

ADVANCED MODERN ENGINEERING MATHEMATICS

Fifth Edition

Glyn James & Phil Dyke



Pearson



Advanced Modern Engineering Mathematics

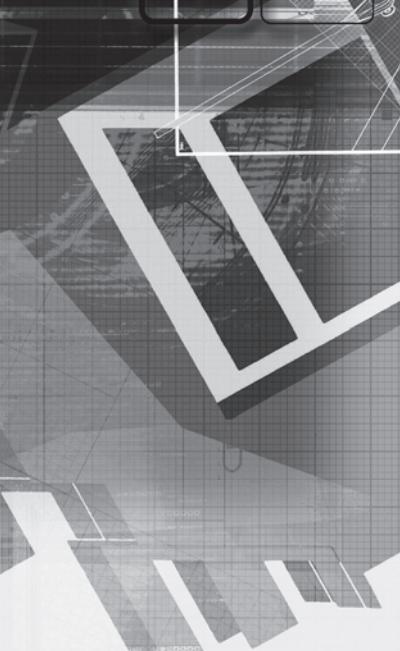
Fifth Edition



We work with leading authors to develop the strongest educational materials in mathematics, bringing cutting-edge thinking and best learning practice to a global market.

Under a range of well-known imprints, including Prentice Hall, we craft high-quality print and electronic publications which help readers to understand and apply their content, whether studying or at work.

To find out more about the complete range of our publishing, please visit us on the World Wide Web at:
www.pearsoned.co.uk



Advanced Modern Engineering Mathematics

Fifth Edition

Glyn James

Phil Dyke

and

David Burley

Dick Clements

Matthew Craven

Tim Reis

John Searl

Julian Stander

Nigel Steele

Jerry Wright

Coventry University

University of Plymouth

University of Sheffield

University of Bristol

University of Plymouth

University of Greenwich

University of Edinburgh

University of Plymouth

Coventry University

AT&T



Pearson

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Dubai • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • São Paulo • Mexico City • Madrid • Amsterdam • Munich • Paris • Milan

Pearson Education Limited

KAO Two
KAO Park
Harlow CM17 9NA
United Kingdom
Tel: +44 (0)1279 623623
Web: www.pearson.com/uk

First published 1993 (print)

Second edition published 1999 (print)

Third edition published 2004 (print)

Fourth edition published 2011 (print)

Fifth edition published 2018 (print and electronic)

© Pearson Education Limited 1993, 2011 (print)

© Pearson Education Limited 2018 (print and electronic)

The rights of Glyn James, David Burley, Dick Clements, Matthew Craven, Phil Dyke, Tim Reis, John Searl, Julian Stander, Nigel Steele, and Jerry Wright to be identified as authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

The print publication is protected by copyright. Prior to any prohibited reproduction, storage in a retrieval system, distribution or transmission in any form or by any means, electronic, mechanical, recording or otherwise, permission should be obtained from the publisher or, where applicable, a licence permitting restricted copying in the United Kingdom should be obtained from the Copyright Licensing Agency Ltd, Barnard's Inn, 86 Fetter Lane, London EC4A 1EN.

The ePublication is protected by copyright and must not be copied, reproduced, transferred, distributed, leased, licensed or publicly performed or used in any way except as specifically permitted in writing by the publishers, as allowed under the terms and conditions under which it was purchased, or as strictly permitted by applicable copyright law. Any unauthorised distribution or use of this text may be a direct infringement of the authors' and the publisher's rights and those responsible may be liable in law accordingly.

All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

Pearson Education is not responsible for the content of third-party internet sites.

ISBN: 978-1-292-17434-1 (print)

978-1-292-17582-9 (PDF)

978-1-292-17583-6 (ePub)

British Library Cataloguing-in-Publication Data

A catalogue record for the print edition is available from the British Library

Library of Congress Cataloging-in-Publication Data

Names: James, Glyn, author.

Title: Advanced modern engineering mathematics / Glyn James (Coventry University) [and nine others].

Description: Fifth edition. | Harlow, United Kingdom : Pearson Education, 2018. | Includes index.

Identifiers: LCCN 2018008839 | ISBN 9781292174341 (print) | ISBN 9781292175829 (pdf) | ISBN 9781292175836 (epub)

Subjects: LCSH: Engineering mathematics

Classification: LCC TA330 .A38 2018 | DDC 620.001/51--dc23

LC record available at https://urldefense.proofpoint.com/v2/url?u=https-3A__lccn.loc.gov_2018008839&d=DwIFAg&c=0YLnzTkWODJlub_y7qAx8Q&r=eK0q0-QqUPLJD1OLTc7YiWdHxmNowNBMcvK9N3XeA-U&m=4G2pfE4jTE5-LBbM4fpokbAjBOaA-h39xC4FZ8wxPIU&s=fH6box9GavoU_1p97XMA38XG0TLT-fWxCM6hQneaghGY&e=

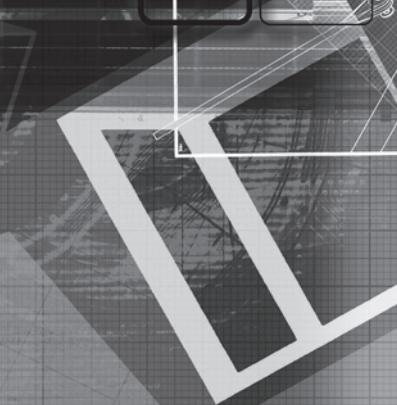
10 9 8 7 6 5 4 3 2 1

22 21 20 19 18

Print edition typeset in 10/12pt Times LT Std by Spi Global

Printed and bound in Malaysia

NOTE THAT ANY PAGE CROSS REFERENCES REFER TO THE PRINT EDITION



Contents

Preface	xix
About the Authors	xxi
Publisher's Acknowledgements	xxiii

Chapter 1 Matrix Analysis	1
1.1 Introduction	2
1.2 Review of matrix algebra	2
1.2.1 Definitions	3
1.2.2 Basic operations on matrices	3
1.2.3 Determinants	5
1.2.4 Adjoint and inverse matrices	5
1.2.5 Linear equations	7
1.2.6 Rank of a matrix	8
1.3 Vector spaces	9
1.3.1 Linear independence	10
1.3.2 Transformations between bases	11
1.3.3 Exercises (1–4)	13
1.4 The eigenvalue problem	13
1.4.1 The characteristic equation	14
1.4.2 Eigenvalues and eigenvectors	16
1.4.3 Exercises (5–6)	22
1.4.4 Repeated eigenvalues	22
1.4.5 Exercises (7–9)	26
1.4.6 Some useful properties of eigenvalues	26
1.4.7 Symmetric matrices	28
1.4.8 Exercises (10–13)	29

1.5	Numerical methods	29
1.5.1	The power method	29
1.5.2	Exercises (14–18)	35
<hr/>		
1.6	Reduction to canonical form	36
1.6.1	Reduction to diagonal form	36
1.6.2	The Jordan canonical form	39
1.6.3	Exercises (19–26)	43
1.6.4	Quadratic forms	44
1.6.5	Exercises (27–33)	50
<hr/>		
1.7	Functions of a matrix	51
1.7.1	Exercises (34–41)	62
<hr/>		
1.8	Singular value decomposition	63
1.8.1	Singular values	65
1.8.2	Singular value decomposition (SVD)	69
1.8.3	Pseudo inverse	72
1.8.4	Exercises (42–49)	78
<hr/>		
1.9	State-space representation	79
1.9.1	Single-input-single-output (SISO) systems	79
1.9.2	Multi-input-multi-output (MIMO) systems	84
1.9.3	Exercises (50–54)	85
<hr/>		
1.10	Solution of the state equation	86
1.10.1	Direct form of the solution	86
1.10.2	The transition matrix	88
1.10.3	Evaluating the transition matrix	89
1.10.4	Exercises (55–60)	91
1.10.5	Spectral representation of response	92
1.10.6	Canonical representation	95
1.10.7	Exercises (61–67)	100
<hr/>		
1.11	Engineering application: Lyapunov stability analysis	101
1.11.1	Exercises (68–72)	103
<hr/>		
1.12	Engineering application: capacitor microphone	104
<hr/>		
1.13	Review exercises (1–19)	108

Chapter 2 Numerical Solution of Ordinary Differential Equations 113

2.1	Introduction	114
2.2	Engineering application: motion in a viscous fluid	114
2.3	Numerical solution of first-order ordinary differential equations	115
2.3.1	A simple solution method: Euler's method	116
2.3.2	Analysing Euler's method	120
2.3.3	Using numerical methods to solve engineering problems	123
2.3.4	Exercises (1–7)	125
2.3.5	More accurate solution methods: multistep methods	126
2.3.6	Local and global truncation errors	132
2.3.7	More accurate solution methods: predictor–corrector methods	134
2.3.8	More accurate solution methods: Runge–Kutta methods	139
2.3.9	Exercises (8–17)	143
2.3.10	Stiff equations	145
2.3.11	Computer software libraries	147
2.4	Numerical methods for systems of ordinary differential equations and higher-order differential equations	149
2.4.1	Numerical solution of coupled first-order equations	149
2.4.2	State-space representation of higher-order systems	154
2.4.3	Exercises (18–23)	158
2.4.4	Boundary-value problems	159
2.4.5	The method of shooting	160
2.5	Engineering application: oscillations of a pendulum	162
2.6	Engineering application: heating of an electrical fuse	167
2.7	Review exercises (1–12)	172

Chapter 3 Vector Calculus 175

3.1	Introduction	176
3.1.1	Basic concepts	177
3.1.2	Exercises (1–10)	184
3.1.3	Transformations	185

3.1.4 Exercises (11–17)	188
3.1.5 The total differential	188
3.1.6 Exercises (18–20)	192
<hr/>	
3.2 Derivatives of a scalar point function	192
3.2.1 The gradient of a scalar point function	192
3.2.2 Exercises (21–30)	195
<hr/>	
3.3 Derivatives of a vector point function	196
3.3.1 Divergence of a vector field	196
3.3.2 Exercises (31–37)	198
3.3.3 Curl of a vector field	199
3.3.4 Exercises (38–45)	202
3.3.5 Further properties of the vector operator ∇	202
3.3.6 Exercises (46–55)	206
<hr/>	
3.4 Topics in integration	206
3.4.1 Line integrals	207
3.4.2 Exercises (56–64)	210
3.4.3 Double integrals	211
3.4.4 Exercises (65–76)	216
3.4.5 Green's theorem in a plane	217
3.4.6 Exercises (77–82)	221
3.4.7 Surface integrals	222
3.4.8 Exercises (83–91)	229
3.4.9 Volume integrals	229
3.4.10 Exercises (92–102)	232
3.4.11 Gauss's divergence theorem	233
3.4.12 Stokes' theorem	236
3.4.13 Exercises (103–112)	239
<hr/>	
3.5 Engineering application: streamlines in fluid dynamics	240
<hr/>	
3.6 Engineering application: heat transfer	242
<hr/>	
3.7 Review exercises (1–21)	246
<hr/>	

Chapter 4 Functions of a Complex Variable **249**

4.1 Introduction	250
<hr/>	
4.2 Complex functions and mappings	251
4.2.1 Linear mappings	253
4.2.2 Exercises (1–8)	260

4.2.3 Inversion	260
4.2.4 Bilinear mappings	265
4.2.5 Exercises (9–19)	271
4.2.6 The mapping $w = z^2$	272
4.2.7 Exercises (20–23)	274
4.3 Complex differentiation	274
4.3.1 Cauchy–Riemann equations	275
4.3.2 Conjugate and harmonic functions	280
4.3.3 Exercises (24–32)	282
4.3.4 Mappings revisited	282
4.3.5 Exercises (33–37)	286
4.4 Complex series	287
4.4.1 Power series	287
4.4.2 Exercises (38–39)	291
4.4.3 Taylor series	291
4.4.4 Exercises (40–43)	294
4.4.5 Laurent series	295
4.4.6 Exercises (44–46)	300
4.5 Singularities and zeros	300
4.5.1 Exercises (47–49)	303
4.6 Engineering application: analysing AC circuits	304
4.7 Engineering application: use of harmonic functions	305
4.7.1 A heat transfer problem	305
4.7.2 Current in a field-effect transistor	307
4.7.3 Exercises (50–56)	310
4.8 Review exercises (1–19)	311

Chapter 5 Laplace Transforms	315
-------------------------------------	------------

5.1 Introduction	316
5.1.1 Definition and notation	316
5.1.2 Other results from MEM	318
5.2 Step and impulse functions	320
5.2.1 The Heaviside step function	320
5.2.2 Laplace transform of unit step function	323
5.2.3 The second shift theorem	325

5.2.4	Inversion using the second shift theorem	328
5.2.5	Differential equations	331
5.2.6	Periodic functions	335
5.2.7	Exercises (1–12)	339
5.2.8	The impulse function	341
5.2.9	The sifting property	342
5.2.10	Laplace transforms of impulse functions	343
5.2.11	Relationship between Heaviside step and impulse functions	346
5.2.12	Exercises (13–18)	351
5.2.13	Bending of beams	352
5.2.14	Exercises (19–21)	356
<hr/>		
5.3	Transfer functions	356
5.3.1	Definitions	356
5.3.2	Stability	359
5.3.3	Impulse response	364
5.3.4	Initial- and final-value theorems	365
5.3.5	Exercises (22–33)	370
5.3.6	Convolution	371
5.3.7	System response to an arbitrary input	374
5.3.8	Exercises (34–38)	378
<hr/>		
5.4	Solution of state-space equations	378
5.4.1	SISO systems	378
5.4.2	Exercises (39–47)	382
5.4.3	MIMO systems	383
5.4.4	Exercises (48–50)	390
<hr/>		
5.5	Engineering application: frequency response	390
<hr/>		
5.6	Engineering application: pole placement	398
5.6.1	Poles and eigenvalues	398
5.6.2	The pole placement or eigenvalue location technique	398
5.6.3	Exercises (51–56)	400
<hr/>		
5.7	Review exercises (1–18)	401
<hr/>		

Chapter 6	The z Transform	407
<hr/>		
6.1	Introduction	408
<hr/>		
6.2	The z transform	409
6.2.1	Definition and notation	409
6.2.2	Sampling: a first introduction	413
6.2.3	Exercises (1–2)	414

6.3	Properties of the z transform	414
6.3.1	The linearity property	415
6.3.2	The first shift property (delaying)	416
6.3.3	The second shift property (advancing)	417
6.3.4	Some further properties	418
6.3.5	Table of z transforms	419
6.3.6	Exercises (3–10)	420
6.4	The inverse z transform	420
6.4.1	Inverse techniques	421
6.4.2	Exercises (11–13)	427
6.5	Discrete-time systems and difference equations	428
6.5.1	Difference equations	428
6.5.2	The solution of difference equations	430
6.5.3	Exercises (14–20)	434
6.6	Discrete linear systems: characterization	435
6.6.1	z transfer functions	435
6.6.2	The impulse response	441
6.6.3	Stability	444
6.6.4	Convolution	450
6.6.5	Exercises (21–29)	454
6.7	The relationship between Laplace and z transforms	455
6.8	Solution of discrete-time state-space equations	456
6.8.1	State-space model	456
6.8.2	Solution of the discrete-time state equation	459
6.8.3	Exercises (30–33)	463
6.9	Discretization of continuous-time state-space models	464
6.9.1	Euler's method	464
6.9.2	Step-invariant method	466
6.9.3	Exercises (34–37)	469
6.10	Engineering application: design of discrete-time systems	470
6.10.1	Analogue filters	471
6.10.2	Designing a digital replacement filter	472
6.10.3	Possible developments	473

6.11	Engineering application: the delta operator and the \mathcal{D} transform	473
6.11.1	Introduction	473
6.11.2	The q or shift operator and the δ operator	474
6.11.3	Constructing a discrete-time system model	475
6.11.4	Implementing the design	477
6.11.5	The \mathcal{D} transform	479
6.11.6	Exercises (38–41)	480
6.12	Review exercises (1–18)	480

Chapter 7 Fourier Series	485
---------------------------------	------------

7.1	Introduction	486
7.1.1	Periodic functions	486
7.1.2	Fourier's theorem	487
7.1.3	Functions of period 2π	488
7.1.4	Functions defined over a finite interval	492
7.1.5	Exercises (1–10)	498
7.2	Fourier series of jumps at discontinuities	499
7.2.1	Exercises (11–12)	502
7.3	Engineering application: frequency response and oscillating systems	502
7.3.1	Response to periodic input	502
7.3.2	Exercises (13–16)	507
7.4	Complex form of Fourier series	508
7.4.1	Complex representation	508
7.4.2	The multiplication theorem and Parseval's theorem	512
7.4.3	Discrete frequency spectra	515
7.4.4	Power spectrum	521
7.4.5	Exercises (17–22)	523
7.5	Orthogonal functions	524
7.5.1	Definitions	524
7.5.2	Generalized Fourier series	526
7.5.3	Convergence of generalized Fourier series	527
7.5.4	Exercises (23–29)	529

7.6	Engineering application: describing functions	532
7.7	Review exercises (1–20)	533

Chapter 8 The Fourier Transform		537
8.1	Introduction	538
8.2	The Fourier transform	539
8.2.1	The Fourier integral	539
8.2.2	The Fourier transform pair	544
8.2.3	The continuous Fourier spectra	548
8.2.4	Exercises (1–10)	551
8.3	Properties of the Fourier transform	552
8.3.1	The linearity property	552
8.3.2	Time-differentiation property	552
8.3.3	Time-shift property	553
8.3.4	Frequency-shift property	554
8.3.5	The symmetry property	555
8.3.6	Exercises (11–16)	557
8.4	The frequency response	558
8.4.1	Relationship between Fourier and Laplace transforms	558
8.4.2	The frequency response	560
8.4.3	Exercises (17–21)	563
8.5	Transforms of the step and impulse functions	563
8.5.1	Energy and power	563
8.5.2	Convolution	572
8.5.3	Exercises (22–27)	574
8.6	The Fourier transform in discrete time	575
8.6.1	Introduction	575
8.6.2	A Fourier transform for sequences	575
8.6.3	The discrete Fourier transform	579
8.6.4	Estimation of the continuous Fourier transform	583
8.6.5	The fast Fourier transform	592
8.6.6	Exercises (28–31)	599

8.7	Engineering application: the design of analogue filters	599
8.8	Engineering application: direct design of digital filters and windows	602
8.8.1	Digital filters	602
8.8.2	Windows	607
8.8.3	Exercises (32–33)	611
8.9	Review exercises (1–25)	611

Chapter 9	Partial Differential Equations	615
------------------	---------------------------------------	------------

9.1	Introduction	616
9.2	General discussion	617
9.2.1	Wave equation	617
9.2.2	Heat-conduction or diffusion equation	620
9.2.3	Laplace equation	623
9.2.4	Other and related equations	625
9.2.5	Arbitrary functions and first-order equations	627
9.2.6	Exercises (1–14)	632
9.3	Solution of the wave equation	634
9.3.1	D'Alembert solution and characteristics	634
9.3.2	Separation of variables	643
9.3.3	Laplace transform solution	648
9.3.4	Exercises (15–27)	651
9.3.5	Numerical solution	653
9.3.6	Exercises (28–31)	659
9.4	Solution of the heat-conduction/diffusion equation	660
9.4.1	Separation of variables	660
9.4.2	Laplace transform method	664
9.4.3	Exercises (32–40)	669
9.4.4	Numerical solution	671
9.4.5	Exercises (41–43)	677
9.5	Solution of the Laplace equation	677
9.5.1	Separation of variables	677
9.5.2	Exercises (44–54)	685
9.5.3	Numerical solution	686
9.5.4	Exercises (55–59)	693

9.6	Finite elements	694
9.6.1	Exercises (60–62)	706
9.7	Integral solutions	707
9.7.1	Separation of variables	707
9.7.2	Use of singular solutions	709
9.7.3	Sources and sinks for the heat-conduction equation	712
9.7.4	Exercises (63–67)	715
9.8	General considerations	716
9.8.1	Formal classification	716
9.8.2	Boundary conditions	718
9.8.3	Exercises (68–74)	723
9.9	Engineering application: wave propagation under a moving load	723
9.10	Engineering application: blood-flow model	726
9.11	Review exercises (1–21)	730

Chapter 10 Optimization 735

10.1	Introduction	736
10.2	Linear programming	739
10.2.1	Introduction	739
10.2.2	Simplex algorithm: an example	741
10.2.3	Simplex algorithm: general theory	745
10.2.4	Exercises (1–11)	752
10.2.5	Two-phase method	753
10.2.6	Equality constraints and variables that are unrestricted in sign	761
10.2.7	Exercises (12–20)	762
10.3	Lagrange multipliers	764
10.3.1	Equality constraints	764
10.3.2	Inequality constraints	768
10.3.3	Exercises (21–28)	768
10.4	Hill climbing	769
10.4.1	Single-variable search	769
10.4.2	Exercises (29–34)	775

10.4.3	Simple multivariable searches: steepest ascent and Newton's method	775
10.4.4	Exercises (35–39)	781
10.4.5	Advanced multivariable searches	782
10.4.6	Least squares	786
10.4.7	Exercises (40–43)	789
<hr/>		
10.5	Engineering application: chemical processing plant	790
<hr/>		
10.6	Engineering application: heating fin	792
<hr/>		
10.7	Review exercises (1–26)	795
<hr/>		
Chapter 11	Applied Probability and Statistics	799
<hr/>		
11.1	Introduction	800
<hr/>		
11.2	Review of basic probability theory	801
11.2.1	The rules of probability	801
11.2.2	Random variables	802
11.2.3	The Bernoulli, binomial and Poisson distributions	804
11.2.4	The normal distribution	805
11.2.5	Sample measures	808
<hr/>		
11.3	Estimating parameters	810
11.3.1	Interval estimates and hypothesis tests	810
11.3.2	Distribution of the sample average	810
11.3.3	Confidence interval for the mean	812
11.3.4	Testing simple hypotheses	815
11.3.5	Other confidence intervals and tests concerning means	817
11.3.6	Interval and test for proportion	821
11.3.7	Exercises (1–13)	824
<hr/>		
11.4	Joint distributions and correlation	825
11.4.1	Joint and marginal distributions	825
11.4.2	Independence	828
11.4.3	Covariance and correlation	829
11.4.4	Sample correlation	833
11.4.5	Interval and test for correlation	835
11.4.6	Rank correlation	838
11.4.7	Exercises (14–24)	840

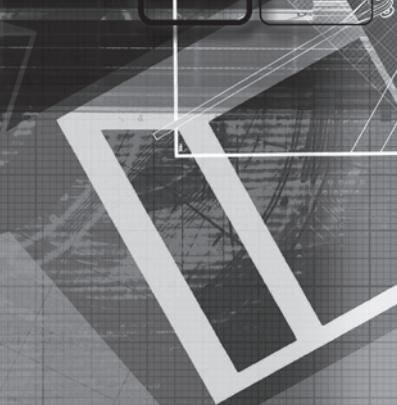
11.5	Regression	841
11.5.1	The method of least squares	842
11.5.2	Residuals	852
11.5.3	Regression and correlation	856
11.5.4	Nonlinear regression	856
11.5.5	Exercises (25–33)	861
<hr/>		
11.6	Goodness-of-fit tests	863
11.6.1	Chi-square distribution and test	863
11.6.2	Contingency tables	867
11.6.3	Exercises (34–42)	873
<hr/>		
11.7	Engineering application: analysis of engine performance data	874
11.7.1	Introduction	874
11.7.2	Difference in mean running times and temperatures	877
11.7.3	Dependence of running time on temperature	880
11.7.4	Test for normality	888
11.7.5	Conclusions	890
<hr/>		
11.8	Engineering application: statistical quality control	891
11.8.1	Introduction	891
11.8.2	Shewhart attribute control charts	891
11.8.3	Shewhart variable control charts	894
11.8.4	Cusum control charts	898
11.8.5	Moving-average control charts	901
11.8.6	Range charts	905
11.8.7	Exercises (43–54)	907
<hr/>		
11.9	Poisson processes and the theory of queues	908
11.9.1	Typical queueing problems	909
11.9.2	Poisson processes	909
11.9.3	Single service channel queue	916
11.9.4	Queues with multiple service channels	921
11.9.5	Queueing system simulation	923
11.9.6	Exercises (55–62)	929
<hr/>		
11.10	Bayes' theorem and its applications	930
11.10.1	Derivation and simple examples	930
11.10.2	Applications in probabilistic inference	933
11.10.3	Bayesian statistical inference	935
11.10.4	Exercises (63–74)	944
<hr/>		
11.11	Review exercises (1–10)	945

Answers to Exercises

949

Index

975



Preface

The first edition of this book appeared in 1993, and it could be assumed, wrongly, that its time has passed as 24 years have now elapsed. It is true that all the original authors apart from myself have retired but, in the intervening years the text has been regularly updated and we have now reached the fifth edition. The words of my colleague and predecessor as editor, Professor Glyn James, still ring true. Here is an excerpt from his preface to the fourth edition (2011):

Throughout the course of history, engineering and mathematics have developed in parallel. All branches of engineering depend on mathematics for their description and there has been a steady flow of ideas and problems from engineering that has stimulated and sometimes initiated branches of mathematics. Thus, it is vital that engineering students receive a thorough grounding in mathematics, with the treatment related to their interests and problems. As with the previous editions, this has been the motivation for the production of this latest edition – a companion text to the fifth edition of *Modern Engineering Mathematics*, this being designed to provide a first-level core studies course in mathematics for undergraduate programmes in all engineering disciplines. Building on the foundations laid in the companion text, this book gives an extensive treatment of some of the more advanced areas of mathematics that have applications in various fields of engineering, particularly as tools for computer-based system modelling, analysis and design. Feedback, from users of the previous editions, on subject content has been highly positive indicating that it is sufficiently broad to provide the necessary second-level, or optional, studies for most engineering programmes, where in each case a selection of the material may be made. Whilst designed primarily for use by engineering students, it is believed that the book is also suitable for use by students of applied mathematics and the physical sciences.

