 - 1)\rho - \frac{9}{8}Br_0^2\rho - \frac{1}{2}E\eta\sin\theta_0.\end{aligned}$$

According to Section 5.4 the behaviour of this system is governed by the roots

$\lambda_1, \lambda_2$  of

$$\begin{aligned} \lambda^2 + \frac{1}{2}\lambda\{D\omega + (E/r_0)\sin\theta_0\} + \frac{D}{4r_0}E\omega\sin\theta_0 \\ - \frac{E}{4r_0}\left(\omega^2 - 1 - \frac{9}{4}Br_0^2\right) = 0. \end{aligned} \quad (13.7.28)$$

Evidently, one of  $\lambda_1, \lambda_2$  is always negative and the other is negative or positive according to the term in (13.7.28) independent of  $\lambda$  is positive or negative. Substitution for  $\theta_0$  from (13.7.24) and (13.7.25) reveals that this term is  $1/64$  times the derivative of the left-hand side of (13.7.27) with respect to  $R$ . When the cubic has a single positive root the derivative must be positive at the root. When there are three positive roots the derivative is positive at the smallest and largest roots but negative at the intermediate one. Thus,  $\lambda_1$  and  $\lambda_2$  are both negative except at the intermediate  $r_0$  of three.

Since  $\lambda_1$  and  $\lambda_2$  both negative corresponds to stability it follows that the limit cycle associated with  $r_0$  is stable unless  $r_0$  is the intermediate value of three when it is unstable. Consequently, the solid curves of Figures 13.7.1 and 13.7.2 indicate stable limit cycles while the dashed curves indicate an absence of stability.

Now, we see that as  $\omega$  increases the system can occupy a stable limit cycle related to the upper solid curve of Figure 13.7.1 until  $\omega$  reaches the point where the solid and dashed curves meet. If  $\omega$  is increased further the system cannot transfer to the limit cycle, which is not stable; so it must jump to the lower solid curve and continue along it. Decreasing  $\omega$  now induces a similar phenomenon but there is a jump upwards when the lower solid curve meets the dashed curve. Translating this back to the perturbed oscillator we expect it to exhibit similar jumps as  $\omega$  is varied. The existence of jumps has been verified experimentally. □

### Example 13.7.3

Consider the general autonomous problem

$$\ddot{x} + x = \epsilon f(x, \dot{x}) \quad (13.7.29)$$

subject to given initial conditions. As in Example 13.7.2 introduce the polar coordinate transformation

$$x(t) = r(t) \cos\{t - \theta(t)\}, \quad (13.7.30)$$

$$\dot{x}(t) = -r(t) \sin\{t - \theta(t)\}. \quad (13.7.31)$$

Substitution into (13.7.29), noting that the derivative of the right-hand side of (13.7.30) must be consistent with the right-hand side of (13.7.31), leads to

$$\begin{aligned} \dot{r} &= -\epsilon \sin\{t - \theta(t)\}f(r(t) \cos\{t - \theta(t)\}, -r(t) \sin\{t - \theta(t)\}), \\ r\dot{\theta} &= \epsilon \cos\{t - \theta(t)\}f(r(t) \cos\{t - \theta(t)\}, -r(t) \sin\{t - \theta(t)\}). \end{aligned} \quad (13.7.32)$$

Now the right-hand sides of these equations are periodic of period  $2\pi$  and, treating  $\theta$  as constant, so the averaged equations are:

$$\dot{r} = -\frac{\epsilon}{2\pi} \int_0^{2\pi} \sin\{t - \theta(t)\} f(r(t) \cos\{t - \theta(t)\}, -r(t) \sin\{t - \theta(t)\}) dt, \quad (13.7.33)$$

$$r\dot{\theta} = \frac{\epsilon}{2\pi} \int_0^{2\pi} \cos\{t - \theta(t)\} f(r(t) \cos\{t - \theta(t)\}, -r(t) \sin\{t - \theta(t)\}) dt. \quad (13.7.34)$$

Since the right-hand sides of these equations are independent of  $\theta$  they can be written in the concise form

$$\begin{aligned} \dot{r} &= -\epsilon f_1(r), \\ \dot{\theta} &= \frac{\epsilon}{r} f_2(r), \end{aligned} \quad (13.7.35)$$

where

$$\begin{aligned} f_1(r) &\equiv \frac{1}{2\pi} \int_0^{2\pi} \sin\{t - \theta(t)\} f(r(t) \cos\{t - \theta(t)\}, -r(t) \sin\{t - \theta(t)\}) dt, \\ f_2(r) &\equiv \frac{1}{2\pi} \int_0^{2\pi} \cos\{t - \theta(t)\} f(r(t) \cos\{t - \theta(t)\}, -r(t) \sin\{t - \theta(t)\}) dt. \end{aligned}$$

Applying the method to **van der Pol's equation**, discussed in Section 5.6,

$$\ddot{x} + x = \epsilon(1 - x^2)\dot{x}$$

one gets the averaged equations

$$\dot{r} = \frac{\epsilon}{2} r(1 - \frac{1}{4}r^2), \quad \dot{\theta} = 0. \quad (13.7.36)$$

Choosing  $r(0) = 2$  shows that  $r(t) = 2, t \geq 0$  and taking  $\theta(0) = 0$  leads to the solution

$$x(t) = 2 \cos t + \mathcal{O}(\epsilon)$$

on a time scale  $1/\epsilon$ . The averaging method also allows one to find an approximation to other solutions by solving equations (13.7.36) to get

$$r(t) = \frac{r(0)e^{\frac{1}{2}\epsilon t}}{\left[1 + \frac{1}{4}r^2(0)(e^{\epsilon t} - 1)\right]^{\frac{1}{2}}}, \quad \theta(t) = \theta(0).$$

Since  $x(t) = r(t) \cos(t - \theta(t))$  it is seen that as  $t$  increases the solutions approach the periodic solution. □

## Exercises

13.1 Show that

$$\begin{aligned}\dot{x} &= x(y - 1), \\ \dot{y} &= \mu - y(x + 1)\end{aligned}$$

has an equilibrium point, which is a stable node, for  $\mu < 1$  that becomes a saddle-point as  $\mu$  passes through the bifurcation point  $\mu = 1$ . Also show that there is an additional equilibrium point when  $\mu > 1$ , which is a stable node.

13.2 Show that

$$\begin{aligned}\dot{x} &= \mu x(2y - 1), \\ \dot{y} &= \mu - y(2x + 1)\end{aligned}$$

has an equilibrium point, which is a stable node, for  $1/2 < \mu < 1$  and a stable focus for  $\mu > 1$ . Prove that  $\mu = 1/2$  is a bifurcation point.

13.3 Show that

$$\ddot{x} + \mu\dot{x} + 2x + x^2\dot{x} + x^3 = 0$$

has a Hopf bifurcation at  $\mu = 0$ .

13.4 Show that the iteration scheme

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

has a stable fixed point  $x_0 = 1$  for  $\mu < 1$  and that  $\mu = 1$  is a bifurcation point where the fixed point  $x_0 = 1/\mu$  appears. Show that period-doubling occurs as soon as  $\mu$  exceeds 3.

13.5 Study the fixed points and period-doubling of

$$x_{n+1} = 1 - \mu x_n^2.$$

13.6 Demonstrate that period-doubling can occur for

$$x_{n+1} = \mu \sin \pi x_n$$

and try to draw a bifurcation diagram for  $\mu$  between 0 and 0.07.

13.7 Show that if the iteration

$$x_{n+1} = \exp(-7.5x_n^2) - 0.9$$

is started with  $x_1 = 0$ ,  $x_n$  tends to 0.067 approximately, but that if  $x_1 = 0.7$ ,  $x_n$  tends to -0.898 approximately.

13.8 Draw a bifurcation diagram for

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

with  $\mu$  ranging from  $-1$  to  $1$ .

13.9 Estimate the Feigenbaum number for

$$x_{n+1} = 1 - \mu x_n^4.$$

13.10 By changing to polar coordinates and using the Poincaré line show that

$$\begin{aligned}\dot{x} &= -y + \frac{1}{2}\mu x - \frac{1}{2}x(x^2 + y^2)^2, \\ \dot{y} &= x + \frac{1}{2}\mu y - \frac{1}{2}y(x^2 + y^2)^2\end{aligned}$$

has a stable limit cycle when  $\mu > 0$ .

13.11 On a Poincaré plane successive points are related by

$$\begin{aligned}x_{n+1} &= \frac{1}{2}y_n, \\ y_{n+1} &= -x_n + \frac{1}{2}\mu y_n - y_n^3.\end{aligned}$$

Show that there is a bifurcation at  $\mu = 3$ . Show that the limit cycle in  $0 < \mu < 3$  is stable and becomes of saddle type when  $\mu$  exceeds 3. Verify that the two other limit cycles for  $\mu$  just above 3 are stable.

13.12 Demonstrate that the solutions of

$$\dot{x} = -\epsilon x \cos^2 t$$

and the averaged equation stay close together if they are initially close.

13.13 A trajectory in the phase plane of

$$\ddot{x} + \epsilon x = -\epsilon x \cos^2 t$$

starts from the same point as a trajectory of the averaged equations. Show that the trajectories stay close together.

13.14 Transform the first order system derived from

$$\ddot{x} + \epsilon(x^2 - 1)\dot{x} + x = \omega\epsilon E \cos \omega t$$

to polar coordinates and then average. If  $E = 1/2$  and  $\mu = (\omega^2 - 1)/\omega\epsilon$  show that the averaged equations have three fixed points, two of which are unstable, when  $\mu = 0$ . Draw a graph of  $r_0$  against  $\mu$  for  $\mu$  going from 0 to 1.

13.15 Show that the equation

$$\ddot{x} + x + \epsilon x^3 = 0,$$

for small positive parameter  $\epsilon$ , has one critical point and that all other solutions are periodic.

Use the method of averaging to construct a first order approximation of the solutions for general ( $\epsilon$  independent) initial conditions.

13.16 Use the method of averaging to show that the equation

$$\ddot{x} + x = \epsilon(1 - ax^2 - b\dot{x}^2)\dot{x}$$

has a periodic solution which, to first order, is approximated by

$$x = \frac{2}{\sqrt{a+3b}} \cos t.$$

# Chapter 14

---

## Numerical Bifurcation Analysis

---

### 14.1 Fixed points and stability

The fixed point is one of the most fundamental features of a dynamical system, and the starting point for most dynamical systems analysis is to find fixed points and determine their stability. For a system of differential equations,

$$\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}), \quad (14.1.1)$$

fixed points  $\mathbf{y}^*$  are found by solving the vector equation

$$\mathbf{f}(\mathbf{y}) = \mathbf{0}. \quad (14.1.2)$$

The method for doing this in MATLAB is to write a function that calculates  $\mathbf{f}(\mathbf{y})$ , just as for solving the differential equations themselves (see Section 5.8). Then the solution can be found using a root-finding method, like Newton's method (see function `newton` in Section 5.9) with a command such as

```
>> y = newton(@myfunc, y0, tol, maxits);
```

where  $y_0$  is an initial approximation to the solution,  $tol$  is a defined convergence tolerance,  $maxits$  is the maximum number of iterations of Newton's method to carry out, and `myfunc` is a function that calculates  $\mathbf{f}(\mathbf{y})$ .

Solving a nonlinear system of equations numerically is not a trivial task and usually requires a reasonable approximation to the solution to use the starting point  $y_0$  for the iterations. In practice, it is sometimes necessary to try a number of different initial approximations by systematically looping through phase space. This process is usually computationally expensive, but can sometimes be made more efficient if some knowledge about which regions of phase space are likely to contain fixed points can be gained, either analytically or from numerical solution of the full time-dependent equations (14.1.1).

Assuming that fixed points  $\mathbf{y}^*$  of a system of the form (14.1.1) have been found, their stability can be determined by evaluating the Jacobian matrix of  $\mathbf{f}(\mathbf{y})$  at  $\mathbf{y} = \mathbf{y}^*$  and calculating its eigenvalues. This can be achieved with the following MATLAB commands (see Section 5.9 for the function `jacobian`).

```
>> [f, J] = jacobian(@myfunc, y_star);
>> lambda = eig(J);
```

If all eigenvalues  $\lambda$  have negative real part, then the fixed point is stable; if at least one eigenvalue has positive real part, then the fixed point is unstable.

### Example 14.1.1

Find the fixed points of the system

$$\dot{y}_1 = y_1 \left( 1 - y_1 - \frac{ay_2(1 - e^{-by_1})}{y_1} \right), \quad (14.1.3)$$

$$\dot{y}_2 = ry_2 \left( 1 - \frac{y_2}{y_1 + c} \right), \quad (14.1.4)$$

and determine their stability, when  $a = 2$ ,  $b = 1$ ,  $c = 0.1$  and  $r = 0.5$ . This is a form of the Volterra–Lotka predator–prey model (see Section 4.5) modified to include more realistic logistic (rather than simply exponential) growth of the prey ( $y_1$ ) and predator ( $y_2$ ) species.

It can easily be seen that  $(y_1, y_2) = (0, 0)$  (extinction of both species),  $(y_1, y_2) = (1, 0)$  (existence of prey species only) and  $(y_1, y_2) = (0, c)$  (existence of predator species only) are fixed points. Furthermore, equation (14.1.4) indicates that any other fixed points must satisfy  $y_2 = y_1 + c$ . However, in this case, there is no closed-form solution to the steady state of equation (14.1.3), and numerical methods become useful. The first step, as always, is to write a MATLAB function that evaluates the right-hand sides of the differential equations as a vector:

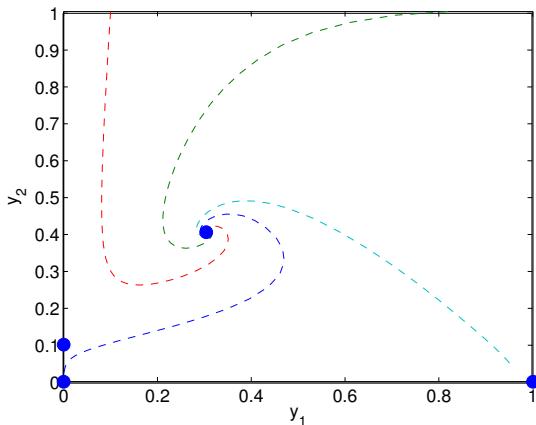
```
function f = lv_logistic(y)
a = 2; b = 1; c = 0.1; r = 0.5;
f = zeros(2, 1);
f(1) = y(1)*(1-y(1)) - a*y(2)*(1-exp(-b*y(1)));
f(2) = r*y(2)*(1-y(2)/(y(1)+c));
```

This function (after modifying the inputs to be  $(t, y)$  rather than just  $y$ ) could be inserted into a differential equation solving routine such as `ode45` to find the solution for a given initial condition. Here, we are interested in the fixed points. To do this via Newton's method, we need an initial estimate of the solution. Let's try  $(1, 1)$ :

```
>> y0 = [1; 1];
>> y_star = newton(@lv_logistic, y0, 1e-10, 100);
```

These commands return  $(y_1, y_2) = (0.3037, 0.4037)$  as the solution (which is easily checked to be a fixed point of the system (14.1.3), (14.1.4)). Hence we have found four fixed points. To determine stability, we calculate the eigenvalues of the Jacobian at each fixed point with the commands:

```
>> [f, J] = jacobian(@lv_logistic, y_star);
>> lambda = eig(J);
```



**FIGURE 14.1.1:** Fixed points (filled circles) and sample orbits (dashed curves) calculated numerically for the modified Volterra–Lotka system (14.1.3), (14.1.4). All the solutions shown tend towards the stable fixed point  $(y_1, y_2) = (0.3, 0.4)$ .

This calculates the eigenvalues  $\lambda$  at each fixed point  $\mathbf{y}^*$ , enabling us to classify their stability:

Fixed point $\mathbf{y}^*$	Eigenvalues $\lambda_1, \lambda_2$	Stability
$(0, 0)$	$0.5, 1$	Unstable
$(1, 0)$	$-1, 0.5$	Saddle
$(0, 0.1)$	$-0.5, 0.8$	Saddle
$(0.3037, 0.4037)$	$-0.3517 \pm 0.4898i$	Stable focus

An alternative approach, in this example, is to work out the Jacobian analytically (rather than numerically using the `jacobian` function):

$$\mathbf{J} = \begin{bmatrix} 1 - 2y_1 - aby_2 e^{-by_1} & a(-1 + e^{-by_1}) \\ \frac{ry_2^2}{(y_1+c)^2} & \frac{r(y_1+c-2y_2)}{y_1+c} \end{bmatrix}.$$

The calculated fixed points may then be substituted into this analytical expression for  $\mathbf{J}$  and the eigenvalues calculated.

Figure 14.1.1 shows the fixed points, together with some sample orbits. It is impossible to be sure that we have located *all* the fixed points of the system (14.1.3), (14.1.4). However, trying initials approximation in Newton's method in the range  $0 < y_1, y_2 < 500$  does not reveal any further fixed points. □

The results of the numerical analysis can be checked by solving the time-dependent problem (see Section 5.8) with an initial condition in the vicinity

of the fixed points. For stable fixed points,  $\mathbf{y}$  should tend towards  $\mathbf{y}^*$ ; for unstable fixed points,  $\mathbf{y}$  should move away from  $\mathbf{y}^*$ .

It should be noted that, although the methods described in this chapter are applicable to continuous dynamical systems (systems of differential equations) many are equally applicable to discrete, iterative systems, or maps (see Sections 13.3 and 14.4). The main differences between a dynamical system based on differential equations and one based on a map is that, for a map defined by  $\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n)$ , fixed points  $\mathbf{x}^*$  satisfy  $\mathbf{x}^* = \mathbf{f}(\mathbf{x}^*)$  rather than equation (14.1.2), and the requirement for stability is that the eigenvalues of the Jacobian evaluated at  $\mathbf{x}^*$  have modulus less than 1 (i.e.,  $|\lambda| < 1$ ).

---

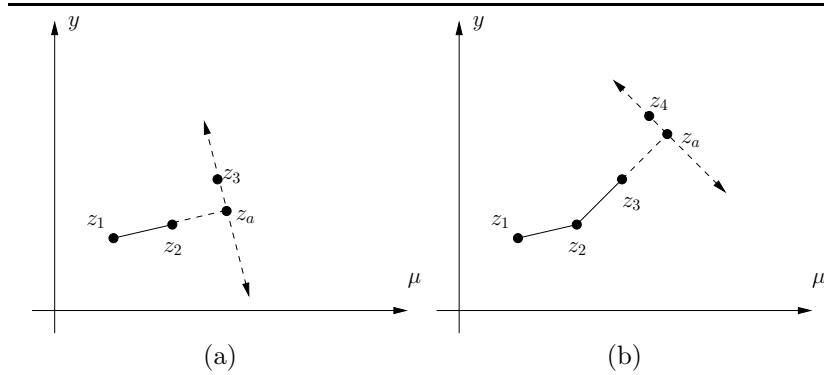
## 14.2 Path-following and bifurcation analysis

Having found the fixed points of a dynamical system and determined their stability, it is often desirable to carry out a bifurcation analysis with respect to some parameter  $\mu$ , as described in Section 13.1. When done numerically, this is known as **path-following**. A naive approach would be simply to repeat the procedure of the previous section for a series of equally spaced values of  $\mu$ . An exhaustive search of phase space for fixed points is usually prohibitively expensive, but this could be avoided because the fixed points already found for the preceding value of  $\mu$  will be close to the fixed points for the current value of  $\mu$ , and hence would be good candidates to try as starting values for Newton's method. The problem with this approach is that it breaks down just where we are most interested in its results: at a bifurcation. Consider for example the fold bifurcation shown in Figure 13.1.3. Starting with a small value of  $\mu$  and working upwards will not work because, for small values of  $\mu$ , there are no fixed points to work with. Starting with a large value of  $\mu$  and trying to follow one of the two fixed points as  $\mu$  decreases will work initially. However, as  $\mu$  approaches the bifurcation point, the method will simply ‘fall’ off the end of the fold bifurcation into the region where there are no fixed points, without finding the other branch of the bifurcation diagram.

Path-following proceeds by treating the parameter  $\mu$  as an additional dependent variable in phase space. Suppose that two nearby points on the bifurcation curve  $\mathbf{z}_1 = (\mathbf{y}_1, \mu_1)$  and  $\mathbf{z}_2 = (\mathbf{y}_2, \mu_2)$  have been found by the naive approach described above (see Figure 14.2.1). We now seek a third point  $\mathbf{z}_3 = (\mathbf{y}_3, \mu_3)$  as the solution to the equation

$$\mathbf{f}(\mathbf{y}, \mu) = \mathbf{0}. \quad (14.2.1)$$

We construct an initial approximation  $\mathbf{z}_a$  (to feed into Newton's method) to this point simply by linearly extrapolating from the two known points:  $\mathbf{z}_a = 2\mathbf{z}_2 - \mathbf{z}_1$  (see Figure 14.2.1). By treating  $\mu$  as a free variable, we have



**FIGURE 14.2.1:** Numerical method for path-following: (a) two existing points  $\mathbf{z}_1$  and  $\mathbf{z}_2$  on the bifurcation diagram are used to generate a third point  $\mathbf{z}_3$ ; (b) the process is repeated to generate  $\mathbf{z}_4$ , and so on.

effectively introduced an extra unknown into the system of equations, so we need an extra equation for the problem to remain well defined. Remember that the goal is to ‘follow’ the fixed point around the fold bifurcation, rather than to find the fixed point for pre-determined values of  $\mu$ . Hence, we obtain an extra equation by constraining the third point in the bifurcation curve to lie on the line that: (i) passes through the initial approximation to the point; (ii) is perpendicular to the line through the first two points (see Figure 14.2.1). This extra equation can be expressed in vector notation as

$$(\mathbf{z}_3 - \mathbf{z}_a) \cdot (\mathbf{z}_a - \mathbf{z}_2) = 0, \quad (14.2.2)$$

where the points  $\mathbf{z}_2$  and  $\mathbf{z}_a$  are known constants. (In higher dimensions, (14.2.2) is the equation of a hyperplane perpendicular to the line through the first two points.) Combining equations (14.2.1) and (14.2.2) gives  $N + 1$  equations in  $N + 1$  unknowns ( $\mathbf{z}_3 = (y_1, \dots, y_n, \mu)$ ). Solving these (via Newton’s method with an initial approximation  $\mathbf{z}_a$ ) gives the point  $\mathbf{z}_3$ . This process can now be repeated using the points  $\mathbf{z}_2$  and  $\mathbf{z}_3$  to construct a new point  $\mathbf{z}_4$ , and so on. The stability of each fixed point found can be determined by looking at the eigenvalues of the Jacobian.

Note that the distance between the initial points  $\mathbf{z}_1$  and  $\mathbf{z}_2$  is the approximate arc length distance between consecutive computed points on the bifurcation curve. These initial points should therefore be chosen to be sufficiently close that a reasonably smooth curve is produced, but not so close that an infeasibly large number of points are required to give the desired portion of the curve.

A MATLAB function implementing this procedure is shown below, and will be used in the following examples

```

function Z = path_follow(dydt, z1, z2, M)
% Function for following the path of a fixed point through
% phase-parameter space
% Inputs: dydt - handle to a function that calculates dy/dt = f(z)
%          z1, z2 - two initial points on the bifurcation curve
%                      where z = [y_1, ..., y_n, mu]
%          M - number of points on the bifurcation curve to
%              calculate
% Outputs: Z - matrix whose columns correspond to points on the
%              bifurcation curve
tol = 1e-10;
ftol = 1e-8;
maxits = 100;
N = length(z1);
Z = zeros(N, M);
Z(:, 1) = z1;
Z(:, 2) = z2;
for I = 1:M-2
    zp = Z(:, I+1);      % most recent known point zp
    za = 2*zp-Z(:, I);  % initial approximation za
    fh = @(z)path_follow_func(dydt, z, za, zp);
    Z(:, I+2) = newton(fh, za, tol, maxits);
    % stop if Newton's method did not find a valid point:
    if ~abs(path_follow_func(dydt, Z(:, I+2), za, zp)) < ftol
        Z = Z(:, 1:I+1);
        return;
    end
end

```

The function `path_follow_func` is a function whose root is the point  $\mathbf{z}$ , i.e., `path_follow_func` returns a vector containing the left-hand sides of equations (14.2.1) and (14.2.2):

```

function f = path_follow_func(dydt, z, za, zp)
N = length(z);
f = zeros(N, 1);
f(1: N-1) = dydt(z);
f(N) = (z-za)'*(za-zp);

```

Also very useful is a function to calculate the eigenvalues of the Jacobian at each point on the bifurcation curve, as this will enable us to distinguish stable and unstable points. All the necessary information is contained in the matrix  $Z$  that is output from `path_follow`: it is just a matter of separating the dependent variables ( $\mathbf{y}$ ) from the parameter  $\mu$ , and treating the function  $f$  as a function of  $\mathbf{y}$  only, rather than  $\mathbf{y}$  and  $\mu$ .

```
function L = calculate_eigenvalues(dydt, Z)
```

```
% Function to calculate eigenvalues of the Jacobian at a number
% of points on a bifurcation curve
% Inputs: dydt - handle to a function that calculates dy/dt=f(z)
%          Z - NxM matrix of points on the bifurcation curve as
%          output from path_follow
% Outputs: L - corresponding (N-1)xM matrix of eigenvalues
[N, M] = size(Z);
L = zeros(N-1, M);
for I = 1:M % go through each point on the bifurcation curve
    y = Z(1:N-1, I); % extract y from the Ith column of Z
    mu = Z(N, I); % extract mu from the Ith column of Z
    % calculate Jacobian of dydt (treating mu as a constant):
    [f, J] = jacobian(@(y)dydt([y; mu]), y);
    L(:, I) = eig(J); % store eigenvalues in Ith column of L
end
```

**Example 14.2.1**

Consider the differential equation

$$\dot{y} = f(y) = y - y^4 + \mu, \quad (14.2.3)$$

where  $\mu$  is a parameter. Find the fixed points of the system when  $\mu = 0$  and follow the fixed points as  $\mu$  varies to produce the bifurcation diagram.