Although the pace of the book is at a somewhat more advanced level than the companion text, the philosophy of learning by doing is retained with continuing emphasis on the development of students' ability to use mathematics with understanding to solve engineering problems. Recognizing the increasing importance of mathematical modelling in engineering practice, many of the worked examples and exercises incorporate mathematical models that are designed both to provide relevance and to reinforce the role of mathematics in various branches of engineering. In addition, each chapter contains specific sections on engineering applications, and these form an ideal framework for individual, or group, study assignments, thereby helping to reinforce the skills of mathematical modelling, which are seen as essential if engineers are to tackle the increasingly complex systems they are being called upon to analyse and design. The importance of numerical methods in problem solving is also recognized, and its treatment is integrated with the analytical work throughout the book.

The position of software use is an important aspect of engineering education. The decision has been taken to use mainly MATLAB but also MAPLE. Students are encouraged to make intelligent use of software and, where appropriate, codes are included, but there is a health warning. The pace of technology shows little signs of lessening, and so in the space of six years, the likely time lapse before a new edition of this text, it is probable that software will continue to be updated, probably annually. There is therefore a real risk that much coding though correct and working at the time of publication could be broken by these updates. Therefore, in this edition the decision has been made not to over-emphasise specific code but to direct students to the companion website or to general principles instead. The software packages, particularly MAPLE, have become easier to use without the need for programming skills. Much is menu driven these days. Here's more from Glyn on the subject that is still true:

Much of the feedback from users relates to the role and use of software packages, particularly symbolic algebra packages. Without making it an essential requirement the authors have attempted to highlight throughout the text situations where the user could make effective use of software. This also applies to exercises and, indeed, a limited number have been introduced for which the use of such a package is essential. Whilst any appropriate piece of software can be used, the authors recommend the use of MATLAB and/or MAPLE. In this edition reference to the use of these two packages is made throughout the text, with commands or codes introduced and illustrated. When indicated, students are strongly recommended to use these packages to check their solutions to exercises. This is not only to help develop proficiency in their use, but also to enable students to appreciate the necessity of having a sound knowledge of the underpinning mathematics if such packages are to be used effectively. Throughout the book two icons are used:

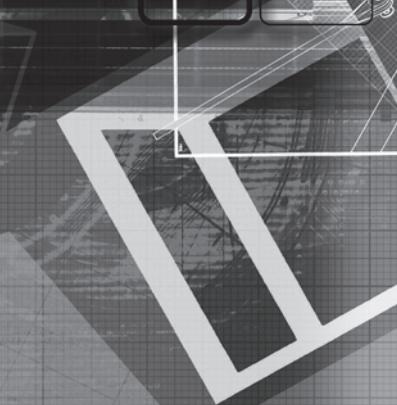
- An open screen  indicates that the use of a software package would be useful (e.g. for checking solutions) but not essential.
- A closed screen  indicates that the use of a software package is essential or highly desirable.

Specific changes in this fifth edition are an improvement in many of the diagrams, taking advantage of present day software, and modernization of the examples and language. Also, the chapter on Applied Probability and Statistics has been significantly modernized by interfacing the presentation with the very powerful software package R. Simply search for 'R Software' and it is a free download. I have been much aided in getting this edition ready for publication by my hardworking colleagues Matthew, Tim and Julian who have joined the editorial team.

Acknowledgements

The authoring team is extremely grateful to all the reviewers and users of the text who have provided valuable comments on previous editions of this book. Most of this has been highly constructive and very much appreciated. The team has continued to enjoy the full support of a very enthusiastic production team at Pearson Education and wishes to thank all those concerned.

Phil Dyke and Glyn James



About the Authors

A new set of authors, Matthew Craven, Tim Reis and Julian Stander under the new editor, one of the original authors, Phil Dyke, have taken on the task of producing this, the fifth edition of *Advanced Modern Engineering Mathematics*.

Phil Dyke is Professor of Applied Mathematics at the University of Plymouth. He was a Head of School for 22 years, 18 of these as Head of Mathematics and Statistics. He has over 45 years teaching and research experience in higher education, much of this teaching engineering students not only mathematics but also marine and coastal engineering. Apart from his contribution to both *Modern Engineering Mathematics* and *Advanced Modern Engineering Mathematics* he is the author of 11 other textbooks ranging in topic from mathematical methods to mechanics and marine physics. He is now semi-retired, but still teaches and writes.

Matthew Craven is a Lecturer in Applied Mathematics at the University of Plymouth. For fifteen years, he has taught foundation year, postgraduate and everything in between. He has research interests in computational simulation, operational research, high performance computing and algebraic systems.

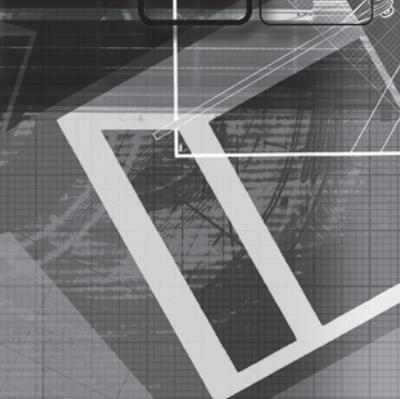
Tim Reis is a Senior Lecturer in Mathematics at the University of Greenwich. He is an applied mathematician with interests in fluid dynamics, numerical analysis, and mathematical modelling. His doctoral thesis was awarded the Vernon Harrison prize by the British Society of Rheology and he continues to conduct research into the modelling and simulation of complex flows. Tim teaches a range of mathematical subjects at undergraduate and postgraduate level and he is actively involved in promoting mathematics to the wider community. He is a fellow of the Institute of Mathematics and its Applications (IMA).

Julian Stander is Associate Professor (Reader) in Mathematics and Statistics at the University of Plymouth. He has over twenty years' cross-disciplinary teaching experience in mathematics and statistics at both the undergraduate and postgraduate level. His research and consultancy interests are in computational statistics, data science, extreme value and dependence modelling, and social media information extraction.

The original editor is **Glyn James** who retired as Dean of the School of Mathematical and Information Sciences at Coventry University in 2001 and is now Emeritus Professor in Mathematics at the University. He graduated from the University College of Wales, Cardiff in the late 1950s, obtaining first class honours degrees in both Mathematics and

Chemistry. He obtained a PhD in Engineering Science in 1971 as an external student of the University of Warwick. He has been employed at Coventry since 1964 and held the position of the Head of Mathematics Department prior to his appointment as Dean in 1992. His research interests are in control theory and its applications to industrial problems. He also has a keen interest in mathematical education, particularly in relation to the teaching of engineering mathematics and mathematical modelling. He was co-chairman of the European Mathematics Working Group established by the European Society for Engineering Education (SEFI) in 1982, a past chairman of the Education Committee of the Institute of Mathematics and its Applications (IMA), and a member of the Royal Society Mathematics Education Subcommittee. In 1995 he was chairman of the Working Group that produced the report ‘Mathematics Matters in Engineering’ on behalf of the professional bodies in engineering and mathematics within the UK. He is also a member of the editorial/advisory board of three international journals. He has published numerous papers and is co-editor of five books on various aspects of mathematical modelling. He is a past Vice-President of the IMA and has also served a period as Honorary Secretary of the Institute. He is a Chartered Mathematician and a Fellow of the IMA.

The original authors are David Burley, Dick Clements, John Searl, Nigel Steele, Jerry Wright together with Phil Dyke. Their short biographies can be found in the previous editions.



Publisher's Acknowledgements

Extracts in section 7.4.1, 8.5, 8.6 from Signal Processing Communication, ISBN 1898563233, 1 ed., Woodhead Publishing Ltd (Chapman, N, Goodhall, D, Steele, N).



1

Matrix Analysis

Chapter 1 Contents

1.1	Introduction	2
1.2	Review of matrix algebra	2
1.3	Vector spaces	9
1.4	The eigenvalue problem	13
1.5	Numerical methods	29
1.6	Reduction to canonical form	36
1.7	Functions of a matrix	51
1.8	Singular value decomposition	63
1.9	State-space representation	79
1.10	Solution of the state equation	86
1.11	Engineering application: Lyapunov stability analysis	101
1.12	Engineering application: capacitor microphone	104
1.13	Review exercises (1–19)	108

1.1

Introduction

In this chapter we turn our attention again to matrices, first considered in Chapter 5 of *Modern Engineering Mathematics* (MEM), and their applications in engineering. At the outset of the chapter we review the basic results of matrix algebra and briefly introduce vector spaces.

As the reader will be aware, matrices are arrays of real or complex numbers, and have a special, but not exclusive, relationship with systems of linear equations. Such systems occur quite naturally in the process of numerical solution of ordinary differential equations used to model everyday engineering processes. In Chapter 9 we shall see that they also occur in numerical methods for the solution of partial differential equations, for example those modelling the flow of a fluid or the transfer of heat. Systems of linear first-order differential equations with constant coefficients are at the core of the **state-space** representation of linear system models. Identification, analysis and indeed design of such systems can conveniently be performed in the state-space representation, with this form assuming a particular importance in the case of multivariable systems.

In all these areas it is convenient to use a matrix representation for the systems under consideration, since this allows the system model to be manipulated following the rules of matrix algebra. A particularly valuable type of manipulation is **simplification** in some sense. Such a simplification process is an example of a system transformation, carried out by the process of matrix multiplication. At the heart of many transformations are the **eigenvalues** and **eigenvectors** of a square matrix. In addition to providing the means by which simplifying transformations can be deduced, system eigenvalues provide vital information on system stability, fundamental frequencies, speed of decay and long-term system behaviour. For this reason, we devote a substantial amount of space to the process of their calculation, both by hand and by numerical means when necessary. Our treatment of numerical methods is intended to be purely indicative rather than complete, because a comprehensive matrix algebra computational tool kit, such as MATLAB, is now part of the essential armoury of all serious users of mathematics.



In addition to developing the use of matrix algebra techniques, we also demonstrate the techniques and applications of matrix analysis, focusing on the state-space system model widely used in control and systems engineering. Here we encounter the idea of a function of a matrix, in particular the matrix exponential, and we see again the role of the eigenvalues in its calculation. This edition also includes a section on singular value decomposition and the pseudo inverse, together with a brief section on Lyapunov stability of linear systems using quadratic forms.

1.2

Review of matrix algebra

This section contains a summary of the definitions and properties associated with matrices and determinants. A full account can be found in chapters of MEM or elsewhere. It is assumed that readers, prior to embarking on this chapter, have a fairly thorough understanding of the material summarized in this section.

1.2.1 Definitions

- (a) An array of real numbers

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

is called an $m \times n$ **matrix** with m rows and n columns. The a_{ij} is referred to as the (ij) th **element** and denotes the element in the i th row and j th column. If $m = n$ then \mathbf{A} is called a **square matrix** of order n . If the matrix has one column or one row then it is called a **column vector** or a **row vector** respectively.

- (b) In a square matrix \mathbf{A} of order n the diagonal containing the elements $a_{11}, a_{22}, \dots, a_{nn}$ is called the **principal** or **leading** diagonal. The sum of the elements in this diagonal is called the **trace** of \mathbf{A} , that is

$$\text{trace } \mathbf{A} = \sum_{i=1}^n a_{ii}$$

- (c) A **diagonal matrix** is a square matrix that has its only non-zero elements along the leading diagonal. A special case of a diagonal matrix is the **unit** or **identity matrix** \mathbf{I} for which $a_{11} = a_{22} = \dots = a_{nn} = 1$.
- (d) A **zero** or **null** matrix $\mathbf{0}$ is a matrix with every element zero.
- (e) The **transposed matrix** \mathbf{A}^T is the matrix \mathbf{A} with rows and columns interchanged, its i, j th element being a_{ji} .
- (f) A square matrix \mathbf{A} is called a **symmetric matrix** if $\mathbf{A}^T = \mathbf{A}$. It is called **skew symmetric** if $\mathbf{A}^T = -\mathbf{A}$.

1.2.2 Basic operations on matrices

In what follows the matrices \mathbf{A} , \mathbf{B} and \mathbf{C} are assumed to have the i, j th elements a_{ij} , b_{ij} and c_{ij} respectively.

Equality

The matrices \mathbf{A} and \mathbf{B} are **equal**, that is $\mathbf{A} = \mathbf{B}$, if they are of the same order $m \times n$ and

$$a_{ij} = b_{ij}, \quad 1 \leq i \leq m, \quad 1 \leq j \leq n$$

Multiplication by a scalar

If λ is a scalar then the matrix $\lambda\mathbf{A}$ has elements λa_{ij} .

Addition

We can only add an $m \times n$ matrix \mathbf{A} to another $m \times n$ matrix \mathbf{B} and the elements of the sum $\mathbf{A} + \mathbf{B}$ are

$$a_{ij} + b_{ij}, \quad 1 \leq i \leq m; \quad 1 \leq j \leq n$$

Properties of addition

- (i) commutative law: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- (ii) associative law: $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- (iii) distributive law: $\lambda(\mathbf{A} + \mathbf{B}) = \lambda\mathbf{A} + \lambda\mathbf{B}$, λ scalar

Matrix multiplication

If \mathbf{A} is an $m \times p$ matrix and \mathbf{B} a $p \times n$ matrix then we define the product $\mathbf{C} = \mathbf{AB}$ as the $m \times n$ matrix with elements

$$c_{ij} = \sum_{k=1}^p a_{ik}b_{kj}, \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n$$

Properties of multiplication

- (i) The commutative law is **not satisfied** in general; that is, in general $\mathbf{AB} \neq \mathbf{BA}$. Order does matter and we distinguish between \mathbf{AB} and \mathbf{BA} by the terminology: **pre-multiplication** of \mathbf{B} by \mathbf{A} to form \mathbf{AB} and **post-multiplication** of \mathbf{B} by \mathbf{A} to form \mathbf{BA} .
- (ii) Associative law: $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$
- (iii) If λ is a scalar then

$$(\lambda\mathbf{A})\mathbf{B} = \mathbf{A}(\lambda\mathbf{B}) = \lambda\mathbf{AB}$$
- (iv) Distributive law over addition:

$$(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC} \quad \mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$$

Note the importance of maintaining order of multiplication as in property (i).
- (v) If \mathbf{A} is an $m \times n$ matrix and if \mathbf{I}_m and \mathbf{I}_n are the unit matrices of order m and n respectively then

$$\mathbf{I}_m\mathbf{A} = \mathbf{A}\mathbf{I}_n = \mathbf{A}$$

Properties of the transpose

If \mathbf{A}^T is the transposed matrix of \mathbf{A} then

- (i) $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
- (ii) $(\mathbf{A}^T)^T = \mathbf{A}$
- (iii) $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$

1.2.3 Determinants

The determinant of a square $n \times n$ matrix \mathbf{A} is denoted by $\det \mathbf{A}$ or $|\mathbf{A}|$.

If we take a determinant of a matrix and delete row i and column j then the determinant remaining is called the **minor** M_{ij} of the (ij) th element. In general we can take any row i (or column) and evaluate an $n \times n$ determinant $|\mathbf{A}|$ as

$$|\mathbf{A}| = \sum_{j=1}^n (-1)^{i+j} a_{ij} M_{ij}$$

A minor multiplied by the appropriate sign is called the **cofactor** A_{ij} of the (ij) th element so $A_{ij} = (-1)^{i+j} M_{ij}$ and thus

$$|\mathbf{A}| = \sum_{j=1}^n a_{ij} A_{ij}$$

Some useful properties

- (i) $|\mathbf{A}^T| = |\mathbf{A}|$
- (ii) $|\mathbf{AB}| = |\mathbf{A}| |\mathbf{B}|$
- (iii) A square matrix \mathbf{A} is said to be **non-singular** if $|\mathbf{A}| \neq 0$ and **singular** if $|\mathbf{A}| = 0$.

1.2.4 Adjoint and inverse matrices

Adjoint matrix

The **adjoint** of a square matrix \mathbf{A} is the transpose of the matrix of cofactors, so for a 3×3 matrix \mathbf{A}

$$\text{adj } \mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}^T$$

Properties

- (i) $\mathbf{A} (\text{adj } \mathbf{A}) = |\mathbf{A}| \mathbf{I}$
- (ii) $|\text{adj } \mathbf{A}| = |\mathbf{A}|^{n-1}$, where n is the order of \mathbf{A}
- (iii) $\text{adj}(\mathbf{AB}) = (\text{adj } \mathbf{B})(\text{adj } \mathbf{A})$

Inverse matrix

Given a square matrix \mathbf{A} if we can construct a square matrix \mathbf{B} such that

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

then we call \mathbf{B} the inverse of \mathbf{A} and write it as \mathbf{A}^{-1} .

Properties

- (i) If \mathbf{A} is non-singular then $|\mathbf{A}| \neq 0$ and $\mathbf{A}^{-1} = (\text{adj } \mathbf{A})/|\mathbf{A}|$.
- (ii) If \mathbf{A} is singular then $|\mathbf{A}| = 0$ and \mathbf{A}^{-1} does not exist.
- (iii) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.



All the basic matrix operations may be implemented in MATLAB using simple commands. In MATLAB a matrix is entered as an array, with row elements separated by spaces (or commas) and each row of elements separated by a semicolon(;), or the return key to go to a new line. Thus, for example,

```
A=[1 2 3; 4 0 5; 7 4 2]
```

gives

```
A=
1 2 3
4 0 5
7 4 2
```

Having specified the two matrices \mathbf{A} and \mathbf{B} the operations of addition, subtraction and multiplication are implemented using respectively the commands

```
C=A+B, C=A-B, C=A*B
```

The trace of the matrix \mathbf{A} is determined by the command `trace(A)`, and its determinant by `det(A)`.

Multiplication of a matrix \mathbf{A} by a scalar is carried out using the command `*`, while raising \mathbf{A} to a given power is carried out using the command `^`. Thus, for example, $3\mathbf{A}^2$ is determined using the command `C=3*A^2`.

The transpose of a real matrix \mathbf{A} is determined using the apostrophe ' key; that is `C=A'` (to accommodate complex matrices the command `C=A.'` should be used). The inverse of \mathbf{A} is determined by `C=inv(A)`.

For matrices involving algebraic quantities, or when exact arithmetic is desirable use of the Symbolic Math Toolbox is required; in which matrices must be expressed in symbolic form using the `sym` command. The command `A=sym(A)` generates the symbolic form of \mathbf{A} . For example, for the matrix

$$\mathbf{A} = \begin{bmatrix} 2.1 & 3.2 & 0.6 \\ 1.2 & 0.5 & 3.3 \\ 5.2 & 1.1 & 0 \end{bmatrix}$$

the commands

```
A=[2.1 3.2 0.6; 1.2 0.5 3.3; 5.2 1.1 0];
A=sym(A)
```

generate

```
A=
[21/10, 16/5, 3/5]
[6/5, 1/2, 33/10]
[26/5, 11/10, 0]
```

Symbolic manipulation can also be undertaken in MATLAB using the MuPAD version of Symbolic Math Toolbox.

Such operations may be performed in Python. Details are not given here, but the interested reader is directed to, for example, *Beginning Python* by Lie Hethand (Springer, 2005). The `numPy` package should be loaded.

1.2.5 Linear equations

In this section we reiterate some definitive statements about the solution of the system of simultaneous linear equations

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

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

 \vdots
 \vdots

$$a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n$$

or, in matrix notation,

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

that is,

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (1.1)$$

where \mathbf{A} is the matrix of coefficients and \mathbf{x} is the vector of unknowns. If $\mathbf{b} = \mathbf{0}$ the equations are called **homogeneous**, while if $\mathbf{b} \neq \mathbf{0}$ they are called **nonhomogeneous** (or **inhomogeneous**). Considering individual cases:

Case (i): If $\mathbf{b} \neq \mathbf{0}$ and $|\mathbf{A}| \neq 0$ then we have a unique solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

Case (ii): If $\mathbf{b} = \mathbf{0}$ and $|\mathbf{A}| \neq 0$ we have the trivial solution $\mathbf{x} = \mathbf{0}$.

Case (iii): If $\mathbf{b} \neq \mathbf{0}$ and $|\mathbf{A}| = 0$ then we have two possibilities: **either** the equations are inconsistent and we have no solution **or** we have infinitely many solutions.

Case (iv): If $\mathbf{b} = \mathbf{0}$ and $|\mathbf{A}| = 0$ then we have infinitely many solutions.

Case (iv) is one of the most important, since from it we can deduce the important result that **the homogeneous equation $\mathbf{A}\mathbf{x} = \mathbf{0}$ has a non-trivial solution if and only if $|\mathbf{A}| = 0$** .



Provided that a solution to (1.1) exists it may be determined in MATLAB using the command $\mathbf{x} = \mathbf{A} \backslash \mathbf{b}$. For example, the system of simultaneous equations

$$x + y + z = 6, \quad x + 2y + 3z = 14, \quad x + 4y + 9z = 36$$

may be written in the matrix form

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 4 & 9 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \\ 36 \end{bmatrix}$$

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

Entering \mathbf{A} and \mathbf{b} and using the command $\mathbf{x} = \mathbf{A} \backslash \mathbf{b}$ provides the answer $x = 1, y = 2, z = 3$.

1.2.6 Rank of a matrix

We adopt the following constructive definition of the **rank**, rank \mathbf{A} of a matrix \mathbf{A} . First, using elementary row operations, the matrix \mathbf{A} is reduced to **echelon form**

$$\left[\begin{array}{ccccccccc} * & & & & & & & & \\ \swarrow & \searrow & & & & & & & \\ * & & & & & & & & \\ & \swarrow & \searrow & & & & & & \\ * & & & & & & & & \\ & \swarrow & \searrow & & & & & & \\ * & & & & & & & & \\ & \swarrow & \searrow & & & & & & \\ * & & & & & & & & \end{array} \right]$$

in which all the entries below the line are zero, and the leading element, marked $*$, in each row above the line is non-zero. Then the number of non-zero rows in the echelon form is equal to rank \mathbf{A} . These are equivalent definitions.

When considering the solution of (1.1) we saw that provided the determinant of the matrix \mathbf{A} was not zero we could obtain explicit solutions in terms of the inverse matrix. However, when we looked at cases with zero determinant the results were much less clear. The idea of the rank of a matrix helps to make these results more precise. Defining the **augmented matrix** $(\mathbf{A} : \mathbf{b})$ for (1.1) as the matrix \mathbf{A} with the column \mathbf{b} added to it then we can state the results of cases (iii) and (iv) of Section 1.2.5 more clearly as follows:

If \mathbf{A} and $(\mathbf{A} : \mathbf{b})$ have different rank then we have no solution to (1.1). If the two matrices have the same rank then a solution exists, and furthermore the solution will contain $n - \text{rank } \mathbf{A}$ free parameters.



In MATLAB the rank of the matrix \mathbf{A} is generated using the command `rank(A)`. For example, if

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

the commands

```
A=[-1 2 2; 0 0 1; -1 2 0];
rank(A)
```

generate

```
ans=2
```

In MAPLE the command is also `rank(A)`.

1.3

Vector spaces

Vectors and matrices form part of a more extensive formal structure called a **vector space**. The theory of vector spaces underpins many approaches to numerical methods and the approximate solution of many equations that arise in engineering analysis. In this section we shall, briefly, introduce some basic ideas of vector spaces necessary for later work in this chapter.

Definition

A **real vector space** V is a set of objects called **vectors** together with rules for addition and multiplication by real numbers. For any three vectors \mathbf{a} , \mathbf{b} and \mathbf{c} in V and any real numbers α and β the sum $\mathbf{a} + \mathbf{b}$ and the product $\alpha\mathbf{a}$ also belong to V and satisfy the following axioms:

- (a) $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$
- (b) $\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$
- (c) there exists a zero vector $\mathbf{0}$ such that

$$\mathbf{a} + \mathbf{0} = \mathbf{a}$$
- (d) for each \mathbf{a} in V there is an element $-\mathbf{a}$ in V such that

$$\mathbf{a} + (-\mathbf{a}) = \mathbf{0}$$
- (e) $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$
- (f) $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$
- (g) $(\alpha\beta)\mathbf{a} = \alpha(\beta\mathbf{a})$
- (h) $1\mathbf{a} = \mathbf{a}$

It is clear that the real numbers form a vector space. The properties given are also satisfied by vectors and by $m \times n$ matrices so vectors and matrices also form vector spaces. The space of all quadratics $a + bx + cx^2$ forms a vector space (check the axioms, (a)–(h)). Many other common sets of objects also form vector spaces. If we can obtain useful information from the general structure then this will be of considerable use in specific cases.

1.3.1 Linear independence

The idea of linear dependence is a general one for any vector space. The vector \mathbf{x} is said to be **linearly dependent** on $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$ if it can be written as

$$\mathbf{x} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_m \mathbf{x}_m$$

for some scalars $\alpha_1, \dots, \alpha_m$. The **set of vectors** $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m$ is said to be **linearly independent** if and only if

$$\beta_1 \mathbf{y}_1 + \beta_2 \mathbf{y}_2 + \cdots + \beta_m \mathbf{y}_m = 0$$

implies that $\beta_1 = \beta_2 = \cdots = \beta_m = 0$.

Let us now take a linearly independent set of vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$ in V and construct a set consisting of all vectors of the form

$$\mathbf{x} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \cdots + \alpha_m \mathbf{x}_m$$

We shall call this set $S(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m)$. It is clearly a vector space, since all the axioms are satisfied.

Example 1.1

Show that

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

form a linearly independent set and describe $S(\mathbf{e}_1, \mathbf{e}_2)$ geometrically.

Solution We have that

$$0 = \alpha \mathbf{e}_1 + \beta \mathbf{e}_2 = \begin{bmatrix} \alpha \\ \beta \\ 0 \end{bmatrix}$$

is only satisfied if $\alpha = \beta = 0$, and hence \mathbf{e}_1 and \mathbf{e}_2 are linearly independent.

$S(\mathbf{e}_1, \mathbf{e}_2)$ is the set of all vectors of the form $\begin{bmatrix} \alpha \\ \beta \\ 0 \end{bmatrix}$, which is just the (x_1, x_2)

plane and is a subset of three-dimensional Euclidean space.

If we can find a set B of linearly independent vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ in V such that

$$S(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = V$$

then B is called a **basis** of the vector space V . Such a basis forms a crucial part of the theory, since every vector \mathbf{x} in V can be written *uniquely* as

$$\mathbf{x} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \dots + \alpha_n \mathbf{x}_n$$

The definition of B implies that \mathbf{x} must take this form. To establish uniqueness, let us assume that we can also write \mathbf{x} as

$$\mathbf{x} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \dots + \beta_n \mathbf{x}_n$$

Then, on subtracting,

$$0 = (\alpha_1 - \beta_1) \mathbf{x}_1 + \dots + (\alpha_n - \beta_n) \mathbf{x}_n$$

and since $\mathbf{x}_1, \dots, \mathbf{x}_n$ are linearly independent, the only solution is $\alpha_1 = \beta_1, \alpha_2 = \beta_2, \dots$; hence the two expressions for \mathbf{x} are the same.

It can also be shown that any other basis for V must also contain n vectors and that any $n + 1$ vectors must be linearly dependent. Such a vector space is said to have **dimension n** (or **infinite dimension** if no finite n can be found). In a three-dimensional Euclidean space

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

form an obvious basis, in fact the standard basis, and

$$\mathbf{d}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{d}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{d}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

is also a perfectly good basis. While the basis can change, the number of vectors in the basis, three in this case, is an intrinsic property of the vector space. If we consider the vector space of quadratics then the sets of functions $\{1, x, x^2\}$ and $\{1, x - 1, x(x - 1)\}$ are both bases for the space, since every quadratic can be written as $a + bx + cx^2$ or as $A + B(x - 1) + Cx(x - 1)$. This space is three-dimensional.

1.3.2 Transformations between bases

Since any basis of a particular space contains the same number of vectors, we can look at transformations from one basis to another. We shall consider a three-dimensional space, but the results are equally valid in any number of dimensions. Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ and $\mathbf{e}'_1, \mathbf{e}'_2, \mathbf{e}'_3$ be two bases of a space. From the definition of a basis, the vectors $\mathbf{e}'_1, \mathbf{e}'_2$ and \mathbf{e}'_3 can be written in terms of $\mathbf{e}_1, \mathbf{e}_2$ and \mathbf{e}_3 as

$$\left. \begin{aligned} \mathbf{e}'_1 &= a_{11}\mathbf{e}_1 + a_{21}\mathbf{e}_2 + a_{31}\mathbf{e}_3 \\ \mathbf{e}'_2 &= a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2 + a_{32}\mathbf{e}_3 \\ \mathbf{e}'_3 &= a_{13}\mathbf{e}_1 + a_{23}\mathbf{e}_2 + a_{33}\mathbf{e}_3 \end{aligned} \right\} \quad (1.2)$$

Taking a typical vector \mathbf{x} in V , which can be written both as

$$\mathbf{x} = x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 + x_3 \mathbf{e}_3 \quad (1.3)$$

and as

$$\mathbf{x} = x'_1 \mathbf{e}'_1 + x'_2 \mathbf{e}'_2 + x'_3 \mathbf{e}'_3$$

we can use the transformation (1.2) to give

$$\begin{aligned} \mathbf{x} &= x'_1(a_{11}\mathbf{e}_1 + a_{21}\mathbf{e}_2 + a_{31}\mathbf{e}_3) + x'_2(a_{12}\mathbf{e}_1 + a_{22}\mathbf{e}_2 + a_{32}\mathbf{e}_3) + x'_3(a_{13}\mathbf{e}_1 + a_{23}\mathbf{e}_2 + a_{33}\mathbf{e}_3) \\ &= (x'_1 a_{11} + x'_2 a_{12} + x'_3 a_{13})\mathbf{e}_1 + (x'_1 a_{21} + x'_2 a_{22} + x'_3 a_{23})\mathbf{e}_2 + (x'_1 a_{31} + x'_2 a_{32} + x'_3 a_{33})\mathbf{e}_3 \end{aligned}$$

On comparing with (1.3) we see that

$$x_1 = a_{11}x'_1 + a_{12}x'_2 + a_{13}x'_3$$

$$x_2 = a_{21}x'_1 + a_{22}x'_2 + a_{23}x'_3$$

$$x_3 = a_{31}x'_1 + a_{32}x'_2 + a_{33}x'_3$$

or

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

Thus changing from one basis to another is equivalent to transforming the coordinates by multiplication by a matrix, and we thus have another interpretation of matrices. Successive transformations to a third basis will just give $\mathbf{x}'' = \mathbf{B}\mathbf{x}'$, and hence the composite transformation is $\mathbf{x} = (\mathbf{AB})\mathbf{x}''$ and is obtained through the standard matrix rules.

For convenience of working it is usual to take mutually orthogonal vectors as a basis, so that $\mathbf{e}_i^T \mathbf{e}_j = \delta_{ij}$ and $\mathbf{e}'_i^T \mathbf{e}'_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Using (1.2) and multiplying out these orthogonality relations, we have

$$\mathbf{e}'_i^T \mathbf{e}'_j = \sum_k a_{ki} \mathbf{e}_k^T \sum_p a_{pj} \mathbf{e}_p = \sum_k \sum_p a_{ki} a_{pj} \mathbf{e}_k^T \mathbf{e}_p = \sum_k \sum_p a_{ki} a_{pj} \delta_{kp} = \sum_k a_{ki} a_{kj}$$

Hence

$$\sum_k a_{ki} a_{kj} = \delta_{ij}$$

or in matrix form

$$\mathbf{A}^T \mathbf{A} = \mathbf{I}$$

Note that such a matrix \mathbf{A} with $\mathbf{A}^{-1} = \mathbf{A}^T$ is called an **orthogonal matrix**.

1.3.3 Exercises

- 1 Which of the following sets form a basis for the three-dimensional Euclidean space \mathbb{R}^3 ?

(a) $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ (b) $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 2 \\ 5 \end{bmatrix}$

(c) $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$

- 2 Given the unit vectors

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

find the transformation that takes these to the vectors

$$\mathbf{e}'_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}'_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{e}'_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Under this, how does the vector $\mathbf{x} = x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 + x_3 \mathbf{e}_3$ transform and what is the geometrical interpretation? What lines transform into scalar multiples of themselves?

- 3 Show that the set of all cubic polynomials forms a vector space. Which of the following sets of functions are bases of that space?

- (a) $\{1, x, x^2, x^3\}$
 (b) $\{1 - x, 1 + x, 1 - x^3, 1 + x^3\}$
 (c) $\{1 - x, 1 + x, x^2(1 - x), x^2(1 + x)\}$
 (d) $\{x(1 - x), x(1 + x), 1 - x^3, 1 + x^3\}$
 (e) $\{1 + 2x, 2x + 3x^2, 3x^2 + 4x^3, 4x^3 + 1\}$

- 4 Describe the vector space

$$S(x + 2x^3, 2x - 3x^5, x + x^3)$$

What is its dimension?

1.4 The eigenvalue problem

A problem that leads to a concept of crucial importance in many branches of mathematics and its applications is that of seeking non-trivial solutions $\mathbf{x} \neq \mathbf{0}$ to the matrix equation

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

This is referred to as the *eigenvalue problem*; values of the scalar λ for which non-trivial solutions exist are called **eigenvalues** and the corresponding solutions $\mathbf{x} \neq \mathbf{0}$ are called the **eigenvectors**. Such problems arise naturally in many branches of engineering. For example, in vibrations the eigenvalues and eigenvectors describe the frequency and mode of vibration respectively, while in mechanics they represent principal stresses and the principal axes of stress in bodies subjected to external forces. In Section 1.11, and later in Section 5.4.1, we shall see that eigenvalues also play an important role in the stability analysis of dynamical systems.

For continuity some of the introductory material on eigenvalues and eigenvectors, contained in Chapter 5 of MEM, is first revisited.

1.4.1 The characteristic equation

The set of simultaneous equations

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad (1.4)$$

where \mathbf{A} is an $n \times n$ matrix and $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$ is an $n \times 1$ column vector can be written in the form

$$(\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = 0 \quad (1.5)$$

where \mathbf{I} is the identity matrix. The matrix equation (1.5) represents simply a set of homogeneous equations, and we know that a non-trivial solution exists if

$$c(\lambda) = |\lambda\mathbf{I} - \mathbf{A}| = 0 \quad (1.6)$$

Here $c(\lambda)$ is the expansion of the determinant and is a polynomial of degree n in λ , called the **characteristic polynomial** of \mathbf{A} . Thus

$$c(\lambda) = \lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \dots + c_1\lambda + c_0$$

and the equation $c(\lambda) = 0$ is called the **characteristic equation** of \mathbf{A} . We note that this equation can be obtained just as well by evaluating $|\mathbf{A} - \lambda\mathbf{I}| = 0$; however, the form (1.6) is preferred for the definition of the characteristic equation, since the coefficient of λ^n is then always +1.

In many areas of engineering, particularly in those involving vibration or the control of processes, the determination of those values of λ for which (1.5) has a non-trivial solution (that is, a solution for which $\mathbf{x} \neq 0$) is of vital importance. These values of λ are precisely the values that satisfy the characteristic equation, and are called the **eigenvalues** of \mathbf{A} .

Example 1.2

Find the characteristic equation for the matrix

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

Solution By (1.6), the characteristic equation for \mathbf{A} is the cubic equation

$$c(\lambda) = \begin{vmatrix} \lambda-1 & -1 & 2 \\ 1 & \lambda-2 & -1 \\ 0 & -1 & \lambda+1 \end{vmatrix} = 0$$

Expanding the determinant along the first column gives

$$\begin{aligned} &= (\lambda-1) \begin{vmatrix} \lambda-2 & -1 \\ -1 & \lambda+1 \end{vmatrix} - (-1) \begin{vmatrix} 1 & -1 \\ 0 & \lambda+1 \end{vmatrix} + 2 \begin{vmatrix} 1 & \lambda-2 \\ 0 & 1 \end{vmatrix} \\ &= (\lambda-1)[(\lambda-2)(\lambda+1)-1] + \lambda+1 + 2(-1) \\ &= (\lambda-1)(\lambda^2-\lambda-3) + \lambda-1 \\ &= (\lambda-1)(\lambda^2-\lambda-2) \end{aligned}$$

Thus, after simplification,

$$c(\lambda) = \lambda^3 - 2\lambda^2 - \lambda + 2 = 0$$

is the required characteristic equation.

For matrices of large order, determining the characteristic polynomial by direct expansion of $|\lambda I - \mathbf{A}|$ is unsatisfactory in view of the large number of terms involved in the determinant expansion. Alternative procedures are available to reduce the amount of calculation, and that due to Dmitry Konstantinovich Faddeev (1907–1989) may be stated as follows.

The method of Faddeev

If the characteristic polynomial of an $n \times n$ matrix \mathbf{A} is written as

$$\lambda^n - p_1\lambda^{n-1} - \cdots - p_{n-1}\lambda - p_n$$

then the coefficients p_1, p_2, \dots, p_n can be computed using

$$p_r = \frac{1}{r} \text{tr}(\mathbf{A}_r) \quad (r = 1, 2, \dots, n)$$

where

$$\mathbf{A}_r = \begin{cases} \mathbf{A} & (r = 1) \\ \mathbf{AB}_{r-1} & (r = 2, 3, \dots, n) \end{cases}$$

and

$$\mathbf{B}_r = \mathbf{A}_r - p_r \mathbf{I}, \quad \text{where } \mathbf{I} \text{ is the } n \times n \text{ identity matrix}$$

The calculations may be checked using the result that

$$\mathbf{B}_n = \mathbf{A}_n - p_n \mathbf{I} \quad \text{must be the zero matrix}$$

Example 1.3

Using the method of Faddeev, obtain the characteristic equation of the matrix \mathbf{A} of Example 1.2.

Solution

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

We have $n = 3$, so let the characteristic equation be

$$c(\lambda) = \lambda^3 - p_1\lambda^2 - p_2\lambda - p_3$$

Then, following the procedure described above,

$$p_1 = \text{tr}(\mathbf{A}) = (1 + 2 - 1) = 2$$

$$\mathbf{B}_1 = \mathbf{A} - 2\mathbf{I} = \begin{bmatrix} -1 & 1 & -2 \\ -1 & 0 & 1 \\ 0 & 1 & -3 \end{bmatrix}$$

$$\mathbf{A}_2 = \mathbf{AB}_1 = \begin{bmatrix} -2 & -1 & 5 \\ -1 & 0 & 1 \\ -1 & -1 & 4 \end{bmatrix}$$

$$p_2 = \frac{1}{2}\text{tr}(\mathbf{A}_2) = \frac{1}{2}(-2 + 0 + 4) = 1$$

$$\mathbf{B}_2 = \mathbf{A}_2 - \mathbf{I} = \begin{bmatrix} -3 & -1 & 5 \\ -1 & -1 & 1 \\ -1 & -1 & 3 \end{bmatrix}$$

$$\mathbf{A}_3 = \mathbf{AB}_2 = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -2 \end{bmatrix}$$

$$p_3 = \frac{1}{3}\text{tr}(\mathbf{A}_3) = \frac{1}{3}(-2 - 2 - 2) = -2$$

Then, the characteristic polynomial of \mathbf{A} is

$$c(\lambda) = \lambda^3 - 2\lambda^2 - \lambda + 2$$

in agreement with the result of Example 1.2. In this case, however, a check may be carried out on the computation, since

$$\mathbf{B}_3 = \mathbf{A}_3 + 2\mathbf{I} = 0$$

as required.

1.4.2 Eigenvalues and eigenvectors

The roots of the characteristic equation (1.6) are called the **eigenvalues** of the matrix \mathbf{A} (the terms latent roots, proper roots and characteristic roots are also sometimes used). By the Fundamental Theorem of Algebra, a polynomial equation of degree n has exactly n roots, so that the matrix \mathbf{A} has exactly n eigenvalues λ_i , $i = 1, 2, \dots, n$. These eigenvalues may be real or complex, and not necessarily distinct. Corresponding to each eigenvalue λ_i , there is a non-zero solution $\mathbf{x} = \mathbf{e}_i$ of (1.5); \mathbf{e}_i is called the **eigenvector** of \mathbf{A} corresponding to the eigenvalue λ_i . We note that if $\mathbf{x} = \mathbf{e}_i$ satisfies (1.5) then any scalar multiple $\beta_i \mathbf{e}_i$ of \mathbf{e}_i also satisfies (1.5), so that the eigenvector \mathbf{e}_i may only be determined to within a scalar multiple.

Example 1.4

Determine the eigenvalues and eigenvectors for the matrix \mathbf{A} of Example 1.2.

Solution

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

The eigenvalues λ_i of \mathbf{A} satisfy the characteristic equation $c(\lambda) = 0$, and this has been obtained in Examples 1.2 and 1.3 as the cubic

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

which can be solved to obtain the eigenvalues λ_1 , λ_2 and λ_3 . Alternatively, it may be possible, using the determinant form $|\lambda\mathbf{I} - \mathbf{A}|$, or indeed (as we often do when seeking the eigenvalues) the form $|\mathbf{A} - \lambda\mathbf{I}|$, by carrying out suitable row and/or column operations to factorize the determinant. In this case

$$|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} 1 - \lambda & 1 & -2 \\ -1 & 2 - \lambda & 1 \\ 0 & 1 & -1 - \lambda \end{vmatrix}$$

and adding column 1 to column 3 gives

$$\begin{vmatrix} 1 - \lambda & 1 & -1 - \lambda \\ -1 & 2 - \lambda & 0 \\ 0 & 1 & -1 - \lambda \end{vmatrix} = -(1 + \lambda) \begin{vmatrix} 1 - \lambda & 1 & 1 \\ -1 & 2 - \lambda & 0 \\ 0 & 1 & 1 \end{vmatrix}$$

Subtracting row 3 from row 1 gives

$$-(1 + \lambda) \begin{vmatrix} 1 - \lambda & 0 & 0 \\ -1 & 2 - \lambda & 0 \\ 0 & 1 & 1 \end{vmatrix} = -(1 + \lambda)(1 - \lambda)(2 - \lambda)$$

Setting $|\mathbf{A} - \lambda\mathbf{I}| = 0$ gives the eigenvalues as $\lambda_1 = 2$, $\lambda_2 = 1$ and $\lambda_3 = -1$. The order in which they are written is arbitrary, but for consistency we shall adopt the convention of taking $\lambda_1, \lambda_2, \dots, \lambda_n$ in decreasing order.

Having obtained the eigenvalues λ_i ($i = 1, 2, 3$), the corresponding eigenvectors \mathbf{e}_i are obtained by solving the appropriate homogeneous equations

$$(\mathbf{A} - \lambda_i \mathbf{I}) \mathbf{e}_i = 0 \quad (1.7)$$

When $i = 1$, $\lambda_1 = \lambda_1 = 2$ and (1.7) is

$$\begin{bmatrix} -1 & 1 & -2 \\ -1 & 0 & 1 \\ 0 & 1 & -3 \end{bmatrix} \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \end{bmatrix} = 0$$

that is,

$$-e_{11} + e_{12} - 2e_{13} = 0$$

$$-e_{11} + 0e_{12} + e_{13} = 0$$

$$0e_{11} + e_{12} - 3e_{13} = 0$$

leading to the solution

$$\frac{e_{11}}{-1} = \frac{-e_{12}}{3} = \frac{e_{13}}{-1} = \beta_1$$

where β_1 is an arbitrary non-zero scalar. Thus the eigenvector e_1 corresponding to the eigenvalue $\lambda_1 = 2$ is

$$e_1 = \beta_1 [1 \quad 3 \quad 1]^T$$

As a check, we can compute

$$\mathbf{A}e_1 = \beta_1 \begin{bmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = \beta_1 \begin{bmatrix} 2 \\ 6 \\ 2 \end{bmatrix} = 2\beta_1 \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = \lambda_1 e_1$$

and thus conclude that our calculation was correct.

When $i = 2$, $\lambda_i = \lambda_2 = 1$ and we have to solve

$$\begin{bmatrix} 0 & 1 & -2 \\ -1 & 1 & 1 \\ 0 & 1 & -2 \end{bmatrix} \begin{bmatrix} e_{21} \\ e_{22} \\ e_{23} \end{bmatrix} = 0$$

that is,

$$0e_{21} + e_{22} - 2e_{23} = 0$$

$$-e_{21} + e_{22} + e_{23} = 0$$

$$0e_{21} + e_{22} - 2e_{23} = 0$$

leading to the solution

$$\frac{e_{21}}{-1} = \frac{-e_{22}}{2} = \frac{e_{23}}{-1} = \beta_2$$

where β_2 is an arbitrary scalar. Thus the eigenvector e_2 corresponding to the eigenvalue $\lambda_2 = 1$ is

$$e_2 = \beta_2 [3 \quad 2 \quad 1]^T$$

Again a check could be made by computing $\mathbf{A}e_2$.

Finally, when $i = 3$, $\lambda_i = \lambda_3 = -1$ and we obtain from (1.7)

$$\begin{bmatrix} 2 & 1 & -2 \\ -1 & 3 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} e_{31} \\ e_{32} \\ e_{33} \end{bmatrix} = 0$$

that is,

$$\begin{aligned} 2e_{31} + e_{32} - 2e_{33} &= 0 \\ -e_{31} + 3e_{32} + e_{33} &= 0 \\ 0e_{31} + e_{32} + 0e_{33} &= 0 \end{aligned}$$

and hence

$$\frac{e_{31}}{-1} = \frac{e_{32}}{0} = \frac{e_{33}}{-1} = \beta_3$$

Here again β_3 is an arbitrary scalar, and the eigenvector e_3 corresponding to the eigenvalue λ_3 is

$$e_3 = \beta_3 [1 \quad 0 \quad 1]^T$$

The calculation can be checked as before. Thus we have found that the eigenvalues of the matrix A are 2, 1 and -1 , with corresponding eigenvectors

$$\beta_1 [1 \quad 3 \quad 1]^T, \quad \beta_2 [3 \quad 2 \quad 1]^T \quad \text{and} \quad \beta_3 [1 \quad 0 \quad 1]^T$$

respectively.

Since in Example 1.4 the β_i , $i = 1, 2, 3$, are arbitrary, it follows that there are an infinite number of eigenvectors, scalar multiples of each other, corresponding to each eigenvalue. It is convenient to scale the eigenvectors according to some convention. A convention frequently adopted is to **normalize** the eigenvectors so that they are uniquely determined up to a scale factor of ± 1 . The normalized form of an eigenvector $e = [e_1 \quad e_2 \quad \dots \quad e_n]^T$ is denoted by \hat{e} and is given by

$$\hat{e} = \frac{e}{|e|}$$

where

$$|e| = \sqrt{(e_1^2 + e_2^2 + \dots + e_n^2)}$$

For example, for the matrix A of Example 1.4, the normalized forms of the eigenvectors are

$$\hat{e}_1 = [1/\sqrt{11} \quad 3/\sqrt{11} \quad 1/\sqrt{11}]^T, \quad \hat{e}_2 = [3/\sqrt{14} \quad 2/\sqrt{14} \quad 1/\sqrt{14}]^T$$

and

$$\hat{e}_3 = [1/\sqrt{2} \quad 0 \quad 1/\sqrt{2}]^T$$

However, throughout the text, unless otherwise stated, the eigenvectors will always be presented in their ‘simplest’ form, so that for the matrix of Example 1.4 we take $\beta_1 = \beta_2 = \beta_3 = 1$ and write

$$e_1 = [1 \quad 3 \quad 1]^T, \quad e_2 = [3 \quad 2 \quad 1]^T \quad \text{and} \quad e_3 = [1 \quad 0 \quad 1]^T$$



For a $n \times n$ matrix \mathbf{A} the MATLAB command `p=poly(A)` generates an $n + 1$ element row vector whose elements are the coefficients of the characteristic polynomial of \mathbf{A} , the coefficients being ordered in descending powers. The eigenvalues of \mathbf{A} are the roots of the polynomial and are generated using the command `roots(p)`. The command

```
[M, S]=eig(A)
```

generates the normalized eigenvectors of \mathbf{A} as the columns of the matrix \mathbf{M} and its corresponding eigenvalues as the diagonal elements of the diagonal matrix \mathbf{S} (\mathbf{M} and \mathbf{S} are called respectively the modal and spectral matrices of \mathbf{A} and we shall return to discuss them in more detail in Section 1.6.1). In the absence of the left-hand arguments, the command `eig(A)` by itself simply generates the eigenvalues of \mathbf{A} .