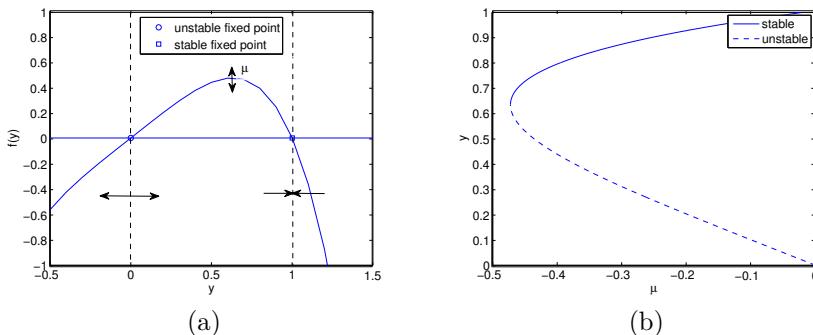
Figure 14.2.2(a) shows a graph of  $f(y)$  against  $y$  when  $\mu = 0$ . Varying  $\mu$  simply shifts this graph in the vertical direction. When  $\mu = 0$ , the fixed points satisfy  $y(1 - y^3) = 0$ . There are therefore two fixed points  $y^* = 0$  and  $y^* = 1$ . We now use the path-following routine to follow the first of these two fixed points as  $\mu$  decreases. We already have one point on the bifurcation curve  $\mathbf{z}_1 = [y_1, \mu_1] = [0, 0]$ . A second point  $\mathbf{z}_2$  can be obtained by numerically solving the steady-state equation when say  $\mu = -\delta\mu$ , where  $\delta\mu$  is a suitably small number, say  $10^{-3}$ . Note that  $\delta\mu$  is not a prescribed step size for  $\mu$ , merely the approximate distance between consecutive computed points on the bifurcation curve.

We again need a function that calculates the right-hand side of the differential equation. In line with the path-following routine described above, this is regarded as a function of the parameter as well as the independent variable, and hence the input to the function is  $\mathbf{z} = (y, \mu)$ .

```
function f = fold_bif(z)
f = z(1)-z(1)^4+z(2);
```

We are now in a position to compute the bifurcation curve, and calculate the corresponding eigenvalues with the following commands

```
>> Z = path_follow(@fold_bif, z1, z2, 1000);
>> L = calculate_eigenvalues(@fold_bif, Z);
```



**FIGURE 14.2.2:** (a) Graph of the function  $f(y)$  for the differential equation (14.2.3) against  $y$  when  $\mu = 0$ , showing the two fixed points, their stability and how the parameter  $\mu$  alters the system. (b) Bifurcation diagram calculated by numerical path-following. The system has a fold bifurcation at  $\mu = -\frac{3}{4^{4/3}} \approx -0.47$ : as  $\mu$  decreases towards this value, the stable and the unstable fixed point collide and destroy each other.

Note that, in the one-dimensional case, the Jacobian is just a  $1 \times 1$  matrix simply  $df/dy$ . Hence the stability of any fixed point can easily be determined analytically (by the sign of  $df/dy$ ). Nevertheless, the command is shown here to illustrate the general method for  $n$ -dimensional systems.

To plot the bifurcation curve, in a one-dimensional example such as this one, we simply plot the first row of the matrix  $Z$  (i.e., the values of  $y^*$ ) against the second row (i.e., the corresponding values of  $\mu$ ):

```
>> plot(Z(2, :), Z(1, :))
```

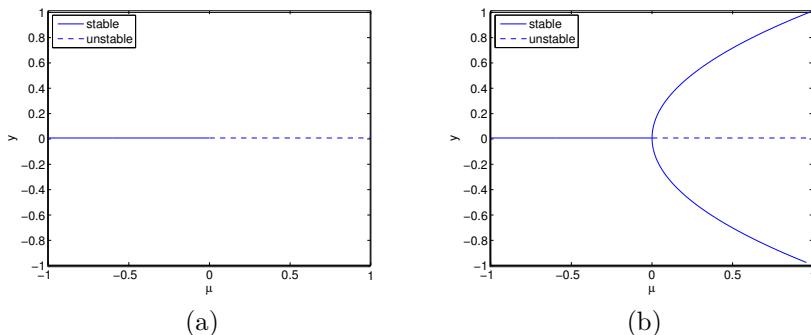
This produces the graph shown in Figure 14.2.2(b), which also indicates the stability of each part of the curve (as determined by the contents of the eigenvalue matrix  $L$ ). The system has a **fold bifurcation** in which the stable and the unstable fixed point collide and destroy each other. The exact value of  $\mu$  at which this occurs can be calculated, as it is the value at which the maximum of the function  $f(y)$  (see Figure 14.2.2(a)) is zero. Standard calculus techniques show that the bifurcation point is  $\mu = -\frac{3}{4^{4/3}}$ . □

### Example 14.2.2

Plot the bifurcation diagram for the differential equation

$$\dot{y} = f(y) = \mu y - y^3, \quad (14.2.4)$$

where  $\mu$  is a parameter.



**FIGURE 14.2.3:** Bifurcation diagram for the differential equation (14.2.4): (a) the curve produced by following the fixed point  $(y, \mu) = (0, -1)$ ; (b) the curve produced by following the fixed point  $(y, \mu) = (1, 1)$ . The system has a pitchfork bifurcation at  $\mu = 0$ : as  $\mu$  increases through 0, the stable fixed point becomes unstable, and two new stable fixed points are created.

The differential equation (14.2.4) can be dealt with analytically using the methods of Chapter 13, and is the prototypical form of a system with a **pitchfork bifurcation**. Nevertheless, we will implement the numerical path-following method to illustrate how this type of bifurcation can be found. Before we begin, we need two points on the bifurcation curve, which can be obtained by finding the fixed point when, say,  $\mu = -1$  and  $\mu = -1 + \delta\mu$ . For  $\mu < 0$ , there is a unique fixed point at  $y = 0$ , so our first two points are defined by

```
>> z1 = [0; -1];
>> z2 = [0; -1+delta_mu];
```

We also need a MATLAB function that evaluates  $f(y, \mu)$ :

```
function f = pitchfork_bif(z)
f = z(2)*z(1)-z(1)^3;
```

We can now call `path_follow` as above, and plot the bifurcation curve, which is shown in Figure 14.2.3(a). The change of stability at  $\mu = 0$  indicates a bifurcation, and there may be something else on the bifurcation diagram that the path-following routine has not found. We therefore check if there are any other fixed points when  $\mu > 0$ . At  $\mu = 1$ , there are fixed points at  $y = \pm 1$ . Let's try and follow the fixed point at  $y = 1$  as  $\mu$  decreases towards the bifurcation point. This gives another branch of the bifurcation diagram, shown in Figure 14.2.3(b). The bifurcation diagram now clearly shows the pitchfork bifurcation, in which a single stable fixed point bifurcates into one unstable and two stable fixed points.



### 14.3 Following stable limit cycles

Fixed points are relatively easy objects to deal with, as they are the solution of a known system of algebraic equations, and their stability is easily determined by the Jacobian. Periodic orbits (i.e., solutions  $\mathbf{y}(t)$  to the system of differential equations that satisfy  $\mathbf{y}(t+T) = \mathbf{y}(t)$  for some fixed, positive value of  $T$ ) are more difficult to analyse. Even with a numerical solution of the differential equations, periodic orbits can be difficult to identify.

Consider the case of a **Hopf bifurcation**, in which a fixed point loses stability and a stable limit cycle is simultaneously created. This is indicated by a complex pair of eigenvalues crossing the imaginary axis in the complex plane. The procedure outlined in Section 14.2 should follow the fixed point through the bifurcation, but what about the limit cycle? As with any attractor, the limit cycle can be found by numerically solving the system of differential equations for a suitable initial condition, in this case a point near to the unstable fixed point. Disregarding the transient behaviour should yield the limit cycle. This cycle can now be ‘followed’ as the parameter  $\mu$  changes: change  $\mu$  by a small amount and use the final value of  $\mathbf{y}$  from the previous solution as the initial condition for the current value of  $\mu$ . In order to produce a meaningful bifurcation diagram, it is usually desirable to plot one or more scalar values representing the limit cycle for each value of  $\mu$ . Typically, these might be the maximum and minimum values of  $|y|$  over the course of the cycle, or the value of  $|y|$  on a suitably defined Poincaré section (see Section 13.5). If the limit cycle loses stability as  $\mu$  changes, the numerical solution of the differential equations will cease to converge to the limit cycle.

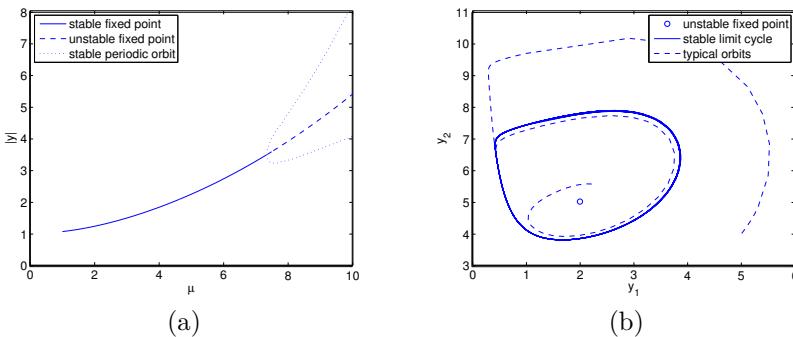
#### **Example 14.3.1**

Consider the system of differential equations

$$\dot{y}_1 = \mu - y_1 - \frac{4y_1y_2}{y_1^2 + 1}, \quad (14.3.1)$$

$$\dot{y}_2 = y_1 \left( 1 - \frac{y_2}{y_1^2 + 1} \right). \quad (14.3.2)$$

This system is a simplified model of the chlorine dioxide–iodine–malonic acid reaction (comparable to the Field–Noyes model of the Belousov–Zhabotinskii reaction seen in Chapter 4). Here,  $y_1$  and  $y_2$  represent the dimensionless concentrations of iodine and chlorine dioxide respectively, and  $\mu > 0$  is a param-



**FIGURE 14.3.1:** (a) Bifurcation diagram for the system (14.3.1), (14.3.2). The system has a Hopf bifurcation at  $\mu = \mu_c \approx 7.3$ : as  $\mu$  increases through  $\mu_c$ , the stable fixed point becomes unstable, and a stable limit cycle is created. (b) Phase portrait when  $\mu > \mu_c$ , showing the unstable fixed point, stable limit cycle and typical orbits.

eter representing one of the reaction rates. Plot the bifurcation diagram for the system (14.3.1), (14.3.2).

We start, as in the previous section, by looking for fixed points of the system. Finding the fixed point at  $\mu = 1$  and then following it as  $\mu$  increases, as described in the previous section, produces the fixed point curve shown in Figure 14.3.1(a). The fixed point loses stability at  $\mu = \mu_c \approx 7.3$ : at this value of  $\mu$ , a complex conjugate pair of eigenvalues pass through the imaginary axis, which is the signature of a Hopf bifurcation. (In this example, the fixed point and its stability can also be found analytically, using the techniques described in Chapter 13.) How can we test for limit cycles and include them in the bifurcation diagram? The following MATLAB program implements the procedure outlined above and, for each value of  $\mu$ , records the maximum and minimum values of  $|y|$  for plotting.

```

mu = mu_min:delta_mu:mu_max;
nudge = 1e-6; % small perturbation to steady state
t_trans = 100; % enough time to remove transients
t_cycle = 100; % enough time to go round cycle at least once
M = length(mu);
maxy = zeros(1, M);
miny = maxy;
fh = @(t, y)myrhs([y; mu(1)]);
% solve the DEs to remove transients:
[t, Y] = ode45(fh, [0, t_trans], (1+nudge)*y_star);
for I = 1:M % go through each value of mu
    maxy(I) = max(Y);
    miny(I) = min(Y);
end

```

```
% solve the DEs, starting with the most recent value of y:
fh = @(t, y)myrhs([y; mu(I)]);
y0 = Y(size(Y, 1), :);
[t, Y] = ode45(fh, [0, t_cycle], y0);
% calculate the norm of y at each point on the orbit:
P = size(Y, 1);
y_norm = zeros(P, 1);
for J = 1:P;
    y_norm(J) = norm(Y(J, :));
end
% record the maximum and minimum values of the norm:
maxy(I) = max(y_norm);
miny(I) = min(y_norm);
end
plot(mu, miny, mu, maxy)
```

Here, the minimum and maximum values of  $\mu$  (`mu_min` and `mu_max`), the increment in  $\mu$  (`delta_mu`) and the fixed point at the minimum value of  $\mu$  (`y_star`) need to be specified. `myrhs` is the usual function for calculating  $\dot{y}$ :

```
function f = myrhs(z)
f = zeros(2, 1);
f(1) = z(3)-z(1)-4*z(1)*z(2)/(z(1)^2+1);
f(2) = z(1)*(1-z(2)/(z(1)^2+1));
```

(As before, the last entry `z(3)` in the vector `z` is used to store the value of the parameter  $\mu$ .) This program produces the stable limit cycle curve shown on the bifurcation diagram in Figure 14.3.1(a), confirming that the system does indeed have a Hopf bifurcation at  $\mu = \mu_c$ . Figure 14.3.1(b) shows the phase portrait when  $\mu > \mu_c$ , including the unstable fixed point, some typical orbits and the stable limit cycle (found by calculating an orbit and disregarding the transients).

□

## 14.4 Bifurcation in discrete systems

In Section 13.3, we saw examples of how a discrete, iterative scheme of the form

$$x_{n+1} = f(x_n) \quad (14.4.1)$$

can undergo bifurcations as a parameter  $\mu$  of the function  $f$  is varied. Bifurcation diagrams of the type shown in Section 13.3 can be easily constructed

simply by iterating equation (14.4.1) and discarding transient behaviour to plot only stable fixed and periodic points. As in Section 14.3, we are following stable objects: in Section 14.3, these objects were limit cycles of a system of differential equations; here, they are fixed or periodic points of the iterative equation (14.4.1). The technique is therefore very similar: each time we move to a new value of the parameter  $\mu$ , we use the final value of  $x$  from the previous set of iterations as the initial condition for the next set (it is also wise to perturb the initial condition slightly to avoid getting caught on an unstable solution). The following MATLAB function produces a matrix of  $x$  values suitable for plotting in a bifurcation diagram. Each column of the matrix contains all the stable points of the iterative scheme for the corresponding value of  $\mu$ . As usual, the user needs to supply a function that calculates  $f(x)$  as in equation (14.4.1).

```

function Xs = bif_discrete(func, mu, x1)
% Function to produce the bifurcation diagram for an iterative
% map
% Inputs: func - handle to the iteration function f(x, mu)
%          mu - vector of values of the parameter mu
%          x1 - initial condition for x
% Outputs: Xs - matrix containing the stable fixed/periodic
%          points for each value of mu
Ntrans = 100; % iterations to get past transient behaviour
Nplot = 50;      % number of points to plot for each mu
nudge = 1e-6; % small perturbation to each initial condition
M = length(mu);
Xs = zeros(Nplot, M);
x = x1;           % set x to initial condition
for J = 1:M
    % Do Ntrans iterations with Jth value of mu:
    for K = 1:Ntrans
        x = func(x, mu(J));
    end
    % Do Nplot iterations, storing results in Jth column of X:
    Xs(1, J) = x;
    for K = 1:Nplot-1
        Xs(K+1, J) = func(Xs(K, J), mu(J));
    end
    % perturb final value slightly for next set of iterations:
    x = Xs(Nplot, J)+nudge;
end

```

The number of iterations (`Ntrans`) required to get past the transient behaviour, and the number of stable  $x$  points (`Nplot`) to store for each value of  $\mu$  can be adjusted as required.

**Example 14.4.1**

Use the MATLAB function `bif_discrete` to produce the bifurcation diagrams for the logistic map (13.3.2) shown in Figures 13.3.1 and 13.4.1.

First, we need a function that calculates  $f(x)$  for a given value of  $\mu$ :

```
function f = logistic_map(x, mu)
f = mu*x*(1-x);
```

Next we need to set up a suitable vector of  $\mu$  values. The range for  $\mu$  in Figure 13.3.1 is between 3 and 4. If we use a step size of 0.001, then the appropriate command is

```
>> mu = 3:0.001:4;
```

Finally, we just need to call `bif_discrete` to generate the  $x$  values, and then plot them to produce the diagram:

```
>> Xs = bif_discrete(@logistic_map, mu, 0.4);
>> plot(mu, Xs, 'k.', 'MarkerSize', 1)
```

(The '`k.`' option tells MATLAB to plot the graph using disconnected points of the same colour (black), rather than a connected line, and reducing the `MarkerSize` to 1 uses smaller points to produce a higher-quality plot.) This produces the diagram shown in Figure 13.3.1. Figure 13.4.1 can be reproduced simply by changing the range for  $\mu$ . □

## 14.5 Strange attractors and chaos

Fixed points and limit cycles are some of the fundamental objects of dynamical systems. However, as seen in Chapter 13, more complex types of dynamics can also occur in nonlinear systems. We have seen how attracting (stable) solutions (including fixed points and limit cycles) can be found by solving a system of differential equations numerically and disregarding any transient behaviour. Having found an attractor in this way, we may wish to know whether we are seeing regular, periodic behaviour or something more complex, such as chaos. One means of making this distinction is to examine the **sensitivity to initial conditions**, i.e., the tendency for two nearby initial conditions to result in quite different solutions. This sensitivity may be quantified by the **Lyapunov exponents** of the system, which measure the rate of divergence (or convergence) of nearby orbits. If the divergence/convergence for two initial conditions  $\mathbf{y}_1(0)$  and  $\mathbf{y}_2(0)$  is exponential in time then we may write

$$|\mathbf{y}_2(t) - \mathbf{y}_1(t)| = |\mathbf{y}_2(0) - \mathbf{y}_1(0)| e^{\lambda t}.$$

The constant  $\lambda$  is the **maximal Lyapunov exponent** (for an  $n$ -dimensional system, there are in fact  $n$  Lyapunov exponents, but it is the largest one that controls the rate of separation of orbits). If  $\lambda > 0$  then the orbits separate as  $t$  increases; if  $\lambda < 0$  then the orbits become closer together as  $t$  increases. Rearranging for  $\lambda$  and letting  $t$  tend to zero gives the following definition of the maximal Lyapunov exponent:

$$\lambda = \lim_{t \rightarrow 0} \frac{1}{t} \ln \frac{|\mathbf{y}_2(t) - \mathbf{y}_1(t)|}{|\mathbf{y}_2(0) - \mathbf{y}_1(0)|}.$$

If the maximal Lyapunov exponent of an attractor is positive, this is indicative of chaotic dynamics and the attractor is called a **strange attractor**.

The maximal Lyapunov exponent of an attractor can be calculated numerically as follows.

1. Start with an initial condition  $\mathbf{y}_1(0)$  that is in the attractor's basin of attraction (i.e., that tends towards the attractor). Calculate the numerical solution for sufficient time  $t_0$  that the solution  $\mathbf{y}_1(t_0)$  is on the attractor. This may require some knowledge of the system, and the characteristic timescale for any transients.
2. Choose a point  $\mathbf{y}_2(t_0)$  in phase space that is separated from  $\mathbf{y}_1(t_0)$  by a small distance  $d_0$ , i.e., such that  $|\mathbf{y}_2(t_0) - \mathbf{y}_1(t_0)| = d_0$ .
3. Advance both solutions by a short time  $\delta t$  and calculate the new separation:

$$d_1 = |\mathbf{y}_2(t_0 + \delta t) - \mathbf{y}_1(t_0 + \delta t)|.$$

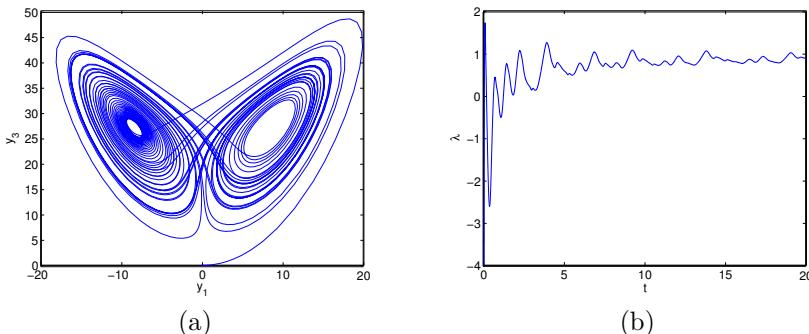
4. Evaluate

$$\lambda = \frac{1}{\delta t} \ln \frac{d_1}{d_0}.$$

5. Readjust the orbit  $\mathbf{y}_2$  so that its separation from  $\mathbf{y}_1$  is equal to  $d_0$ , without changing the direction from  $\mathbf{y}_1$  to  $\mathbf{y}_2$  in phase space, which is the direction of maximum expansion. This may be done by setting

$$\mathbf{y}_{2(\text{new})} = \mathbf{y}_1 + \frac{d_0}{d_1} (\mathbf{y}_{2(\text{old})} - \mathbf{y}_1).$$

6. Repeat steps 3–5 and take the average of the values of  $\lambda$  calculated in step 4. After a sufficient number of repetitions, this average should settle down to the maximal Lyapunov exponent, which should be reasonably independent of the initial condition  $\mathbf{y}_1(0)$  (provided it converges to the desired attractor), and the values of  $d_0$  and  $\delta t$ .



**FIGURE 14.5.1:** (a) Graph of  $y_1(t)$  against  $y_3(t)$  for the Lorenz equations (14.5.1)–(14.5.3) with  $\sigma = 10$ ,  $r = 28$ ,  $b = 8/3$  and  $\mathbf{y}(0) = (0, 1, 0)$ . The solution quickly converges to the Lorenz attractor, with its characteristic butterfly shape. (b) Maximal Lyapunov exponent  $\lambda$  for the Lorenz attractor computed numerically. The running average of  $\lambda$  converges to approximately 0.9. Since  $\lambda > 0$ , orbits on the attractor are sensitive to initial conditions, indicating chaotic dynamics.

**Example 14.5.1**

## The Lorenz equations

$$\dot{y}_1 = \sigma(y_2 - y_1), \quad (14.5.1)$$

$$\dot{y}_2 = ry_1 - y_2 - y_1 y_3, \quad (14.5.2)$$

$$\dot{y}_3 = y_1 y_2 - b y_3, \quad (14.5.3)$$

are a simplified model of convection rolls in the atmosphere. (They also describe the motion of a waterwheel that consists of a number of leaky buckets, with a steady supply of water from the top.) When  $\sigma = 10$ ,  $r = 28$  and  $b = 8/3$ , find the attractor for the system (14.5.1)–(14.5.3) with the initial condition  $\mathbf{y}(0) = (0, 1, 0)$ , and determine its maximal Lyapunov exponent.

To find the attractor, we simply find the numerical solution of the differential equations with the specified initial condition. Plotting  $y_1(t)$  against  $y_3(t)$  gives the graph shown in Figure 14.5.1(a). Note that the orbit appears to cross itself in numerous places. However, if we were to plot the orbit in its true, three-dimensional phase space  $(y_1, y_2, y_3)$ , we would see that in fact this never happens.