For the matrix \mathbf{A} of Example 1.4 the commands

```
A=[1 1 -2; -1 2 1; 0 1 -1];
[M, S]=eig(A)
```

generate the output

```
0.3015 -0.8018 0.7071
M=0.9045 -0.5345 0.0000
0.3015 -0.2673 0.7071

2.0000 0 0
S=0 1.0000 0
0 0 -1.0000
```

These concur with our calculated answers, with $\beta_1 = 0.3015$, $\beta_2 = -0.2673$ and $\beta_3 = 0.7071$.

Using the Symbolic Math Toolbox in MATLAB we saw earlier that the matrix \mathbf{A} may be converted from numeric into symbolic form using the command `A=sym(A)`. Then its symbolic eigenvalues and eigenvectors are generated using the sequence of commands

```
A=[1 1 -2; -1 2 1; 0 1 -1];
A=sym(A);
[M, S]=eig(A)
```

as

```
M=[3, 1, 1]
[2, 3, 0]
[1, 1, 1]

S=[1, 0, 0]
[0, 2, 0]
[0, 0, -1]
```

In MAPLE the command `Eigenvalues(A);` returns a vector of eigenvalues. The command `Eigenvectors(A);` returns both a vector of eigenvalues as before and a matrix containing the eigenvalues, so that the i th column is an eigenvector corresponding to the eigenvalue in the i th entry of the preceding vector. Thus the commands:

```

with(LinearAlgebra),
A:=Matrix([[1,1,-2], [-1,2,1], [0,1,-1]]);
Eigenvalues(A);

return


$$\begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}$$


and the command

Eigenvectors(A);

returns


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


```

Example 1.5

Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$

Solution Now

$$\begin{aligned} |\lambda I - \mathbf{A}| &= \begin{vmatrix} \lambda - \cos \theta & \sin \theta \\ -\sin \theta & \lambda - \cos \theta \end{vmatrix} \\ &= \lambda^2 - 2\lambda \cos \theta + \cos^2 \theta + \sin^2 \theta = \lambda^2 - 2\lambda \cos \theta + 1 \end{aligned}$$

So the eigenvalues are the roots of

$$\lambda^2 - 2\lambda \cos \theta + 1 = 0$$

that is,

$$\lambda = \cos \theta \pm j \sin \theta$$

Solving for the eigenvectors as in Example 1.4, we obtain

$$\mathbf{e}_1 = [1 \ -j]^T \text{ and } \mathbf{e}_2 = [1 \ j]^T$$



This example may be done in MATLAB as

```

syms t;
A=[cos(t) -sin(t); sin(t) cos(t)];
[M,S]=eig(A)
symplify(M)

```

We see that eigenvalues can be complex numbers, and that the eigenvectors may have complex components. This situation arises when the characteristic equation has complex (conjugate) roots.

1.4.3 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 5 Using the method of Faddeev, obtain the characteristic polynomials of the matrices

$$(a) \begin{bmatrix} 3 & 2 & 1 \\ 4 & 5 & -1 \\ 2 & 3 & 4 \end{bmatrix}$$

$$(b) \begin{bmatrix} 2 & -1 & 1 & 2 \\ 0 & 1 & 1 & 0 \\ -1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$(c) \begin{bmatrix} 1 & 0 & -4 \\ 0 & 5 & 4 \\ -4 & 4 & 3 \end{bmatrix}$$

$$(d) \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & 2 \\ -1 & 1 & 3 \end{bmatrix}$$

$$(e) \begin{bmatrix} 5 & 0 & 6 \\ 0 & 11 & 6 \\ 6 & 6 & -2 \end{bmatrix}$$

$$(f) \begin{bmatrix} 1 & -1 & 0 \\ 1 & 2 & 1 \\ -2 & 1 & -1 \end{bmatrix}$$

- 6 Find the eigenvalues and corresponding eigenvectors of the matrices

$$(a) \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

$$(b) \begin{bmatrix} 1 & 2 \\ 3 & 2 \end{bmatrix}$$

$$(g) \begin{bmatrix} 4 & 1 & 1 \\ 2 & 5 & 4 \\ -1 & -1 & 0 \end{bmatrix}$$

$$(h) \begin{bmatrix} 1 & -4 & -2 \\ 0 & 3 & 1 \\ 1 & 2 & 4 \end{bmatrix}$$

1.4.4 Repeated eigenvalues

In the examples considered so far the eigenvalues λ_i ($i = 1, 2, \dots$) of the matrix \mathbf{A} have been distinct, and in such cases the corresponding eigenvectors can be found and are linearly independent. The matrix \mathbf{A} is then said to have a full set of linearly independent eigenvectors. It is clear that the roots of the characteristic polynomial $c(\lambda)$ may not all be distinct; and when $c(\lambda)$ has $p \leq n$ distinct roots, $c(\lambda)$ may be factorized as

$$c(\lambda) = (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \dots (\lambda - \lambda_p)^{m_p}$$

indicating that the root $\lambda = \lambda_i$, $i = 1, 2, \dots, p$, is a root of order m_i , where the integer m_i is called the **algebraic multiplicity** of the eigenvalue λ_i . Clearly $m_1 + m_2 + \dots + m_p = n$. When a matrix \mathbf{A} has repeated eigenvalues, the question arises as to whether it is possible to obtain a full set of linearly independent eigenvectors for \mathbf{A} . We first consider two examples to illustrate the situation.

Example 1.6

Determine the eigenvalues and corresponding eigenvectors of the matrix

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

Solution We find the eigenvalues from

$$\begin{vmatrix} 3 - \lambda & -3 & 2 \\ -1 & 5 - \lambda & -2 \\ -1 & 3 & -\lambda \end{vmatrix} = 0$$

as $\lambda_1 = 4$, $\lambda_2 = \lambda_3 = 2$.

The eigenvectors are obtained from

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{e}_i = 0 \quad (1.8)$$

and when $\lambda = \lambda_1 = 4$, we obtain from (1.8)

$$\mathbf{e}_1 = [1 \quad -1 \quad -1]^T$$

When $\lambda = \lambda_2 = \lambda_3 = 2$, (1.8) becomes

$$\begin{bmatrix} 1 & -3 & 2 \\ -1 & 3 & -2 \\ -1 & 3 & -2 \end{bmatrix} \begin{bmatrix} e_{21} \\ e_{22} \\ e_{23} \end{bmatrix} = 0$$

so that the corresponding eigenvector is obtained from the single equation

$$e_{21} - 3e_{22} + 2e_{23} = 0 \quad (1.9)$$

Clearly we are free to choose any two of the components e_{21} , e_{22} or e_{23} at will, with the remaining one determined by (1.9). Suppose we set $e_{22} = \alpha$ and $e_{23} = \beta$; then (1.9) means that $e_{21} = 3\alpha - 2\beta$, and thus

$$\mathbf{e}_2 = [3\alpha - 2\beta \quad \alpha \quad \beta]^T = \alpha \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} \quad (1.10)$$

Now $\lambda = 2$ is an eigenvalue of multiplicity 2, and we seek, if possible, two linearly independent eigenvectors defined by (1.10). Setting $\alpha = 1$ and $\beta = 0$ yields

$$\mathbf{e}_2 = [3 \quad 1 \quad 0]^T$$

and setting $\alpha = 0$ and $\beta = 1$ gives a second vector

$$\mathbf{e}_3 = [-2 \quad 0 \quad 1]^T$$

These two vectors are linearly independent and of the form defined by (1.10), and it is clear that many other choices are possible. However, any other choices of the form (1.10) will be linear combinations of \mathbf{e}_2 and \mathbf{e}_3 as chosen above. For example, $\mathbf{e} = [1 \quad 1 \quad 1]$ satisfies (1.10), but $\mathbf{e} = \mathbf{e}_2 + \mathbf{e}_3$.

In this example, although there was a repeated eigenvalue of algebraic multiplicity 2, it was possible to construct two linearly independent eigenvectors corresponding to this eigenvalue. Thus the matrix \mathbf{A} has three and only three linearly independent eigenvectors.



Repeating the above, the MATLAB commands

```
A=[3 -3 2; -1 5 -2; -1 3 0];
[M, S]=eig(A)
```

generate

```
0.5774 -0.5774 -0.9633
M=-0.5774 -0.5774 -0.2075
-0.5774 -0.5774 0.1704
4.0000 0 0
S= 0 2.0000 0
0 0 2.0000
```

Clearly the first column of \mathbf{M} (corresponding to the eigenvalue $\lambda_1 = 4$) is a scalar multiple of \mathbf{e}_1 . The second and third columns of \mathbf{M} (corresponding to the repeated eigenvalue $\lambda_2 = \lambda_3 = 2$) are not scalar multiples of \mathbf{e}_2 and \mathbf{e}_3 . However, both satisfy (1.10) and are equally acceptable as a pair of linearly independent eigenvectors corresponding to the repeated eigenvalue. It is left as an exercise to show that both are linear combinations of \mathbf{e}_2 and \mathbf{e}_3 .

Check that in symbolic form the commands

```
A=sym(A);
[M, S]=eig(A)
```

generate

```
M=[-1, 3, -2]
[1, 1, 0]
[1, 0, 1]

S=[4, 0, 0]
[0, 2, 0]
[0, 0, 2]
```

In MAPLE the command `Eigenvectors(A)` ; produces corresponding results. Thus the commands

```
with(LinearAlgebra):
A:=Matrix([[3,-3,2],[-1,5,-2],[-1,3,0]]);
Eigenvectors(A);

return
```

$$\begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix} \begin{bmatrix} -2 & 3 & -1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Example 1.7

Determine the eigenvalues and corresponding eigenvectors for the matrix

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

Solution Solving $|\mathbf{A} - \lambda\mathbf{I}| = 0$ gives the eigenvalues as $\lambda_1 = \lambda_2 = 2$, $\lambda_3 = 1$. The eigenvector corresponding to the non-repeated or simple eigenvalue $\lambda_3 = 1$ is easily found as

$$\mathbf{e}_3 = [1 \quad 1 \quad -1]^T$$

When $\lambda = \lambda_1 = \lambda_2 = 2$, the corresponding eigenvector is given by

$$(\mathbf{A} - 2\mathbf{I})\mathbf{e}_1 = 0$$

that is, as the solution of

$$-e_{11} + 2e_{12} + 2e_{13} = 0 \tag{i}$$

$$e_{13} = 0 \tag{ii}$$

$$-e_{11} + 2e_{12} = 0 \tag{iii}$$

From (ii) we have $e_{13} = 0$, and from (i) and (ii) it follows that $e_{11} = 2e_{12}$. We deduce that there is only one linearly independent eigenvector corresponding to the repeated eigenvalue $\lambda = 2$, namely

$$\mathbf{e}_1 = [2 \quad 1 \quad 0]^T$$

and in this case the matrix \mathbf{A} does not possess a full set of linearly independent eigenvectors.

We see from Examples 1.6 and 1.7 that if an $n \times n$ matrix \mathbf{A} has repeated eigenvalues then a full set of n linearly independent eigenvectors may or may not exist. The number of linearly independent eigenvectors associated with a repeated eigenvalue λ_i of algebraic multiplicity m_i is given by the **nullity** q_i of the matrix $\mathbf{A} - \lambda_i \mathbf{I}$, where

$$q_i = n - \text{rank } (\mathbf{A} - \lambda_i \mathbf{I}), \quad \text{with } 1 \leq q_i \leq m_i \quad (1.11)$$

q_i is sometimes referred to as the **degeneracy** of the matrix $\mathbf{A} - \lambda_i \mathbf{I}$ or the **geometric multiplicity** of the eigenvalue λ_i , since it determines the dimension of the space spanned by the corresponding eigenvector(s) \mathbf{e}_i .

Example 1.8

Confirm the findings of Examples 1.6 and 1.7 concerning the number of linearly independent eigenvectors found.

Solution

In Example 1.6, we had an eigenvalue $\lambda_2 = 2$ of algebraic multiplicity 2. Correspondingly,

$$\mathbf{A} - \lambda_2 \mathbf{I} = \begin{bmatrix} 3-2 & -3 & 2 \\ -1 & 5-2 & -2 \\ -1 & 3 & -2 \end{bmatrix} = \begin{bmatrix} 1 & -3 & 2 \\ -1 & 3 & -2 \\ -1 & 3 & -2 \end{bmatrix}$$

and performing the row operation of adding row 1 to rows 2 and 3 yields

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

Adding 3 times column 1 to column 2 followed by subtracting 2 times column 1 from column 3 gives finally

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

indicating a rank of 1. Then from (1.11) the nullity $q_2 = 3 - 1 = 2$, confirming that corresponding to the eigenvalue $\lambda = 2$ there are two linearly independent eigenvectors, as found in Example 1.6.

In Example 1.7 we again had a repeated eigenvalue $\lambda_1 = 2$ of algebraic multiplicity 2. Then

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

Performing row and column operations as before produces the matrix

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

this time indicating a rank of 2. From (1.11) the nullity $q_1 = 3 - 2 = 1$, confirming that there is one and only one linearly independent eigenvector associated with this eigenvalue, as found in Example 1.7.

1.4.5 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 7 Obtain the eigenvalues and corresponding eigenvectors of the matrices

$$(a) \begin{bmatrix} 2 & 2 & 1 \\ 1 & 3 & 1 \\ 1 & 2 & 2 \end{bmatrix} \quad (b) \begin{bmatrix} 0 & -2 & -2 \\ -1 & 1 & 2 \\ -1 & -1 & 2 \end{bmatrix}$$

$$(c) \begin{bmatrix} 4 & 6 & 6 \\ 1 & 3 & 2 \\ -1 & -5 & -2 \end{bmatrix} \quad (d) \begin{bmatrix} 7 & -2 & -4 \\ 3 & 0 & -2 \\ 6 & -2 & -3 \end{bmatrix}$$

- 8 Given that $\lambda = 1$ is a three-times repeated eigenvalue of the matrix

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

using the concept of rank, determine how many linearly independent eigenvectors correspond to this value of λ . Determine a corresponding set of linearly independent eigenvectors.

- 9 Given that $\lambda = 1$ is a twice-repeated eigenvalue of the matrix

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

how many linearly independent eigenvectors correspond to this value of λ ? Determine a corresponding set of linearly independent eigenvectors.

1.4.6 Some useful properties of eigenvalues

The following basic properties of the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ of an $n \times n$ matrix \mathbf{A} are sometimes useful. The results are readily proved either from the definition of eigenvalues as the values of λ satisfying (1.4), or by comparison of corresponding characteristic polynomials (1.6). Consequently, the proofs are left to Exercise 10.

Property 1.1

The sum of the eigenvalues of \mathbf{A} is

$$\sum_{i=1}^n \lambda_i = \text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}$$

Property 1.2

The product of the eigenvalues of \mathbf{A} is

$$\prod_{i=1}^n \lambda_i = \det(\mathbf{A})$$

where $\det(\mathbf{A})$ denotes the determinant of the matrix \mathbf{A} .

Property 1.3

The eigenvalues of the inverse matrix \mathbf{A}^{-1} , provided it exists, are

$$\frac{1}{\lambda_1}, \quad \frac{1}{\lambda_2}, \quad \dots, \quad \frac{1}{\lambda_n}$$

Property 1.4

The eigenvalues of the transposed matrix \mathbf{A}^T are

$$\lambda_1, \quad \lambda_2, \quad \dots, \quad \lambda_n$$

as for the matrix \mathbf{A} .

Property 1.5

If k is a scalar then the eigenvalues of $k\mathbf{A}$ are

$$k\lambda_1, \quad k\lambda_2, \quad \dots, \quad k\lambda_n$$

Property 1.6

If k is a scalar and \mathbf{I} the $n \times n$ identity (unit) matrix then the eigenvalues of $\mathbf{A} \pm k\mathbf{I}$ are respectively

$$\lambda_1 \pm k, \quad \lambda_2 \pm k, \quad \dots, \quad \lambda_n \pm k$$

Property 1.7

If k is a positive integer then the eigenvalues of \mathbf{A}^k are

$$\lambda_1^k, \lambda_2^k, \dots, \lambda_n^k$$

1.4.7 Symmetric matrices

A square matrix \mathbf{A} is said to be **symmetric** if $\mathbf{A}^T = \mathbf{A}$. Such matrices form an important class and arise in a variety of practical situations. Two important results concerning the eigenvalues and eigenvectors of such matrices are

- (a) the eigenvalues of a real symmetric matrix are real;
- (b) for an $n \times n$ real symmetric matrix it is always possible to find n linearly independent eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ that are mutually orthogonal so that $\mathbf{e}_i^T \mathbf{e}_j = 0$ for $i \neq j$.

If the orthogonal eigenvectors of a symmetric matrix are normalized as

$$\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_n$$

then the **inner (scalar) product** is

$$\hat{\mathbf{e}}_i^T \hat{\mathbf{e}}_j = \delta_{ij} \quad (i, j = 1, 2, \dots, n)$$

where δ_{ij} is the Kronecker delta defined in Section 1.3.2.

The set of normalized eigenvectors of a symmetric matrix therefore forms an orthonormal set (that is, it forms a mutually orthogonal normalized set of vectors).

Example 1.9

Obtain the eigenvalues and corresponding orthogonal eigenvectors of the symmetric matrix

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

and show that the normalized eigenvectors form an orthonormal set.

Solution The eigenvalues of \mathbf{A} are $\lambda_1 = 6$, $\lambda_2 = 3$ and $\lambda_3 = 1$, with corresponding eigenvectors

$$\mathbf{e}_1 = [1 \ 2 \ 0]^T, \quad \mathbf{e}_2 = [0 \ 0 \ 1]^T, \quad \mathbf{e}_3 = [-2 \ 1 \ 0]^T$$

which in normalized form are

$$\hat{\mathbf{e}}_1 = [1 \ 2 \ 0]^T / \sqrt{5}, \quad \hat{\mathbf{e}}_2 = [0 \ 0 \ 1]^T, \quad \hat{\mathbf{e}}_3 = [-2 \ 1 \ 0]^T / \sqrt{5}$$

Evaluating the inner products, we see that, for example,

$$\hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_1 = \frac{1}{5} + \frac{4}{5} + 0 = 1, \quad \hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_3 = -\frac{2}{5} + \frac{2}{5} + 0 = 0$$

and that, in general,

$$\hat{\mathbf{e}}_i^T \hat{\mathbf{e}}_j = \delta_{ij} \quad (i, j = 1, 2, 3)$$

confirming that the eigenvectors form an orthonormal set.

1.4.8 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

10 Verify Properties 1.1–1.7 of Section 1.4.6.

11 Given that the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 1 & 1 \\ 2 & 5 & 4 \\ -1 & -1 & 0 \end{bmatrix}$$

are 5, 3 and 1:

- (a) confirm Properties 1.1–1.4 of Section 1.4.6;
- (b) taking $k = 2$, confirm Properties 1.5–1.7 of Section 1.4.6.

12 Determine the eigenvalues and corresponding eigenvectors of the symmetric matrix

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

and verify that the eigenvectors are mutually orthogonal.

13 The 3×3 symmetric matrix \mathbf{A} has eigenvalues 6, 3 and 2. The eigenvectors corresponding to the eigenvalues 6 and 3 are $[1 \ 1 \ 2]^T$ and $[1 \ 1 \ -1]^T$ respectively. Find an eigenvector corresponding to the eigenvalue 2.

1.5 Numerical methods

In practice we may well be dealing with matrices whose elements are decimal numbers or with matrices of high orders. In order to determine the eigenvalues and eigenvectors of such matrices, it is necessary that we have numerical algorithms at our disposal.

1.5.1 The power method

Consider a matrix \mathbf{A} having n distinct eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ and corresponding n linearly independent eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. Taking this set of vectors as the basis, we can write any vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$ as a linear combination in the form

$$\mathbf{x} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \dots + \alpha_n \mathbf{e}_n = \sum_{i=1}^n \alpha_i \mathbf{e}_i$$

Then, since $\mathbf{A}\mathbf{e}_i = \lambda_i \mathbf{e}_i$ for $i = 1, 2, \dots, n$,

$$\mathbf{Ax} = \mathbf{A} \sum_{i=1}^n \alpha_i \mathbf{e}_i = \sum_{i=1}^n \alpha_i \lambda_i \mathbf{e}_i$$

and, for any positive integer k ,

$$\mathbf{A}^k \mathbf{x} = \sum_{i=1}^n \alpha_i \lambda_i^k \mathbf{e}_i$$

or, equivalently,

$$\mathbf{A}^k \mathbf{x} = \lambda_1^k \left[\alpha_1 \mathbf{e}_1 + \sum_{i=2}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{e}_i \right] \quad (1.12)$$

Assuming that the eigenvalues are ordered such that

$$|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$$

and that $\alpha_1 \neq 0$, we have from (1.12)

$$\lim_{k \rightarrow \infty} \mathbf{A}^k \mathbf{x} = \lambda_1^k \alpha_1 \mathbf{e}_1 \quad (1.13)$$

since all the other terms inside the square brackets tend to zero. The eigenvalue λ_1 and its corresponding eigenvector \mathbf{e}_1 are referred to as the **dominant** eigenvalue and eigenvector respectively. The other eigenvalues and eigenvectors are called **subdominant**.

Thus if we introduce the iterative process

$$\mathbf{x}^{(k+1)} = \mathbf{A} \mathbf{x}^{(k)} \quad (k = 0, 1, 2, \dots)$$

starting with some arbitrary vector $\mathbf{x}^{(0)}$ not orthogonal to \mathbf{e}_1 , it follows from (1.13) that

$$\mathbf{x}^{(k)} = \mathbf{A}^k \mathbf{x}^{(0)}$$

will converge to the dominant eigenvector of \mathbf{A} .

A clear disadvantage with this scheme is that if $|\lambda_1|$ is large then $\mathbf{A}^k \mathbf{x}^{(0)}$ will become very large, and computer overflow can occur. This can be avoided by scaling the vector $\mathbf{x}^{(k)}$ after each iteration. The standard approach is to make the largest element of $\mathbf{x}^{(k)}$ unity using the scaling factor $\max(\mathbf{x}^{(k)})$, which represents the element of $\mathbf{x}^{(k)}$ having the largest modulus.

Thus in practice we adopt the iterative process

$$\begin{aligned} \mathbf{y}^{(k+1)} &= \mathbf{A} \mathbf{x}^{(k)} \\ \mathbf{x}^{(k+1)} &= \frac{\mathbf{y}^{(k+1)}}{\max(\mathbf{y}^{(k+1)})} \quad (k = 0, 1, 2, \dots) \end{aligned} \quad (1.14)$$

and it is common to take $\mathbf{x}^{(0)} = [1 \quad 1 \quad \dots \quad 1]^T$.

Corresponding to (1.12), we have

$$\mathbf{x}^{(k)} = R \lambda_1^k \left[\alpha_1 \mathbf{e}_1 + \sum_{i=2}^n \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k \mathbf{e}_i \right]$$

where

$$R = [\max(y^{(1)}) \max(y^{(2)}) \dots \max(y^{(k)})]^{-1}$$

Again we see that $\mathbf{x}^{(k)}$ converges to a multiple of the dominant eigenvector \mathbf{e}_1 . Also, since $\mathbf{A}\mathbf{x}^{(k)} \rightarrow \lambda_1 \mathbf{x}^{(k)}$, we have $\mathbf{y}^{(k+1)} \rightarrow \lambda_1 \mathbf{x}^{(k)}$, and since the largest element of $\mathbf{x}^{(k)}$ is unity, it follows that the scaling factors $\max(\mathbf{y}^{(k+1)})$ converge to the dominant eigenvalue λ_1 . The **rate of convergence** depends primarily on the ratios

$$\left| \frac{\lambda_2}{\lambda_1} \right|, \left| \frac{\lambda_3}{\lambda_1} \right|, \dots, \left| \frac{\lambda_n}{\lambda_1} \right|$$

The smaller these ratios, the faster the rate of convergence. The iterative process represents the simplest form of the **power method**, and a pseudocode for the basic algorithm is given in Figure 1.1.

Figure 1.1 Outline pseudocode program for power method to calculate the maximum eigenvalue.

```
{read in  $\mathbf{x}^T = [x_1 \ x_2 \ \dots \ x_n]$ }
  m ← 0
  repeat
    m_old ← m
    {evaluate  $\mathbf{y} = \mathbf{A}\mathbf{x}$ }
    {find m = max( $y_i$ )}
    { $\mathbf{x}^T = [y_1/m \ y_2/m \ \dots \ y_n/m]$ }
  until  $abs(m - m_{old}) < tolerance$ 
{write (results)}
```

Example 1.10

Use the power method to find the dominant eigenvalue and the corresponding eigenvector of the matrix

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

Solution Taking $\mathbf{x}^{(0)} = [1 \ 1 \ 1]^T$ in (1.14), we have

$$\mathbf{y}^{(1)} = \mathbf{A}\mathbf{x}^{(0)} = \begin{bmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix} = 2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}; \quad \lambda_1^{(1)} = 2$$

$$\mathbf{x}^{(1)} = \frac{1}{2}\mathbf{y}^{(1)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{y}^{(2)} = \mathbf{A}\mathbf{x}^{(1)} = \begin{bmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = 2 \begin{bmatrix} 0.5 \\ 1 \\ 0.5 \end{bmatrix}; \quad \lambda_2^{(2)} = 2$$

$$\mathbf{x}^{(2)} = \frac{1}{2}\mathbf{y}^{(2)} = \begin{bmatrix} \frac{1}{2} \\ 1 \\ \frac{1}{2} \end{bmatrix}$$

$$\mathbf{y}^{(3)} = \mathbf{A}\mathbf{x}^{(2)} = \begin{bmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} \\ 1 \\ \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 2 \\ \frac{1}{2} \end{bmatrix} = 2 \begin{bmatrix} 0.25 \\ 1 \\ 0.25 \end{bmatrix}; \quad \lambda_3^{(2)} = 2$$

$$\mathbf{x}^{(3)} = \begin{bmatrix} 0.25 \\ 1 \\ 0.25 \end{bmatrix}$$

Continuing with the process, we have

$$\mathbf{y}^{(4)} = 2[0.375 \ 1 \ 0.375]^T, \quad \mathbf{y}^{(5)} = 2[0.312 \ 1 \ 0.312]^T$$

$$\mathbf{y}^{(6)} = 2[0.344 \ 1 \ 0.344]^T, \quad \mathbf{y}^{(7)} = 2[0.328 \ 1 \ 0.328]^T$$

$$\mathbf{y}^{(8)} = 2[0.336 \ 1 \ 0.336]^T$$

Clearly $\mathbf{y}^{(k)}$ is approaching the vector $2[\frac{1}{3} \ 1 \ \frac{1}{3}]^T$, so that the dominant eigenvalue is 2 and the corresponding eigenvector is $[\frac{1}{3} \ 1 \ \frac{1}{3}]^T$, which conforms to the answer obtained in Example 1.4.

Example 1.11

Find the dominant eigenvalue of

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

Solution Starting with $\mathbf{x}^{(0)} = [1 \ 1 \ 1 \ 1]^T$, the iterations give the following:

Iteration k	1	2	3	4	5	6	7
Eigenvalue	-3	2.6667	3.3750	3.0741	3.2048	3.1636	3.1642
$\mathbf{x}_1^{(k)}$	1 0	-0.3750	-0.4074	-0.4578	-0.4549	-0.4621	-0.4621
$\mathbf{x}_2^{(k)}$	1 0.6667	0.6250	0.4815	0.4819	0.4624	0.4621	0.4621
$\mathbf{x}_3^{(k)}$	1 1	1	1	1	1	1	1
$\mathbf{x}_4^{(k)}$	1 0	0.3750	0.1852	0.2651	0.2293	0.2403	0.2401

This indicates that the dominant eigenvalue is approximately 3.16, with corresponding eigenvector $[-0.46 \ 0.46 \ 1 \ 0.24]^T$.

The power method is suitable for obtaining the dominant eigenvalue and corresponding eigenvector of a matrix \mathbf{A} having real distinct eigenvalues. The smallest eigenvalue,

provided it is non-zero, can be obtained by using the same method on the inverse matrix \mathbf{A}^{-1} when it exists. This follows since if $\mathbf{Ax} = \lambda x$ then $\mathbf{A}^{-1}\mathbf{x} = \lambda^{-1}\mathbf{x}$. To find the subdominant eigenvalue using this method the dominant eigenvalue must first be removed from the matrix using **deflation methods**. We shall illustrate such a method for symmetric matrices only.

Let \mathbf{A} be a symmetric matrix having real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$. Then, by result (b) of Section 1.4.7, it has n corresponding mutually orthogonal normalized eigenvectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_n$ such that

$$\hat{\mathbf{e}}_i^T \hat{\mathbf{e}}_j = \delta_{ij} \quad (i, j = 1, 2, \dots, n)$$

Let λ_1 be the dominant eigenvalue and consider the matrix

$$\mathbf{A}_1 = \mathbf{A} - \lambda_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1^T$$

which is such that

$$\mathbf{A}_1 \hat{\mathbf{e}}_1 = (\mathbf{A} - \lambda_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1^T) \hat{\mathbf{e}}_1 = \mathbf{A} \hat{\mathbf{e}}_1 - \lambda_1 \hat{\mathbf{e}}_1 (\hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_1) = \lambda_1 \hat{\mathbf{e}}_1 - \lambda_1 \hat{\mathbf{e}}_1 = 0$$

$$\mathbf{A}_1 \hat{\mathbf{e}}_2 = \mathbf{A} \hat{\mathbf{e}}_2 - \lambda_1 \hat{\mathbf{e}}_1 (\hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_2) = \lambda_2 \hat{\mathbf{e}}_2$$

$$\mathbf{A}_1 \hat{\mathbf{e}}_3 = \mathbf{A} \hat{\mathbf{e}}_3 - \lambda_1 \hat{\mathbf{e}}_1 (\hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_3) = \lambda_3 \hat{\mathbf{e}}_3$$

⋮

$$\mathbf{A}_1 \hat{\mathbf{e}}_n = \mathbf{A} \hat{\mathbf{e}}_n - \lambda_1 \hat{\mathbf{e}}_1 (\hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_n) = \lambda_n \hat{\mathbf{e}}_n$$

Thus the matrix \mathbf{A}_1 has the same eigenvalues and eigenvectors as the matrix \mathbf{A} , except that the eigenvalue corresponding to λ_1 is now zero. The power method can then be applied to the matrix \mathbf{A}_1 to obtain the subdominant eigenvalue λ_2 and its corresponding eigenvector \mathbf{e}_2 . By repeated use of this technique, we can determine all the eigenvalues and corresponding eigenvectors of \mathbf{A} .

Example 1.12

Given that the symmetric matrix

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

has a dominant eigenvalue $\lambda_1 = 6$ with corresponding normalized eigenvector $\hat{\mathbf{e}}_1 = [1 \ 2 \ 0]^T / \sqrt{5}$ find the subdominant eigenvalue λ_2 and corresponding eigenvector $\hat{\mathbf{e}}_2$.

Solution Following the above procedure,

$$\mathbf{A}_1 = \mathbf{A} - \lambda_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1^T$$

$$= \begin{bmatrix} 2 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix} - \frac{6}{5} \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \end{bmatrix} = \begin{bmatrix} \frac{4}{5} & -\frac{2}{5} & 0 \\ -\frac{2}{5} & \frac{1}{5} & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Applying the power method procedure (1.14), with $\mathbf{x}^{(0)} = [1 \quad 1 \quad 1]^T$, gives

$$\mathbf{y}^{(1)} = \mathbf{A}_1 \mathbf{x}^{(0)} = \begin{bmatrix} \frac{2}{5} \\ -\frac{1}{5} \\ 3 \end{bmatrix} = 3 \begin{bmatrix} \frac{2}{15} \\ -\frac{1}{15} \\ 1 \end{bmatrix}; \quad \lambda_2^{(1)} = 3$$

$$\mathbf{x}^{(1)} = \begin{bmatrix} \frac{2}{15} \\ -\frac{1}{15} \\ 1 \end{bmatrix} = \begin{bmatrix} 0.133 \\ -0.133 \\ 1 \end{bmatrix}$$

$$\mathbf{y}^{(2)} = \mathbf{A}_1 \mathbf{x}^{(1)} = \begin{bmatrix} \frac{2}{15} \\ -\frac{1}{15} \\ 3 \end{bmatrix} = 3 \begin{bmatrix} \frac{2}{45} \\ -\frac{2}{45} \\ 1 \end{bmatrix}; \quad \lambda_2^{(2)} = 3$$

$$\mathbf{x}^{(2)} = \begin{bmatrix} \frac{2}{45} \\ -\frac{2}{45} \\ 1 \end{bmatrix} = \begin{bmatrix} 0.044 \\ -0.044 \\ 1 \end{bmatrix}$$

$$\mathbf{y}^{(3)} = \mathbf{A}_1 \mathbf{x}^{(2)} = \begin{bmatrix} \frac{2}{45} \\ -\frac{2}{45} \\ 3 \end{bmatrix} = 3 \begin{bmatrix} \frac{2}{135} \\ -\frac{2}{135} \\ 1 \end{bmatrix}; \quad \lambda_2^{(3)} = 3$$

$$\mathbf{x}^{(3)} = \begin{bmatrix} 0.015 \\ -0.015 \\ 1 \end{bmatrix}$$

Clearly the subdominant eigenvalue of \mathbf{A} is $\lambda_2 = 3$, and a few more iterations confirm the corresponding normalized eigenvector as $\hat{\mathbf{e}}_2 = [0 \quad 0 \quad 1]^T$. This is confirmed by the solution of Example 1.9. Note that the third eigenvalue may then be obtained using Property 1.1 of Section 1.4.6, since

$$\text{tr}(\mathbf{A}) = 10 = \lambda_1 + \lambda_2 + \lambda_3 = 6 + 3 + \lambda_3$$

giving $\lambda_3 = 1$. Alternatively, λ_3 and $\hat{\mathbf{e}}_3$ can be obtained by applying the power method to the matrix $\mathbf{A}_2 = \mathbf{A}_1 - \lambda_2 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_2^T$.



Although it is good as an illustration of the principles underlying iterative methods for evaluating eigenvalues and eigenvectors, the power method is of little practical importance, except possibly when dealing with large sparse matrices. In order to evaluate all the eigenvalues and eigenvectors of a matrix, including those with repeated eigenvalues, more sophisticated methods are required. Many of the numerical methods available, such as the **Jacobi** and **Householder methods**, are only applicable to symmetric matrices, and involve reducing the matrix to a special form so that its eigenvalues can be readily calculated. Analogous methods for non-symmetric matrices

are the **LR** and **QR methods**. It is methods such as these, together with others based on the inverse iterative method, that form the basis of the algorithms that exist in modern software packages such as MATLAB. Such methods will not be pursued further here, and the interested reader is referred to specialist texts on numerical analysis.

1.5.2 Exercises

- 14 Use the power method to estimate the dominant eigenvalue and its corresponding eigenvector for the matrix

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

Stop when you consider the eigenvalue estimate is correct to 2 decimal places.

- 15 Repeat Exercise 14 for the matrices

$$(a) \mathbf{A} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \quad (b) \mathbf{A} = \begin{bmatrix} 3 & 0 & 1 \\ 2 & 2 & 2 \\ 4 & 2 & 5 \end{bmatrix}$$

$$(c) \mathbf{A} = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

- 16 The symmetric matrix

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

has dominant eigenvector $\mathbf{e}_1 = [1 \ 1 \ 2]^T$.

Obtain the matrix

$$\mathbf{A}_1 = \mathbf{A} - \lambda_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1^T$$

where λ_1 is the eigenvalue corresponding to the eigenvector \mathbf{e}_1 . Using the deflation method, obtain the subdominant eigenvalue λ_2 and corresponding eigenvector \mathbf{e}_2 correct to 2 decimal places, taking $[1 \ 1 \ 1]^T$ as a first approximation to \mathbf{e}_2 . Continue the process to obtain the third eigenvalue λ_3 and its corresponding eigenvector \mathbf{e}_3 .

- 17 Show that the characteristic equation of the matrix

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

is

$$f(\lambda) = \lambda^3 - 15\lambda^2 + 51\lambda - 17 = 0$$

Using the Newton–Raphson iterative procedure

$$\lambda_{n+1} = \lambda_n - \frac{f(\lambda_n)}{f'(\lambda_n)}$$

with a suitable initial value in the interval $9 < \lambda < 11$, determine the eigenvalue in this interval correct to 3 decimal places.

Using Properties 1.1 and 1.2 of Section 1.4.6, determine the other two eigenvalues of \mathbf{A} to the same approximation.

- 18 (a) If the eigenvalues of the $n \times n$ matrix \mathbf{A} are

$$\lambda_1 > \lambda_2 > \lambda_3 \dots \lambda_n \geq 0$$

show that the eigenvalue λ_n can be found by applying the power method to the matrix $k\mathbf{I} - \mathbf{A}$, where \mathbf{I} is the identity matrix and $k \geq \lambda_1$.

- (b) Show that the eigenvalues of the matrix

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

satisfy the inequality

$$0 \leq \lambda \leq 4$$

Hence, using the result proved in (a), determine the smallest modulus eigenvalue of \mathbf{A} correct to 2 decimal places.

1.6

Reduction to canonical form

In this section we examine the process of reduction of a matrix to **canonical form**. Specifically, we examine methods by which certain square matrices can be reduced or transformed into diagonal form. The process of transformation can be thought of as a change of system coordinates, with the new coordinate axes chosen in such a way that the system can be expressed in a simple form. The simplification may, for example, be a transformation to principal axes or a decoupling of system equations.

We will see that not all matrices can be reduced to diagonal form. In some cases we can only achieve the so-called Jordan canonical form, but many of the advantages of the diagonal form can be extended to this case as well.

The transformation to diagonal form is just one example of a **similarity transform**. Other such transforms exist, but, in common with the transformation to diagonal form, their purpose is usually that of simplifying the system model in some way.

1.6.1 Reduction to diagonal form

For an $n \times n$ matrix \mathbf{A} possessing a full set of n linearly independent eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ we can write down a **modal matrix** \mathbf{M} having the n eigenvectors as its columns:

$$\mathbf{M} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \mathbf{e}_3 \quad \dots \quad \mathbf{e}_n]$$

The diagonal matrix having the eigenvalues of \mathbf{A} as its diagonal elements is called the **spectral matrix** corresponding to the modal matrix \mathbf{M} of \mathbf{A} , often denoted by Λ . That is,

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ & & \ddots \\ 0 & & \lambda_n \end{bmatrix}$$

with the (ij) th element being given by $\lambda_i \delta_{ij}$, where δ_{ij} is the Kronecker delta and $i, j = 1, 2, \dots, n$. It is important that the pair of matrices \mathbf{M} and Λ are written down correctly. If the i th column of \mathbf{M} is the eigenvector \mathbf{e}_i then the element in the (i, i) position in Λ must be λ_i , the eigenvalue corresponding to the eigenvector \mathbf{e}_i .



We saw in Section 1.4.2 that in MATLAB the command

`[M, S] = eig(A)`

generates the modal and spectral matrices for the matrix \mathbf{A} . (Note: For convenience \mathbf{S} is used to represent Λ when using MATLAB; whilst both are produced by the command `Eigenvalues(A)` in MAPLE.)

Example 1.13

Obtain a modal matrix and the corresponding spectral matrix for the matrix \mathbf{A} of Example 1.4.

Solution

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

having eigenvalues $\lambda_1 = 2$, $\lambda_2 = 1$ and $\lambda_3 = -1$, with corresponding eigenvectors

$$\mathbf{e}_1 = [1 \ 3 \ 1]^T, \quad \mathbf{e}_2 = [3 \ 2 \ 1]^T, \quad \mathbf{e}_3 = [1 \ 0 \ 1]^T$$

Choosing as modal matrix $\mathbf{M} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3]^T$ gives

$$\mathbf{M} = \begin{bmatrix} 1 & 3 & 1 \\ 3 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

The corresponding spectral matrix is

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

Returning to the general case, if we premultiply the matrix \mathbf{M} by \mathbf{A} , we obtain

$$\begin{aligned} \mathbf{AM} &= \mathbf{A}[\mathbf{e}_1 \ \mathbf{e}_2 \ \dots \ \mathbf{e}_n] = [\mathbf{A}\mathbf{e}_1 \ \mathbf{A}\mathbf{e}_2 \ \dots \ \mathbf{A}\mathbf{e}_n] \\ &= [\lambda_1\mathbf{e}_1 \ \lambda_2\mathbf{e}_2 \ \dots \ \lambda_n\mathbf{e}_n] \quad (\text{by definition}) \end{aligned}$$

so that

$$\mathbf{AM} = \mathbf{MA} \tag{1.15}$$

Since the n eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$ are linearly independent, the matrix \mathbf{M} is non-singular, and so \mathbf{M}^{-1} exists. Thus premultiplying by \mathbf{M}^{-1} gives

$$\mathbf{M}^{-1}\mathbf{AM} = \mathbf{M}^{-1}\mathbf{MA} = \mathbf{A} \tag{1.16}$$

indicating that the similarity transformation $\mathbf{M}^{-1}\mathbf{AM}$ reduces the matrix \mathbf{A} to the **diagonal or canonical form Λ** . Thus a matrix \mathbf{A} possessing a full set of linearly independent eigenvectors is reducible to diagonal form, and the reduction process is often referred to as the **diagonalization** of the matrix \mathbf{A} . Since

$$\mathbf{A} = \mathbf{M}\Lambda\mathbf{M}^{-1} \tag{1.17}$$

it follows that \mathbf{A} is uniquely determined once the eigenvalues and corresponding eigenvectors are known. Note that knowledge of the eigenvalues and eigenvectors alone is not sufficient: in order to structure \mathbf{M} and Λ correctly, the association of eigenvalues and the *corresponding* eigenvectors must also be known.

Example 1.14

Verify results (1.16) and (1.17) for the matrix \mathbf{A} of Example 1.13.

Solution Since

$$\mathbf{M} = \begin{bmatrix} 1 & 3 & 1 \\ 3 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad \text{we have} \quad \mathbf{M}^{-1} = \frac{1}{6} \begin{bmatrix} -2 & 2 & 2 \\ 3 & 0 & -3 \\ -1 & -2 & 7 \end{bmatrix}$$

Taking

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

matrix multiplication confirms the results

$$\mathbf{M}^{-1} \mathbf{A} \mathbf{M} = \mathbf{A}, \quad \mathbf{A} = \mathbf{M} \mathbf{A} \mathbf{M}^{-1}$$

For an $n \times n$ symmetric matrix \mathbf{A} it follows, from result (b) of Section 1.4.7, that to the n real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ there correspond n linearly independent normalized eigenvectors $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_n$ that are mutually orthogonal so that

$$\hat{\mathbf{e}}_i^T \hat{\mathbf{e}}_j = \delta_{ij} \quad (i, j = 1, 2, \dots, n)$$

The corresponding modal matrix

$$\hat{\mathbf{M}} = [\hat{\mathbf{e}}_1 \quad \hat{\mathbf{e}}_2 \quad \dots \quad \hat{\mathbf{e}}_n]$$

is then such that

$$\begin{aligned} \hat{\mathbf{M}}^T \hat{\mathbf{M}} &= \begin{bmatrix} \hat{\mathbf{e}}_1^T \\ \hat{\mathbf{e}}_2^T \\ \vdots \\ \hat{\mathbf{e}}_n^T \end{bmatrix} \begin{bmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 & \dots & \hat{\mathbf{e}}_n \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_2 & \dots & \hat{\mathbf{e}}_1^T \hat{\mathbf{e}}_n \\ \hat{\mathbf{e}}_2^T \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2^T \hat{\mathbf{e}}_2 & \dots & \hat{\mathbf{e}}_2^T \hat{\mathbf{e}}_n \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{e}}_n^T \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_n^T \hat{\mathbf{e}}_2 & \dots & \hat{\mathbf{e}}_n^T \hat{\mathbf{e}}_n \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \mathbf{I} \end{aligned}$$

That is, $\hat{\mathbf{M}}^T \hat{\mathbf{M}} = \mathbf{I}$ and so $\hat{\mathbf{M}}^T = \mathbf{M}^{-1}$. Thus $\hat{\mathbf{M}}$ is an **orthogonal matrix** (the term **orthonormal matrix** would be more appropriate, but the nomenclature is long established).

It follows from (1.16) that a symmetric matrix \mathbf{A} can be reduced to diagonal form \mathbf{A} using the orthogonal transformation

$$\hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} = \mathbf{A} \quad (1.18)$$

Example 1.15

For the symmetric matrix \mathbf{A} considered in Example 1.9 write down the corresponding orthogonal modal matrix $\hat{\mathbf{M}}$ and show that $\hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} = \mathbf{A}$ where \mathbf{A} is the spectral matrix.

Solution From Example 1.9 the eigenvalues are $\lambda_1 = 6$, $\lambda_2 = 3$ and $\lambda_3 = 1$, with corresponding normalized eigenvectors

$$\hat{\mathbf{e}}_1 = [1 \ 2 \ 0]^T / \sqrt{5}, \quad \hat{\mathbf{e}}_2 = [0 \ 0 \ 1]^T, \quad \hat{\mathbf{e}}_3 = [-2 \ 1 \ 0]^T / \sqrt{5}$$

The corresponding modal matrix is

$$\hat{\mathbf{M}} = \begin{bmatrix} \sqrt{\frac{1}{5}} & 0 & -2\sqrt{\frac{1}{5}} \\ 2\sqrt{\frac{1}{5}} & 0 & \sqrt{\frac{1}{5}} \\ 0 & 1 & 0 \end{bmatrix}$$

and, by matrix multiplication,

$$\hat{\mathbf{M}}^T \hat{\mathbf{M}} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{A}$$

1.6.2 The Jordan canonical form

If an $n \times n$ matrix \mathbf{A} does not possess a full set of linearly independent eigenvectors then it cannot be reduced to diagonal form using the similarity transformation $\mathbf{M}^{-1} \mathbf{A} \mathbf{M}$. In such a case, however, it is possible to reduce \mathbf{A} to a **Jordan canonical form** (or **Jordan normal form**), making use of ‘generalized’ eigenvectors.

As indicated in (1.11), if a matrix \mathbf{A} has an eigenvalue λ_i of algebraic multiplicity m_i and geometric multiplicity q_i , with $1 \leq q_i \leq m_i$, then there are q_i linearly independent eigenvectors corresponding to λ_i . Consequently, we need to generate $m_i - q_i$ generalized eigenvectors in order to produce a full set. To obtain these, we first obtain the q_i linearly independent eigenvectors by solving

$$(\mathbf{A} - \lambda_i \mathbf{I}) \mathbf{e}_i = \mathbf{0}$$

Then for each of these vectors we try to construct a generalized eigenvector \mathbf{e}_i^* such that

$$(\mathbf{A} - \lambda_i \mathbf{I}) \mathbf{e}_i^* = \mathbf{e}_i$$

If the resulting vector \mathbf{e}_i^* is linearly independent of all the eigenvectors (and generalized eigenvectors) already found then it is a valid additional generalized eigenvector. If further generalized eigenvectors corresponding to λ_i are needed, we then repeat the process using

$$(\mathbf{A} - \lambda_i \mathbf{I}) \mathbf{e}_i^{**} = \mathbf{e}_i^*$$

and so on until sufficient vectors are found.

Example 1.16

Obtain a generalized eigenvector corresponding to the eigenvalue $\lambda = 2$ of Example 1.7.

Solution For

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

we found in Example 1.7 that corresponding to the eigenvalue $\lambda_i = 2$ there was only one linearly independent eigenvector

$$\mathbf{e}_1 = [2 \quad 1 \quad 0]^T$$

and we need to find a generalized eigenvector to produce a full set. To obtain the generalized eigenvector \mathbf{e}_1^* , we solve

$$(\mathbf{A} - 2\mathbf{I})\mathbf{e}_1^* = \mathbf{e}_1$$

that is, we solve

$$\begin{bmatrix} -1 & 2 & 2 \\ 0 & 0 & 1 \\ -1 & 2 & 0 \end{bmatrix} \begin{bmatrix} e_{11}^* \\ e_{12}^* \\ e_{13}^* \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

At once, we have $e_{13}^* = 1$ and $e_{11}^* = 2e_{12}^*$, and so

$$\mathbf{e}_1^* = [2 \quad 1 \quad 1]^T$$

Thus, by including generalized eigenvectors, we have a full set of eigenvectors for the matrix \mathbf{A} given by

$$\mathbf{e}_1 = [2 \quad 1 \quad 0]^T, \quad \mathbf{e}_2 = [2 \quad 1 \quad 1]^T, \quad \mathbf{e}_3 = [1 \quad -1 \quad 1]^T$$

If we include such generalized eigenvectors, it is always possible to obtain for an $n \times n$ matrix \mathbf{A} a modal matrix \mathbf{M} with n linearly independent columns $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n$. Corresponding to (1.15), we have

$$\mathbf{AM} = \mathbf{MJ}$$

where \mathbf{J} is called the **Jordan normal form** of \mathbf{A} . Premultiplying by \mathbf{M}^{-1} then gives

$$\mathbf{M}^{-1}\mathbf{AM} = \mathbf{J} \tag{1.19}$$

The process of reducing \mathbf{A} to \mathbf{J} is known as the **reduction** of \mathbf{A} to its Jordan normal, or canonical, form. This is named after Marie Ennemond Camille Jordan (1838–1922) who was particularly known for his work on analysis and group theory.