The procedure for computing the maximal Lyapunov exponent of this attractor can be implemented by the following MATLAB function:

```
function mle = lyapunov(dydt, y0, t_trans, d0, delta_t, t_max)
% Function to calculate the maximal Lyapunov exponent of
```

```
% an attractor of a system dy/dt=f(y, t) (it is implicitly
% assumed that f(y, t) is independent of t)
% Inputs: dydt - handle to a function that calculates dy/dt
%          y0 - initial condition in the basin of attraction
%          t_trans - time for transients to disappear
%          d0 - initial separation of two orbits
%          delta_t - time step
%          t_max - length of time (after t_trans) to integrate for
%                  (steps 3-5 are repeated until this time is reached)
% Outputs: mle- running average of the maximal Lyapunov
%           exponent at each time step

% integrate to get rid of transient behaviour:
[t, Y] = ode45(dydt, [0, t_trans], y0);
y1 = Y(size(Y, 1), :)'; % final value of solution is y1
y2 = y1+[d0; zeros(length(y1)-1, 1)]; % perturb by d0 to get y2
N_steps = t_max/delta_t; % number of steps required
sl = zeros(1, N_steps);
sum = 0;
for I = 1:N_steps
    % integrate both orbits by a time delta_t:
    y1 = rk4step(dydt, [0, delta_t], y1);
    y2 = rk4step(dydt, [0, delta_t], y2);
    d1 = norm(y2-y1); % new separation
    lambda = log(d1/d0)/delta_t; % Lyapunov exponent
    sum = sum+lambda; % running sum of Lyapunov exponents
    sl(I) = sum;
    y2 = y1+(y2-y1)*d0/d1; % renormalise y2 so separation is d0
end
% divide running sum by number of iterations to get running avg:
mle = sl./(1:N_steps);
```

Here, rather than using one of MATLAB's built-in solvers, which can give inaccurate results in chaotic systems, `rk4step` is a simple function for calculating one step of the fourth-order Runge-Kutta method for solving differential equations, with a pre-defined step size:

```
function y = rk4step(dydt, tspan, y0)
% Function for calculating one step of the fourth-order
% Runge-Kutta method on a system dy/dt = f(t, y)
% Inputs: dydt - handle to a function that calculates dy/dt
%          tspan - vector containing the starting and
%                  finishing values of t
%          y0 - initial condition vector
% Outputs: y - solution vector after one RK4 step
delta_t = tspan(2)-tspan(1);
```

```

k1 = dydt(tspan(1), y0);
k2 = dydt(0.5*(tspan(1)+tspan(2)), y0+0.5*delta_t*k1);
k3 = dydt(0.5*(tspan(1)+tspan(2)), y0+0.5*delta_t*k2);
k4 = dydt(tspan(2), y0+delta_t*k3);
y = y0+delta_t/6*(k1+2*k2+2*k3+k4);

```

As usual, we require a function for calculating the right-hand sides of our differential equations (14.5.1)–(14.5.3) (since we are not doing a bifurcation analysis here, the value of the parameters  $\sigma$ ,  $r$  and  $b$  are fixed):

```

function f = lorenz(t, y)
sigma = 10; r = 28; b = 8/3;
f = zeros(3,1);
f(1) = sigma*(y(2)-y(1));
f(2) = r*y(1)-y(2)-y(1)*y(3);
f(3) = y(1)*y(2)-b*y(3);

```

In addition, the function `lyapunov` requires the initial condition  $y(0)$ , the transient time (i.e., the time required for the orbit to settle down onto the attractor), the values of  $d_0$  and  $\delta t$ , and the total length of time for which to repeat steps 3–5 of the procedure above. These can be supplied with the following MATLAB command.

```
>> lambda = lyapunov(@lorenz, [0; 1; 0], 100, 1e-8, 1e-4, 20);
```

This returns into the variable `lambda` a running average of the maximal Lyapunov exponent calculated after successive repetitions of steps 3–5. Plotting this running average against  $t$  gives the graphs shown in Figure 14.5.1(b). As the graph shows,  $\lambda$  converges to a value of approximately 0.9. Since  $\lambda$  is positive, orbits on the attractor are sensitive to initial conditions, demonstrating that it is a strange attractor, with associated chaotic dynamics. □

## 14.6 Stability analysis of partial differential equations

A complete account of the numerous numerical methods for solving partial differential equations is beyond the scope of this book. However, this section contains an outline of a method for analysing the stability of steady states of partial differential equations. Given a partial differential equation

$$\frac{\partial u}{\partial t} = \mathcal{L}(u), \quad (14.6.1)$$

where  $\mathcal{L}$  is some differential operator, steady solutions of the equation satisfy

$$0 = \mathcal{L}(u). \quad (14.6.2)$$

Equation (14.6.2) may be solved numerically by approximating the solution  $u(x)$  to (14.6.2) by a vector  $\mathbf{u} = (u_0, \dots, u_n)$  which satisfies

$$\mathbf{0} = \mathbf{f}(\mathbf{u}), \quad (14.6.3)$$

where  $\mathbf{f}(\mathbf{u})$  is the approximation to  $\mathcal{L}(u)$  obtained by replacing derivatives by finite differences (see Section 3.8). Combining (14.6.3) with the boundary conditions on  $u$  gives a system of algebraic equations that, given a sufficiently accurate initial approximation, may be solved by Newton's method as in Section 5.9.

How can we determine the stability of the steady state? One approach would be to start with an initial condition that is close to the steady state, then numerically solve (14.6.1) to see if the solution converges towards the steady state, or moves away from it. However, this may be computationally expensive. A more efficient approach is to examine the eigenvalues of the Jacobian matrix that is calculated during the Newton's method routine. Just as with a fixed point of a system of ordinary differential equations, if all the eigenvalues of the Jacobian evaluated at the steady state have negative real part, then the steady state is stable, otherwise it is unstable. The reason for this is as follows. Just as we replaced  $\mathcal{L}(u)$  by a finite difference approximation  $\mathbf{f}(\mathbf{u})$  in the steady-state equation (14.6.2), we may do likewise in the full partial differential equation (14.6.1). This gives a system of  $n - 1$  ordinary differential equations (where  $n$  is the number of grid points being used), of the type studied in Section 14.1:

$$\frac{d\mathbf{u}}{dt} = \mathbf{f}(\mathbf{u}). \quad (14.6.4)$$

Clearly, the steady state of this system of differential equations is given by the solution  $\mathbf{u}^*$  to equation (14.6.3), and the stability of this steady state is determined by the Jacobian of  $\mathbf{f}(\mathbf{u})$  evaluated at  $\mathbf{u}^*$ .

In approximating the solution  $u(x, t)$  by a time-dependent vector  $\mathbf{u}(t)$ , we are approximating the continuous spectrum of eigenvalues of the partial differential equation by the  $n$  discrete eigenvalues of the corresponding system of ordinary differential equations. As we decrease the space step  $\delta x$  and hence increase the number  $n$  of mesh points, the set of eigenvalues should tend towards a continuous curve in the complex plane. If this curve lies entirely to the left of the imaginary axis, the steady state is stable, otherwise it is unstable.

## 14.7 Notes

An excellent introduction to dynamical systems is S.H. Strogatz, *Nonlinear dynamics and chaos*, Westview Press, MA, 1994, and several of the examples in this Chapter are based on material in this book.

The model in Example 14.3.1 is due to I. Lengyel, G. Rabai and I.R. Epstein. Experimental and modelling study of oscillations in the chlorine dioxide–iodine–malonic acid reaction. *J. Am. Chem. Soc.* **112**, 9104–9110, 1990.

The method for numerical calculation of the maximal Lyapunov exponent is based on that of J.C. Sprott, *Chaos and time-series analysis*, Oxford University Press, 2003 (section 5.6).

The Lorenz equations were originally analysed in E. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141, 1963.

For more mathematical detail on path-following methods, and possible improvements to the basic method described here, see J. Brindley, C. Kaa-Petersen and A. Spence. Path-following methods in bifurcation problems. *Physica D* **34**, 456–461, 1989.

## AUTO

Although relatively simple systems of the type studied in this chapter may be analysed numerically by writing simple programs in any computer language such as MATLAB, there are dedicated software packages available for more complex problems. The original numerical bifurcation analysis software is a package called AUTO, which is freely available, along with complete documentation, from: <http://indy.cs.concordia.ca/auto>.

---

## Exercises

- 14.1 Consider the following model for a population of a prey ( $y_1$ ) and predator ( $y_2$ ) species

$$\begin{aligned}\dot{y}_1 &= y_1 \left( 1 - y_1 - \frac{ay_2(1 - e^{-y_1})}{y_1 + c} \right), \\ \dot{y}_2 &= ry_2 \left( 1 - \frac{y_2}{y_1 + c} \right).\end{aligned}$$

Find the fixed points (equilibria) of the system when  $a = 2$ ,  $c = 0.1$ ,  $r = 0.5$  by solving the steady-state equations numerically. Determine the stability of the fixed points. Check your answers by finding the solution of the differential equations for initial conditions near to each of the fixed points.

- 14.2 For the system (13.2.1)–(13.2.2), verify the existence of a stable limit cycle near the origin for  $\mu = 0.01$  by finding the solution numerically for a suitable initial condition. Follow the stable limit cycle as the value of  $\mu$  increases and draw a bifurcation diagram that plots the amplitude of the limit cycle against  $\mu$ .

- 14.3 A population of animals grows logistically at rate  $r > 0$  and is subject to external predation. If the intensity of predation decays as the population size grows, then the total size  $y$  of the population can be modelled by

$$\dot{y} = ry(1 - y) - \mu y e^{-y/2},$$

where  $\mu \geq 0$  is a parameter describing the level of predation.

- (a) Find the fixed points of the equation when  $\mu = 0$ , and determine their stability.
- (b) Fix a fixed value of  $r = 5$ , construct the bifurcation diagram for  $0 \leq \mu < 5$  and identify any bifurcation points.

- 14.4 Consider the system

$$\begin{aligned}\dot{y}_1 &= \mu - y_1 - \frac{8y_1y_2}{y_1^2 + 2}, \\ \dot{y}_2 &= y_1 \left( 1 - \frac{y_2}{y_1^2 + 2} \right),\end{aligned}$$

where  $\mu$  is the control parameter. Show that the system has a bifurcation at which a fixed point changes from stable to unstable. Draw the bifurcation diagram and determine the bifurcation point. Show that there is a stable limit cycle for parameter values for which the fixed point is unstable.

- 14.5 For the system in Exercise 14.1, produce a bifurcation diagram with respect to the parameter  $a$  by following each fixed point as the value of  $a$  changes in the range  $0 < a < 20$ , and plotting  $\sqrt{y_1^2 + y_2^2}$  against  $a$ . Assume all other parameter values are as in Exercise 14.1.

- 14.6 Plot the bifurcation diagram for the iteration scheme

$$x_{n+1} = \mu \sin(\pi x_n)$$

for values of  $\mu$  in the range 0 to 1.

- 14.7 Consider the iteration scheme

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

- (a) Find the largest value of  $\mu$  such that  $f(x)$  is always between 0 and 1 provided  $x$  is between 0 and 1.
- (b) Plot the bifurcation diagram for a suitable range of values of  $\mu$ .

- 14.8 (a) Consider the Lorenz equations (14.5.1)–(14.5.3). Find the fixed points of the system analytically. Write down the Jacobian matrix for the system and substitute the expression for each fixed point into the Jacobian matrix.

- (b) Write a simple MATLAB program that calculates the Jacobian matrix you wrote down in part (a), and its eigenvalues, for each fixed point, for  $\sigma = 10$ ,  $b = 8/3$  and  $r = 15$ . Now get your program to repeat this task for a range of values of  $r$  from 15 to 35. For each fixed point, plot the maximum real part of the eigenvalues against  $r$ . What does your numerical analysis tell you about the stability of the fixed points?
- 14.9 Use the MATLAB code provided in Section 14.5 to calculate the maximal Lyapunov exponent for the Lorenz equations (14.5.1)–(14.5.3), for a range of values of the parameter  $r$  between  $r = 15$  and  $r = 35$ . How do your findings relate to your answer to Exercise 14.8?

# **Chapter 15**

---

## **Growth of Tumours**

---

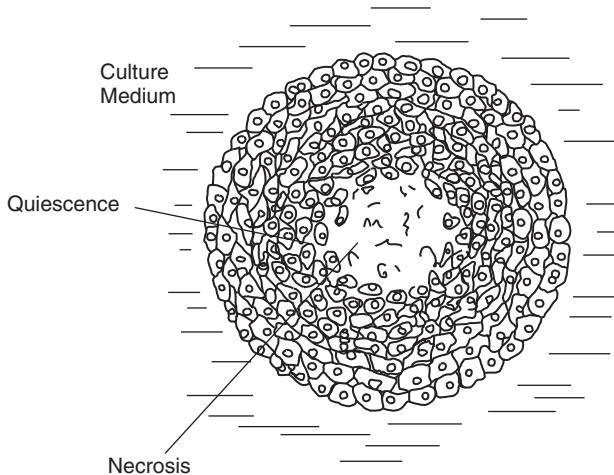
### **15.1 Introduction**

Tumours can arise from the cells of nearly all types of body tissue and this diversity of origin is largely responsible for the wide variety of the structural appearance of tumours. In this chapter we describe the main features of tumour growth and ways in which to model the processes involved. During the course of discussion we shall draw attention to some of the recent ideas and developments. In the earliest stages of development, tumour growth seems to be regulated by direct diffusion of nutrients and wastes from and to surrounding tissue. When a tumour is very small, every cell receives nourishment by simple diffusion and the growth rate is exponential in time. However, this stage cannot be sustained because, as a nutrient is consumed, its concentration must decrease towards the centre of the tumour. Eventually the concentration of a vital nutrient near the centre will fall below a critical level insufficient to sustain cell life. A central necrotic core then develops. The rate of growth of the tumour then falls away and it becomes more difficult to obtain nourishment and to dispose of wastes solely by diffusion.

Unfortunately this is not the end of the process. A majority of tumours exhibit the phenomenon of angiogenesis marking the transition from the relatively harmless and localised avascular state described above to the more dangerous, vascular state, wherein the tumour develops the ability to proliferate, invade surrounding tissue and metastasize, i.e., spread, to distant parts of the body.

In its early stages the tumour grows only a few millimetres in diameter and consists of an outer shell several cell layers thick, which grows and proliferates (see Figure 15.1.1). As we move into the interior of the tumour, the proliferation of cells decreases markedly until we reach a region of quiescent nondividing cells and further inward until we reach the central core of necrotic debris in various stages of disintegration.

We have discussed various processes of diffusion before, but here we are faced with a completely new situation, namely that the boundary of the tumour is moving and is unknown except in the initial stages of growth. Indeed the basic problem to be addressed is to formulate mathematical models that enable us to focus attention on the movement of the tumour outer cell layer,



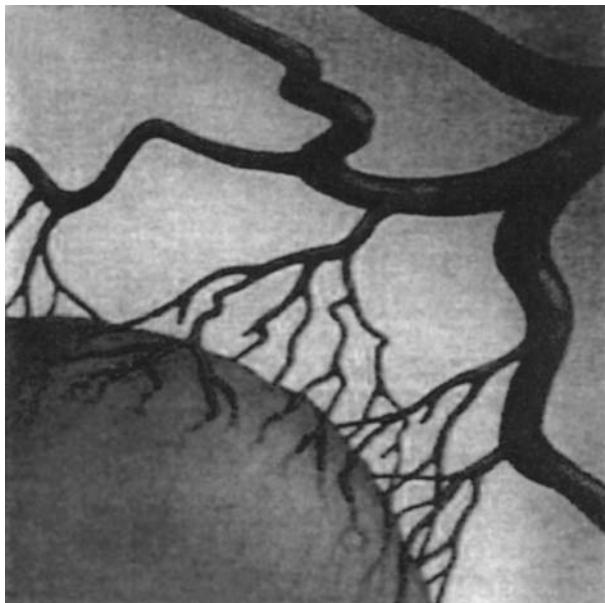
**FIGURE 15.1.1:** An avascular tumour cell colony.

to track its movement and to examine its stability.

The question of stability is extremely important; it is central to determining whether a tumour is nonmalignant and growth limited or, like cancer cells, exhibits unlimited growth and becomes malignant. Most tumours involve some vascularisation due to angiogenesis. This is the process by which tumours induce blood vessels from the host tissue to sprout capillaries which migrate towards and ultimately penetrate the tumour, providing it with a circulating blood supply (see Figure 15.1.2).

To be able to control vascularisation in cancer cells is of paramount importance. It is well documented that cancer cells produce a variety of chemical growth factors called **tumour angiogenesis factors** (or TAF) that stimulate the formation of new capillaries. That is, as the tumour approaches its diffusion-limited size, the TAF initiates angiogenesis. The malignancy becomes vascularised and perfusion replaces diffusion as the dominant mechanism for the supply of nutrients and the removal of wastes. Once the tumour connects with the circulatory system all constraints imposed by diffusion are eliminated and subsequent growth is explosive. If one could block the chemical messages for vascularisation sent from the tumour to surrounding tissue, it might be possible to maintain the tumour indefinitely in its dormant or non-malignant state or even to kill the tumour completely by cutting off its blood supply. Currently there are drugs undergoing clinical trials that are designed specifically to starve malignant tumours. These are called anti-angiogenesis drugs and include angiostatin and endostatin.

Mathematical modelling of tumour growth and the processes of angiogenesis are being actively pursued and hopefully will help biomedical scientists and clinicians to develop strategies with which to better understand and combat



**FIGURE 15.1.2:** A vascularised tumour as the result of angiogenesis.

this life-threatening disease.

The two mathematical models we develop here are concerned with avascular tumour growth.

The first model, referred to as **Model I**, is the earliest mathematical model and is based on the following simplifying assumptions.

- The cell colony and surrounding medium are essentially in a diffusive equilibrium state at all times. The tumour has a three-layer structure comprising an outer layer of live proliferating cells enveloping a thin inner layer of quiescent nonproliferating cells, which, in turn, envelopes a large core of necrotic debris.
- Cells proliferate as long as the available concentration of nutrient supply denoted by  $\sigma(x, y, z, t)$  remains above a critical level  $\sigma_1$ . Cells die when  $\sigma$  falls below a critical level  $\sigma_2$ . In the quiescent region  $\sigma_2 < \sigma < \sigma_1$ . The thickness  $h$  of the layer of live proliferating cells depends on  $\sigma_1$  and the value of  $\sigma$  at the outer surface of the tumour. Experimental evidence suggests that

$$h = \begin{cases} \nu\sqrt{\sigma - \sigma_1}, & \text{for } \sigma > \sigma_1, \\ 0 & \text{for } \sigma < \sigma_1, \end{cases} \quad (15.1.1)$$

where  $\nu$  is a positive constant.

- (c) If  $dA$  is an element of surface area of the tumour, then the incremental volume of live cells  $dV = h dA$  creates new cell volume at the rate  $\beta h dA$  where  $\beta$  is a constant. Nutrient is consumed by this volume at the rate  $\gamma h dA$  where  $\gamma$  is another constant.
- (d) Proliferating cells become quiescent when the nutrient supply  $\sigma$  lies in the region  $\sigma_2 < \sigma < \sigma_1$  and the rate of gain of quiescent mass per unit volume is constant.
- (e) Necrotic debris disintegrates continually into simpler compounds. The rate of loss of necrotic mass per unit volume is constant.
- (f) A surface tension force  $T$  proportional to the mean curvature  $\kappa$  of the boundary keeps the tumour a compact and continuous mass.
- (g) The birth or death of cells produces internal pressure differentials which cause the motion of cellular material. This is assumed to be governed by

$$\mathbf{q} = -\nabla P \quad (15.1.2)$$

where  $\mathbf{q}(x, y, z, t)$  is the cell velocity and  $P(x, y, z, t)$  is proportional to internal pressure. Indeed the colony is assumed to behave like an incompressible fluid composed of cells and cellular debris.

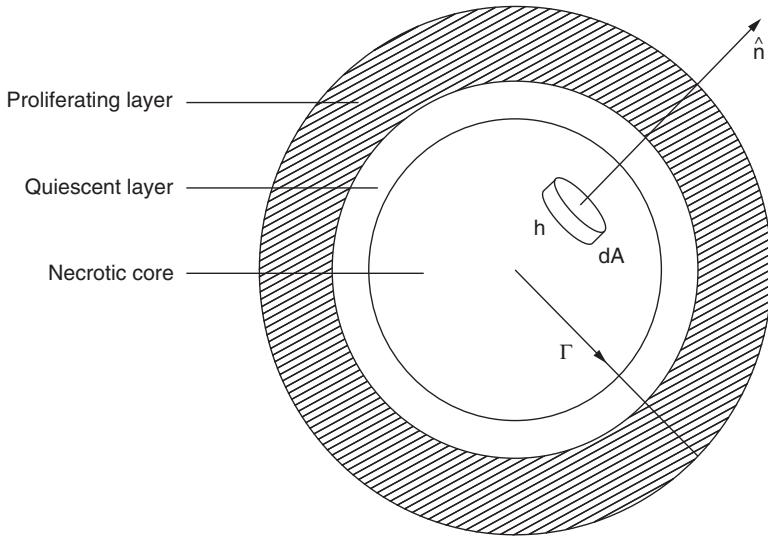
The second model, referred to as **Model II**, while being inspired by the insights gained from **Model I**, concentrates more on nutrient supply in contrast to pressure profiles. In this model assumptions (a), (d), (e) and (g) are retained. Assumption (b) is largely retained except for the assumption relating the thickness of the proliferating layer to nutrient supply through equation (15.1.1). Assumption (f) is replaced by the “Gibbs–Thomson” relation. This states that, on the boundary of live proliferating cells, the nutrient concentration is less than that in the surrounding tissue by a factor proportional to the mean curvature  $\kappa$  of the boundary, this being the energy needed to maintain intercellular bonds.

## 15.2 Mathematical Model I of tumour growth

In order to develop this mathematical model of tumour growth, we endeavour to combine diffusion processes with the above assumptions, to arrive at a set of equations which allow us to relate the dynamics of the surface of the tumour with variations in the nutrient concentration  $\sigma$  and the internal pressure  $P$ .

Suppose the outer surface is represented by the unknown functional equation

$$\Gamma(x, y, z, t) = 0. \quad (15.2.1)$$



**FIGURE 15.2.1:** Model of a tumour.

Similarly the outer surface of the necrotic core is represented by the unknown functional equations  $\Gamma_N(x, y, z, t) = 0$ .

Apply the law of conservation of mass to the elemental volume shown in Figure 15.2.1. This says that since  $h$  is small, the mass/volume flow out of the surface  $dA$  of the elemental volume  $dV$ , namely  $(\mathbf{q}_+ \cdot \hat{n} - \mathbf{q}_- \cdot \hat{n})dA$ , equals the rate of mass/volume production within this small volume, namely  $\beta h dA$ . Thus

$$\mathbf{q}_+ \cdot \hat{n} = \mathbf{q}_- \cdot \hat{n} + \beta h, \quad \text{on } \Gamma = 0. \quad (15.2.2)$$

Similarly the rate of nutrient diffusion into  $dV$  (with diffusion coefficient  $k$ ) through the outer surface is  $k\hat{n} \cdot \nabla \sigma dA$ , which is equal to the rate at which nutrient is consumed in this small volume, namely  $\gamma h dA$ . Thus

$$k\hat{n} \cdot \nabla \sigma = \gamma h, \quad \text{on } \Gamma = 0. \quad (15.2.3)$$

Notice that since  $\sigma \leq \sigma_1$  in the quiescent region and the necrotic core, there is no diffusive transport from the interior.

Suppose the proliferation rate of new cells is so large that its product with small quantities, such as outer shell thickness, is of order one, i.e.,

$$\beta h = \beta \nu \sqrt{(\sigma - \sigma_1)} = \lambda \sqrt{(\sigma - \sigma_1)}, \quad \lambda = O(1)$$

and

$$\gamma h = \gamma \nu \sqrt{(\sigma - \sigma_1)} = \mu \sqrt{(\sigma - \sigma_1)}, \quad \mu = O(1).$$

Next we suppose that the pressure  $P$  and tangential velocity components are continuous across each of the surfaces  $\Gamma = \Gamma_N = 0$ .

For example, at the outer surface  $\Gamma = 0$

$$\begin{aligned} P_+ &= P_- = P, \\ \mathbf{q}_+ \times \hat{n} &= \mathbf{q}_- \times \hat{n}. \end{aligned} \quad (15.2.4)$$

From assumption (f) the pressure on the surface of the tumour must equal the surface tension  $T$  and thus in turn is proportional to the mean curvature  $\kappa$ , i.e.,

$$P = \alpha\kappa \quad \text{on } \Gamma = 0, \quad (15.2.5)$$

where  $\alpha$  is a constant.

If a typical point on the outer surface of the tumour is represented by the vector  $\mathbf{r}$ , then the motion of  $\Gamma = 0$  is represented by

$$\frac{d\mathbf{r}}{dt} = \mathbf{q}_+ \quad (15.2.6)$$

where  $\Gamma(x, y, z, t) = 0$  is assumed known.

If we let  $\mathbf{q}$  denote cell velocity within the tumour and  $S(x, y, z, t)$  the cell loss rate at a point inside the tumour then conservation of mass can be written as

$$\nabla \cdot \mathbf{q} = -S. \quad (15.2.7)$$

The cell loss rate  $S$  is modelled in the following way. Cell loss due to apoptosis, programmed cell death, is restricted to the proliferating and quiescent region and occurs at the constant rate  $S_1$ . Cell loss due to necrosis is assumed to occur at the constant rate  $S_2$ . In terms of the **Heaviside** step function  $H$  we can write  $S$  in the concise form

$$S(x, y, z, t) = S_1 H(|\mathbf{r}| - |\mathbf{r}_N|) + S_2 H(|\mathbf{r}_N| - |\mathbf{r}|) \quad (15.2.8)$$

where  $\mathbf{r}_N$  is a point on the surface of the necrotic region  $\Gamma_N = 0$ .