If \mathbf{A} has p distinct eigenvalues then the matrix \mathbf{J} is of the block-diagonal form

$$\mathbf{J} = [\mathbf{J}_1 \quad \mathbf{J}_2 \quad \cdots \quad \mathbf{J}_p]$$

where each submatrix \mathbf{J}_i ($i = 1, 2, \dots, p$) is associated with the corresponding eigenvalue λ_i . The submatrix \mathbf{J}_i will have λ_i as its leading diagonal elements, with zeros elsewhere except on the diagonal above the leading diagonal. On this diagonal the entries will have the value 1 or 0, depending on the number of generalized eigenvectors used and how they

were generated. To illustrate this, suppose that \mathbf{A} is a 7×7 matrix with eigenvalues $\lambda_1 = 1$, $\lambda_2 = 2$ (occurring twice), $\lambda_3 = 3$ (occurring four times), and suppose that the number of linearly independent eigenvectors generated in each case is

$$\lambda_1 = 1, \quad 1 \text{ eigenvector}$$

$$\lambda_2 = 2, \quad 1 \text{ eigenvector}$$

$$\lambda_3 = 3, \quad 2 \text{ eigenvectors}$$

with one further generalized eigenvector having been determined for $\lambda_2 = 2$ and two more for $\lambda_3 = 3$.

Corresponding to $\lambda_1 = 1$, the Jordan block \mathbf{J}_1 will be just [1], while that corresponding to $\lambda_2 = 2$ will be

$$\mathbf{J}_2 = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

Corresponding to $\lambda_3 = 3$, the Jordan block \mathbf{J}_3 can take one of the two forms

$$\mathbf{J}_{3,1} = \begin{bmatrix} \lambda_3 & 1 & 0 & \cdots & 0 \\ 0 & \lambda_3 & 1 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_3 \end{bmatrix} \quad \text{or} \quad \mathbf{J}_{3,2} = \begin{bmatrix} \lambda_3 & 1 & & 0 & 0 \\ 0 & \lambda_3 & & 0 & 0 \\ 0 & 0 & \lambda_3 & 1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_3 \end{bmatrix}$$

depending on how the generalized eigenvectors are generated. Corresponding to $\lambda_3 = 3$, we had two linearly independent eigenvectors $\mathbf{e}_{3,1}$ and $\mathbf{e}_{3,2}$. If both generalized eigenvectors are generated from *one* of these vectors then \mathbf{J}_3 will take the form $\mathbf{J}_{3,1}$, whereas if one generalized eigenvector has been generated from each eigenvector then \mathbf{J}_3 will take the form $\mathbf{J}_{3,2}$.

Example 1.17

Obtain the Jordan canonical form of the matrix \mathbf{A} of Example 1.16, and show that $\mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \mathbf{J}$ where \mathbf{M} is a modal matrix that includes generalized eigenvectors.

Solution

For

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

from Example 1.16 we know that the eigenvalues of \mathbf{A} are $\lambda_1 = 2$ (twice) and $\lambda_3 = 1$. The eigenvector corresponding to $\lambda_3 = 1$ has been determined as $\mathbf{e}_3 = [1 \ 1 \ -1]^T$ in Example 1.7 and corresponding to $\lambda_1 = 2$ we found one linearly independent eigenvector $\mathbf{e}_1 = [2 \ 1 \ 0]^T$ and a generalized eigenvector $\mathbf{e}_1^* = [2 \ 1 \ 1]^T$. Thus the modal matrix including this generalized eigenvector is

$$\mathbf{M} = \begin{bmatrix} 2 & 2 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & -1 \end{bmatrix}$$

and the corresponding Jordan canonical form is

$$\mathbf{J} = \begin{bmatrix} 2 & 1 & : & 0 \\ 0 & 2 & : & 0 \\ \cdots & \cdots & \ddots & \cdots \\ 0 & 0 & : & 1 \end{bmatrix}$$

To check this result, we compute \mathbf{M}^{-1} as

$$\mathbf{M}^{-1} = \begin{bmatrix} 2 & -3 & -1 \\ -1 & 2 & 1 \\ -1 & 2 & 0 \end{bmatrix}$$

and, forming $\mathbf{M}^{-1}\mathbf{A}\mathbf{M}$, we obtain \mathbf{J} as expected.



In MATLAB the command `J=jordan(A)` computes the Jordan form of \mathbf{A} ; including the case when \mathbf{J} is diagonal and all the eigenvectors of \mathbf{A} are linearly independent. The command

`[M, J] =jordan (A)`

also computes the similarity transformation or modal matrix \mathbf{M} that may include generalized eigenvectors.

Numerical calculation of the Jordan form is very sensitive to round-off errors and so on. This makes it very difficult to compute the Jordan form reliably and almost any change in \mathbf{A} causes it to be diagonal.

For the matrix \mathbf{A} in Example 1.17 the sequence of commands

```
A=[1 2 2; 0 2 1; -1 2 2];
[M, J]=jordan (A)
```

returns

```
-1 -2 2
M=-1 -1 1
1 0 -1
1 0 0
J= 0 2 1
0 0 2
```

which is equally acceptable to the solution given in Example 1.17. (This can be checked by evaluating $\mathbf{M}^{-1}\mathbf{A}\mathbf{M}$.)

Using the Symbolic Math Toolbox in MATLAB the sequence of commands

```
A=[1 2 2; 0 2 1; -1 2 2];
AS=sym (A)
[M, J]=jordan (AS)
```

returns the same output as above. In practice, this sequence of commands is only really effective when the elements of the matrix \mathbf{A} are integers or ratios of small integers.

1.6.3 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 19 Show that the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 6 & -12 \\ 0 & -13 & 30 \\ 0 & -9 & 20 \end{bmatrix}$$

are 5, 2 and -1 . Obtain the corresponding eigenvectors. Write down the modal matrix \mathbf{M} and spectral matrix \mathbf{A} . Evaluate \mathbf{M}^{-1} and show that $\mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \mathbf{A}$.

- 20 Using the eigenvalues and corresponding eigenvectors of the symmetric matrix

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

obtained in Example 1.9, verify that $\hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} = \mathbf{A}$ where $\hat{\mathbf{M}}$ and \mathbf{A} are respectively a normalized modal matrix and a spectral matrix of \mathbf{A} .

- 21 Given

$$\mathbf{A} = \begin{bmatrix} 5 & 10 & 8 \\ 10 & 2 & -2 \\ 8 & -2 & 11 \end{bmatrix}$$

find its eigenvalues and corresponding eigenvectors. Normalize the eigenvectors and write down the corresponding normalized modal matrix $\hat{\mathbf{M}}$. Write down $\hat{\mathbf{M}}^T$ and show that $\hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} = \mathbf{A}$, where \mathbf{A} is the spectral matrix of \mathbf{A} .

- 22 Determine the eigenvalues and corresponding eigenvectors of the matrix

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

Write down the modal matrix \mathbf{M} and spectral matrix \mathbf{A} . Confirm that $\mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \mathbf{A}$ and that $\mathbf{A} = \mathbf{M}\mathbf{A}\mathbf{M}^{-1}$.

- 23

Determine the eigenvalues and corresponding eigenvectors of the symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 3 & -2 & 4 \\ -2 & -2 & 6 \\ 4 & 6 & -1 \end{bmatrix}$$

Verify that the eigenvectors are orthogonal, and write down an orthogonal matrix \mathbf{L} such that $\mathbf{L}^T \mathbf{A} \mathbf{L} = \mathbf{A}$, where \mathbf{A} is the spectral matrix of \mathbf{A} .

- 24

A 3×3 symmetric matrix \mathbf{A} has eigenvalues 6, 3 and 1. The eigenvectors corresponding to the eigenvalues 6 and 1 are $[1 \ 2 \ 0]^T$ and $[-2 \ 1 \ 0]^T$ respectively. Find the eigenvector corresponding to the eigenvalue 3, and hence determine the matrix \mathbf{A} .

- 25

Given that $\lambda = 1$ is a three times-repeated eigenvalue of the matrix

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

use the nullity, given by (1.11), of a suitable matrix to show that there is only one corresponding linearly independent eigenvector. Obtain two further generalized eigenvectors, and write down the corresponding modal matrix \mathbf{M} . Confirm that $\mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \mathbf{J}$, where \mathbf{J} is the appropriate Jordan matrix.

- 26

Show that the eigenvalues of the matrix

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

are $-2, -2, 4$ and 4 . Using the nullity, given by (1.11), of appropriate matrices, show that there are two linearly independent eigenvectors corresponding to the repeated eigenvalue -2 and only one corresponding to the repeated eigenvalue 4 . Obtain a further generalized eigenvector corresponding to the eigenvalue 4 . Write down the Jordan canonical form of \mathbf{A} .

1.6.4 Quadratic forms

A **quadratic form** in n independent variables x_1, x_2, \dots, x_n is a homogeneous second-degree polynomial of the form

$$\begin{aligned} V(x_1, x_2, \dots, x_n) &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + \dots + a_{1n}x_1x_n \\ &\quad + a_{21}x_2x_1 + a_{22}x_2^2 + \dots + a_{2n}x_2x_n \\ &\quad \vdots \\ &\quad + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \dots + a_{nn}x_n^2 \end{aligned} \tag{1.20}$$

Defining the vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$ and the matrix

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

the quadratic form (1.20) may be written in the form

$$V(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} \tag{1.21}$$

The matrix \mathbf{A} is referred to as the matrix of the quadratic form and the determinant of \mathbf{A} is called the **discriminant** of the quadratic form.

Now a_{ij} and a_{ji} in (1.20) are both coefficients of the term $x_i x_j$ ($i \neq j$), so that for $i \neq j$ the coefficient of the term $x_i x_j$ is $a_{ij} + a_{ji}$. By defining new coefficients a'_{ij} and a'_{ji} for $x_i x_j$ and $x_j x_i$ respectively, such that $a'_{ij} = a'_{ji} = \frac{1}{2}(a_{ij} + a_{ji})$, the matrix \mathbf{A} associated with the quadratic form $V(\mathbf{x})$ may be taken to be symmetric. Thus for real quadratic forms we can, without loss of generality, consider the matrix \mathbf{A} to be a symmetric matrix.

Example 1.18

Find the real symmetric matrix corresponding to the quadratic form

$$V(x_1, x_2, x_3) = x_1^2 + 3x_2^2 - 4x_3^2 - 3x_1x_2 + 2x_1x_3 - 5x_2x_3$$

Solution If $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$, then by comparing the coefficients of (1.20) and the above expression, we find that

$$V(x_1, x_2, x_3) = [x_1 \ x_2 \ x_3] \begin{bmatrix} 1 & -\frac{3}{2} & \frac{2}{2} \\ -\frac{3}{2} & 3 & -\frac{5}{2} \\ \frac{2}{2} & -\frac{5}{2} & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

where the matrix of the quadratic form is

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

In Section 1.6.1 we saw that a real symmetric matrix \mathbf{A} can always be reduced to the diagonal form

$$\hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} = \Lambda$$

where K is the normalized orthogonal modal matrix of \mathbf{A} and Λ is its spectral matrix. Thus for a real quadratic form we can specify a change of variables

$$\mathbf{x} = \hat{\mathbf{M}}\mathbf{y}$$

where $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_n]^T$, such that

$$V = \mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{y}^T \hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} \mathbf{y} = \mathbf{y}^T \Lambda \mathbf{y}$$

giving

$$V = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_n y_n^2 \quad (1.22)$$

Hence the quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ may be reduced to the sum of squares by the transformation $\mathbf{x} = \hat{\mathbf{M}}\mathbf{y}$, where $\hat{\mathbf{M}}$ is the normalized modal matrix of \mathbf{A} . The resulting form given in (1.22) is called the **canonical form** of the quadratic form V given in (1.21). The reduction of a quadratic form to its canonical form has many applications in engineering, particularly in stress analysis.

Example 1.19

Find the canonical form of the quadratic form

$$V = 2x_1^2 + 5x_2^2 + 3x_3^2 + 4x_1x_2$$

Can V take negative values for any values of x_1 , x_2 and x_3 ?

Solution At once, we have

$$V = \mathbf{x}^T \begin{bmatrix} 2 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix} \mathbf{x} = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

where

$$\mathbf{x} = [x_1 \ x_2 \ x_3]^T, \quad \mathbf{A} = \begin{bmatrix} 2 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

The real symmetric matrix \mathbf{A} is the matrix of Example 1.15, where we found the normalized orthogonal modal matrix K and spectral matrix Λ to be

$$\hat{\mathbf{M}} = \begin{bmatrix} \sqrt{\frac{1}{5}} & 0 & -2\sqrt{\frac{1}{5}} \\ 2\sqrt{\frac{1}{5}} & 0 & \sqrt{\frac{1}{5}} \\ 0 & 1 & 0 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

such that $\hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} = \mathbf{A}$. Thus, setting $\mathbf{x} = \hat{\mathbf{M}}\mathbf{y}$, we obtain

$$V = \mathbf{y}^T \hat{\mathbf{M}}^T \mathbf{A} \hat{\mathbf{M}} \mathbf{y} = \mathbf{y}^T \begin{bmatrix} 6 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{y} = 6y_1^2 + 3y_2^2 + y_3^2$$

as the required canonical form.

Clearly V is non-negative for all y_1, y_2 and y_3 . Since $\mathbf{x} = \hat{\mathbf{M}}\mathbf{y}$ and $\hat{\mathbf{M}}$ is an orthogonal matrix it follows that $\mathbf{y} = \hat{\mathbf{M}}^T \mathbf{x}$, so for all \mathbf{x} there is a corresponding \mathbf{y} . It follows that V cannot take negative values for any values of x_1, x_2 and x_3 .

The quadratic form of Example 1.19 was seen to be non-negative for any vector \mathbf{x} , and is positive provided that $\mathbf{x} \neq \mathbf{0}$. Such a quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ is called a **positive-definite** quadratic form, and, by reducing to canonical form, we have seen that this property depends only on the eigenvalues of the real symmetric matrix \mathbf{A} . This leads us to classify quadratic forms $V = \mathbf{x}^T \mathbf{A} \mathbf{x}$, where $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$ in the following manner.

- (a) V is **positive-definite** (that is $V > 0$ for all vectors \mathbf{x} except $\mathbf{x} = 0$) if and only if all the eigenvalues of \mathbf{A} are positive.
- (b) V is **positive-semidefinite** (that is $V \geq 0$ for all vectors \mathbf{x} and $V = 0$ for at least one vector $\mathbf{x} \neq 0$) if and only if all the eigenvalues of \mathbf{A} are non-negative and at least one of the eigenvalues is zero.
- (c) V is **negative-definite** if $-V$ is positive-definite, with a corresponding condition on the eigenvalues of $-\mathbf{A}$.
- (d) V is **negative-semidefinite** if $-V$ is positive-semidefinite, with a corresponding condition on the eigenvalues of $-\mathbf{A}$.
- (e) V is **indefinite** (that is V takes at least one positive value and at least one negative value) if and only if the matrix \mathbf{A} has both positive and negative eigenvalues.

Since the classification of a real quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ depends entirely on the location of the eigenvalues of the symmetric matrix \mathbf{A} , it may be viewed as a property of \mathbf{A} itself. For this reason, it is common to talk of positive-definite, positive-semidefinite, and so on, symmetric matrices without reference to the underlying quadratic form.

Example 1.20

Classify the following quadratic forms:

- (a) $3x_1^2 + 2x_2^2 + 3x_3^2 - 2x_1x_2 - 2x_2x_3$
- (b) $7x_1^2 + x_2^2 + x_3^2 - 4x_1x_2 - 4x_1x_3 + 8x_2x_3$
- (c) $-3x_1^2 - 5x_2^2 - 3x_3^2 + 2x_1x_2 + 2x_2x_3 - 2x_1x_3$
- (d) $4x_1^2 + x_2^2 + 15x_3^2 - 4x_1x_2$

Solution (a) The matrix corresponding to the quadratic form is

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

The eigenvalues of \mathbf{A} are 4, 3 and 1, so the quadratic form is positive-definite.

(b) The matrix corresponding to the quadratic form is

$$\mathbf{A} = \begin{bmatrix} 7 & -2 & -2 \\ -2 & 1 & 4 \\ -2 & 4 & 1 \end{bmatrix}$$

The eigenvalues of \mathbf{A} are 9, 3 and -3 , so the quadratic form is indefinite.

(c) The matrix corresponding to the quadratic form is

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

The eigenvalues of \mathbf{A} are -6 , -3 and -2 , so the quadratic form is negative-definite.

(d) The matrix corresponding to the quadratic form is

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 15 \end{bmatrix}$$

The eigenvalues of \mathbf{A} are 15, 5 and 0, so the quadratic form is positive-semidefinite.

In Example 1.20 classifying the quadratic forms involved determining the eigenvalues of \mathbf{A} . If \mathbf{A} contains one or more parameters then the task becomes difficult, if not impossible, even with the use of a symbolic algebra computer package. Frequently in engineering, particularly in stability analysis, it is necessary to determine the range of values of a parameter k , say, for which a quadratic form remains definite or at least semi-definite in sign. J. J. Sylvester determined criteria for the classification of quadratic forms (or the associated real symmetric matrix) that do not require the computation of the eigenvalues. These criteria are known as **Sylvester's conditions**, which we shall briefly discuss without proof.

In order to classify the quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$ Sylvester's conditions involve consideration of the principal minors of \mathbf{A} . A **principal minor** P_i of order i ($i = 1, 2, \dots, n$) of an $n \times n$ square matrix \mathbf{A} is the determinant of the submatrix, of order i , whose principal diagonal is part of the principal diagonal of \mathbf{A} . Note that when $i = n$ the principal minor is $\det \mathbf{A}$. In particular, the **leading principal minors** of \mathbf{A} are

$$D_1 = |a_{11}|, \quad D_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad D_3 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \quad \dots, \quad D_n = \det \mathbf{A}$$

Example 1.21

Determine all the principal minors of the matrix

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

and indicate which are the leading principal minors.

Solution

- (a) The principal minor of order three is

$$P_3 = \det \mathbf{A} = 5(2 - k^2) \quad (\text{leading principal minor } D_3)$$

- (b) The principal minors of order two are

- (i) deleting row 1 and column 1,

$$P_{21} = \begin{vmatrix} 2 & 0 \\ 0 & 5 \end{vmatrix} = 10$$

- (ii) deleting row 2 and column 2,

$$P_{22} = \begin{vmatrix} 1 & 0 \\ 0 & 5 \end{vmatrix} = 5$$

- (iii) deleting row 3 and column 3,

$$P_{23} = \begin{vmatrix} 1 & k \\ k & 2 \end{vmatrix} = 2 - k^2 \quad (\text{leading principal minor } D_2)$$

- (c) The principal minors of order one are

- (i) deleting rows 1 and 2 and columns 1 and 2,

$$P_{11} = |5| = 5$$

- (ii) deleting rows 1 and 3 and columns 1 and 3,

$$P_{12} = |2| = 2$$

- (iii) deleting rows 2 and 3 and columns 2 and 3,

$$P_{13} = |1| = 1 \quad (\text{leading principal minor } D_1)$$

Sylvester's conditions: These state that the quadratic form $\mathbf{x}^T \mathbf{A} \mathbf{x}$, where \mathbf{A} is an $n \times n$ real symmetric matrix, is

- (a) **positive-definite** if and only if all the leading principal minors of \mathbf{A} are positive; that is, $D_i > 0$ ($i = 1, 2, \dots, n$);
- (b) **negative-definite** if and only if the leading principal minors of \mathbf{A} alternate in sign with $a_{11} < 0$; that is, $(-1)^i D_i > 0$ ($i = 1, 2, \dots, n$);
- (c) **positive-semidefinite** if and only if $\det \mathbf{A} = 0$ and *all* the principal minors of \mathbf{A} are non-negative; that is, $\det \mathbf{A} = 0$ and $P_i \geq 0$ for *all* principal minors;
- (d) **negative-semidefinite** if and only if $\det \mathbf{A} = 0$ and $(-1)^i P_i \geq 0$ for *all* principal minors.

Example 1.22 For what values of k is the matrix \mathbf{A} of Example 1.21 positive-definite?

Solution We need for all leading principal minors of \mathbf{A} to be positive. These are

$$D_1 = 1, \quad D_2 = 2 - k^2, \quad D_3 = 5(2 - k^2)$$

These will be positive provided that $2 - k^2 > 0$, so the matrix will be positive-definite if $k^2 < 2$, that is $-\sqrt{2} < k < \sqrt{2}$.

Example 1.23 Using Sylvester's conditions, confirm the conclusions of Example 1.20.

Solution (a) The matrix of the quadratic form is

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

and its leading principal minors are

$$3, \quad \begin{vmatrix} 3 & -1 \\ -1 & 2 \end{vmatrix} = 5, \quad \det \mathbf{A} = 12$$

Thus, by Sylvester's condition (a), the quadratic form is positive-definite.

(b) The matrix of the quadratic form is

$$\mathbf{A} = \begin{bmatrix} 7 & -2 & -2 \\ -2 & 1 & 4 \\ -2 & 4 & 1 \end{bmatrix}$$

and its leading principal minors are

$$7, \quad \begin{vmatrix} 7 & -2 \\ -2 & 1 \end{vmatrix} = 3, \quad \det \mathbf{A} = -81$$

Thus none of Sylvester's conditions can be satisfied, and the quadratic form is indefinite.

(c) The matrix of the quadratic form is

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

and its leading principal minors are

$$-3, \quad \begin{vmatrix} -3 & 1 \\ 1 & -5 \end{vmatrix} = 14, \quad \det \mathbf{A} = -36$$

Thus, by Sylvester's condition (b), the quadratic form is negative-definite.

- (d) The matrix of the quadratic form is

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 15 \end{bmatrix}$$

and its leading principal minors are

$$4, \begin{vmatrix} 4 & -2 \\ -2 & 1 \end{vmatrix} = 0, \quad \det \mathbf{A} = 0$$

We therefore need to evaluate all the principal minors to see if the quadratic form is positive-semidefinite. The principal minors are

$$4, 1, 15, \begin{vmatrix} 4 & -2 \\ -2 & 1 \end{vmatrix} = 0, \begin{vmatrix} 1 & 0 \\ 0 & 15 \end{vmatrix} = 15, \begin{vmatrix} 4 & 0 \\ 0 & 15 \end{vmatrix} = 60, \quad \det \mathbf{A} = 0$$

Thus, by Sylvester's condition (c), the quadratic form is positive-semidefinite.

1.6.5 Exercises

- 27 Reduce the quadratic form

$$2x_1^2 + 5x_2^2 + 2x_3^2 + 4x_2x_3 + 2x_3x_1 + 4x_1x_2$$

to the sum of squares by an orthogonal transformation.

- 28 Classify the quadratic forms

- (a) $x_1^2 + 2x_2^2 + 7x_3^2 - 2x_1x_2 + 4x_1x_3 - 2x_2x_3$
- (b) $x_1^2 + 2x_2^2 + 5x_3^2 - 2x_1x_2 + 4x_1x_3 - 2x_2x_3$
- (c) $x_1^2 + 2x_2^2 + 4x_3^2 - 2x_1x_2 + 4x_1x_3 - 2x_2x_3$

- 29 (a) Show that $ax_1^2 - 2bx_1x_2 + cx_2^2$ is positive-definite if and only if $a > 0$ and $ac > b^2$.
 (b) Find inequalities that must be satisfied by a and b to ensure that $2x_1^2 + ax_2^2 + 3x_3^2 - 2x_1x_2 + 2bx_2x_3$ is positive-definite.

- 30 Evaluate the definiteness of the matrix

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

- (a) by obtaining the eigenvalues;
- (b) by evaluating the principal minors.

- 31 Determine the exact range of k for which the quadratic form

$$Q(x, y, z) = k(x^2 + y^2) + 2xy + z^2 + 2xz - 2yz$$

is positive-definite in x, y and z . What can be said about the definiteness of Q when $k = 2$?

- 32 Determine the minimum value of the constant a such that the quadratic form

$$\mathbf{x}^T \begin{bmatrix} 3+a & 1 & 1 \\ 1 & a & 2 \\ 1 & 2 & a \end{bmatrix} \mathbf{x}$$

where $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$, is positive-definite.

- 33 Express the quadratic form

$$Q = x_1^2 + 4x_1x_2 - 4x_1x_3 - 6x_2x_3 + \lambda(x_2^2 + x_3^2)$$

in the form $\mathbf{x}^T \mathbf{A} \mathbf{x}$, where $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$ and \mathbf{A} is a symmetric matrix. Hence determine the range of values of λ for which Q is positive-definite.

1.7 Functions of a matrix

Let \mathbf{A} be an $n \times n$ constant square matrix, so that

$$\mathbf{A}^2 = \mathbf{AA}, \mathbf{A}^3 = \mathbf{AA}^2 = \mathbf{A}^2\mathbf{A}, \quad \text{and so on}$$

are all defined. We can then define a function $f(\mathbf{A})$ of the matrix \mathbf{A} using a power series representation. For example,

$$f(\mathbf{A}) = \sum_{r=0}^p \beta_r \mathbf{A}^r = \beta_0 \mathbf{I} + \beta_1 \mathbf{A} + \cdots + \beta_p \mathbf{A}^p \quad (1.23)$$

where we have interpreted \mathbf{A}^0 as the $n \times n$ identity matrix \mathbf{I} .

Example 1.24

Given the 2×2 square matrix

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

determine $f(\mathbf{A}) = \sum_{r=0}^2 \beta_r \mathbf{A}^r$ when $\beta_0 = 1$, $\beta_1 = -1$ and $\beta_2 = 3$.

Solution

Now

$$\begin{aligned} f(\mathbf{A}) &= \beta_0 \mathbf{I} + \beta_1 \mathbf{A} + \beta_2 \mathbf{A}^2 = 1 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 1 \begin{bmatrix} 1 & -1 \\ 2 & 3 \end{bmatrix} + 3 \begin{bmatrix} -1 & -4 \\ 8 & 7 \end{bmatrix} \\ &= \begin{bmatrix} -3 & -11 \\ 22 & 19 \end{bmatrix} \end{aligned}$$

Note that \mathbf{A} is a 2×2 matrix and $f(\mathbf{A})$ is another 2×2 matrix.

Suppose that in (1.23) we let $p \rightarrow \infty$, so that

$$f(\mathbf{A}) = \sum_{r=0}^{\infty} \beta_r \mathbf{A}^r$$

We can attach meaning to $f(\mathbf{A})$ in this case if the matrices

$$f_p(\mathbf{A}) = \sum_{r=0}^p \beta_r \mathbf{A}^r$$

tend to a constant $n \times n$ matrix in the limit as $p \rightarrow \infty$.

Example 1.25

For the matrix

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

using a computer and larger and larger values of p , we infer that

$$f(\mathbf{A}) = \lim_{p \rightarrow \infty} \sum_{r=0}^p \frac{\mathbf{A}^r}{r!} \approx \begin{bmatrix} 2.71828 & 0 \\ 0 & 2.71828 \end{bmatrix}$$

indicating that

$$f(\mathbf{A}) = \begin{bmatrix} e & 0 \\ 0 & e \end{bmatrix}$$

What would be the corresponding results if

$$(a) \quad \mathbf{A} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad (b) \quad \mathbf{A} = \begin{bmatrix} -t & 0 \\ 0 & t \end{bmatrix}?$$

Solution (a) The computer will lead to the prediction

$$f(\mathbf{A}) \approx \begin{bmatrix} (2.71828)^{-1} & 0 \\ 0 & 2.71828 \end{bmatrix}$$

indicating that

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

(b) The computer is of little help in this case. However, hand calculation shows that we are generating the matrix

$$f(\mathbf{A}) = \begin{bmatrix} 1 - t + \frac{1}{2}t^2 - \frac{1}{6}t^3 + \dots & 0 \\ 0 & 1 + t + \frac{1}{2}t^2 + \frac{1}{6}t^3 + \dots \end{bmatrix}$$

indicating that

$$f(\mathbf{A}) = \begin{bmatrix} e^{-t} & 0 \\ 0 & e^t \end{bmatrix}$$

By analogy with the definition of the scalar exponential function

$$e^{at} = 1 + at + \frac{a^2 t^2}{2!} + \dots + \frac{a^r t^r}{r!} + \dots = \sum_{r=0}^{\infty} \frac{(at)^r}{r!}$$

it is natural to define the matrix function $e^{\mathbf{A}t}$, where t is a scalar parameter, by the power series

$$f(\mathbf{A}) = \sum_{r=0}^{\infty} \frac{\mathbf{A}^r}{r!} t^r \quad (1.24)$$

In fact the matrix in part (b) of Example 1.25 illustrates that this definition is reasonable.

In Example 1.25 we were able to spot the construction of the matrix $f(\mathbf{A})$, but this will not be the case when \mathbf{A} is a general $n \times n$ square matrix. In order to overcome this limitation and generate a method that will not rely on our ability to ‘spot’ a closed form of the limiting matrix, we make use of the Cayley–Hamilton theorem, which may be stated as follows.

Theorem 1.1

Cayley–Hamilton theorem

A square matrix \mathbf{A} satisfies its own characteristic equation; that is, if

$$\lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_1\lambda + c_0 = 0$$

is the characteristic equation of an $n \times n$ matrix \mathbf{A} then

$$\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + c_{n-2}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{A} + c_0\mathbf{I} = \mathbf{0} \quad (1.25)$$

where \mathbf{I} is the $n \times n$ identity matrix.

[end of theorem](#)

The proof of this theorem is not trivial, and is not included here. We shall illustrate the theorem using a simple example.

[The interested reader may consult the original proof in G. Frobenius. Über Lineare Substitutionen und Bilineare Formen. J. für die Reine U. Angew. Math., 84:1–63, 1878.]

Example 1.26

Verify the Cayley–Hamilton theorem for the matrix

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

Solution

The characteristic equation of \mathbf{A} is

$$\begin{vmatrix} 3 - \lambda & 4 \\ 1 & 2 - \lambda \end{vmatrix} = 0 \quad \text{or} \quad \lambda^2 - 5\lambda + 2 = 0$$

Since

$$\mathbf{A}^2 = \begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 13 & 20 \\ 5 & 8 \end{bmatrix}$$

we have

$$\mathbf{A}^2 - 5\mathbf{A} + 2\mathbf{I} = \begin{bmatrix} 13 & 20 \\ 5 & 8 \end{bmatrix} - 5 \begin{bmatrix} 3 & 4 \\ 1 & 2 \end{bmatrix} + 2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{0}$$

thus verifying the validity of the Cayley–Hamilton theorem for this matrix.

In the particular case when \mathbf{A} is a 2×2 matrix with characteristic equation

$$c(\lambda) = \lambda^2 + a_1\lambda + a_2 = 0 \quad (1.26)$$

it follows from the Cayley–Hamilton theorem that

$$c(\mathbf{A}) = \mathbf{A}^2 + a_1\mathbf{A} + a_2\mathbf{I} = \mathbf{0}$$

The significance of this result for our present purposes begins to appear when we rearrange to give

$$\mathbf{A}^2 = -a_1\mathbf{A} - a_2\mathbf{I}$$

This means that \mathbf{A}^2 can be written in terms of \mathbf{A} and $\mathbf{A}^0 = \mathbf{I}$. Moreover, multiplying by \mathbf{A} gives

$$\mathbf{A}^3 = -a_1\mathbf{A}^2 - a_2\mathbf{A} = -a_1(-a_1\mathbf{A} - a_2\mathbf{I}) - a_2\mathbf{A}$$

Thus \mathbf{A}^3 can also be expressed in terms of \mathbf{A} and $\mathbf{A}^0 = \mathbf{I}$; that is, in terms of powers of \mathbf{A} less than $n = 2$, the order of the matrix \mathbf{A} in this case. It is clear that we could continue the process of multiplying by \mathbf{A} and substituting \mathbf{A}^2 for as long as we could manage the algebra. However, we can quickly convince ourselves that for any integer $r \geq n$

$$\mathbf{A}^r = \alpha_0\mathbf{I} + \alpha_1\mathbf{A} \quad (1.27)$$

where α_0 and α_1 are constants whose values will depend on r .

This is a key result deduced from the Cayley–Hamilton theorem, and the determination of the α_i ($i = 0, 1$) is not as difficult as it might appear. To see how to perform the calculations, we use the characteristic equation of \mathbf{A} itself. If we assume that the eigenvalues λ_1 and λ_2 of \mathbf{A} are distinct then it follows from (1.26) that

$$c(\lambda_i) = \lambda_i^2 + a_1\lambda_i + a_2 = 0 \quad \text{for } i = 1, 2$$

Thus we can write

$$\lambda_i^2 = -a_1\lambda_i - a_2$$

in which a_1 and a_2 are the same constants as in (1.26). Then, for $i = 1, 2$,

$$\lambda_i^3 = -a_1\lambda_i^2 - a_2\lambda_i = -a_1(-a_1\lambda_i - a_2) - a_2\lambda_i$$

Proceeding in this way, we deduce that for each of the eigenvalues λ_1 and λ_2 we can write

$$\lambda_i^r = \alpha_0 + \alpha_1\lambda_i$$

with the same α_0 and α_1 as in (1.27). This therefore provides us with a procedure for the calculation of \mathbf{A}^r when $r \geq n$ (the order of the matrix) is an integer.

Example 1.27

Given that the matrix

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

has eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -2$ calculate \mathbf{A}^5 and \mathbf{A}^r , where r is an integer greater than 2.

Solution Since \mathbf{A} is a 2×2 square matrix, it follows from (1.27) that

$$\mathbf{A}^5 = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A}$$

and for each eigenvalue λ_i ($i = 1, 2$) α_0 and α_1 satisfy

$$\lambda_i^5 = \alpha_0 + \alpha_1 \lambda_i$$

Substituting $\lambda_1 = -1$ and $\lambda_2 = -2$ leads to the following pair of simultaneous equations:

$$(-1)^5 = \alpha_0 + \alpha_1(-1), \quad (-2)^5 = \alpha_0 + \alpha_1(-2)$$

which can be solved for α_0 and α_1 to give

$$\alpha_0 = 2(-1)^5 - (-2)^5, \quad \alpha_1 = (-1)^5 - (-2)^5$$

Then

$$\begin{aligned} \mathbf{A}^5 &= [2(-1)^5 - (-2)^5] \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + [(-1)^5 - (-2)^5] \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \\ &= \begin{bmatrix} 2(-1)^5 - (-2)^5 & (-1)^5 - (-2)^5 \\ (-2)((-1)^5 - (-2)^5) & 2(-2)^5 - (-1)^5 \end{bmatrix} = \begin{bmatrix} 30 & 31 \\ -62 & -63 \end{bmatrix} \end{aligned}$$

Replacing the exponent 5 by the general value r , the algebra is identical, and it is easy to see that

$$\mathbf{A}^r = \begin{bmatrix} 2(-1)^r - (-2)^r & (-1)^r - (-2)^r \\ -2((-1)^r - (-2)^r) & 2(-2)^r - (-1)^r \end{bmatrix}$$

To evaluate α_0 and α_1 in (1.24), we assumed that the matrix \mathbf{A} had distinct eigenvalues λ_1 and λ_2 , leading to a pair of simultaneous equations for α_0 and α_1 . What happens if the 2×2 matrix \mathbf{A} has a repeated eigenvalue so that $\lambda_1 = \lambda_2 = \lambda$, say? We shall apparently have just a single equation to determine the two constants α_0 and α_1 . However, we can obtain a second equation by differentiating with respect to λ , as illustrated in Example 1.28.

Example 1.28

Given that the matrix

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

has eigenvalues $\lambda_1 = \lambda_2 = -1$, determine \mathbf{A}^r , where r is an integer greater than 2.

Solution Since \mathbf{A} is a 2×2 matrix, it follows from (1.27) that

$$\mathbf{A}^r = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A}$$

with α_0 and α_1 satisfying

$$\lambda^r = \alpha_0 + \alpha_1 \lambda \tag{1.28}$$

Since in this case we have only one value of λ , namely $\lambda = -1$, we differentiate (1.28) with respect to λ , to obtain

$$r\lambda^{r-1} = \alpha_1 \quad (1.29)$$

Substituting $\lambda = -1$ in (1.28) and (1.29) leads to

$$\alpha_1 = (-1)^{r-1}r, \quad \alpha_0 = (-1)^r + \alpha_1 = (1-r)(-1)^r$$

giving

$$\begin{aligned} \mathbf{A}^r &= (1-r)(-1)^r \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - r(-1)^r \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} \\ &= \begin{bmatrix} (1-r)(-1)^r & -r(-1)^r \\ r(-1)^r & (1+r)(-1)^r \end{bmatrix} \end{aligned}$$

Having found a straightforward way of expressing any positive integer power of the 2×2 square matrix \mathbf{A} we see that the same process could be used for each of the terms in (1.23) for $r \geq 2$. Thus, for a 2×2 matrix \mathbf{A} and some α_0 and α_1 ,

$$f(\mathbf{A}) = \sum_{r=0}^p \beta_r \mathbf{A}^r = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A}$$

If, as $p \rightarrow \infty$,

$$f(\mathbf{A}) = \lim_{p \rightarrow \infty} \sum_{r=0}^p \beta_r \mathbf{A}^r$$

exists, that is, it is a 2×2 matrix with finite entries independent of p , then we may write

$$f(\mathbf{A}) = \sum_{r=0}^{\infty} \beta_r \mathbf{A}^r = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A} \quad (1.30)$$

We are now in a position to check the results of our computer experiment with the matrix

$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ of Example 1.25. We have defined

$$f(\mathbf{A}) = e^{\mathbf{A}t} = \sum_{r=0}^{\infty} \frac{\mathbf{A}^r}{r!} t^r$$

so we can write

$$e^{\mathbf{A}t} = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A}$$

Since \mathbf{A} has repeated eigenvalue $\lambda = 1$, we adopt the method of Example 1.28 to give

$$e^t = \alpha_0 + \alpha_1, \quad t e^t = \alpha_1$$

leading to

$$\alpha_1 = t e^t, \quad \alpha_0 = (1-t)e^t$$

Thus

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

Setting $t = 1$ confirms our inference in Example 1.25.

Example 1.29

Calculate $e^{\mathbf{A}t}$ and $\sin \mathbf{A}t$ when

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

Solution Again \mathbf{A} has repeated eigenvalues, with $\lambda_1 = \lambda_2 = 1$. Thus for $e^{\mathbf{A}t}$ we have

$$e^{\mathbf{A}t} = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A}$$

with

$$e^t = \alpha_0 + \alpha_1, \quad t e^t = \alpha_1$$

leading to

$$e^{\mathbf{A}t} = \begin{bmatrix} e^t & -t e^t \\ 0 & e^t \end{bmatrix}$$

Similarly,

$$\sin \mathbf{A}t = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A}$$

with

$$\sin t = \alpha_0 + \alpha_1, \quad t \cos t = \alpha_1$$

leading to

$$\sin \mathbf{A}t = \begin{bmatrix} \sin t & -t \cos t \\ 0 & \sin t \end{bmatrix}$$

Although we have worked so far with 2×2 matrices, nothing in our development restricts us to this case. The Cayley–Hamilton theorem allows us to express positive integer powers of any $n \times n$ square matrix \mathbf{A} in terms of powers of \mathbf{A} up to $n - 1$. That is, if \mathbf{A} is an $n \times n$ matrix and $p \geq n$ then

$$\mathbf{A}^p = \sum_{r=0}^{n-1} \beta_r \mathbf{A}^r = \beta_0 \mathbf{I} + \beta_1 \mathbf{A} + \cdots + \beta_{n-1} \mathbf{A}^{n-1}$$

From this we can deduce that for an $n \times n$ matrix \mathbf{A} we may write

$$f(\mathbf{A}) = \sum_{r=0}^{\infty} \beta_r \mathbf{A}^r$$

as

$$f(\mathbf{A}) = \sum_{r=0}^{n-1} \alpha_r \mathbf{A}^r \quad (1.31a)$$

which generalizes the result (1.30). Again the coefficients $\alpha_0, \alpha_1, \dots, \alpha_{n-1}$ are obtained by solving the n equations

$$f(\lambda_i) = \sum_{r=0}^{n-1} \alpha_r \lambda_i^r \quad (i = 1, 2, \dots, n) \quad (1.31b)$$

where $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of \mathbf{A} . If \mathbf{A} has repeated eigenvalues, we differentiate as before, noting that if λ_i is an eigenvalue of multiplicity m then the first $m - 1$ derivatives

$$\frac{d^k}{d\lambda_i^k} f(\lambda_i) = \frac{d^k}{d\lambda_i^k} \sum_{r=0}^{n-1} \alpha_r \lambda_i^r \quad (k = 1, 2, \dots, m-1)$$

are also satisfied by λ_i .

Sometimes it is advantageous to use an alternative approach to evaluate

$$f(\mathbf{A}) = \sum_{r=0}^p \beta_r \mathbf{A}^r$$

If \mathbf{A} possesses n linearly independent eigenvectors then there exists a modal matrix \mathbf{M} and spectral matrix \mathbf{A} such that

$$\mathbf{M}^{-1} \mathbf{A} \mathbf{M} = \mathbf{A} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$$

Now

$$\begin{aligned} \mathbf{M}^{-1} f(\mathbf{A}) \mathbf{M} &= \sum_{r=0}^p \beta_r (\mathbf{M}^{-1} \mathbf{A}^r \mathbf{M}) = \sum_{r=0}^p \beta_r (\mathbf{M}^{-1} \mathbf{A} \mathbf{M})^r \\ &= \sum_{r=0}^p \beta_r \mathbf{A} = \sum_{r=0}^p \beta_r \text{diag}(\lambda_1^r, \lambda_2^r, \dots, \lambda_n^r) \\ &= \text{diag} \left(\sum_{r=0}^p \beta_r \lambda_1^r, \sum_{r=0}^p \beta_r \lambda_2^r, \dots, \sum_{r=0}^p \beta_r \lambda_n^r \right) \\ &= \text{diag}(f(\lambda_1), f(\lambda_2), \dots, f(\lambda_n)) \end{aligned}$$

This gives us a second method of computing functions of a square matrix, since we see that

$$f(\mathbf{A}) = \mathbf{M} \text{diag}(f(\lambda_1), f(\lambda_2), \dots, f(\lambda_n)) \mathbf{M}^{-1} \quad (1.32)$$

Example 1.30

Using the result (1.32), calculate \mathbf{A}^k for the matrix

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

of Example 1.27, i.e. show that $\mathbf{A}^k = \mathbf{M}\text{diag}(\lambda_1^k, \lambda_2^k)\mathbf{M}^{-1}$.

Solution \mathbf{A} has eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -2$ with corresponding eigenvectors

$$\mathbf{e}_1 = [1 \quad -1]^T, \quad \mathbf{e}_2 = [1 \quad -2]^T$$

Thus a modal matrix \mathbf{M} and corresponding spectral matrix Λ are

$$\mathbf{M} = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$$

Clearly

$$\mathbf{M}^{-1} = \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix}$$

Taking $f(\mathbf{A}) = \mathbf{A}^k$, we have

$$\text{diag}(f(-1), f(-2)) = \text{diag}((-1)^k, (-2)^k)$$

Thus, from (1.32),

$$f(\mathbf{A}) = \mathbf{M} \begin{bmatrix} (-1)^k & 0 \\ 0 & (-2)^k \end{bmatrix} \mathbf{M}^{-1} = \begin{bmatrix} 2(-1)^k - (-2)^k & (-1)^k - (-2)^k \\ 2((-2)^k - (-1)^k) & 2(-2)^k - (-1)^k \end{bmatrix}$$

as determined in Example 1.27.

Example 1.30 demonstrates a second approach to the calculation of a function of a matrix. There is little difference in the labour associated with each method, so perhaps the only comment we should make is that each approach gives a different perspective on the construction of the matrix function either from powers of the matrix itself or from its spectral and modal matrices.

Later in this chapter we need to make use of some properties of the exponential matrix $e^{\mathbf{A}t}$, where \mathbf{A} is a constant $n \times n$ square matrix. These are now briefly discussed.

(i) Considering the power series definition given in (1.24)

$$e^{\mathbf{A}t} = \mathbf{I} + \mathbf{A}t + \frac{1}{2!} \mathbf{A}^2 t^2 + \frac{1}{3!} \mathbf{A}^3 t^3 + \dots$$

term-by-term differentiation gives

$$\frac{d}{dt} e^{\mathbf{A}t} = \mathbf{A} + \frac{2}{2!} \mathbf{A}^2 t + \frac{3}{3!} \mathbf{A}^3 t^2 + \dots = \mathbf{A}[\mathbf{I} + \mathbf{A}t + \frac{1}{2!} \mathbf{A}^2 t^2 + \dots]$$

so that

$$\frac{d}{dt} (e^{\mathbf{A}t}) = \mathbf{A} e^{\mathbf{A}t} = e^{\mathbf{A}t} \mathbf{A} \tag{1.33}$$

(ii) Likewise, term-by-term integration of the power series gives

$$\begin{aligned}\int_0^t e^{\mathbf{A}\tau} d\tau &= \mathbf{I} \int_0^t d\tau + \mathbf{A} \int_0^t \tau d\tau + \frac{1}{2!} \mathbf{A}^2 \int_0^t \tau^2 d\tau + \dots \\ &= \mathbf{I}t + \frac{1}{2!} \mathbf{A}t^2 + \frac{1}{3!} \mathbf{A}^2 t^3 + \dots\end{aligned}$$

so that

$$\mathbf{A} \int_0^t e^{\mathbf{A}\tau} d\tau + \mathbf{I} = e^{\mathbf{A}t}$$

giving

$$\int_0^t e^{\mathbf{A}\tau} d\tau = \mathbf{A}^{-1} [e^{\mathbf{A}t} - \mathbf{I}] = [e^{\mathbf{A}t} - \mathbf{I}] \mathbf{A}^{-1} \quad (1.34)$$

provided the inverse exists.

$$(iii) \quad e^{\mathbf{A}(t_1+t_2)} = e^{\mathbf{A}t_1} e^{\mathbf{A}t_2} \quad (1.35)$$

Although this property is true in general we shall illustrate its validity for the particular case when \mathbf{A} has n linearly independent eigenvectors. Then, from (1.32),

$$e^{\mathbf{A}t_1} = \mathbf{M} \text{diag}(e^{\lambda_1 t_1}, e^{\lambda_2 t_1}, \dots, e^{\lambda_n t_1}) \mathbf{M}^{-1}$$

$$e^{\mathbf{A}t_2} = \mathbf{M} \text{diag}(e^{\lambda_1 t_2}, e^{\lambda_2 t_2}, \dots, e^{\lambda_n t_2}) \mathbf{M}^{-1}$$

so that

$$e^{\mathbf{A}t_1} e^{\mathbf{A}t_2} = \mathbf{M} \text{diag}(e^{\lambda_1(t_1+t_2)}, e^{\lambda_2(t_1+t_2)}, \dots, e^{\lambda_n(t_1+t_2)}) \mathbf{M}^{-1} = e^{\mathbf{A}(t_1+t_2)}$$

(iv) It is important to note that in general

$$e^{\mathbf{A}t} e^{\mathbf{B}t} \neq e^{(\mathbf{A}+\mathbf{B})t}$$

It follows from the power series definition that

$$e^{\mathbf{A}t} e^{\mathbf{B}t} = e^{(\mathbf{A}+\mathbf{B})t} \quad (1.36)$$

if and only if the matrices \mathbf{A} and \mathbf{B} commute; that is, if $\mathbf{AB} = \mathbf{BA}$.

To conclude this section we consider the derivative and integral of a matrix $\mathbf{A}(t) = [a_{ij}(t)]$, whose elements $a_{ij}(t)$ are functions of t . The derivative and integral of $\mathbf{A}(t)$ are defined respectively by

$$\frac{d}{dt} \mathbf{A}(t) = \left[\frac{d}{dt} a_{ij}(t) \right] \quad (1.37a)$$

and

$$\int \mathbf{A}(t) dt = \left[\int a_{ij}(t) dt \right] \quad (1.37b)$$

that is, each element of the matrix is differentiated or integrated as appropriate.

Example 1.31 Evaluate $d\mathbf{A}/dt$ and $\int \mathbf{A} dt$ for the matrix

$$\begin{bmatrix} t^2 + 1 & t - 3 \\ 2 & t^2 + 2t - 1 \end{bmatrix}$$

Solution Using (1.37a),

$$\frac{d\mathbf{A}}{dt} = \begin{bmatrix} \frac{d}{dt}(t^2 + 1) & \frac{d}{dt}(t - 3) \\ \frac{d}{dt}(2) & \frac{d}{dt}(t^2 + 2t - 1) \end{bmatrix} = \begin{bmatrix} 2t & 1 \\ 0 & 2t + 2 \end{bmatrix}$$

Using (1.37b),

$$\begin{aligned} \int \mathbf{A} dt &= \begin{bmatrix} \int (t^2 + 1) dt & \int (t - 3) dt \\ \int 2 dt & \int (t^2 + 2t - 1) dt \end{bmatrix} = \begin{bmatrix} \frac{1}{3}t^3 + t + c_{11} & \frac{1}{2}t^2 - 3t + c_{12} \\ 2t + c_{21} & \frac{1}{3}t^3 + t^2 - t + c_{22} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{3}t^3 + t & \frac{1}{2}t^2 - 3t \\ 2t & \frac{1}{3}t^3 + t^2 - t \end{bmatrix} + \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{3}t^3 + t & \frac{1}{2}t^2 - 3t \\ 2t & \frac{1}{3}t^3 + t^2 - t \end{bmatrix} + \mathbf{C} \end{aligned}$$

where \mathbf{C} is a constant matrix.