The equation for nutrient concentration  $\sigma$ , which is assumed to be in diffusive equilibrium, is

$$\nabla^2 \sigma = 0, \quad (15.2.9)$$

outside and within the tumour colony.

There are several problems that can be investigated with this model, including the effect of a nearby source of nutrient or the presence of another tumour cell colony. It is also of importance to examine the effect of the presence of an impermeable wall (e.g., artery). Here we shall consider the surrounding medium to be large in comparison with the tumour size and that there is a constant supply of the nutrient, i.e.,

$$\sigma \rightarrow \sigma_\infty \quad \text{as } |\mathbf{r}| \rightarrow \infty. \quad (15.2.10)$$

It is convenient to bring our mathematical model together in the following collection of equations and boundary conditions:

$$\begin{aligned} \nabla^2 P &= S \text{ inside } \Gamma = 0, \\ \nabla^2 \sigma &= 0 \text{ in the tumour and the surrounding medium,} \end{aligned} \quad (15.2.11)$$

where the first of equations (15.2.11) comes from combining (15.1.2) and (15.2.7).

On the boundary  $\Gamma = 0$  of the tumour we have

$$P = \alpha\kappa, \quad (15.2.12)$$

$$\mathbf{q}_+ \cdot \hat{n} = -\hat{n} \cdot \nabla P + \lambda\sqrt{(\sigma - \sigma_1)}, \quad (15.2.13)$$

$$\mathbf{q}_+ \times \hat{n} = -\nabla P \times \hat{n}, \quad (15.2.14)$$

$$\hat{n} \cdot \nabla \sigma = \mu\sqrt{(\sigma - \sigma_1)}. \quad (15.2.15)$$

The boundary surface is defined by

$$\frac{d\mathbf{r}}{dt} = \mathbf{q}_+ \quad (15.2.16)$$

and the initial configuration is given by

$$\mathbf{r} = \mathbf{a}, \quad (15.2.17)$$

at  $t = 0$  where  $\mathbf{a}$  is assumed known.

Furthermore  $P$  and  $\sigma$  are continuous together with their normal derivatives across the surface

$$\Gamma_N = 0.$$

Finally

$$\sigma = \sigma_2 \quad \text{on } \Gamma_N = 0. \quad (15.2.18)$$

The set of equations (15.2.11)–(15.2.18) constitutes a **moving boundary problem** which is usually very difficult to solve both analytically and computationally. However, for some prescribed geometric configurations we can solve the equations exactly. This is the subject of the following section.

### 15.3 Spherical tumour growth based on Model I

Suppose, as is commonly observed *in vitro*, that the tumour is initially a sphere (of radius  $a$ ) and continues to grow as a sphere. In this situation an exact solution may be obtained. Because the tumour maintains its spherical shape the equation of the outer surface of proliferating cells can be represented by

$$r = R(t), \quad (15.3.1)$$

where  $R(t)$  denotes the radius of the tumour at time  $t$ . Clearly  $R(0) = a$ . Equations (15.2.11) expressed in spherical polar coordinates reduce, because

of spherical symmetry, to

$$\begin{aligned} \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial P}{\partial r} \right) &= S, \quad r \leq R(t), \\ \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \sigma}{\partial r} \right) &= 0. \end{aligned} \quad (15.3.2)$$

In order to proceed with the solution we have to construct appropriate solutions in each of the proliferating cell regions and the necrotic core, respectively.

Region:  $R_N < r \leq R$ :

Here we have to solve the equations

$$\begin{aligned} \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial P}{\partial r} \right) &= S_1, \quad R_N < r \leq R(t), \\ \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \sigma}{\partial r} \right) &= 0. \end{aligned} \quad (15.3.3)$$

From these we find that

$$P = S_1 \frac{r^2}{6} + \frac{A_1}{r} + B_1 \quad (15.3.4)$$

and

$$\sigma = \frac{C_1}{r} + D_1, \quad (15.3.5)$$

where  $A_1, B_1, C_1$  and  $D_1$  are constants to be determined by the boundary conditions. Using the condition (15.2.12) we obtain from (15.3.3)

$$S_1 \frac{R^2}{6} + \frac{A_1}{R} + B_1 = \frac{\alpha}{R}, \quad (15.3.6)$$

on noting that the mean curvature  $\kappa = 1/R$ . From (15.2.10) and (15.3.5) we find

$$D_1 = \sigma_\infty. \quad (15.3.7)$$

Region:  $0 < r \leq R_N$ : In this region we find

$$\begin{aligned} P &= S_2 \frac{r^2}{6} + \frac{A_2}{r} + B_2, \\ \sigma &= \frac{C_2}{r} + D_2. \end{aligned} \quad (15.3.8)$$

Here we require both  $P$  and  $\sigma$  to exist at  $r = 0$  and so we must take  $A_2 = C_2 = 0$ . On the boundary  $r = R_N$  we demand that

$$\begin{aligned} S_2 \frac{R_N^2}{6} + B_2 &= S_1 \frac{R_N^2}{6} + \frac{A_1}{R_N} + B_1, \\ S_2 \frac{R_N}{3} &= S_1 \frac{R_N}{3} - \frac{A_1}{R_N^2} \end{aligned} \quad (15.3.9)$$

and also

$$\begin{aligned} C_1 &= (\sigma_2 - \sigma_\infty)R_N, \\ D_2 &= \sigma_2. \end{aligned}$$

In principle we now have enough information to solve for all the unknown quantities  $A_i, B_i, (i = 1, 2)$ , provided we know the boundary radius  $R_N$ . Since  $A_2 = 0$  the three equations (15.3.6) and (15.3.9) can be solved for  $A_1, B_1, B_2$ .

The all important question now is to determine how the boundary of the tumour,  $r = R(t)$ , evolves with time. To do this we need to use equations (15.2.13) and (15.2.16). First of all combine (15.2.13) and (15.2.15) to get

$$q = -\frac{\partial P}{\partial r} + \frac{\lambda}{\mu} \frac{\partial \sigma}{\partial r}$$

on  $r = R(t)$  which when combined with (15.2.16) gives the evolution

$$\frac{dr}{dt} = -\frac{\partial P}{\partial r} + \frac{\lambda}{\mu} \frac{\partial \sigma}{\partial r}. \quad (15.3.10)$$

We now carry out the task of solving for all the unknown quantities  $A_1, B_1, B_2$ . From (15.3.9) we can solve directly for  $A_1$  to get

$$A_1 = (S_1 - S_2) \frac{R_N^3}{3}. \quad (15.3.11)$$

Knowing  $A_1$  we can use equation (15.3.6) to find  $B_1$ , namely,

$$B_1 = \frac{\alpha}{R} - S_1 \frac{R^2}{6} - (S_1 - S_2) \frac{R_N^3}{3R}. \quad (15.3.12)$$

With  $A_1, B_1$  known, the first of equations (15.3.9) can be used to get  $B_2$ , i.e.,

$$\begin{aligned} B_2 &= \frac{\alpha}{R} - S_1 \frac{R^2}{6} - (S_1 - S_2) \frac{R_N^3}{3R} + (S_1 - S_2) \frac{R_N^2}{2}, \\ &= \frac{\alpha}{R} - S_1 \frac{R^2}{6} + (S_1 - S_2)(3R - 2R_N) \frac{R_N^2}{6R}. \end{aligned} \quad (15.3.13)$$

Having determined all the quantities  $A_i, B_i, C_i, D_i, i = 1, 2$  we summarise the pressure and nutrient distributions in each of the two regions as follows:

Region:  $R_N < r \leq R$ :

$$\begin{aligned} P &= \frac{\alpha}{R} - \frac{S_1}{6}(R^2 - r^2) + (S_1 - S_2) \frac{R_N^3}{3} \left( \frac{1}{r} - \frac{1}{R} \right), \\ \sigma &= \frac{(\sigma_2 - \sigma_\infty)}{r} R_N + \sigma_\infty. \end{aligned} \quad (15.3.14)$$

Region:  $0 < r \leq R_N$ :

$$\begin{aligned} P &= \frac{\alpha}{R} + S_2 \frac{r^2}{6} - S_1 \frac{R^2}{6} + (S_1 - S_2) \frac{R_N^2}{6R} (3R - 2R_N), \\ \sigma &= \sigma_2. \end{aligned} \quad (15.3.15)$$

To determine the evolution of the outer boundary,  $R(t)$ , of the tumour we substitute (15.3.14) in (15.3.10) to get

$$\frac{dR}{dt} = -S_1 \frac{R}{3} + (S_1 - S_2) \frac{R_N^3}{3R^2} - \frac{\lambda}{\mu} (\sigma_2 - \sigma_\infty) \frac{R_N}{R^2}. \quad (15.3.16)$$

This nonlinear ordinary differential equation can be solved numerically provided  $R_N$  is known. We can find  $R_N$  in terms of  $R$  by using (15.3.14) in (15.2.15) to obtain the quadratic equation

$$(\sigma_\infty - \sigma_2)^2 R_N^2 + \mu^2 (\sigma_\infty - \sigma_2) R_N - \mu^2 (\sigma_\infty - \sigma_1) R^4 = 0, \quad (15.3.17)$$

which has the positive solution

$$R_N = \frac{-\mu^2 R^3 + \mu R^2 \sqrt{\mu^2 R^2 + 4(\sigma_\infty - \sigma_1)}}{2(\sigma_\infty - \sigma_2)}. \quad (15.3.18)$$

This expression for  $R_N$  can now be substituted into the ordinary differential equation (15.3.16) for the determination of  $R(t)$ . Figure 15.3.1 shows the numerical solution of (15.3.16), (15.3.18) and also the outer boundary of the quiescent layer,  $R_Q$ , determined by calculating the depth of the thin layer  $h$  of proliferating cells, using (15.1.1), for the parameter values  $S_1 = 60$ ,  $S_2 = 100$  together with  $\lambda = 1$ ,  $\mu = 10$ ,  $\nu = 0.002$  and  $\sigma_1 = 0.7$ ,  $\sigma_2 = 0.5$ ,  $\sigma_\infty = 1$ .

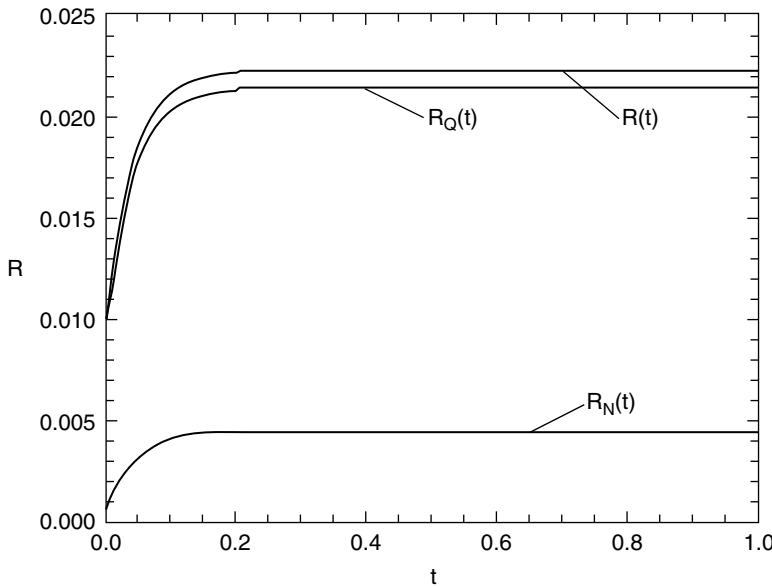
In the opening section of this chapter we gave a brief account of the processes of angiogenesis whereby a tumour may become vascularised and receive its own blood supply. The growth of the tumour is then freed of its diffusion limiting mechanism and tumour invasion and metastases occurs. Our model of avascular tumour growth can be modified to give some insight into this important phenomenon as follows.

If the tumour has become vascularised then we may assume that both the quiescent layer and the necrotic core are absent. In other words we may assume  $R_Q = R_N = 0$ . However assumption b) and equation (15.1.1) no longer apply. In other words, cell proliferation may take place throughout the tumour. If we assume cell proliferation due to mitosis occurs at the constant rate  $S_3$  then (15.2.8) may be replaced by

$$S(x, y, z, t) = (S_1 - S_3) H(|\mathbf{r}|). \quad (15.3.19)$$

This time the pressure and nutrient concentration within the tumour is given by

$$\begin{aligned} P &= \frac{\alpha}{R} + \frac{(S_3 - S_1)}{6} (R^2 - r^2), \\ \sigma &= \sigma_\infty. \end{aligned} \quad (15.3.20)$$



**FIGURE 15.3.1:** Evolution of spherical tumour with time.

Notice that we retain the cell loss due to apoptosis by including the rate  $S_1$ . The rate of growth of the tumour is now governed by the simple first order differential equation

$$\frac{dR}{dt} = \frac{(S_3 - S_1)}{3} R, \quad (15.3.21)$$

subject to the initial condition  $R(0) = a$ .

Thus

$$R(t) = a \exp((S_3 - S_1)t/3). \quad (15.3.22)$$

From this simple result we deduce that if the rate of cell production exceeds the rate of apoptosis ( $S_3 > S_1$ ) then the tumour grows exponentially fast, whereas if apoptosis exceeds cell mitosis ( $S_3 < S_1$ ) the tumour decays exponentially with time.

## 15.4 Stability of tumour growth based on Model I

Surface tension, as we have already mentioned, plays an important role in maintaining the compactness of the tumour. Any perturbation that allows surface tension to be overcome by pressure forces could be an important feature

in determining whether the tumour becomes vascularised and consequently malignant.

In other words, the question of whether the spherical tumour is stable with respect to small deviations from its spherical shape is of fundamental importance. The models described above can be examined for stability under such perturbations and, in the case of instability, lead to some quite dramatic deviations from the initial spherical shape. In some situations the tumour may even break into two or more pieces.

A detailed analysis of stability is beyond the scope of this book. Nevertheless we can indicate how such an analysis can be performed. To do this we motivate the ideas by considering a further modification of our general model. Specifically we consider a tumour with a very small proliferating outer layer enclosing a large necrotic core. In fact this model is the one originally developed by H.P. Greenspan (see the notes at the end of this chapter). In other words we take the cell loss rate to be

$$S(x, y, z, t) = S_2. \quad (15.4.1)$$

Carrying out the by now familiar analysis we find

$$P = \frac{S_2}{6}(r^2 - R^2) + \frac{\alpha}{R}, \quad r \leq R(t) \quad (15.4.2)$$

and

$$\sigma = \frac{D}{r} + \sigma_\infty, \quad r \geq R(t), \quad (15.4.3)$$

where the constant  $D$  is determined from the boundary condition (15.2.15) to be

$$D = \frac{1}{2}\mu R^2\{\mu R - \sqrt{[\mu^2 R^2 + 4(\sigma_\infty - \sigma_2)]}\}. \quad (15.4.4)$$

Notice that since there is no quiescent layer we have set  $\sigma_1 = \sigma_2$ . The rate of growth of the radius of the tumour is determined from (15.2.13) and (15.2.16) as the solution to the ordinary differential equation

$$\frac{dR}{dt} = -\frac{1}{2}S_2R + \lambda\sqrt{(\sigma_\infty - \sigma_2) + \frac{1}{2}\mu R[\mu R - (\mu^2 R^2 + 4(\sigma_\infty - \sigma_2))^{1/2}]}. \quad (15.4.5)$$

Without solving this equation we can estimate the ultimate size of a *stable* spherical tumour by setting  $\frac{dR}{dt} = 0$  to get the limiting size

$$R_\infty = \frac{3\lambda}{S_2} \left( \frac{\sigma_\infty - \sigma_2}{3\lambda\mu/S_2 + 1} \right)^{1/2}. \quad (15.4.6)$$

To indicate how a stability analysis should proceed let  $\bar{P}(r, t)$ ,  $\bar{\sigma}(r, t)$  and  $\bar{R}(r, t)$  be the quantities determined by (15.4.2), (15.4.3) and the solution to (15.4.5). Suppose these quantities are perturbed by amounts  $\epsilon\bar{P}(r, \theta, t)$ ,

$\epsilon\tilde{\sigma}(r, \theta, t)$  and  $\epsilon\tilde{\xi}(\theta, t)$  independent of the azimuthal angle  $\phi$  and where  $\epsilon$  is a small parameter. Under these perturbations the total pressure and nutrient concentration are represented by

$$\begin{aligned} P(r, \theta, t) &= P(r, t) + \epsilon\tilde{P}(r, \theta, t), \\ \sigma(r, \theta, t) &= \sigma(r, t) + \epsilon\tilde{\sigma}(r, \theta, t) \end{aligned} \quad (15.4.7)$$

and the surface of the tumour is given at any time by

$$\Gamma(r, \theta, t) = r - R(t) - \epsilon\xi(\theta, t) = 0. \quad (15.4.8)$$

On substituting (15.4.7) and (15.4.8) into equations (15.2.11)–(15.2.17) and equating coefficients of  $\epsilon$  to zero, we have the following problem to solve:

$$\begin{aligned} \nabla^2\tilde{P} &= 0, \quad r \leq R(t), \\ \nabla^2\tilde{\sigma} &= 0, \quad r \geq R(t), \end{aligned} \quad (15.4.9)$$

with  $\tilde{\sigma} \rightarrow 0$  as  $r \rightarrow \infty$ , while on  $r = R(t)$ ,

$$\begin{aligned} \frac{\partial\xi}{\partial t} &= -\left(\frac{\partial^2\tilde{P}}{\partial r^2}\xi + \frac{\partial\tilde{P}}{\partial r}\right) + \frac{\lambda}{2\sqrt{\tilde{\sigma}}}\left(\xi\frac{\partial\tilde{\sigma}}{\partial r} + \tilde{\sigma}\right), \\ \frac{\partial^2\tilde{\sigma}}{\partial r^2}\xi + \frac{\partial\tilde{\sigma}}{\partial r} &= \frac{\mu}{2\sqrt{\tilde{\sigma}}}\left(\xi\frac{\partial\tilde{\sigma}}{\partial r} + \tilde{\sigma}\right), \\ \frac{\partial\tilde{P}}{\partial r}\xi + \tilde{P} &= -\frac{\alpha}{2R^2}\left(\frac{\partial}{\partial\eta}(1-\eta^2)\frac{\partial\xi}{\partial\eta} + 2\xi\right), \end{aligned} \quad (15.4.10)$$

where  $\eta = \cos\theta$ . The solutions of (15.4.10) are time-dependent multiples of harmonic functions from which the equation for  $\xi(t)$  can be solved to show, for a range of parameter values, that if the harmonics are of sufficiently high order then  $\xi(t)$  has exponential growth. The tumour is then unstable and radically departs from its original spherical shape no matter how small  $\epsilon$  may be.

## 15.5 Mathematical Model II of tumour growth

The tumour colony, as in **Model I**, is assumed to have a multilayered structure with proliferating cells in the outer layer, quiescent cells in the adjacent middle layer and necrotic cells confined to the central core. Cells towards the centre of the tumour are deprived of sufficient nutrients and cease to proliferate. As the tumour grows, the proportion of each cell type changes. Assuming there is no increase in oxygen and nutrients available to the tumour then, as it enlarges, there is an increase in the quiescent layer, which in turn leads to

a further decrease in the minimum nutrient concentration towards the centre. More cells towards the centre of the tumour, being starved of nutrients, become necrotic.

In order to develop the mathematical model we again make some simplifying assumptions, in addition to assumptions (a)-(g) of Section 15.1.

- (h) Local nutrient concentration provides an accurate measure of cell density and type, with high nutrient levels indicating cell proliferation and lower levels indicating quiescence and then necrosis.
- (j) The tumour maintains its multilayered structure as it grows.
- (k) The nutrient concentration satisfies the Gibbs–Thomson condition on the outer boundary of live proliferating cells.

Now we consider the situation wherein tumour growth is controlled by three factors:

- i an externally supplied nutrient,  $\sigma$ , which is sufficient to promote cell proliferation;
- ii an internally supplied inhibitor,  $\omega$ , which may be regarded as a growth inhibitor produced by the tumour;
- iii an externally supplied inhibitor,  $\beta$ , which could be an anti-cancer drug.

Reaction–diffusion kinetics are used to model the evolution of  $\sigma$ ,  $\omega$  and  $\beta$ , with  $F_\sigma$ ,  $F_\omega$  and  $F_\beta$  their respective reaction rates. It is known that a typical chemical diffusion time scale is much shorter than a typical tumour cell doubling time so that, as the tumour grows,  $\sigma$ ,  $\omega$  and  $\beta$  rapidly redistribute throughout the new volume. We therefore assume that these quantities are in diffusive steady state. In other words the governing equations are:

$$0 = \nabla^2\sigma + F_\sigma = \nabla^2\omega + F_\omega = \nabla^2\beta + F_\beta, \quad (15.5.1)$$

in which the assumed diffusion coefficients have been absorbed into the reaction terms.

Regarding the tumour as an incompressible fluid, local changes in the cell population due to proliferation and apoptosis (programmed cell death) of cells will induce motion of neighbouring cells. Similarly to **model I** we denote by  $S(\sigma, \omega, \beta)$  the loss rate at a point inside the tumour and denote the cell velocity by  $\mathbf{q}$ .

By applying the law of mass conservation we obtain the equation

$$\nabla \cdot \mathbf{q} = -S(\sigma, \omega, \beta). \quad (15.5.2)$$

As before let the outer boundary of the tumour be represented by  $\Gamma(x, y, z, t) = 0$ . Then the equation on motion of a point on  $\Gamma(x, y, z, t) = 0$  is given by

$$\hat{\mathbf{n}} \cdot \frac{d\mathbf{r}}{dt} = \mathbf{q} \cdot \hat{\mathbf{n}}, \quad (15.5.3)$$

where  $\hat{\mathbf{n}}$  is the unit outward normal to the surface. In a similar manner we define the boundaries  $\Gamma_Q(x, y, z, t) = 0$  and  $\Gamma_N(x, y, z, t) = 0$  to delineate quiescent and necrotic layers. These interfaces are defined implicitly and arise when the nutrient  $\sigma$  passes through prescribed critical values just as we did in the case of **Model I**. Thus quiescence occurs when  $\sigma = \sigma_1$  and necrosis when  $\sigma = \sigma_2 < \sigma_1$ . Consequently cells cease to proliferate if  $\sigma_2 < \sigma < \sigma_1$  and cells death leading to necrosis occurs when  $\sigma < \sigma_2$ .

The next step is to use Darcy's law which relates internal pressure  $p$  to the velocity  $\mathbf{q}$  by

$$\mathbf{q} = -\mu \nabla \mathbf{p} \quad (15.5.4)$$

where  $\mu$  denotes the motility of tumour cells. By eliminating  $\mathbf{q}$  between equations (15.5.2) and (15.5.4) we obtain the system

$$\mu \nabla^2 p = S(\sigma, \omega, \beta) \quad (15.5.5)$$

$$\hat{\mathbf{n}} \cdot \frac{d\mathbf{r}}{dt} = -\mu \nabla p \cdot \hat{\mathbf{n}}, \text{ on } \Gamma(x, y, z, t) = 0. \quad (15.5.6)$$

To complete the model we need to impose boundary and initial conditions. To ensure that  $\sigma$ ,  $\omega$  and  $\beta$  remain bounded at the centre of the tumour we demand

$$\nabla \sigma = \nabla \omega = \nabla \beta = \mathbf{0} \quad \text{at } (x, y, z) = (0, 0, 0). \quad (15.5.7)$$

On the boundary of  $\Gamma(x, y, z, t) = 0$  we impose

$$\sigma = \sigma_\infty - 2\alpha\kappa, \beta = \beta_\infty, \omega = 0, p = p_\infty. \quad (15.5.8)$$

Here  $\sigma_\infty$ ,  $\beta_\infty$  and  $p_\infty$  denote the values of  $\sigma$ ,  $\beta$  and  $p$  in the outer tissue. The condition on the nutrient  $\sigma$  realises the assumption that it satisfies the Gibbs–Thomson condition and where  $\kappa$  denotes the mean curvature on the outer boundary of the tumour. By assuming that the internally produced inhibitor does not spread beyond the tumour we have set  $\omega = 0$  on  $\Gamma(x, y, z, t) = 0$ .

To ensure continuity of  $\sigma$ ,  $\omega$ ,  $\beta$  and  $p$ , together with their first derivatives across the interior layers, we specify:  $\sigma, \omega, \beta, p$  and  $\nabla \sigma, \nabla \omega, \nabla \beta, \nabla p$  are continuous across  $\Gamma_Q(x, y, z, t) = 0$  and  $\Gamma_N(x, y, z, t) = 0$ .

Finally the interior layers are specified implicitly by

$$\sigma = \sigma_1, \text{ on } \Gamma_Q(x, y, z, t) = 0, \quad (15.5.9)$$

$$\sigma = \sigma_2, \text{ on } \Gamma_N(x, y, z, t) = 0, \quad (15.5.10)$$

and the initial condition

$$\Gamma(x, y, z, 0) = 0 \quad (15.5.11)$$

is assumed to be given.

## 15.6 Spherical tumour growth based on Model II

The assumption of a spherical tumour, as in the case of **Model I**, leads to considerable mathematical simplifications and allows for explicit analysis together with some further understanding of avascular tumour growth.

With radial symmetry the tumour grows as a sphere with radius  $R(t)$  and so the mean curvature  $\kappa = 1/R(t)$ . The model system (15.5.1), (15.5.5), (15.5.6) is

$$0 = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \sigma}{\partial r} \right) + F_\sigma = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \omega}{\partial r} \right) + F_\omega = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \beta}{\partial r} \right) + F_\beta, \quad (15.6.1)$$

$$0 = \frac{\mu}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial p}{\partial r} \right) - S(\sigma, \omega, \beta) \quad (15.6.2)$$

$$\frac{dR}{dt} = -\mu \frac{\partial p}{\partial r}, \text{ on } \Gamma(x, y, z, t) = r - R(t) = 0. \quad (15.6.3)$$

The boundary and initial conditions become

$$\frac{\partial \sigma}{\partial r} = \frac{\partial \omega}{\partial r} = \frac{\partial \beta}{\partial r} = 0, \text{ at } r = 0, \quad (15.6.4)$$

$$\sigma = \sigma_\infty - 2\alpha/R(t), \beta = \beta_\infty, \omega = 0, p = p_\infty \text{ on } r = R(t), \quad (15.6.5)$$

$$\sigma = \sigma_1, \text{ on } r = R_Q(t), \quad (15.6.6)$$

$$\sigma = \sigma_2, \text{ on } r = R_N(t), \quad (15.6.7)$$

$$r = R(0) \text{ is prescribed.} \quad (15.6.8)$$

If we integrate (15.6.2) over the tumour volume we get

$$\mu R^2 \frac{\partial p}{\partial r} = \mu \int_0^R \frac{\partial}{\partial r} \left( r^2 \frac{\partial p}{\partial r} \right) dr = \int_0^R S(\sigma, \omega, \beta) r^2 dr. \quad (15.6.9)$$

Finally, if we eliminate  $p$  from this equation using (15.6.3), we see that the rate of growth of the tumour is governed by

$$R^2 \frac{dR}{dt} = - \int_0^R S(\sigma, \omega, \beta) r^2 dr, \quad (15.6.10)$$

together with boundary and initial conditions (15.6.4)–(15.6.8).

We now model the reaction terms  $F_\sigma$ ,  $F_\omega$  and  $F_\beta$  as well as the nutrient loss term  $S$ . For simplicity we choose

$$F_\sigma = -(\lambda_0 + \lambda_1 \beta) H(r - R_N), \quad (15.6.11)$$

$$F_\beta = -\lambda_2, \quad (15.6.12)$$

$$F_\omega = \lambda_3 H(R_N - r) - \lambda_4, \quad (15.6.13)$$

$$S(\sigma, \omega, \beta) = s\bar{\sigma}H(r - R_N) + s\omega H(R_N - r) - s\sigma H(r - R_Q). \quad (15.6.14)$$

As before  $H$  denotes the Heaviside step function. The form of  $F_\sigma$  assumes that  $\sigma$  is taken up by proliferating and quiescent cells at the constant rate  $\lambda_0$  and degraded by the externally supplied inhibitor at the rate  $\lambda_1\beta$ . At the same time  $\beta$  decays throughout the tumour at the constant rate  $\lambda_2$ . For  $F_\omega$  we assume that  $\omega$  is produced at the constant rate  $\lambda_3$  by the necrotic cells and decays at the constant rate  $\lambda_4$ . The loss rate  $S$  consists of three factors: cell loss due to apoptosis which is restricted to the proliferating and quiescent layers where it occurs at the constant rate  $s\bar{\sigma}$  (where  $s$  is a proportionality constant), cell loss due to necrosis which occurs at the rate  $s\omega$  and a production term due to mitosis, restricted to the outer layer, at the rate  $s\sigma$ .

The model developed above is very flexible and can be used to investigate several aspects of avascular tumour growth. We describe two particular cases; namely uniform growth where there is no quiescence or necrosis and then when quiescence is included.

In the case of uniform growth we assume  $R_Q = R_N = 0$  and that there are no growth inhibitors present, i.e.,  $\beta = \omega = 0$ . In this case the first of equations (15.6.1) reduces to

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \sigma}{\partial r} \right) - \lambda_0 = 0,$$

which on applying the boundary conditions (15.6.4), (15.6.5) can be solved to give

$$\sigma = \frac{\lambda_0}{6} (r^2 - R^2) + \sigma_\infty - \frac{2\alpha}{R}. \quad (15.6.15)$$

In a similar manner (15.6.10) reduces to

$$R^2 \frac{dR}{dt} = s \int_0^R (\sigma - \bar{\sigma}) r^2 dr,$$

which may be integrated to give

$$\frac{dR}{dt} = -\frac{s}{3} \left( \lambda_0 \frac{R^3}{15} + (\bar{\sigma} - \sigma_\infty) R + 2\alpha \right). \quad (15.6.16)$$

If we impose the constraint  $\sigma(0, t) > \sigma_1$  then we get an expression which defines the range of tumour radii for which nonquiescent solutions are obtained, i.e.,

$$\sigma_\infty - \sigma_1 > \frac{\lambda_0 R^2}{6} + \frac{2\alpha}{R}.$$

When  $\alpha = 0$  this expression gives an upper bound on  $R$  with quiescence predicted when  $R^2 \geq 6(\sigma_\infty - \sigma_1)/\lambda_0$ . When  $\alpha > 0$  there is a nonzero lower bound for  $R$  which can be interpreted as a nucleation radius. By setting  $d/dt = 0$  in equation (15.6.16) we obtain an expression which shows how the steady state tumour radius depends on the system parameters. It is important therefore to determine the stability or otherwise of these steady states and thus to assess whether the tumour becomes malignant or not.

Suppose we now consider the presence of an externally supplied inhibitor. In this situation  $\beta$  and  $\sigma$  are given by

$$\beta = \beta_\infty - \frac{\lambda_2}{6}(R^2 - r^2), \quad (15.6.17)$$

$$\sigma = \sigma_\infty - \frac{2\alpha}{R} - \frac{\lambda_0}{6}(R^2 - r^2) - \frac{\lambda_1}{6} \left( \beta_\infty - \frac{\lambda_2 R^2}{6} \right) (R^2 - r^2) - \frac{\lambda_1 \lambda_2}{120}(R^4 - r^4) \quad (15.6.18)$$

and the ordinary differential equation governing  $R$  becomes

$$\frac{dR}{dt} = -\frac{sR}{3} \left( \frac{\lambda_0 R^2}{15} - (\sigma_\infty - \bar{\sigma}) + \frac{2\alpha}{R} \right) - \frac{s\lambda_1 R^3}{45} \left( \beta_\infty - \frac{2\lambda_2 R^2}{21} \right). \quad (15.6.19)$$

Once again, for valid nonquiescence we require  $\sigma(0, t) > \sigma_1$  and noting the expressions for  $\sigma$ ,  $\beta$  and  $dR/dt$  we see that

$$\sigma(0, t) |_{\beta>0} = \sigma(0, t) |_{\beta=0} - \frac{\lambda_1 R^2}{6} \left( \beta(0, t) + \frac{\lambda_2 R^2}{20} \right) < \sigma(0, t) |_{\beta=0}$$

and

$$\left( \frac{3}{sR} \frac{dR}{dt} \right)_{\beta>0} = \left( \frac{3}{sR} \frac{dR}{dt} \right)_{\beta=0} - \frac{\lambda_1 R^2}{15} \left( \beta(0, t) + \frac{\lambda_2 R^2}{14} \right) < \left( \frac{3}{sR} \frac{dR}{dt} \right)_{\beta=0}.$$

Consequently we deduce that the minimum nutrient concentration is reduced when the external inhibitor is present and so quiescence is more rapidly initiated.

Now consider the case when the tumour consists of an outer proliferating layer and an inner quiescent layer. Without the externally supplied inhibitor,  $\beta = 0$ , we find that  $\sigma(r, t)$  satisfies (15.6.15) and  $R(t)$  satisfies the ordinary differential equation

$$\frac{R^2}{s} \frac{dR}{dt} = -\frac{1}{3} \left( \frac{\lambda_0 R^2}{15} - (\sigma_\infty - \bar{\sigma}) + \frac{2\alpha}{R} \right) (R^3 - R_Q^3) + \frac{\lambda_0 R_Q^3}{30} (R^2 - R_Q^2) - \frac{\bar{\sigma} R_Q^3}{3}, \quad (15.6.20)$$

with the condition  $\sigma(0, t) > \sigma_2$  to ensure nonnecrotic solutions. The radius of the quiescent  $R_Q(t)$  is determined implicitly by the condition  $\sigma(R_Q, t) = \sigma_2$ . That is

$$\sigma_2 = \sigma_\infty - \frac{2\alpha}{R} - \frac{\lambda_0}{6}(R^2 - R_Q^2).$$

## 15.7 Stability of tumour growth based on Model II

As we did in the case of **Model I** we give a flavour of the analysis of the stability of a uniformly proliferating tumour colony. Setting  $\beta = 0 = R_Q =$

$R_N$  and assuming the spherical tumour undergoes a small perturbation away from the purely radial, then, as in **Model I** equations (15.4.7) and (15.4.8), we assume

$$\begin{aligned} p(r, \theta, t) &= p(r, t) + \epsilon \tilde{p}(r, \theta, t), \\ \sigma(r, \theta, t) &= \sigma(r, t) + \epsilon \tilde{\sigma}(r, \theta, t) \end{aligned} \quad (15.7.1)$$

and the surface of the tumour is given at any time by

$$\Gamma(r, \theta, t) = r - R(t) - \epsilon \xi(\theta, t) = 0, \quad (15.7.2)$$

where (15.6.2) is solved to give

$$p(r, t) = p_\infty + \frac{s}{6\mu} \left( \sigma_\infty - \bar{\sigma} - \frac{2\alpha}{R} - \frac{\lambda_0 R^2}{6} \right) + \frac{s\lambda_0}{120\mu} (R^4 - r^4). \quad (15.7.3)$$

The first-order terms are seen to satisfy the system:

$$0 = \nabla^2 \tilde{\sigma} = \mu \nabla^2 \tilde{p} + s \tilde{\sigma}, \quad (15.7.4)$$

$$\frac{\partial \xi}{\partial t} = -\mu \left( \frac{\partial \tilde{p}}{\partial t} + \xi \frac{\partial^2 p}{\partial r^2} \right) |_{r=R(t)}, \quad (15.7.5)$$

subject to the conditions

$$\begin{aligned} \frac{\partial \tilde{p}}{\partial r} &= 0 = \frac{\partial \tilde{\sigma}}{\partial r} \\ \tilde{\sigma} &= \frac{\alpha}{R^2} \left( 2\xi + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \xi}{\partial \theta} \right) \right) - \xi \frac{\partial \sigma}{\partial r} |_{r=R(t)} \\ \tilde{p} &= -\xi \frac{\partial p}{\partial r} |_{r=R(t)} \\ \xi(\theta, 0) &= \xi_0(\theta), \text{ given.} \end{aligned}$$

The solutions of the above system of equation can be solved, as in the case of **Model I**, in terms of time-dependent multiples of harmonic functions.

In the absence of cell-cell adhesion, ( $\alpha = 0$ ), it can shown that radially symmetric steady state solutions are generally unstable. However when  $\alpha \neq 0$  there is a range of harmonic modes for which there are both stable and unstable modes.

There are many problems relating to stability and mode configurations that may be investigated as well as their consequences for tumour development.

## 15.8 Notes

**Model I** of tumour growth discussed in this chapter is inspired by the model due to H.P. Greenspan, On the growth and stability of cell cultures

and solid tumors, *J. Theor. Biol.*, **56**, 229–242, 1976 and is the outcome of a previous article by the same author, Models for the growth of a solid tumor by diffusion, *Stud. Appl. Math.*, **51**, 317–340, 1972.

In the former paper, Greenspan carries out a detailed stability analysis along the lines briefly described in Section 15.4 and also shows that tumour cell colonies which share the same nutrient supply repel each other and move apart. The model is quite versatile and can be modified to describe several important problems; for example, that of describing the movement of tumours in the presence of solid boundaries or the effect on growth with changes in the nutrient supply. Similar problems can be posed for the models developed in this chapter.

Since the work of Greenspan there has been a considerable development in our understanding of growth of solid avascular tumours both in *in vitro* and in *in vivo* studies. This has led to other mathematical models being developed. Among such models is **Model II** which focuses on nutrient supply and diffusion processes. This model is due to H.M. Byrne and M.A.J. Chaplain, Free boundary value problems associated with the growth and development of multicellular spheroids, *Eur. J. Appl. Math.*, **8**, 639–658, 1997.

Research into avascular tumour growth is increasingly being carried out across the world, with the developed countries investing vast sums of money in order to find cures and in improving treatments. At the same time mathematics is making a growing contribution to understanding the biochemical and physical mechanisms underlying avascular tumour growth. A recent review which provides an up-to-date and comprehensive account of existing models is given in the article T. Roose, S. J. Chapman and P. K. Maini, Mathematical models of Avascular Tumor growth, *SIAM Review*, **49**, 179–208, 2007.

We have mentioned in the introduction to this chapter the important role of angiogenesis, the process whereby the tumour is able to acquire its own blood supply from host tissue and nearby blood vessels. Many of the fundamental discoveries associated with tumour angiogenesis were made by J. Folkman and his colleagues (see the biography *Dr. Folkman's War* by Robert Cooke, Random House, New York, 2000). Mathematical models of angiogenesis of varying degrees of complexity have and continue to be developed, see for example: M.A.J. Chaplain and A.R.A. Anderson, Modelling the growth and form of capillary networks in *On Growth and Form*, M.A.J. Chaplain, G.D. Singh, and J.C. McLachlan, Eds. John Wiley & Sons, 1999; H.A. Levine, S. Pamuk, B.D. Sleeman and M. Nilsen-Hamilton, Mathematical modeling of capillary formation and development in tumor angiogenesis: Penetration into the stroma, *Bull. Math. Biol.*, **63**, 801–863, 2001; M.J. Plank, and B.D. Sleeman, Lattice and non-lattice models of tumour angiogenesis, *Bull. Math. Biol.*, **66**, 1785–1819, 2004.

## Further reading

- Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, 3rd ed., Garland, New York and London, 1994.
- Folkman, J., The vascularisation of tumors. *Sci. Am.*, **234**, 58–64, 1976.
- King, R. J. B., *Cancer Biology*, 2nd ed., Harlow, London and NY: Prentice Hall, 2000.
- Paweverz, N. and M. Knierim, Tumor related angiogenesis, *Crit. Rev. Oncol. Hematol.*, **9**, 197–242, 1989.
- Plank, M. J. and Sleeman, B. D. Tumour induced angiogenesis: a review, *J. Theor. Med.*, **5**, 137–153, 2003.
- Sherratt, J. A. and M. A. J. Chaplain, A new mathematical model for avascular tumour growth. *J. Math. Biol.*, **43**, 291–312, 2001.
- Sleeman, B. D., Mathematical aspects of modelling tumour angiogenesis, in *Nonlinear dynamics and evolution equation*, Eds H. Brunner, Xiao-Qiang, and Xinfu Zou, Fields Institute Communications, vol. **48**, 257–278, 2007.

## Exercises

- 15.1 Write a MATLAB program to solve the differential equation (15.3.16) for the evolution of the outer boundary,  $R(t)$ , of a tumour subject to the initial condition  $R(0)=1$  and where  $R_N$  is given by (15.3.18). Experiment by choosing various values of  $S_i, i = 1, 2$  and  $\sigma_j, j = 1, 2$  and  $\sigma_\infty$  to investigate the effects of proliferation, quiescence and necrosis.
- 15.2 Verify the steady state radius given by (15.4.6).
- 15.3 A tumour colony is cultured in a circular dish of radius  $A$ . Assume the nutrient concentration  $\sigma$  has the constant value  $\sigma_A$  at the edge of the dish. If the tumour colony is initially a circle of radius  $a$  and is assumed to grow radially, formulate a mathematical model of type I of colony growth under the same assumptions as for the growth of a spherical tumour colony.

*Hint:* formulate the model using polar coordinates.

- 15.4 In Exercise 15.3 show that the outer boundary of the colony satisfies the differential equation

$$\frac{dR}{dt} = -S_1 \frac{R}{2} - (S_2 - S_1) \frac{R_N^2}{2R} + \frac{\lambda}{\mu} \frac{(\sigma_A - \sigma_2)}{R \ln(A/R_N)},$$

where  $R(t)$  and  $R_N(t)$  are related by

$$\frac{(\sigma_A - \sigma_2)^2}{R^2} = \mu^2[(\sigma_A - \sigma_2) \ln(R/A) \ln(A/R_N) + (\sigma_A - \sigma_1)(\ln(A/R_N))^2].$$

15.5 Investigate the growth of a fully vascularised tumour colony growing in a circular dish, using **Model I**.

15.6 A vascularised tumour colony in a laboratory experiment grows along a straight narrow tube. Using **Model I** and by ignoring any curvature at the growing boundary, determine the rate of growth of the tumour colony.

15.7 In equation (15.6.16) set  $dR/dt = 0$  and draw a graph of  $R$  versus  $\sigma_\infty - \bar{\sigma}$  to show how the steady state varies with the parameter  $\alpha$ .

15.8 Linearise (15.6.16) about the steady state solution and determine its stability. Show that the steady state is locally stable provided the tumour radius  $R$  satisfies

$$R^2 > \frac{5}{\lambda_0}(\sigma_\infty - \bar{\sigma}).$$

15.9 For **Model II** with uniform growth, in the presence of an externally supplied inhibitor, show that  $\beta$  and  $\sigma$  are given by equations (15.6.17) and (15.6.18).

Show that the tumour radius  $R$  satisfies the ordinary differential equation (15.6.19) and write a MATLAB program to solve it.

15.10 Suppose a tumour comprises an outer proliferating rim and an inner quiescent region. Use **Model II** with  $\beta = 0$  to show that  $\sigma$  satisfies (15.6.15) and that the tumour radius  $R(t)$  satisfies the ordinary differential equation

$$\begin{aligned} \frac{R^2}{s} \frac{dR}{dt} &= -\frac{1}{3} \left( \frac{\lambda_0 R^2}{15} + (\bar{\sigma} - \sigma_\infty) + \frac{2\alpha}{R} \right) (R^3 - R_Q^3) + \frac{\lambda_0 R_Q^3}{30} (R^2 - R_Q^2) \\ &\quad - \frac{\bar{\sigma} R_Q^3}{3}, \end{aligned}$$

where  $R_Q$  is defined explicitly in terms of  $R$  by the condition  $\sigma = \sigma_1$ :

$$\sigma_1 = \sigma_\infty - \frac{2\alpha}{R} - \frac{\lambda_0}{6} (R^2 - R_Q^2)$$

and  $\sigma(0, t) > \sigma_2$  must be satisfied for nonnecrotic solutions.

# Chapter 16

---

## Epidemics

---

### 16.1 The Kermack–McKendrick model

The problem of epidemics is to assess how a group of individuals with a communicable infection spreads the disease to a population able to catch it. The model constructed depends upon the assumption made about the disease and the behaviour of the population. One possibility has been discussed (Section 5.2) and the aim is now to amplify the study, as well as to take account of more characteristics of the process of infection.

In the Kermack–McKendrick model, the population is presumed to be constant in size and to be divided into three classes. There are  $I$  infected individuals who can pass on the disease to others and  $s$  susceptibles who have yet to contract the disease and become infectious. The remaining group contains  $r$  members who have been infected but cannot transmit the disease for some reason, e.g., they have been isolated from the rest of the population. These types of models are often referred to as **SIR models**. The governing equations will be taken as

$$\dot{s} = -asI, \quad (16.1.1)$$

$$\dot{I} = asI - bI, \quad (16.1.2)$$

$$\dot{r} = bI \quad (16.1.3)$$

with  $a$  and  $b$  positive constants.

The basis for (16.1.1) is that the susceptibles become infected at a rate that is proportional to the number of contacts between individuals of  $s$  and  $I$ , assuming that contact depends only on the numbers of each group, i.e., there is uniform mixing of the population. The assumption in (16.1.3) is that the rate at which individuals become unable to transmit the disease is proportional to the number infected. It represents some kind of average of the process in which particular individuals take different lengths of time to reach a state in which they neither contract nor pass on the infection.

If now we add on the hypothesis of the constancy of the population, we have

$$s(t) + I(t) + r(t) = N \quad (16.1.4)$$

where  $N$  is the population's invariable size. A derivative of (16.1.4) then adds (16.1.2) to (16.1.1) and (16.1.3).

At the beginning,  $t = 0$  and  $r = 0$  so that

$$s(0) + I(0) = N. \quad (16.1.5)$$

There must be some infected and some available for infection at the start, so  $I(0) > 0$  and  $s(0) > 0$ .

Only solutions in which  $s$ ,  $I$  and  $r$  are nonnegative are of concern. Therefore, (16.1.1) implies that  $\dot{s} < 0$  when both infected and susceptibles are present. Hence  $s$  decreases steadily and

$$s(t) < s(t_1) < s(0) \quad (16.1.6)$$

for  $t > t_1 > 0$ . The constant diminution of  $s$  and the fact that  $s$  must be nonnegative means that, as  $t \rightarrow \infty$ ,  $s$  must tend to a limit (which may be zero), i.e.,  $s(\infty) = \lim_{t \rightarrow \infty} s(t)$  exists.

From (16.1.2),  $\dot{I} < 0$  if  $as < b$ . In view of (16.1.6) it follows that, if  $as(0) < b$ ,  $\dot{I} < 0$  for all  $t$  and the infection is eventually wiped out. Thus there is a critical level, or **threshold value**, of  $b/a$  which the initial population of susceptibles must exceed if the epidemic is to spread. The threshold will be low if  $b \ll a$ , i.e., the rate at which immunity is conferred is small enough for the disease to prosper.

According to (16.1.3),  $r$  increases monotonically while (16.1.4) forces  $r(t) \leq N$ . Hence  $r(\infty) = \lim_{t \rightarrow \infty} r(t)$  exists. It follows from (16.1.4) and what has been proved about  $s$  that  $I(\infty) = \lim_{t \rightarrow \infty} I(t)$  also exists. The quantity  $\{I(\infty) - r(\infty)\}/N$  is a measure of the extent to which the infection swept through the population and so now we attempt to estimate the limits.

From (16.1.1) and (16.1.3)

$$\frac{ds}{dr} = -\frac{as}{b}$$

so that

$$s = s(0)e^{-ar/b}. \quad (16.1.7)$$

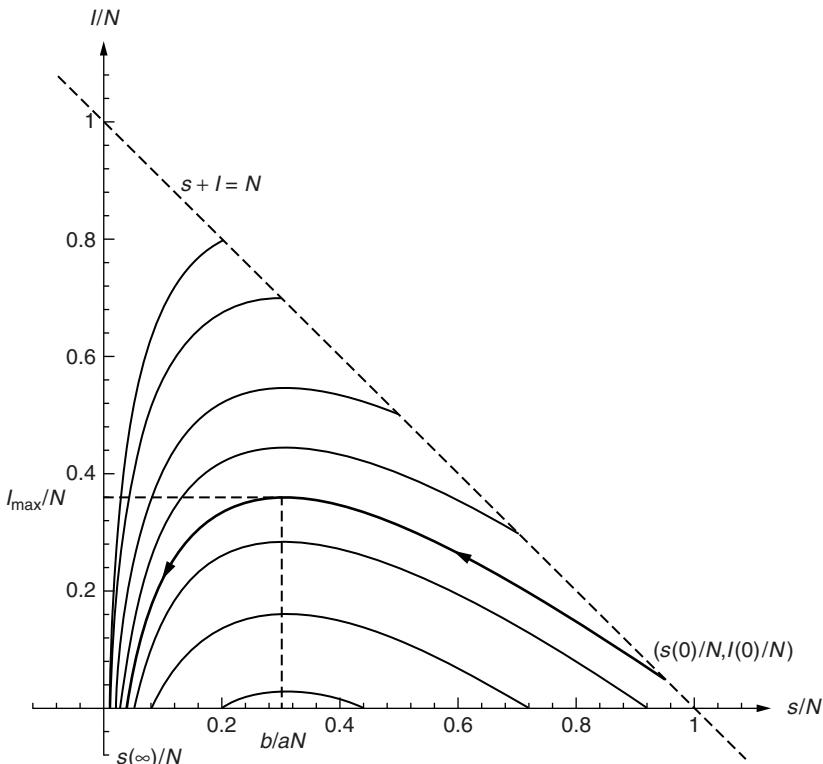
Since  $r \leq N$ , we are forced to have  $s \geq s(0)e^{-aN/b}$  and therefore  $s(\infty) > 0$ , i.e., there are always susceptibles available. Thus *some individuals never suffer from the disease*; there are still susceptibles available when the disease stops spreading.

The trajectories in the phase plane for  $s$  and  $I$  can be drawn. They are solutions of, according to (16.1.1) and (16.1.2),

$$\frac{dI}{ds} = -1 + \frac{b}{as}$$

whence, on account of (16.1.5),

$$I = N - s + (b/a) \ln\{s/s(0)\}.$$



**FIGURE 16.1.1:** The trajectories for an epidemic.

They are plotted in Figure 16.1.1, the direction of the arrows being dictated by the steady decrease of  $s$ . Evidently,  $s(t) - s(\infty) > 0$  and  $I(t) \rightarrow 0$  as  $t \rightarrow \infty$  (there are no critical points off  $I = 0$ ). The consequence of this, (16.1.7) and (16.1.4), is that

$$s(\infty) = s(0) \exp[-a\{N - s(\infty)\}/b], \quad (16.1.8)$$

a transcendental equation to determine  $s(\infty)$ . The equation is satisfied by only one positive value of  $s(\infty)$  less than  $b/a$ . Once  $s(\infty)$  is known,  $r(\infty)$  can be calculated from  $r(\infty) = N - s(\infty)$  and the spread of the infection measured by  $r(\infty)/N$  or  $1 - s(\infty)/N$ .

## 16.2 Vaccination

In the preceding section, it has been implicitly assumed that the individuals who are naturally immune are small enough in number to be neglected and we shall continue to adopt this hypothesis. However, if a calculation by the

previous model suggests that the infection will spread to unacceptable levels, we may decide to introduce some immunity by means of vaccination.

For a simple model assume that vaccination removes an individual instantaneously from  $s$  or  $r$  without joining the group  $I$  and that a device is available that prevents the infected from being vaccinated. There is then a group  $v$  of vaccinated and we have

$$\dot{s} = -asI - \alpha(t), \quad (16.2.1)$$

$$\dot{v} = \alpha(t) \quad (16.2.2)$$

where (16.2.1) replaces (16.1.1) but (16.1.2) and (16.1.3) are retained,  $\alpha$  being the vaccination rate.

A decision on the function to be selected as the vaccination rate is not at all easy. Any vaccination program involves cost through the employment of people, equipment and supplies. Against this must be set the damage that is caused to society by the infection. It may be desirable to vaccinate in order to limit the total number who are infected in a given time interval, or to keep the peak of those infected at any particular instant below some level, or both. To ensure that no more than  $N_1$  of the population contract the disease for  $0 \leq t \leq T$  we want

$$r(T) + I(T) \leq N_1 \quad (16.2.3)$$

and, to force the peak of infection below  $N_2$  in the same time interval, we need

$$\max_{0 \leq t \leq T} I(t) \leq N_2. \quad (16.2.4)$$

To control the disease in this way,  $\alpha(t)$  would have to be chosen so that (16.2.3) and (16.2.4) were satisfied while the cost of vaccination was kept to a minimum. This is a problem in **dynamic programming** for which there are available techniques, but they are beyond the scope of this book.

Before pursuing the model too far for a particular disease, it would be necessary to verify that it complied with the two hypotheses on vaccination, for vaccination may not grant immediate immunity and there may be no acceptable way of determining those already infected.

### 16.3 An incubation model

In the Kermack–McKendrick model of Section 16.1, an individual is capable of passing on the infection as soon as he or she has succumbed to it. For some diseases there is an incubation period during which an individual has become infected but cannot communicate the disease to others. The incorporation of this feature into the model of an epidemic will be considered in this

section. At the same time, the possibility that an individual may be infectious for not more than a finite time will be allowed for.

Let  $s(t)$  be the number of those who are susceptible but who have not yet been exposed to the disease and let  $E(t)$  be the number who have been exposed to but cannot yet transmit the infection. The groups  $I$  and  $r$  are defined as before.

Continue to assume that there is homogeneous mixing of the population so that the rate of exposure is proportional to the number of contacts between individuals of  $s$  and  $I$ . Then

$$\dot{s} = -asI. \quad (16.3.1)$$

Thus (16.1.6) is still valid and  $s(\infty)$  exists. Also a solution of (16.3.1) satisfies

$$s(t) = s(0) \exp \left\{ -a \int_0^t I(u) du \right\}. \quad (16.3.2)$$

Suppose that the epidemic is started by a number of infectious individuals entering a population which has been unexposed previously to the disease. Assume that we know how many of these infectious individuals will remain infectious as time varies. Until the population produces members who are infectious

$$I(t) = I_0(t) \quad (16.3.3)$$

where  $I_0(t)$  is a known function.

Assume now that an individual who contracts the infection at time  $t$  does not become infectious until time  $t + T$  so that  $T$  specifies the incubation period. Then, for  $0 \leq t \leq T$ ,

$$E(t) = s(0) - s(t) \quad (16.3.4)$$

and, for  $t \geq T$ ,

$$E(t) = s(t - T) - s(t). \quad (16.3.5)$$

Keeping the population constant entails

$$s(t) + E(t) + I(t) + r(t) = N. \quad (16.3.6)$$

While  $t \leq T$ , incubation prevents any new infectious individuals from appearing. Therefore, (16.3.3) is valid and  $s(t)$  can be found from (16.3.2). Then  $E(t)$  follows from (16.3.4) and  $r(t)$  from (16.3.6). Thus, all quantities are known during the initial incubation period.

When  $t \geq T$ , some of those who have been incubating the disease will have become infectious. In fact, all those who were infected up to the time  $t - T$  will be infectious. Consequently, (16.3.3) must be modified and replaced by

$$I(t) = I_0(t) + s(0) - s(t - T). \quad (16.3.7)$$

Substitution of (16.3.7) into (16.3.1) leads to a differential-difference equation for  $s(t)$  but its solution will not be required for our purpose.

Now introduce the assumption that an individual remains infectious only for the time interval  $\sigma$ . A necessary consequence is that  $I_0(t) = 0$  for  $t > \sigma$  because all those who started the epidemic were infectious from the beginning. Since none of those infected in the initial incubation period cease being infectious for  $t \leq T + \sigma$ , the equation (16.3.7) is valid for  $T \leq t \leq T + \sigma$ . However, when  $t \geq T + \sigma$ , some of the infected are no longer infectious; they are the ones who were infected up to the time  $t - T - \sigma$ . Therefore, their numbers must be removed from the infectious and (16.3.7) must be replaced by

$$I(t) = I_0(t) + s(t - T - \sigma) - s(t - T). \quad (16.3.8)$$

Actually, the term  $I_0(t)$  could be dropped because it is zero for  $t > \sigma$  and, a fortiori, for  $t \geq T + \sigma$ .

To determine  $s(\infty)$  the integral in (16.3.2) has to be calculated. Now, for  $t > T + \sigma$ ,

$$\int_0^t I(u)du = \int_0^T I(u)du + \int_T^{T+\sigma} I(u)du + \int_{T+\sigma}^t I(u)du.$$

Insert (16.3.3), (16.3.7) and (16.3.8), respectively, into the three integrals. Then, after a change of variable of integration where  $s$  is involved,

$$\int_0^t I(u)du = \int_0^\sigma I_0(u)du + \sigma s(0) - \int_{t-T-\sigma}^{t-T} s(u)du$$

since  $I_0(u)$  vanishes for  $u > \sigma$ . Let  $t \rightarrow \infty$  so that both limits in the integral of  $s(u)$  become large. But it has been pointed out that  $s(u) \rightarrow s(\infty)$  as  $u \rightarrow \infty$  so that  $s(u)$  can be replaced effectively by  $s(\infty)$  in the integral, i.e.,

$$\int_{t-T-\sigma}^{t-T} s(u)du \sim s(\infty) \int_{t-T-\sigma}^{t-T} du = \sigma s(\infty)$$

as  $t \rightarrow \infty$ . Hence

$$\int_0^\infty I(u)du = \int_0^\sigma I_0(u)du + \sigma\{s(0) - s(\infty)\}.$$

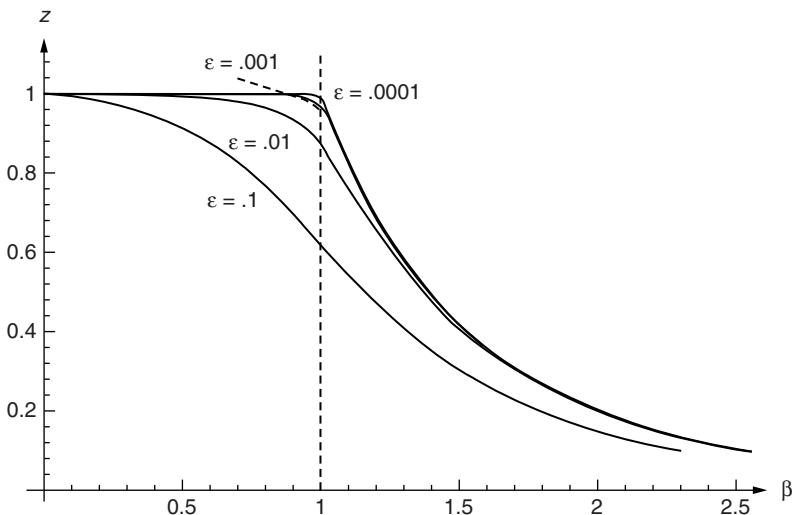
Consequently, we deduce from (16.3.2) that

$$s(\infty) = s(0) \exp \left[ -a \int_0^\sigma I_0(v)dv - a\sigma\{s(0) - s(\infty)\} \right], \quad (16.3.9)$$

which constitutes an equation to determine  $s(\infty)$ .

Equation (16.3.9) can be expressed in the form

$$z = \exp\{\beta(z - 1 - \epsilon)\}, \quad (16.3.10)$$



**FIGURE 16.3.1:** Variation of  $z = s(\infty)/s(0)$  with  $\beta$ .

where  $z = s(\infty)/s(0)$ ,  $\beta = a\sigma s(0)$  and

$$\epsilon = (a/\beta) \int_0^\sigma I_0(v)dv.$$

The left-hand side of (16.3.10) is 0 at  $z = 0$  and 1 at  $z = 1$ , whereas the right-hand side is positive and less than 1 (since  $\epsilon > 0$ ), respectively, at the two points. Hence there is a positive root with  $z \leq 1$ , i.e.,  $s(\infty)/s(0) \leq 1$  as there should be. The root does not differ by much from  $z = 1$  when  $\beta\epsilon$  is fairly small. Some graphs of the value of  $z$  for various values of  $\beta$  and  $\epsilon$  are shown in Figure 16.3.1. They are knee-shaped with a distinct kick near  $\beta = 1$  when  $\epsilon$  is very small. Interpreting  $\epsilon$  as a measure of the infectiousness of the initial infectives, and  $\beta$  as the number of susceptibles likely to be infected by each infective, we see that, when the initial infectives are only a very small proportion of the population, the infection tends to die out without altering the susceptibles much if  $\beta < 1$ , but to originate a substantial epidemic if  $\beta > 1$ . This is a kind of threshold effect. The effect becomes less pronounced for higher values of  $\epsilon$ , indicating that when there are sufficient initial infectives there is a tendency for a significant epidemic to occur.

Various refinements to the foregoing model have been proposed, some of which are in the exercises. One possibility is that the parameter  $a$  should be made a function of time in order to permit seasonal and cyclic changes. For example, measles has relatively high and low incidences in alternate years, whereas chickenpox recurs annually. It is more probable that the low incidence of measles is due to periodic alteration of  $a$  than to a threshold effect.

The model can be generalised to the case of two populations as when parasites transmit disease (e.g., mosquitoes infecting humans with malaria) or when an infection is transferred from one sex to another, but details will be omitted.

---

## 16.4 Spreading in space

A basic hypothesis in the preceding sections is that the population is mixing thoroughly, so that there is no distinction between the individuals in one place and those in another. When this is not so, the disease may spread faster in some parts than in others and it is necessary to allow the variables to depend on space as well as time. For simplicity, we shall assume that only one space variable is involved, though most of the ideas carry over to more general cases.

Take  $x$  as the single space variable. Then, in the Kermack–McKendrick model,  $s$  and  $I$  will be functions of  $t$  and  $x$  giving the distributions of susceptibles and infected, respectively, at time  $t$  at the position  $x$ . Moreover, it is necessary to allow for the effect of the infected at one position on the susceptibles at another place. Therefore, replace (16.1.1) by

$$\frac{\partial s}{\partial t} = - \left\{ \int_{-\infty}^{\infty} F(x, x') I(t, x') dx' \right\} s(t, x) \quad (16.4.1)$$

where the nonnegative function  $F(x, x')$  is a measure of the influence of the infected at the position  $x'$  on the susceptibles at  $x$ . The integration permits the possibility of all those infected at time  $t$  affecting the susceptibles at  $x$ . The corresponding replacements for (16.1.2) and (16.1.3) are

$$\frac{\partial I}{\partial t} = \left\{ \int_{-\infty}^{\infty} F(x, x') I(t, x') dx' \right\} s(t, x) - g(t, x) I(t, x) \quad (16.4.2)$$

and

$$\partial r / \partial t = g(t, x) I(t, x), \quad (16.4.3)$$

respectively.

The presence of  $F$  in the system (16.4.1)–(16.4.3) makes the system highly flexible and offers the ability to cover many different situations. Unfortunately, this very flexibility makes the system awkward from an experimental point of view. Unless the structure of  $F$  is very simple it is unlikely that its functional behaviour can be determined by experimental observations alone. The best approach is to make plausible assumptions about  $F$  based on any information available and compare the consequent predictions with experiment. For instance, it is known that certain epidemics have a propensity for spreading like a wave travelling in space. One possible supposition on  $F$  that reproduces such behaviour will be discussed now.

A plausible hypothesis for some diseases is to imagine that the transmission of infection depends only on the distance between individuals. Furthermore, it seems reasonable to suppose that the transmission will be relatively poor unless the individuals are close to one another. This suggests trying

$$F(x, x') = \begin{cases} f(|x - x'|) & (|x - x'| < \delta) \\ 0 & (|x - x'| \geq \delta) \end{cases}$$

where  $\delta$  is a small quantity. Then

$$\begin{aligned} \int_{-\infty}^{\infty} F(x, x') I(t, x') dx' &= \int_{x-\delta}^{x+\delta} f(|x - x'|) I(t, x') dx' \\ &= \int_{-\delta}^{\delta} f(|u|) I(t, x + u) du \end{aligned}$$

on putting  $x' = x + u$ . On account of the smallness of  $\delta$  introduce the approximation

$$I(t, x + u) = I(t, x) + u \frac{\partial}{\partial x} I(t, x) + \frac{1}{2} u^2 \frac{\partial^2}{\partial x^2} I(t, x).$$

As a result

$$\int_{-\infty}^{\infty} F(x, x') I(t, x') dx' = I(t, x) \int_{-\delta}^{\delta} f(|u|) du + \frac{1}{2} \frac{\partial^2 I}{\partial x^2} \int_{-\delta}^{\delta} u^2 f(|u|) du$$

since  $uf(|u|)$  is an odd function of  $u$ . With this approximation equations (16.4.1) and (16.4.2) convert to

$$\frac{\partial s}{\partial t} = -s \left( \theta I + \phi \frac{\partial^2 I}{\partial x^2} \right), \quad (16.4.4)$$

$$\frac{\partial I}{\partial t} = -\frac{\partial s}{\partial t} - gI \quad (16.4.5)$$

where  $\theta$  and  $\phi$  are positive constants defined by

$$\theta = \int_{-\delta}^{\delta} f(|u|) du, \quad \phi = \frac{1}{2} \int_{-\delta}^{\delta} u^2 f(|u|) du.$$

The partial differential equations (16.4.4) and (16.4.5) constitute the **diffusion approximation** to the problem of the spatial spread of an epidemic.

Let us examine whether the diffusion approximation is capable of sustaining a wave-like epidemic. If it is, then there must be solutions of (16.4.4) and (16.4.5) which are travelling waves. According to Section 10.7 a travelling wave is a function of  $x - ct$  where  $c$  is the speed of the wave. Consequently, solutions of the form

$$s(t, x) = s_1(x - ct), \quad I(t, x) = I_1(x - ct)$$

are sought in which the constant  $c$  is real and positive. Whether  $c$  can be chosen freely or is subject to limitations will have to be investigated.

Write  $\zeta = x - ct$  and substitute the assumed progressive waveforms for  $s$  and  $I$  into the partial differential equations. There results

$$cs'_1 = (\theta I_1 + \phi I''_1)s_1, \quad (16.4.6)$$

$$cI'_1 = -cs'_1 + gI_1 \quad (16.4.7)$$

where the primes signify derivatives with respect to  $\zeta$ .

Divide (16.4.6) by  $s_1$  and integrate with respect to  $\zeta$  starting from  $\zeta = A$ . Then

$$\begin{aligned} c \ln\{s_1(\zeta)/s_1(A)\} &= \int_A^\zeta \{\theta I_1(u) + \phi I''_1(u)\} du \\ &= \theta \int_A^\zeta I_1(u) du + \phi \{I'_1(\zeta) - I'_1(A)\}. \end{aligned} \quad (16.4.8)$$

Integration of (16.4.7) gives, when  $g$  is a constant,

$$c\{I_1(\zeta) + s_1(\zeta) - I_1(A) - s_1(A)\} = g \int_A^\zeta I_1(u) du. \quad (16.4.9)$$

Elimination of the integral from (16.4.8) and (16.4.9) leads to

$$\begin{aligned} \phi I'_1(\zeta) &= \phi I'_1(A) + c \ln\{s_1(\zeta)/s_1(A)\} \\ &\quad - \{\theta c/g\}\{I_1(\zeta) + s_1(\zeta) - I_1(A) - s_1(A)\}, \end{aligned} \quad (16.4.10)$$

which is a differential equation of the first order for  $I_1$ . Insertion of (16.4.10) into (16.4.7) gives

$$\begin{aligned} s'_1(\zeta) &= (g/c)I_1(\zeta) - I'_1(A) - (c/\phi) \ln\{s_1(\zeta)/s_1(A)\} \\ &\quad + \{\theta c/g\phi\}\{I_1(\zeta) + s_1(\zeta) - I_1(A) - s_1(A)\}. \end{aligned} \quad (16.4.11)$$

In this way (16.4.6) and (16.4.7) are transformed into the first-order system (16.4.10) and (16.4.11).

Since the spread of the epidemic is dictated by the motion of the wave we do not expect any individuals to be infected at the position  $x$  until the wave arrives. The number of susceptibles at that position before the advent of the wave is fixed by the initial state of the susceptibles. Since  $t \rightarrow -\infty$  corresponds to the time before the wave gets going these conditions are met by requiring:  $s_1(\infty)$  to be some positive constant and  $I_1(\infty) = 0$ ,  $I'_1(\infty) = 0$  assuming that the wave starts smoothly. Letting  $A \rightarrow \infty$  in (16.4.10) and (16.4.11) we have

$$\begin{aligned} I'_1(\zeta) &= (c/\phi) \ln\{s_1(\zeta)/s_1(\infty)\} \\ &\quad - \{\theta c/\phi g\}(I_1(\zeta) + s_1(\zeta) - s_1(\infty)), \end{aligned} \quad (16.4.12)$$

$$\begin{aligned} s'_1(\zeta) &= (g/c)I_1(\zeta) - (c/\phi) \ln\{s_1(\zeta)/s_1(\infty)\} \\ &\quad + \{\theta c/g\phi\}(I_1(\zeta) + s_1(\zeta) - s_1(\infty)). \end{aligned} \quad (16.4.13)$$

The discussion of (16.4.12) and (16.4.13) is simplified somewhat by switching to the variables  $X(\zeta) = s_1(\zeta)/s_1(\infty)$  and  $Y(\zeta) = I_1(\zeta)/s_1(\infty)$ . This has the advantage that only values of  $X(\zeta)$  between 0 and 1 need be considered. In terms of the new variables, (16.4.12) and (16.4.13) become

$$X'(\zeta)/\alpha = -\ln X(\zeta) + \beta\{X(\zeta) - 1\} + (\beta + \gamma)Y(\zeta), \quad (16.4.14)$$

$$Y'(\zeta)/\alpha = \ln X(\zeta) - \beta\{X(\zeta) - 1\} - \beta Y(\zeta) \quad (16.4.15)$$

where

$$\alpha = c/\phi s_1(\infty), \quad \beta = \theta s_1(\infty)/g, \quad \gamma = \phi g s_1(\infty)/c^2. \quad (16.4.16)$$

Note that the constants  $\alpha$ ,  $\beta$  and  $\gamma$  are all positive.

Once the wave has passed a particular position there should be no infected there and the number of susceptibles should have settled down to a steady level (which may be zero). This will be the state at a point as  $t \rightarrow \infty$  and so the conditions  $Y(-\infty) = 0$  and  $X'(-\infty) = 0$  are imposed. It follows from (16.4.14) that the final number of susceptibles, obtained from  $X(-\infty)$ , is determined by

$$\ln X(-\infty) = \beta\{X(-\infty) - 1\}. \quad (16.4.17)$$

The exponential of (16.4.17) is the same as (16.3.10) with  $\epsilon = 0$ , which enables us to say that, when  $\beta \leq 1$ , the only solution of (16.4.17) in the permitted range of  $X$  is  $X(-\infty) = 1$ . But, in that case, the susceptibles have been unaffected and there has been no epidemic. Therefore, in order for an epidemic to occur, it is necessary that  $\beta > 1$ , i.e.,

$$s_1(\infty) > g/\theta \quad (16.4.18)$$

from (16.4.16). Thus (16.4.18) is a necessary condition for the existence of a wave and sets a lower bound on the initial density of susceptibles in order for the population to support a progressing epidemic. Hence there is a *threshold* below which the susceptibles are inadequate for a wave-like epidemic.

When (16.4.18) is satisfied there is a solution of (16.4.17) with  $X(-\infty) < 1$  and, furthermore,  $X(-\infty) < 1/\beta$  (see Exercise 16.5). Consequently,

$$s_1(-\infty) < g/\theta. \quad (16.4.19)$$

Comparison with (16.4.18) indicates that, after the passage of the wave, the number of susceptibles is insufficient to sustain the wave again. In other words, *according to this model, once a wave epidemic has swept through a population it cannot do so again until more susceptibles have been added to the population.*

So far nothing has been said about the speed at which the epidemic travels through the population. Some information can be garnered from the phase plane for  $X$  and  $Y$  on the assumption that (16.4.18) and (16.4.19) hold. Equations (16.4.14) and (16.4.15) possess then only two critical points at  $(X(-\infty), 0)$  and  $(1, 0)$ , respectively. Near the point  $(X(-\infty), 0)$  put  $X(\zeta) =$

$X(-\infty) + \xi$ ,  $Y(\zeta) = \eta$  and linearise the differential equations (Section 5.3) to give

$$\begin{aligned}\xi'/\alpha &= \left\{ \beta - \frac{1}{X(-\infty)} \right\} \xi + (\beta + \gamma)\eta, \\ \eta'/\alpha &= \left\{ \frac{1}{X(-\infty)} - \beta \right\} \xi - \beta\eta.\end{aligned}$$

The equation for the exponents (see Section 5.4) is

$$\lambda^2 + \frac{\alpha\lambda}{X(-\infty)} - \alpha^2\gamma \left\{ \frac{1}{X(-\infty)} - \beta \right\} = 0, \quad (16.4.20)$$

which corresponds to a saddle-point since  $\beta X(-\infty) < 1$ . The dividing lines of Figure 5.4.3 have equations

$$X(-\infty)(\beta + \gamma)\eta = \left[ \frac{1}{2} - \beta X(-\infty) \pm \frac{1}{2} \{1 + 4\gamma X(-\infty)(1 - \beta X(-\infty))\}^{1/2} \right] \xi$$

with the upper sign for the line which leaves the saddle-point. Evidently, the line with the upper sign has positive slope whereas that with the lower sign has negative slope. Consequently, there is just one trajectory from  $(X(-\infty), 0)$  entering the first quadrant of the phase plane in  $X(\zeta) > X(-\infty)$ .

The linearised equations for the critical point  $(1, 0)$  are the same except for  $X(-\infty)$  being replaced by unity. Hence (16.4.20) continues to hold with the same replacement. Since  $\beta > 1$  the critical point is either a stable node or a stable focus. A focus is unacceptable because a trajectory in its neighbourhood would spiral around the critical point  $(1, 0)$  and values of  $X(\zeta)$  greater than 1 would occur on it. Since  $X(\zeta)$  is not permitted to exceed 1 the critical point cannot be a focus. In order that it be a stable node it is necessary that

$$4(\beta - 1)\gamma \leq 1. \quad (16.4.21)$$

Expressed in terms of the original quantities via (16.4.16) this implies that

$$c \geq 2[\phi s_1(\infty)\{\theta s_1(\infty) - g\}]^{1/2}, \quad (16.4.22)$$

which sets a lower bound on the speed of a possible wave.

To confirm that there is a wave with a speed satisfying (16.4.22) we must show that there is a trajectory connecting the two critical points subject to (16.4.18), (16.4.19) and (16.4.21). Since it has been shown already that there is a trajectory leaving  $(X(-\infty), 0)$  on which  $X(\zeta)$  increases it will be sufficient to verify that this trajectory must end up at  $(1, 0)$ . On  $Y = 0$  we see from (16.4.14) and (16.4.15) that  $Y' > 0$  and  $X' < 0$  between the critical points. Hence no trajectory can leave the first quadrant by crossing  $Y = 0$  between the critical points. On  $X = X(-\infty)$  it is clear that  $X' > 0$  and  $Y' < 0$  so that no trajectory in  $X > X(-\infty)$  can depart through the line  $X = X(-\infty)$ .

Now consider what happens on the line

$$(\beta + \gamma)Y = \left(\frac{1}{2} - \beta\right)(X - 1). \quad (16.4.23)$$

Since  $\ln X \leq X - 1$  for  $X > 0$

$$\begin{aligned} X'/\alpha &\geq -\frac{1}{2}(X - 1) > 0, \\ Y'/\alpha &\leq \left(\frac{1}{2}\beta + \gamma - \gamma\beta\right)(X - 1)/(\beta + \gamma). \end{aligned}$$

Accordingly, on a trajectory

$$\frac{dY}{dX} \leq \frac{2\gamma(\beta - 1) - \beta}{\beta + \gamma} \leq \frac{1/2 - \beta}{\beta + \gamma}$$

by virtue of (16.4.21). The conclusion is that, since  $\beta > 1$ , no trajectory can pass through the line (16.4.23) from below to above. Thus the sides of the triangle bounded by  $Y = 0$ ,  $X = X(-\infty)$  and (16.4.23) are impassable barriers to a trajectory in the interior. It follows from the Poincaré–Bendixson theorem (Section 5.6) that a trajectory from  $(X(-\infty), 0)$  must go to either a limit cycle or the other critical point. If there were a limit cycle the observation at the end of Section 5.6 indicates that there would have to be a critical point interior to it and, therefore, in the interior of the triangle. But that is contrary to what has been demonstrated already. Any trajectory inside the triangle must go more or less directly to the stable node at  $(1, 0)$ . Therefore, the trajectory entering from  $(X(-\infty), 0)$  must go to  $(1, 0)$ , i.e., there is a unique trajectory starting from  $(X(-\infty), 0)$  and ending at  $(1, 0)$ .

The trajectory is traversed from start to finish as  $\zeta$  increases. Increasing  $\zeta$  corresponds to decreasing  $t$ . Therefore, as  $t$  increases,  $\zeta$  falls and progress along the trajectory must be in the opposite direction. In other words, as  $t$  grows with  $x$  fixed,  $s_1$  changes continuously from  $s_1(\infty)$  to  $s_1(-\infty)$ . An epidemic wave can exist that alters the number of susceptibles from  $s_1(\infty)$  before the arrival of the wave to  $s_1(-\infty)$  after the wave has finally died away provided that (16.4.18), (16.4.19) and (16.4.22) are valid. The only restriction on the wave speed is that of (16.4.22). For every value of  $c$  which satisfies (16.4.22) there is a trajectory in the phase plane which accomplishes the transition of susceptibles from start to finish.

In summary the predictions of this model are as follows:

- (i) There can be no epidemic wave unless there are sufficient initial susceptibles as specified by (16.4.18).
- (ii) If there are enough susceptibles initially for a wave the number of susceptibles left behind is insufficient for another wave without regeneration.
- (iii) There can be no wave with a speed less than

$$2[\phi s_1(\infty)\{\theta s_1(\infty) - g\}]^{1/2}$$

and the larger  $s_1(\infty)$  the faster the wave must travel, i.e., the more susceptibles there are initially the quicker the epidemic must spread. There can be epidemic waves with speeds higher than the minimum. Altering  $c$  changes  $\gamma$  (but not  $\beta$  or  $X(-\infty)$ ) and thereby the pattern of trajectories but there is still one from  $(X(-\infty), 0)$  to  $(1, 0)$ .

---

## Exercises

- 16.1 Show that (16.1.8) determines one and only one positive value of  $s(\infty)$  less than  $b/a$ .
- 16.2 In another version of the incubation model, a variable time to become infectious is allowed for by saying that an individual becomes infectious at time  $t$  if exposed to the disease at the earlier time  $\tau(t)$  so that  $\tau(t) \leq t$ . Assume that  $\tau(0) = 0$  and that  $\tau(t)$  increases steadily with  $t$ . Show that the equations for  $I(t)$  in Section 16.3 are replaced by

$$\begin{aligned} I(t) &= I_0(t) + s(0) - s\{\tau(t)\} \quad (0 \leq t \leq \sigma), \\ I(t) &= I_0(t) + s\{\tau(t-\sigma)\} - s\{\tau(t)\} \quad (t \geq \sigma) \end{aligned}$$

when an individual remains infectious for the time  $\sigma$ .

Deduce that

$$\int_0^\infty I(v)dv = \int_0^\sigma I_0(v)dv + \sigma\{s(0) - s(\tau(\infty))\}.$$

Show that (16.3.10) is still satisfied with  $z = s(\tau(\infty))/s(0)$  and

$$\epsilon = \frac{1}{\beta} \left[ a \int_0^\sigma I_0(v)dv - \ln \left\{ \frac{s(\tau(\infty))}{s(\infty)} \right\} \right].$$

Prove that  $\epsilon > 0$ .

- 16.3 To allow for a nonconstant time during which an individual is infectious in Exercise 16.2 let  $p(x)$  be the proportion of individuals who are still infectious after a time  $x$  has elapsed from becoming infectious. Assume that  $p(x)$  is positive, decreases steadily as  $x$  increases and  $p(x) = 0$  if  $x > \sigma$ . Show that

$$I(t) = I_0(t) - \int_0^t p(t-u) \frac{d}{du} s\{\tau(u)\} du.$$

Prove that

$$\int_0^\infty I(v)dv = \int_0^\sigma I_0(v)dv + \{s(0) - s(\tau(\infty))\} \int_0^\sigma p(u)du$$

and that the other conclusions to Exercise 16.2 still hold, but with

$$\beta = as(0) \int_0^\sigma p(u)du.$$

- 16.4 Show that, if  $\epsilon > 0$  in (16.3.10) and  $\beta > 1$ , there is a unique positive root with  $z < 1$  and it satisfies  $0 < z < 1/\beta$ .

[Hint:  $\ln x \leq x - 1$  for  $x > 0$ .]

- 16.5 If  $\epsilon = 0$  in (16.3.10) show that there is one positive root with  $z < 1$  if, and only if,  $\beta > 1$  and that then  $0 < z < 1/\beta$ .

- 16.6 Prove that (16.4.1)–(16.4.3) imply that

$$s(t, x) + I(t, x) + r(t, x) = N(x)$$

where  $N(x)$  is the initial population at point  $x$ , and deduce that

$$\lim_{t \rightarrow \infty} r(t, x) = R(x)$$

exists.

- 16.7 If  $g(t, x) = h(x)$  in (16.4.1)–(16.4.3), prove that

$$N(x) = R(x) + s(0, x) \exp \left\{ - \int_{-\infty}^{\infty} F(x, x') R(x') dx' / h(x') \right\}$$

where  $N$  and  $R$  are defined in Exercise 16.6.

- 16.8 Generalise (16.4.1)–(16.4.3) to two space dimensions  $x, y$  and show that, if

$$F(x, y, x', y') = f_1(|x - x'|)f_2(|y - y'|)$$

in the small region  $|x - x'| < \delta_1, |y - y'| < \delta_2$  but is otherwise zero, the associated diffusion approximation includes a partial differential equation of the form

$$\frac{\partial s}{\partial t} = -s \left( \theta I + \phi_1 \frac{\partial^2 I}{\partial x^2} + \phi_2 \frac{\partial^2 I}{\partial y^2} \right).$$

- 16.9 Substitute for  $\partial s / \partial t$  in (16.4.5) from (16.4.4). Show that the resulting partial differential equation for  $I$  has a solution

$$I(t, x) = \frac{e^{(\theta s - g)t}}{t^{1/2}} e^{-x^2/4\phi t s}$$

when  $s$  and  $g$  are constants. If  $a$  is positive and  $\int_a^\infty I(t, x)dx = C(t)$  find the formula for  $C(t)$ . Deduce that, as  $t \rightarrow 0$  with  $a$  fixed,

$$aC(t) = 2\phi t^{1/2} s \exp\{(\theta s - g)t - a^2/4\phi ts\}$$

approximately.

---

# Answers to Selected Exercises

---

## Chapter 1

1.1  $r_0 - \alpha t$ .

1.2  $p_0(n)^{t/T}$ .

1.4 The curve drops steadily, without any point of inflexion, from an initial value of  $p(0)$  to an eventual value of  $N_0/a$ . This demonstrates that, if the population could be started above  $N_0/a$  by some means, crowding would force it down to  $N_0/a$  in the end.

1.6 (a)  $y = Dte^t$ ; (b)  $y^2 + 3y = Dt$ ; (c)  $y^2 + 1 = Dt(1+t)$ .

1.7 (a)  $Dt = e^{y^2/2t^2}$ ; (b)  $\{(y/t) - 1\}e^{y/t} = C + \ln t$ ; (c)  $y = t/(C + \ln t)$ .

1.8 (a) New equation is  $dy/dt = (t+y)/(t-y)$ , which is of homogeneous type (with solution  $\tan^{-1} z - \frac{1}{2} \ln(1+z^2) - \ln t = C$ ,  $z = y/t$ ).

(b) New equation is  $dy/dt = (at+by)/(a't+b'y)$ , which is of homogeneous type.

1.9 (a)  $y = (t+C)(1-t^2)^{-1/2}$ ; (b)  $y = Ce^{t^2/2} - e^{2t}$ ; (c)  $w = \pm t(C+t^2)^{1/2}$ .

1.10  $y = (1+Ce^{t^2/2})^{-1}$ .

1.11 (a)  $y^2 = C(t+1)^2 - 2(t+1)$ ; (b)  $t^2 = Ce^{2y} + 2y$ .

1.12  $y = \frac{2}{t} + \frac{1}{Ct^8 - 2t/7}$ .

1.13 (a)  $y = \frac{2}{t} + \frac{1}{Ct^5 - t/4}$ ; (b)  $y = t + \frac{1}{t+C}$ .

1.14 (a)  $\left\{ \frac{t-1}{t+1} \right\}^{1/2} y = \frac{1}{2} \ln(t+1) + C$ ;

(b)  $\sinh^{-1}(2y/t) = \ln t + C$ ;

(c)  $e^t + e^{-2y} = C$ ;

(d)  $\frac{1}{2} \ln\{4(t+1)^2 + (2y+3)^2\} + \tan^{-1}\left(\frac{y+3/2}{t+1}\right) = C$ ;

- (e)  $\frac{1}{2} \ln\left(\frac{1+2t-y}{1-2t+y}\right) = 2t + C;$   
 (f)  $(t+a)^{-3}y = (t+a)^2 + C;$   
 (g)  $\tan(2y/t) + \ln t = C;$   
 (h)  $\frac{1}{8}(2y-2t+5)^2 + 2t = C;$   
 (i)  $(2y-1)\cos^2 2t = C(2y+1);$   
 (j)  $(1-t)^{1/2}y = 2(1+t)^{1/2} + C;$   
 (k)  $2t - 2y + \ln\{(2t+2y+1)^2 + 1\} = C;$   
 (l)  $te^y = \tan y + C.$

1.16 40 min.

1.17 200 h.

1.18 98.1%.

1.19  $c = (2D/W)e^{-t/6}$ ; about 8 hours; about  $7\frac{1}{2}$  hours.

1.20 25–100 mg/h.

1.21  $p = p_0 \exp\{-K(t-t_0)/R\}; p = RI_0 + (p_0 - RI_0) \exp\{-K(t-t_0)/R\}.$

## Chapter 2

- 2.1 (a)  $C_1 e^{3t} + C_2 e^{5t};$   
 (b)  $(C_1 + C_2 t) e^{4t};$   
 (c)  $(C_1 \cos 2t + C_2 \sin 2t) e^{-t} + t^2 + \frac{1}{5}t - \frac{12}{25};$   
 (d)  $C_1 e^{2t} + C_2 e^{-t} + 10te^{2t};$   
 (e)  $C_1 \cos \omega t + C_2 \sin \omega t + \cos \Omega t / (\omega^2 - \Omega^2);$   
 (f)  $C_1 \cos \omega t + C_2 \sin \omega t + (t/2\omega) \sin \omega t;$   
 (g)  $(C_1 \cos \frac{1}{2}\sqrt{3}t + C_2 \sin \frac{1}{2}\sqrt{3}t) e^{-t/2} + \frac{3}{2} + \frac{9}{26} \cos 2t - \frac{3}{13} \sin 2t.$
- 2.2 (a)  $C_1 e^{-2t} + C_2 e^{-t} + 2t^2 - 7t + 17/2;$   
 (b)  $C_1 e^{(\sqrt{3}-1)t} + C_2 e^{-(\sqrt{3}+1)t} - \frac{3}{10}t^5 - \frac{3}{2}t^4 - 9t^3 - 36t^2 - 99t - 135;$   
 (c)  $C_1 + C_2 e^{-3t} + \frac{1}{2}t^4 - \frac{2}{3}t^3 + \frac{7}{6}t^2 + \frac{2}{9}t;$   
 (d)  $(C_1 \cos \frac{1}{2}\sqrt{3}t + C_2 \sin \frac{1}{2}\sqrt{3}t) e^{-t/2} + 2e^t(2 \sin t - 3 \cos t);$   
 (e)  $C_1 + C_2 e^{-3t} + \frac{5}{3}t - \frac{3}{2}(3 \cos t + \sin t);$   
 (f)  $(C_1 + C_2 t) e^t + \cos t + 3t^2 e^t;$   
 (g)  $C_1 \cos 3t + C_2 \sin 3t + (\sin 3t - \frac{1}{2} \cos 3t)t.$

- 2.3 (a)  $\{(At + B) \cos t + (Ct + D) \sin t\}t + Et + F;$   
 (b)  $At^2 + Bt + C + (Dt + E)t^2 e^{2t} + (Ft + G) \cos 2t + (Ht + I) \sin 2t;$   
 (c)  $(At^2 + Bt + C)e^t \sin 2t + (Et^2 + Ft + G)e^t \cos 2t + e^t(H \cos t + I \sin t) + Ke^t;$   
 (d)  $(At + B)e^t + \{(Ct + D) \cos 2t + (Et + F) \sin 2t\}t.$
- 2.4  $(C_1 + C_2 \ln t + \ln^2 t)t.$
- 2.5  $C_1 e^{2t} + C_2 e^{-t} + 5te^{2t}.$
- 2.6  $C_1 \cos t + C_2 \sin t + 3t.$
- 2.7  $\frac{1}{5}e^t - \frac{1}{4}t \cos 2t.$
- 2.8 (a)  $\frac{4}{3}e^{-t}(1+t)^{5/2};$   
 (b)  $(C_1 - t) \cos t + (C_2 + \ln \sin t) \sin t.$
- 2.10 (a)  $A \cos t + B \sin t + C \cos 2t + D \sin 2t;$   
 (b)  $(A + Bt)e^{-2t} + Ce^{-3t};$   
 (c)  $A + Bt + (C \cos 3t + D \sin 3t)e^{-t};$   
 (d)  $(C_1 + C_2 t + C_3 t^2 + C_4 t^3)e^t.$
- 2.12 (a)  $C_1 + C_2 e^t + C_3 e^{-t} + \frac{1}{2} \cos t;$   
 (b)  $C_1 \cos t + C_2 \sin t + C_3 e^t + C_4 e^{-t} - 1 - t^2;$   
 (c)  $(t^2 - 5t)e^t + C_1 + C_2 t + C_3 e^t + C_4 e^{-t}.$
- 2.13 (b)  $(At + B)e^{-t^2};$   
 (c)  $(\cos 1 + 2 \sin 1 - 1)/t^2 + (2 + \cos 1 - \sin 1)/t - (t \sin t + 2 \cos t)/t^2.$

## Chapter 3

- 3.1 (a)  $x = C_1 e^t + C_2 e^{-5t} - \frac{6}{7} e^{2t},$   
 $y = -C_1 e^t + C_2 e^{-5t} + \frac{8}{7} e^{2t};$   
 (b)  $x = (C_1 + C_2 t - 8t^{5/2})e^t,$   
 $y = (C_1 - \frac{1}{2}C_3 + C_2 t + 10t^{3/2} - 8t^{5/2})e^t.$
- 3.2  $x = \frac{3}{2} + Ce^{-t/2}, y = -\frac{1}{2} - 3Ce^{-t/2}.$
- 3.3 (a)  $x = C_1 e^{-7t} + 2C_2 e^{-3t},$   
 $y = 2C_1 e^{-7t} + 3C_2 e^{-3t};$

(b)  $x = (C_1 \cos 2t + C_2 \sin 2t)e^{-t},$   
 $y = (C_2 \cos 2t - C_1 \sin 2t)e^{-t};$

(c)  $x = (C_1 + C_2 t)e^t,$   
 $y = C_2 e^t;$

(d)  $x = C_1 + C_3 e^{2t},$   
 $y = C_2 e^{-t} - 2C_3 e^{2t},$   
 $z = C_1 - 2C_2 e^{-t} + C_3 e^{2t};$

(e)  $x = C_1 e^{2t} + \{C_2 + (2+t)C_3\}e^t,$   
 $y = (C_2 + C_3 t)e^t,$   
 $z = C_1 e^{2t} + \{C_2 + (1-t)C_3\}e^t.$

3.4  $x = (2+3t)e^t - t - 1,$   
 $y = (6t+1)e^t - 3t - 1.$

3.5 (a)  $x = e^t - \frac{1}{7}e^{-5t} - \frac{6}{7}e^{2t},$   
 $y = -e^t - \frac{1}{7}e^{-5t} + \frac{8}{7}e^{2t};$

(b)  $x = 2e^{-4t} + e^{3t} + t - 2,$   
 $y = e^{-4t} - 3e^{3t} + 2.$

3.6  $x = t^3 e^{3t} + t,$   
 $y = 1 - 3t,$   
 $z = -6.$

3.7  $t (0 \leq t \leq \pi), \frac{1}{2}\pi e^{\pi-t} - \frac{1}{2}\pi \cos t - (\frac{1}{2}\pi + 1) \sin t (t > \pi).$

3.14  $-\frac{4}{\pi}t^{1/2} \sum_{m=0}^{\infty} \frac{\sin(2m+1)(t-\pi)}{(2m+1)\{(2m+1)^2 + 1\}}.$

---

## Chapter 4

4.1

$$\epsilon \frac{d^2x}{dt^2} + (3x^2 + a) \frac{dx}{dt} + x = x_a.$$

4.4

$$\begin{aligned} \phi'' - c\phi' + \phi(1-\phi)(\phi-a) &= \psi, \\ c\psi' &= b\phi - \gamma\psi. \end{aligned}$$

4.6

$$c = 0.$$

4.8

$$\begin{aligned} a^{-1} &= 1 + \left[ \frac{c}{2} + \sqrt{\left( \frac{c^2}{4} + 1 \right)} \right]^2, \quad \theta = \frac{\pi}{2}; \\ a^{-1} &= 1 + \left( \frac{1 + \left( \frac{c}{2} + \gamma \right) [c + \left( \frac{c}{2} + \gamma \right) p]}{1 + \left( \gamma - \frac{c}{2} \right) ((\gamma - \frac{c}{2}) p - c)} \right)^{1/2} \\ &\quad \times \exp \left[ \frac{c\beta}{2} \cos^{-1} \left( \frac{p - \frac{c^2}{2} - 1}{p + 1} \right) \right], \end{aligned}$$

where

$$\gamma = \sqrt{\left( \frac{c^2}{4} + 1 \right)}, \quad \beta = \sqrt{\left( p - \frac{c^2}{4} \right)}, \quad p = \tan \theta.$$

4.9

$$(0, 0, 0); \quad \left( X_i, \frac{fk_3AX_i}{(k_1A + k_2X_i)}, \frac{k_3AX_i}{k_5} \right),$$

where  $X_i$ ,  $i = 1, 2$ , satisfies

$$2k_2k_4X^2 + [k_2k_3A(f-1) + 2k_1k_4A]X - k_1k_2A^2(f+1) = 0.$$

4.12

$$\begin{aligned} \frac{dX}{dt} &= A - (B+1)X + X^2Y, \\ \frac{dY}{dt} &= BX - X^2Y. \end{aligned}$$

4.14

$$(0, 0); \quad \left( \frac{a}{b}, 0 \right); \quad \left( \frac{e}{f}, \frac{af-be}{cf} \right).$$

4.15

$$\begin{aligned} (0, 0); \quad \left( \frac{a}{b}, 0 \right); \quad &\left( \frac{e}{f} + \bar{X}, \frac{(e+f\bar{X})[af-b(e+f\bar{X})]}{cef} \right), \\ &c \neq 0, af > b(e+f\bar{X}). \end{aligned}$$

$$\begin{aligned} (0, 0); \quad \left( \frac{a}{b}, 0 \right); \quad \left( 0, \frac{e}{g} \right); \quad &\left( \frac{ce-ag}{cf-gb}, \frac{fa-bc}{cf-gb} \right); \\ &\frac{ce-ag}{cf-gb} > 0 \quad \text{and} \quad \frac{fa-bc}{cf-gb} > 0. \end{aligned}$$

## Chapter 5

5.1 (a) None; (b)  $y_0 = 0$ , singular; (c)  $t_0 = 0$ , singular; (d)  $t_0 = 0$ , singular;  
 (e)  $t_0 = 0$ , singular.

5.2 (a) Unstable node, (b) saddle-point, (c) centre, (d) stable node,  
 (e) unstable focus, (f) saddle-point, (g) stable focus, (h) centre.

5.3 Stable node.

5.4  $a < 0$ , saddle-point;  $b^2 > a > 0$  node (stable  $b > 0$ , unstable  $b < 0$ );  
 $b = 0, a > 0$ , centre;  $b^2 < a$  focus (stable  $b > 0$ , unstable  $b < 0$ ).

5.5 Focus.

5.6 The trajectories are  $\frac{1}{2}cy^2 + x^2 + ax^3 + bx^4 = \text{constant}$  and  $x = 0, y = 0$   
 is a centre.

5.8 If  $b > 0$  there is a centre at  $x = 0, y = 0$  and, if  $b < 0$ , a saddle-point. If  
 $b = 0$  either  $x = \text{constant}$ ,  $y = 0$  for all  $t$  or  $y$  remains of one sign.

5.9  $\theta = \pm\pi$  are saddle-points. For  $0 < \mu < 1$  there is a saddle-point at  $\theta = 0$   
 and centres at  $\theta = \pm\cos^{-1}\mu$ . For  $\mu > 1$ , there is a centre at  $\theta = 0$ .

5.10 The trajectories are  $2\lambda^2y^2 = 2C\lambda^2e^{-2\lambda x} + 1 - 2\lambda x$  where  $C$  is a constant.  
 Closed curves for  $-1 < 2\lambda^2C < 0$ .

5.14 Unstable limit cycles.

---

## Chapter 7

7.2 Critical points are  $(u_i, \frac{b}{\gamma}u_i)$ , where  $u_i$ ,  $i = 1, 2, 3$ , satisfies

$$u^3 - u^2(1+a) + \left(a + \frac{b}{\gamma}\right)u - I = 0,$$

$$I = \frac{b}{\gamma}: \quad (u, w) = \left(1, \frac{b}{\gamma}\right), \quad \text{if } a^2 < \frac{4b}{\gamma};$$

$$(u, w) = \left(1, \frac{b}{\gamma}\right), \left(\frac{a \pm \sqrt{(a^2 - 4b/\gamma)}}{2}, \frac{b}{\gamma} \left[\frac{a \pm \sqrt{(a^2 - 4b/\gamma)}}{2}\right]\right),$$

if  $a^2 \geq \frac{4b}{\gamma}$ ;

$$I = \frac{ab}{\gamma}: \quad (u, w) = \left(a, \frac{ab}{\gamma}\right), \quad \text{if } \frac{b}{\gamma} > \frac{1}{4};$$

$$(u, w) = \left( a, \frac{ab}{\gamma} \right), \left( \frac{1 \pm \sqrt{(1 - 4b/\gamma)}}{2}, \frac{b}{\gamma} \left[ \frac{1 \pm \sqrt{(1 - 4b/\gamma)}}{2} \right] \right), \text{ if } \frac{1}{4} \geq \frac{b}{\gamma}.$$

7.4

$$\begin{aligned} c\phi' &= \phi'' + f(\phi) - \psi, \\ c\psi' &= b\phi - \gamma\psi. \end{aligned}$$

Unique rest state if  $f(\phi) = \phi g(\phi)$  and  $g(\phi) \neq b/\gamma$ .

---

## Chapter 9

9.1 47 years; 200 whales.

9.2 1.26 million (in fact 5 million were killed each year in the early 1870s).

9.3 If  $A = ad - cg$  there are three critical points where  $A > 0$ : saddle-points at  $(0, 0)$  and  $(a/g, 0)$  whereas  $(c/d, A/bd)$  is a stable node if  $A < g^2c/4d$  but a stable focus if  $A > g^2c/4d$ . Predator population becomes  $A/bd$  eventually. If  $A < 0$  there is a saddle-point at  $(0, 0)$  and a stable node at  $(a/g, 0)$ ; the predators are eventually wiped out.

9.4 Four critical points  $(0, 0)$ ,  $(a/g, 0)$ ,  $(0, c'/g')$ ,  $\left( \frac{ag' - bc'}{gg' + bd}, \frac{ad + c'g}{gg' + bd} \right)$ . No.

---

## Chapter 10

10.2  $2y + 2(y - x)^2$ .

10.3  $3x$ .

10.4  $4\theta + 2\rho e^{-\theta}$ .

10.5  $3(y - x)(x - 1) + 3(y - x)^2 e^{-2x} + 3(y - x)e^{-x}$ .

10.6  $\ln y - x^2 + y^2 - \frac{1}{2} \ln(y^2 - x^2)$ .

10.7  $\frac{1}{2}x^2 - \frac{1}{4}y^2 + \frac{1}{2}x^2y + \frac{1}{4}$ .

10.8  $x(1 + 2y/x)^2 - 2xy$ .

10.9  $\frac{1}{2}xy + f(y/x)$  where  $f$  is any function such that  $f(1) = 0$ .

10.10 (a) Elliptic everywhere; (b) hyperbolic everywhere; (c) elliptic in  $|x| > 2|y|$ , hyperbolic in  $|x| < 2|y|$ ; (d) hyperbolic in first and third quadrants, elliptic in second and fourth; (e) parabolic everywhere; (f) parabolic everywhere.

$$10.11 \quad 4 \cos^2 \alpha \frac{\partial^2 u}{\partial \xi \partial \eta} = \frac{\partial u}{\partial \xi} + \frac{\partial u}{\partial \eta}.$$

$$10.12 \quad 2 \frac{\partial^2 u}{\partial \xi \partial \eta} = \frac{1}{\xi - \eta} \left( \frac{\partial u}{\partial \xi} - \frac{\partial u}{\partial \eta} \right).$$

$$10.13 \quad 3 \ln y + 3x^2y - x^2y^4 - 6.$$

$$10.14 \quad (Ax + B)^{2\nu/(2\nu-1)}.$$

10.15 (a) One solution; (b) several solutions; (c) either one or none.

10.21 Use the polar version of the two-dimensional animation program of Section 10.10. The command for defining U is:

$$(a) \quad U = \cos(\Omega * t(I) + m * \Theta + R);$$

$$(b) \quad U = \cos(\Omega * t(I) + m * \Theta + \log(R));$$

## Chapter 11

11.2

$$u(x, t) = 1 - \phi(x/\sqrt{4t}), \quad \phi(s) = \frac{2}{\sqrt{\pi}} \int_0^s e^{-t^2} dt.$$

11.3

$$u(x, t) = \sum_{n=1}^{\infty} A_n e^{-n^2 \pi^2 t} \sin n\pi x, \quad A_n = 2 \int_0^1 f(x) \sin n\pi x dx.$$

11.4

$$u(x, t) = \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{e^{-n^2 t}}{n^3} \sin nx.$$

11.5

$$u(x, t) = 2 + 2 \sum_{n=1}^{\infty} \frac{[(-1)^{n+1} - 1]}{(2 + n^2)} e^{-n^2 t} \cos nx.$$

11.6

$$u(x, y, t) = \frac{4}{\pi^2} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \frac{e^{-(m^2 + n^2)\pi^2 t}}{mn} (-1)^{m+n} \sin m\pi x \sin n\pi y.$$

11.8

$$\begin{aligned}
u &= \frac{1}{2} \exp(c\xi), \quad -\infty < \xi < 0, \\
&= \frac{1}{2} + \frac{c}{2\sqrt{(c^2-4)}} \exp\left[\left(\frac{c+\sqrt{(c^2-4)}}{2}\right)\xi\right] \\
&\quad - \frac{c}{2\sqrt{(c^2-4)}} \exp\left[\left(\frac{c-\sqrt{(c^2-4)}}{2}\right)\xi\right], \quad 0 \leq \xi \leq \xi_1 \\
&= 1 - \frac{1}{4} \exp\left[\left(\frac{c-\sqrt{(c^2+4)}}{2}\right)(\xi-\xi_1)\right], \quad \xi_1 < \xi < \infty,
\end{aligned}$$

where  $c > 2$  with  $c$  and  $\xi_1$  satisfying

$$\begin{aligned}
\exp\left(\frac{1}{2}c\xi_1\right) \sinh\left(\frac{\sqrt{(c^2-4)}}{2}\right)\xi_1 &= \frac{\sqrt{(c^2-4)}}{4c}, \\
\coth\left(\frac{\sqrt{(c^2-4)}}{2}\right)\xi_1 &= 4\frac{\sqrt{(c^2+4)}}{c} - 2.
\end{aligned}$$

11.14 No.

## Chapter 12

12.2

$$c(x, t) = \frac{c_0}{(4\pi t)^{\frac{1}{2}}} \int_{-a}^a \exp\left(-\frac{(x-y)^2}{4t}\right) dy.$$

12.3

$$\begin{aligned}
c(r, t) &= \frac{2c_0 a}{r\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin\left(\frac{n\pi r}{a}\right) \exp\left(-\frac{Dn^2\pi^2 t}{a^2}\right), \quad r \leq a, \\
&= c_0 \left(1 - \frac{2a}{r\pi^{\frac{1}{2}}} \int_{(r-a)/(4\pi t)^{1/2}}^{\infty} e^{-x^2} dx\right), \quad r > a.
\end{aligned}$$

12.7

$$(u, v) = \left(\frac{\alpha}{\alpha - \beta}, \frac{\beta}{\alpha - \beta}\right), \quad \alpha > \beta, \quad \alpha > 1$$

$$\delta > \sqrt{\frac{\alpha}{\beta}}(\sqrt{\alpha - \beta} + \sqrt{\alpha}).$$

## Chapter 14

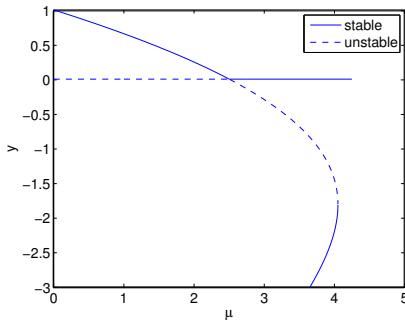
14.1 Fixed points are at  $(0, 0)$  (unstable);  $(0, 0.1)$  (saddle);  $(1, 0)$  (saddle);  $(0.37, 0.47)$  (stable focus).

14.3 (a) When  $\mu = 0$ , fixed points are  $y = 0$  (unstable) and  $y = 1$  (stable).

(b) The function required in MATLAB is

```
function f = myrhs(z)
r = 5;
f = r *z(1)*(1-z(1)) - z(2)*z(1)*(1+exp(-0.5*z(1)));
```

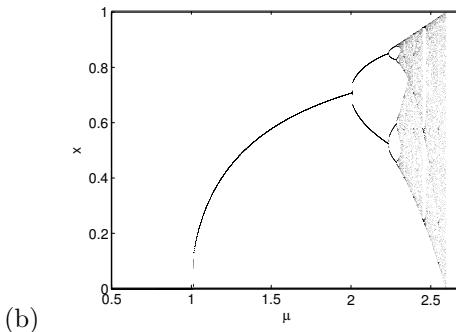
$z(2)$  represents the parameter  $\mu$ . Using this function in conjunction with `newton`, `path_follow` and `calculate_eigenvalues` produces the following bifurcation diagram:



14.4 The system has a Hopf bifurcation at  $\mu_c = 15.1$ . For  $\mu < \mu_c$ , there is a stable fixed point; for  $\mu > \mu_c$ , there is an unstable fixed point and a stable limit cycle.

14.5 The unstable and saddle fixed points do not change location or stability as  $a$  changes; the stable fixed point changes location ( $\sqrt{y_1^2 + y_2^2}$  decreases as  $a$  increases) but does not change stability in the range  $0 < a < 20$ .

14.7 (a)  $\mu = 3\sqrt{3}/2$ .



- 14.8 (a) Fixed points are at  $\mathbf{y}_1^* = (0, 0, 0)$  and  $\mathbf{y}_2^* = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1)$ . The Jacobian matrix at these points is

$$J(\mathbf{y}_1^*) = \begin{bmatrix} -\sigma & \sigma & 0 \\ r & -1 & 0 \\ 0 & 0 & -b \end{bmatrix},$$

$$J(\mathbf{y}_2^*) = \begin{bmatrix} -\sigma & \sigma & 0 \\ 1 & -1 & \mp\sqrt{b(r-1)} \\ \pm\sqrt{b(r-1)} & \pm\sqrt{b(r-1)} & -b \end{bmatrix}.$$

- (b) The fixed point  $\mathbf{y}_1^*$  is unstable for all values of  $r$ ;  $\mathbf{y}_2^*$  is stable for  $r < 24.3$  and unstable for  $r > 24.3$ .

- 14.9 The maximal Lyapunov exponent is negative for  $r < 24.3$  and positive for  $r > 24.3$ .

## Chapter 15

15.5

$$P = \frac{(S_2 - S_1)}{4}(r^2 - R^2) + \frac{\alpha}{R},$$

$$R(t) = R(0) \exp(S_1 - S_2) \frac{t}{2}.$$

15.6

$$X(t) = X(0) \exp(S_1 - S_2)t,$$

where  $X(t)$  is the outer boundary of the tumour colony.



---

# Index

- Activator, 316  
Activator-inhibitor system, 317  
Adrenalin, 94  
Alberts, B., 409  
Anderson, A. R. A., 408  
Angiogenesis, 389, 398, 408  
Animations in MATLAB, 252  
Apoptosis, 394, 399  
Arimoto, S., 111  
Atrium, 92–94  
Attractor, 168  
Autonomous system, 350  
Averaging, 355–363  
Axon, 94–100  
Axoplasm, 94–100
- Béchet, J. J., 111  
Belousov, B. P., 101–103, 110, 197–207, 308–311, 325  
Belousov–Zhabotinskii reaction, 100–103, 110, 197–198, 204, 308–311  
Bernoulli’s differential equation, 29  
Bessel function, 149  
Bessel’s equation, 148  
Bifurcation, 329, 331
  - diagram, 331–333, 335, 340, 343, 354, 370, 373–380
  - discrete, 336–342, 378–380
  - fold, 332, 370, 374
  - Hopf, 335, 346, 376–378
  - numerical analysis, 370–376
  - of limit cycle, 334–336
  - period-doubling, 339–346, 354, 378–380
  - pitchfork, 331, 375
  - point, 277, 329, 331
  - saddle-node, 332
- theory, 277  
Bilipid layer, 295  
Boundary conditions, 81
  - Dirichlet, 86, 236
  - mixed, 237
  - Neumann, 87, 236
  - no flux, 321
  - periodic, 81
- Boundary value problem, 78–87, 275  
Bray, D., 409  
Brindley, J., 386  
Brunner, H., 409  
Byrne, H. M., 408
- Carrying capacity, 10, 107  
Cauchy’s problem, 247–250  
Cells
  - division, 9–11, 391, 402
- Cellular transport model, 295  
Chaos, 340, 342–346, 349, 354, 380–384  
Chaplain, M. A. J., 408, 409  
Characteristic equation, 37, 52
  - repeated roots, 53
- Characteristics, *see* Partial differential equations  
Chemical reactions, 100–105, 110, 308–311  
Chlorine, 94  
Commensalism, 106  
Comparison theorem, 277–288  
Competition, 106
  - model, 109, 143
- Complementary function, 36–40, 52  
Confined set, 325  
Conservation of mass, 393–394  
Conservative system, 132, 330  
Contracting set, 325

- Cooke, R., 408  
 Critical point, *see* Equilibrium  
 Cronin, J., 110
- d'Alembert's solution, 242, 254  
 Damping, 50, 51  
 Darcy's law, 403  
 Diastole, 92, 163, 166  
*Dictyostelium discoideum*, 204  
 Diffusion, 103, 270, 390, 402  
     coefficient, 294  
     driven instability, 311–325  
     equation, 234, 259–264, 293  
     matrix, 293, 313  
     system, 293–325  
     through membranes, 293–299
- Drugs  
     concentration of, 4–8
- Duffing's equation, 161
- Dynamic programming, 414
- Eigenfunction, 79, 82, 263  
     orthogonal, 81
- Eigenvalues, 69, 79, 121, 122, 127,  
     138, 263, 331, 367, 372, 385
- Embryology, 311
- Energy method, 299–303
- Enzymatic reactions, 21, 103–105, 110
- Epidemic, 118–119, 128, 411–424  
     diffusion approximation, 419  
     incubation, 414–418  
     spreading in space, 418–424
- Equilibrium, 95, 118–132, 212–215  
     and bifurcation, 329–334, 367–  
     369  
     discrete system, 337  
     in partial differential equations,  
     311–325, 384–385  
     pseudo-steady state, 105
- Euler's method, 20
- Euler–Darboux equation, 245, 250–  
     251
- Evolutionary equation, 269–277, 313
- Excitability, 177–184
- Existence of solution, 35, 115–118,  
     153–159  
 Existence theorem, 115, 153–159
- Factorisation of the operator, 46–47  
 Feigenbaum's number, 344  
 Fick  
     first law, 294  
     second law, 295, 312
- Field, R. J., 101, 110
- Field–Noyes model, 101–103, 376
- First-order differential equations, 16–  
     19, 153  
     homogeneous, 14–16  
     systems of, 61–68, 157
- Fisher's equation, 198–199
- Fisher, R. A., 198
- Fishing, 213
- FitzHugh, R., 98, 110, 177, 304
- FitzHugh–Nagumo model, 98, 177, 304
- Fixed point, *see* Equilibrium
- Folkman, J., 408
- Forced oscillation, 50–52, 139–143
- Fourier  
     integral theorem, 260–262  
     reciprocal relation, 260  
     series, 263–269  
     transform, 260
- Friedman, A., 289
- Functional response, 108
- Fundamental solution, 262
- Fundamental system, 67–72
- Gauss divergence theorem, 295
- General solution, 34, 55  
     of  $n^{\text{th}}$ -order linear equations, 52  
     of first-order linear equations, 36  
     of second-order linear equations,  
     37  
     of systems of linear equations,  
     61, 66–77
- Gibbs–Thomson relation, 392, 402,  
     403
- Gompertz growth law, 28
- Greenspan, H. P., 400, 407

- Griffiths, D. F., 270, 289  
 Grindrod, P., 207  
 Growth  
     exponential, 2, 107  
     Gompertz, 28  
     logistic, 9–11, 107, 212, 368  
     population, 1–4  
     rate, 2  
     tumours, 389
- Hardy–Weinberg law, 91  
 Hastings, S. P., 325  
 Heart  
     beat, 91–94, 109, 166–174  
     equations, 170  
     model, 163–166
- Heat equation, *see* Diffusion equation
- Hill equation, 21, 26  
 Hodgkin, A. L., 96–100, 109, 304  
 Hodgkin–Huxley equations, 97  
 Holling, C. S., 108, 110  
 Homogeneous differential equations, 14–16  
     associated, 37, 52  
 Huxley, A. F., 96–100, 109, 304
- Immunity, 412  
 Inhibitor, 316, 402  
 Initial value problem, 77–78  
 Integrating factor, 17–19  
 Invariant set, 137  
 Iteration scheme, 337–346, 370, 378–380
- Jacobian matrix, 120, 122, 138, 150, 367, 369, 372, 385  
 John, F., 262, 289  
 Junge, D., 110
- Körös, E., 102, 110  
 Katz, B., 110  
 Keener, J., 207  
 Keller, J. B., 194  
 Kermack–McKendrick model, 411–413, 418
- King, R. J. B., 409  
 Knierim, M., 409  
 Kolmogorov, A., 110, 309
- Laplace’s equation, 234  
 Laplacian operator, 205, 234, 259, 293  
 Law of mass action, 100, 101, 104  
 Levine, H. A., 408  
 Lewis, J., 409
- Limit cycle, 133–139, 184, 205, 218–220, 346–349, 376–378
- bifurcation, 334–336  
 criterion, 137
- Linear differential equations  
      $n^{\text{th}}$  order, 52–54, 64–65  
     constant coefficients, 36–49  
     first-order, 16–19, 35–36  
     systems of, 61–68
- Logistic  
     growth equation, 9–11, 16, 107, 212, 342  
     map, 337–346, 380
- Lorenz equations, 382  
 Lorenz, E., 386  
 Lotka, A. J., 107, 110  
 Lyapunov exponent, 380–384
- MacArthur, R. H., 108, 111  
 Map, *see* Iteration scheme  
 Mathematical modelling, 91  
 Matrix notation, 72–77, 120, 122, 143  
 Maximum principle, 277–287  
 Maynard Smith, J., 110  
 McKean, H. P., 99, 111, 194  
 McLachlan, J. C., 408  
 Michaelis–Menten kinetics, 105, 108  
 Mitchell, A. R., 270, 289  
 Morphogenesis, 312  
 Moving boundary problem, 389–399  
 Murray, J. D., 110, 207, 325  
 Myelinated nerve, 96
- Nagumo, J., 98, 111, 177, 304  
 Necrosis, 391, 394  
 Nerve impulse transmission, 94–100, 109, 177–193, 304–308

- Newton's method, 150, 367, 370, 385  
 Nilsen-Hamilton, M., 408  
 Nodes of Ranvier, 95  
 Nonautonomous system, 350  
 Nonlinear differential equations, 11–16  
     systems of, 118–153  
 Normal form  
     elliptic partial differential equation, 236, 237, 240  
     hyperbolic partial differential equation, 240  
     parabolic partial differential equation, 239, 240  
 Noyes, R. M., 101, 110  
 Nullcline, 145, 147  
 Numerical methods  
     bifurcations, 370–376  
     discrete systems, 378–380  
     equilibrium and stability, 367–370  
     first-order equations, 19–24  
     limit cycles, 376–378  
     linear boundary value problems, 84–87  
     Lyapunov exponents, 380–384  
     nonlinear boundary value problems, 149–153  
     partial differential equations, 384–385  
     systems of equations, 143–147  
 Order of a differential equation, 33  
 Pacemaker, 92, 172–174  
 Pamuk, S., 408  
 Parseval equality, 302  
 Partial differential equations, 33, 103  
     characteristic form, 226, 235  
     characteristics, 223–250  
     elliptic, 234–240  
     first-order, 223–231  
     hyperbolic, 234, 239–251  
     parabolic, 234, 239–240, 259–264  
     plotting solutions, 251–254  
     second-order, 232–251, 259–288, 293–325  
     stability of, 384–385  
 Particular integral, 36, 37, 41–49, 52  
 Partition coefficient, 296  
 Path following, 370–378  
 Pattern and form, 312  
 Pattern formation, 311–325  
     in finite domains, 321  
 Paweletz, N., 409  
 Periodic motion, 125  
 Petrovskii, I., 110, 309  
 Phase plane, 119–121, 168–171, 177–187, 198–199, 212, 412  
     plotting, 144, 146  
 Phase space, 120, 349, 382  
 Piecewise linear model, 99, 190–193  
 Piskounov, N., 110, 309  
 Plank, M. J., 408, 409  
 Plasma membrane, 295  
 Poincaré  
     inequality, 300  
     plane, 350–355  
     section, 346–349, 376  
 Poincaré–Bendixson theorem, 137–139, 180, 183, 219–220, 423  
 Positional information, 312  
 Positive feedback system, 317  
 Potassium ions, 94  
 Predation, 106  
 Predator–prey system, 106–110, 211–220  
 Protter, M. H., 289  
 Pseudo-steady state, 105  
 Quiescence, 391  
 Raff, M., 409  
 Rayleigh's equation, 161  
 Reaction–diffusion, 102, 103, 197, 204, 270, 299, 308, 313, 402  
 Repetitive firing, 177–184  
 Resonance  
     subharmonic, 142  
 Riccati equation, 29

- Rinzel, J., 194  
Roberts, K., 409  
Rosenzweig, M. L., 108, 111  
Rosenzweig–MacArthur model, 108, 217  
Rubinow, S. I., 325  
Runge–Kutta method, 383  
Rybak, B., 111  
  
Schwarz's inequality, 202  
Self-excited oscillation, 140  
Sensitivity to initial conditions, 343, 346, 380  
Separation of variables, 11–13, 25, 105, 298, 314, 323  
partial differential equations, 259, 262–270  
Sherratt, J. A., 409  
Simple harmonic motion, 51, 52  
Singh, G. D., 408  
SIR model, 411  
Sleeman, B. D., 408, 409  
Smoller, J., 325  
Sneyd, J., 207  
Sodium ions, 94  
Solute, 294  
Solvent, 294  
Space clamp, 96  
Sprott, J. C., 386  
Stability, 212–215  
and bifurcation, 329–334, 367–369  
centre, 125, 127  
focus, 126, 127  
global, 121, 128–132  
in diffusion systems, 311–325  
in discrete systems, 337–340  
in partial differential equations, 384–385  
limit cycles, 134–139, 218–220, 346–349, 376–378  
local, 121–128  
node, 122, 127  
saddle-point, 123, 127  
Steady state, *see* Equilibrium  
  
Strange attractor, 381  
Strogatz, S. H., 385  
Sturm–Liouville problem, 79  
Surface plotting in MATLAB, 251  
Surface tension, 392, 394  
Symbolic computation  
    first-order equations, 24–27  
    first-order systems, 147–148  
    higher-order equations, 148–149  
Systole, 166  
  
Taylor series, 150, 314  
Test determinant, 61–64  
Threshold, 93, 95, 166–168, 177, 412  
Transient, 51  
Trapping region, 137, 139  
Trigger, 93, 103, 166  
Tuckwell, H. C., 110  
Tumour  
    angiogenesis factor, 390  
    avascular, 389  
    model, 389–407  
    spherical, 395–401, 404–407  
    stability, 399–401, 406–407  
    vascular, 389  
Turing instability, 311–325  
Turing space, 320  
Turing, A. M., 312, 325  
  
Undetermined coefficients, 41–45  
Uniqueness of solution, 35, 55–56, 115–118, 238  
  
Vaccination, 413–414  
van der Pol  
    equation, 133, 134, 140, 146, 363  
    oscillator, 98  
Variation of parameters, 47–49, 54  
Vector notation, 72–77, 120, 122, 143  
Ventricle, 92  
Verhulst's differential equation, 3  
Volterra, V., 107, 111  
Volterra–Lotka model, 107–108, 143, 215–220, 368  
  
Watson, J. D., 409

## Wave

chemical, 103  
electrochemical, 92, 204  
equation, 240–245, 254  
front, 271  
pulse, 271  
speed, 271, 419  
spiral, 204–207  
train, 271  
travelling, 99, 103, 184–190, 197–  
203, 271, 419

Weinberger, H. F., 289

Wolpert, L., 312, 325

Wronskian, 67, 70

Xiao-Qiang, 409

Xinfu Zou, 409

Yoshizawa, S., 111

Zaikin, A. N., 111

Zeeman, E. C., 94, 109

Zhabotinskii, A. M., 101, 111, 197–  
207, 308–311, 325

Praise for the First Edition

“...concise coverage of a broad range of topics. ...”

—*SIAM Review*, Vol. 46, No. 1

“... a valuable addition to the growing literature in mathematical biology from a deterministic modeling approach. ... topics are carefully chosen and well balanced. ...The book is written by experts in the research fields of dynamical systems and population biology. As such, it presents a clear picture of how applied dynamical systems and theoretical biology interact and stimulate each other ... .”

—*Mathematical Reviews*, Issue 2004g

Applying mathematics to biological and physical situations, **Differential Equations and Mathematical Biology, Second Edition** introduces fundamental modeling and analytical techniques used to understand biological phenomena. In this edition, many of the chapters have been expanded to include new and topical material.

**New to the Second Edition**

- A section on spiral waves
- Recent developments in tumor biology
- More on the numerical solution of differential equations and numerical bifurcation analysis
- MATLAB® files available for download online
- Many additional examples and exercises

The book uses various differential equations to model biological phenomena, including the heartbeat cycle, chemical reactions, electrochemical pulses in the nerve, predator-prey models, tumor growth, and epidemics. It explains how bifurcation and chaotic behavior play key roles in fundamental problems of biological modeling. The authors also present a unique treatment of pattern formation in developmental biology based on Turing's famous idea of diffusion-driven instabilities.



**CRC Press**

Taylor & Francis Group  
an Informa business  
[www.crcpress.com](http://www.crcpress.com)

6000 Broken Sound Parkway, NW  
Suite 300, Boca Raton, FL 33487  
270 Madison Avenue  
New York, NY 10016  
2 Park Square, Milton Park  
Abingdon, Oxon OX14 4RN, UK

C8357

ISBN: 978-1-4200-8357-6

9 0000



9 781420 083576