Using the Symbolic Math Toolbox in MATLAB the derivative and integral of the matrix $\mathbf{A}(t)$ is generated using the commands `diff(A)` and `int(A)` respectively. To illustrate this confirm that the derivative of the matrix $\mathbf{A}(t)$ of Example 1.31 is generated using the sequence of commands

```
syms t
A=[t^2+1 t-3; 2 t^2+2*t-1];
df=diff(A);
pretty(df)
```

and its integral by the additional commands

```
I=int(A);
pretty(I)
```

From the basic definitions, it follows that for constants α and β

$$\frac{d}{dt}(\alpha\mathbf{A} + \beta\mathbf{B}) = \alpha \frac{d\mathbf{A}}{dt} + \beta \frac{d\mathbf{B}}{dt} \quad (1.38)$$

$$\int (\alpha\mathbf{A} + \beta\mathbf{B}) dt = \alpha \int \mathbf{A} dt + \beta \int \mathbf{B} dt \quad (1.39)$$

$$\frac{d}{dt}(\mathbf{AB}) = \mathbf{A} \frac{d\mathbf{B}}{dt} + \frac{d\mathbf{A}}{dt} \mathbf{B} \quad (1.40)$$

Note in (1.40) that order is important, since in general $\mathbf{AB} \neq \mathbf{BA}$.

Note that in general

$$\frac{d}{dt} [\mathbf{A}(t)]^n \neq n \mathbf{A}^{n-1} \frac{d\mathbf{A}}{dt}$$

1.7.1 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

34 Show that the matrix

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

satisfies its own characteristic equation.

35 Given

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

use the Cayley–Hamilton theorem to evaluate

- (a) \mathbf{A}^2 (b) \mathbf{A}^3 (c) \mathbf{A}^4

36 The characteristic equation of an $n \times n$ matrix \mathbf{A} is

$$\lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_1\lambda + c_0 = 0$$

so, by the Cayley–Hamilton theorem,

$$\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + c_{n-2}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{A} + c_0\mathbf{I} = \mathbf{0}$$

If \mathbf{A} is non-singular then every eigenvalue is non-zero, so $c_0 \neq 0$ and

$$\mathbf{I} = -\frac{1}{c_0} (\mathbf{A}^n + c_{n-1}\mathbf{A}^{n-1} + \cdots + c_1\mathbf{A})$$

which on multiplying throughout by \mathbf{A}^{-1} gives

$$\mathbf{A}^{-1} = -\frac{1}{c_0} (\mathbf{A}^{n-1} + c_{n-1}\mathbf{A}^{n-2} + \cdots + c_1\mathbf{I}) \quad (1.41)$$

(a) Using (1.41) find the inverse of the matrix

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

(b) Show that the characteristic equation of the matrix

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

is

$$\lambda^3 - 3\lambda^2 - 7\lambda - 11 = 0$$

Evaluate \mathbf{A}^2 and, using (1.41), determine \mathbf{A}^{-1} .

37 Given

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

compute \mathbf{A}^2 and, using the Cayley–Hamilton theorem, compute

$$\mathbf{A}^7 - 3\mathbf{A}^6 + \mathbf{A}^4 + 3\mathbf{A}^3 - 2\mathbf{A}^2 + 3\mathbf{I}$$

38 Evaluate $e^{\mathbf{A}t}$ for

$$(a) \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad (b) \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$$

39 Given

$$\mathbf{A} = \frac{\pi}{2} \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

show that

$$\sin \mathbf{A} = \frac{4}{\pi} \mathbf{A} - \frac{4}{\pi^2} \mathbf{A}^2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

40 Given

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

evaluate

$$(a) \frac{d\mathbf{A}}{dt} \quad (b) \int_1^2 \mathbf{A} dt$$

41 Given

$$\mathbf{A} = \begin{bmatrix} t^2 + 1 & t - 1 \\ 5 & 0 \end{bmatrix}$$

evaluate \mathbf{A}^2 and show that

$$\frac{d}{dt} (\mathbf{A}^2) \neq 2\mathbf{A} \frac{d\mathbf{A}}{dt}$$

1.8 Singular value decomposition

So far we have been concerned mainly with square matrices, dealing in particular with the inverse matrix, the eigenvalue problem and reduction to canonical form. In this section we consider analogous results for **non-square** (or **rectangular**) matrices, all of which have important applications in engineering.

First we review some definitions associated with non-square matrices:

- (a) A non-square $m \times n$ matrix

$$\mathbf{A} = (a_{ij}), i = 1, 2, \dots, m; j = 1, 2, \dots, n$$

is said to be **diagonal** if all the i, j entries are zero except possibly for $i = j$. For example:

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \\ 0 & 0 \end{bmatrix} \quad \text{is a diagonal } 3 \times 2 \text{ matrix}$$

whilst

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \quad \text{is a diagonal } 2 \times 3 \text{ matrix}$$

- (b) The **row rank** of a $m \times n$ matrix \mathbf{A} denotes the maximum number of linearly independent rows of \mathbf{A} , whilst the **column rank** of \mathbf{A} denotes the maximum number of linearly independent columns of \mathbf{A} . It turns out that these are the same and referred to simply as the rank of the matrix \mathbf{A} and denoted by $r = \text{rank}(\mathbf{A})$. It follows that $r \leq \min(m, n)$. The matrix \mathbf{A} is said to be of **full rank** if $r = \min(m, n)$.

Example 1.32

For the 3×4 matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 7 & 10 \\ 2 & 1 & 3 & 5 \end{bmatrix}$$

confirm that row rank (\mathbf{A}) = column rank (\mathbf{A}).

Solution

Following the process outlined in Section 1.2.6 we reduce the matrix to row (column) echelon form using row (column) elementary operations.

- (a) *Row rank: using elementary row operations*

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 7 & 10 \\ 2 & 1 & 3 & 5 \end{bmatrix}$$

 row 2 $- 3 \times$ row 1, row 3 $- 2 \times$ row 1

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -2 & -2 & -2 \\ 0 & -3 & -3 & -3 \end{bmatrix}$$

\downarrow multiply row 2 by $-\frac{1}{2}$

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 1 & 1 \\ 0 & -3 & -3 & -3 \end{bmatrix}$$

\downarrow row 3 + 3 × row 2

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

which is in row echelon form and indicating that

$$\text{row rank } (\mathbf{A}) = 2$$

- (b) *Column rank:* using elementary column operations

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 7 & 10 \\ 2 & 1 & 3 & 5 \end{bmatrix}$$

\downarrow col2 - 2 × col1, col3 - 3 × col1, col4 - 4 × col1

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

\downarrow col3 - col2, col4 - col2

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

which is in column echelon form and indicating that

$$\text{column rank } (\mathbf{A}) = 2$$

confirming that

$$\text{rank}(\mathbf{A}) = \text{row rank } (\mathbf{A}) = \text{column rank } (\mathbf{A}) = 2$$

Note that in this case the matrix \mathbf{A} is not of full rank.

1.8.1 Singular values

For a $m \times n$ matrix \mathbf{A} the **transposed matrix** \mathbf{A}^T has dimension $n \times m$ so that the product \mathbf{AA}^T is a **square matrix** of dimension $m \times m$. This product is also a **symmetric matrix** since

$$(\mathbf{AA}^T)^T = (\mathbf{A}^T)^T(\mathbf{A}^T) = \mathbf{AA}^T$$

It follows from Section 1.4.7 that the $m \times m$ matrix \mathbf{AA}^T has a full set of m linearly independent eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_m$ that are mutually orthogonal, and which can be normalized to give the orthogonal normalized set (or **orthonormal** set) of eigenvectors

$$\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \dots, \hat{\mathbf{u}}_m$$

with $\hat{\mathbf{u}}_i^T \hat{\mathbf{u}}_j = \delta_{ij}$ ($i, j = 1, 2, \dots, m$), where δ_{ij} is the Kronecker delta defined in Section 1.3.2.

(*Reminder:* As indicated in Section 1.4.2 normalized eigenvectors are uniquely determined up to a scale factor of ± 1 .) We then define the $m \times m$ orthogonal matrix $\hat{\mathbf{U}}$ as a matrix having these normalized set of eigenvectors as its columns:

$$\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \dots, \hat{\mathbf{u}}_m] \quad (1.42)$$

with $\hat{\mathbf{U}}^T \hat{\mathbf{U}} = \hat{\mathbf{U}} \hat{\mathbf{U}}^T = \mathbf{I}_m$. Such a matrix is also called a **unitary matrix**.

Let $\lambda_1, \lambda_2, \dots, \lambda_m$ be the corresponding eigenvalues of \mathbf{AA}^T that is

$$(\mathbf{AA}^T)\hat{\mathbf{u}}_i = \lambda_i \hat{\mathbf{u}}_i, \quad i = 1, 2, \dots, m$$

Considering the square of the length, or norm, of the vector $\mathbf{A}\hat{\mathbf{u}}_i$ then from orthogonality

$$|\mathbf{A}\hat{\mathbf{u}}_i|^2 = (\mathbf{A}\hat{\mathbf{u}}_i)^T(\mathbf{A}\hat{\mathbf{u}}_i) = \hat{\mathbf{u}}_i^T(\mathbf{A}^T\mathbf{A}\hat{\mathbf{u}}_i) = \hat{\mathbf{u}}_i^T \lambda_i \hat{\mathbf{u}}_i = \lambda_i$$

(*Note:* The notation $\|\mathbf{A}\hat{\mathbf{u}}_i\|^2$ is also frequently used.) Since $|\mathbf{A}\hat{\mathbf{u}}_i|^2 > 0$ it follows that the eigenvalues λ_i ($i = 1, 2, \dots, m$) of the matrix \mathbf{AA}^T are **all non-negative** and so can be written in the form

$$\lambda_i = \sigma_i^2, \quad i = 1, 2, \dots, m$$

It is also assumed that they are arranged in a non-increasing order so that

$$\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_m^2 \geq 0$$

Some of these eigenvalues may be zero. The number of non-zero values (accounting for multiplicity) is equal to r the rank of the matrix \mathbf{A} . Thus, if $\text{rank}(\mathbf{A}) = r$ then the matrix \mathbf{AA}^T has eigenvalues

$$\sigma_1^2 \geq \sigma_2^2 \geq \dots \geq \sigma_r^2 > 0 \text{ with } \sigma_{r+1}^2 = \dots = \sigma_m^2 = 0$$

The positive square roots of the non-zero eigenvalues of the matrix \mathbf{AA}^T are called the **singular values** of the matrix \mathbf{A} and play a similar role in general matrix theory that eigenvalues play in the theory of square matrices. If the matrix \mathbf{A} has rank r then it has r singular values

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$$

In practice determining the singular values of a non-square matrix provides a means of determining the rank of the matrix.

Example 1.33

For the matrix

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

- (a) Determine the eigenvalues and corresponding eigenvectors of the matrix \mathbf{AA}^T .
- (b) Normalize the eigenvectors to obtain the corresponding orthogonal matrix $\hat{\mathbf{U}}$ and confirm that $\hat{\mathbf{U}}\hat{\mathbf{U}}^T = \mathbf{I}$.
- (c) What are the singular values of \mathbf{A} ?
- (d) What is the rank of \mathbf{A} ?

Solution (a)
$$\mathbf{AA}^T = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}$$

(Note that \mathbf{AA}^T is a symmetric matrix.)The eigenvalues of \mathbf{AA}^T are given by the solutions of the equation

$$|\mathbf{AA}^T - \lambda \mathbf{I}| = \begin{vmatrix} 10 - \lambda & 0 & 2 \\ 0 & 10 - \lambda & 4 \\ 2 & 4 & 2 - \lambda \end{vmatrix} = 0$$

which reduces to

$$(12 - \lambda)(10 - \lambda)\lambda = 0$$

giving the eigenvalues as

$$\lambda_1 = 12, \lambda_2 = 10, \lambda_3 = 0$$

Solving the homogeneous equations

$$(\mathbf{AA}^T - \lambda_i \mathbf{I})\mathbf{u}_i = 0$$

gives the corresponding eigenvectors as:

$$\mathbf{u}_1 = [1 \ 2 \ 1]^T, \quad \mathbf{u}_2 = [2 \ -1 \ 0]^T, \quad \mathbf{u}_3 = [1 \ 2 \ -5]^T$$

- (b) The corresponding normalized eigenvectors are:

$$\hat{\mathbf{u}}_1 = \left[\frac{1}{\sqrt{6}} \quad \frac{2}{\sqrt{6}} \quad \frac{1}{\sqrt{6}} \right]^T, \quad \hat{\mathbf{u}}_2 = \left[\frac{2}{\sqrt{5}} \quad \frac{-1}{\sqrt{5}} \quad 0 \right]^T, \quad \hat{\mathbf{u}}_3 = \left[\frac{1}{\sqrt{30}} \quad \frac{2}{\sqrt{30}} \quad \frac{-5}{\sqrt{30}} \right]^T$$

giving the corresponding orthogonal matrix

$$\hat{\mathbf{U}} = [\hat{\mathbf{u}}_1 \quad \hat{\mathbf{u}}_2 \quad \hat{\mathbf{u}}_3] = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} 0.04082 & 0.8944 & 0.1826 \\ 0.8165 & -0.4472 & 0.3651 \\ 0.4082 & 0.0000 & -0.9129 \end{bmatrix}$$

By direct multiplication

$$\hat{\mathbf{U}}\hat{\mathbf{U}}^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

confirming that $\hat{\mathbf{U}}\hat{\mathbf{U}}^T = \mathbf{I}$.

- (c) The singular values of \mathbf{A} are the square roots of the non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$. Thus the singular values of \mathbf{A} are $\sigma_1 = \sqrt{12}$ and $\sigma_2 = \sqrt{10}$.
- (d) The rank of \mathbf{A} is equal to the number of singular values, giving $\text{rank}(\mathbf{A}) = 2$. This can be confirmed by reducing \mathbf{A} to echelon form.

Likewise, for a $m \times n$ matrix \mathbf{A} the product $\mathbf{A}^T\mathbf{A}$ is a square $n \times n$ symmetric matrix, having a full set of n orthogonal normalized eigenvectors $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \dots, \hat{\mathbf{v}}_n$ which form the columns of the $n \times n$ orthogonal matrix $\hat{\mathbf{V}}$:

$$\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1 \ \hat{\mathbf{v}}_2 \ \dots \ \hat{\mathbf{v}}_n] \quad (1.43)$$

and having corresponding non-negative eigenvalues $\mu_1, \mu_2, \dots, \mu_n$ with

$$\mu_1 \geq \mu_2 \geq \dots \geq \mu_n \geq 0 \quad (1.44)$$

Again the number of non-zero eigenvalues equals r , the rank of \mathbf{A} , so that the product $\mathbf{A}^T\mathbf{A}$ has eigenvalues

$$\mu_1 \geq \mu_2 \geq \dots \geq \mu_r > 0 \text{ with } \mu_{r+1} = \dots = \mu_n = 0$$

Thus

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{v}}_i = \mu_i \hat{\mathbf{v}}_i, \quad \mu_i > 0 \quad (i = 1, 2, \dots, r) \quad (1.45)$$

Premultiplying by \mathbf{A} gives

$$(\mathbf{A}\mathbf{A}^T)(\mathbf{A}\hat{\mathbf{v}}_i) = \mu_i(\mathbf{A}\hat{\mathbf{v}}_i)$$

so that μ_i and $(\mathbf{A}\hat{\mathbf{v}}_i)$ are an eigenvalue and eigenvector pair of the matrix $\mathbf{A}\mathbf{A}^T$; indicating that the non-zero eigenvalues of the product $\mathbf{A}\mathbf{A}^T$ are the same as the non-zero eigenvalues of the product $\mathbf{A}^T\mathbf{A}$. Thus if \mathbf{A} is of rank r then the eigenvalues (1.44) of the product $\mathbf{A}^T\mathbf{A}$ may be written as

$$\mu_i = \begin{cases} \sigma_i^2, & i = 1, 2, \dots, r \\ 0, & i = r + 1, \dots, n \end{cases}$$

In general the vector $(\mathbf{A}\hat{\mathbf{v}}_i)$ is not a unit vector so

$$\mathbf{A}\hat{\mathbf{v}}_i = k\hat{\mathbf{u}}_i \quad (1.46)$$

and we need to show that $k = \sigma_i$. Taking the norm of $(\mathbf{A}\hat{\mathbf{v}}_i)$ gives

$$\begin{aligned} |\mathbf{A}\hat{\mathbf{v}}_i|^2 &= (\mathbf{A}\hat{\mathbf{v}}_i)^T(\mathbf{A}\hat{\mathbf{v}}_i) = \hat{\mathbf{v}}_i^T \mathbf{A}^T \mathbf{A} \hat{\mathbf{v}}_i \\ &= \hat{\mathbf{v}}_i^T \mu_i \hat{\mathbf{v}}_i \quad \text{from (1.45)} \\ &= \mu_i = \sigma_i^2 \end{aligned}$$

giving

$$|\mathbf{A}\hat{\mathbf{v}}_i| = k = \sigma_i$$

It follows from (1.46) that

$$\mathbf{A}\hat{\mathbf{v}}_i = \begin{cases} \sigma_i \hat{\mathbf{u}}_i, & i = 1, 2, \dots, r \\ 0, & i = r + 1, \dots, m \end{cases} \quad (1.47)$$

Clearly the singular values of \mathbf{A} may be determined by evaluating the eigenvalues of the product \mathbf{AA}^T or the product $\mathbf{A}^T\mathbf{A}$. The eigenvectors $\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \dots, \hat{\mathbf{u}}_m$ of the product \mathbf{AA}^T (that is the columns of $\hat{\mathbf{U}}$) are called the **left singular vectors** of \mathbf{A} and the eigenvectors $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \dots, \hat{\mathbf{v}}_n$ of the product $\mathbf{A}^T\mathbf{A}$ (that is columns of $\hat{\mathbf{V}}$) are called the **right singular vectors** of \mathbf{A} .

Example 1.34

For the matrix

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

- (a) Determine the eigenvalues and corresponding eigenvectors of the product \mathbf{AA}^T .
- (b) Normalize the eigenvectors to obtain the orthogonal matrix $\hat{\mathbf{V}}$.
- (c) What are the singular values of \mathbf{A} ?

Solution (a) $\mathbf{AA}^T = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$

The eigenvalues of \mathbf{AA}^T are given by the solutions of the equation

$$|\mathbf{AA}^T - \mu I| = \begin{vmatrix} 11 - \mu & 1 \\ 1 & 11 - \mu \end{vmatrix} = 0$$

which reduces to

$$(\mu - 12)(\mu - 10) = 0$$

giving the eigenvalues as

$$\mu_1 = 12, \mu_2 = 10$$

Solving the homogeneous equations

$$(\mathbf{AA}^T - \mu_i I) \mathbf{v}_i = 0$$

gives the corresponding eigenvectors as

$$\mathbf{v}_1 = [1 \ 1]^T, \quad \mathbf{v}_2 = [1 \ -1]^T$$

- (b) The corresponding normalized eigenvectors are:

$$\hat{v}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^T, \quad \hat{v}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}^T$$

giving the orthogonal matrix

$$\hat{V} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 0.7071 & 0.7071 \\ 0.7071 & -0.7071 \end{bmatrix}$$

- (c) The singular values of \mathbf{A} are the square roots of the non-zero eigenvalues of $\mathbf{A}^T \mathbf{A}$. Thus the singular values of \mathbf{A} are:

$$\sigma_1 = \sqrt{\mu_1} = \sqrt{12} = 3.4641 \quad \text{and} \quad \sigma_2 = \sqrt{10} = 3.1623$$

in agreement with the values obtained in Example 1.33.

1.8.2 Singular value decomposition (SVD)

For an $m \times n$ matrix \mathbf{A} of rank r the m equations (1.47) can be written in the partitioned form

$$\mathbf{A}[\hat{v}_1 \hat{v}_2 \dots \hat{v}_r | \hat{v}_{r+1} \dots \hat{v}_n] = [\hat{u}_1 \hat{u}_2 \dots \hat{u}_r | \hat{u}_{r+1} \dots \hat{u}_m] \Sigma \quad (1.48)$$

where the matrix Σ has the form

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & | & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & | & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r & | & 0 & \cdots & 0 \\ \hline 0 & 0 & \cdots & 0 & | & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & | & 0 & \cdots & 0 \end{bmatrix}$$

where $\sigma_1, \sigma_2, \dots, \sigma_r$ are the singular values of \mathbf{A} . More precisely (1.48) may be written as

$$\mathbf{A}\hat{V} = \hat{U}\Sigma$$

Using the orthogonality property $\hat{V}\hat{V}^T = \mathbf{I}$ leads to the result

$$\mathbf{A} = \hat{U}\Sigma\hat{V}^T \quad (1.49)$$

Such a decomposition (or factorization) of a non-square matrix \mathbf{A} is called the **singular value decomposition** of \mathbf{A} , commonly abbreviated as SVD of \mathbf{A} . It is analogous to the reduction to canonical (or diagonal) form of a square matrix developed in Section 1.6.

Example 1.35

Find the SVD of the matrix

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

and verify your answer.

Solution

The associated matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ and the singular values of \mathbf{A} were determined in Examples 1.33 and 1.34 as:

$$\hat{\mathbf{U}} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix}, \hat{\mathbf{V}} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}, \sigma_1 = \sqrt{12} \text{ and } \sigma_2 = \sqrt{10}$$

From (1.49) it follows that the SVD of \mathbf{A} is

$$\mathbf{A} = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 \\ 0 & \sqrt{10} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix}$$

Direct multiplication of the right-hand side confirms

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

The decomposition (1.47) can always be done. The non-zero diagonal elements of Σ are uniquely determined as the singular values of \mathbf{A} . The matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ are not unique and it is necessary to ensure that linear combinations of their columns satisfy (1.47). This applies when the matrices have repeated eigenvalues, as illustrated in Example 1.36.

Example 1.36

Find the SVD of the matrix

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

Solution $\mathbf{A}\mathbf{A}^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

The product $\mathbf{A}\mathbf{A}^T$ has eigenvalues $\lambda_1 = 4$, $\lambda_2 = 4$, $\lambda_3 = 1$ and $\lambda_4 = 0$. Normalized eigenvectors corresponding to λ_3 and λ_4 are respectively

$$\hat{\mathbf{u}}_3 = [1 \ 0 \ 0 \ 0]^T \quad \text{and} \quad \hat{\mathbf{u}}_4 = [0 \ 0 \ 0 \ 1]^T$$

Various possibilities exist for the repeated eigenvalues $\lambda_1 = \lambda_2 = 4$. Two possible choices of normalized eigenvectors are

$$\hat{\mathbf{u}}_1 = [0 \ 1 \ 0 \ 0]^T \quad \text{and} \quad \hat{\mathbf{u}}_2 = [0 \ 0 \ 1 \ 0]^T$$

or

$$\hat{\mathbf{u}}'_1 = \frac{1}{\sqrt{2}} [0 \ 1 \ 1 \ 0]^T \quad \text{and} \quad \hat{\mathbf{u}}'_2 = \frac{1}{\sqrt{2}} [0 \ 1 \ -1 \ 0]^T$$

(Note that the eigenvectors $\hat{\mathbf{u}}'_1$ and $\hat{\mathbf{u}}'_2$ are linear combinations of $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$.) Likewise

$$\mathbf{A}^T \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}$$

and has eigenvalues $\mu_1 = 4$, $\mu_2 = 4$ and $\mu_3 = 1$. The normalized eigenvector corresponding to the eigenvalue $\mu_3 = 1$ is

$$\hat{\mathbf{v}}_3 = [1 \ 0 \ 0]^T$$

and two possible choices for the eigenvectors corresponding to the repeated eigenvalue $\mu_1 = \mu_2 = 4$ are

$$\hat{\mathbf{v}}_1 = [0 \ 1 \ 0]^T \quad \text{and} \quad \hat{\mathbf{v}}_2 = [0 \ 0 \ 1]^T$$

or

$$\hat{\mathbf{v}}'_1 = \frac{1}{\sqrt{2}} [0 \ 1 \ 1]^T \quad \text{and} \quad \hat{\mathbf{v}}'_2 = \frac{1}{\sqrt{2}} [0 \ 1 \ -1]^T$$

The singular values of \mathbf{A} are $\sigma_1 = 2$, $\sigma_2 = 2$ and $\sigma_3 = 1$ giving

$$\Sigma = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Considering the requirements (1.47) it is readily confirmed that

$$\mathbf{A}\hat{\mathbf{v}}_1 = \sigma_1 \hat{\mathbf{u}}_1, \mathbf{A}\hat{\mathbf{v}}_2 = \sigma_2 \hat{\mathbf{u}}_2 \text{ and } \mathbf{A}\hat{\mathbf{v}}_3 = \sigma_3 \hat{\mathbf{u}}_3$$

so that

$$\hat{U}_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } \hat{V}_1 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

reduces \mathbf{A} to the SVD form $\mathbf{A} = \hat{U}_1 \Sigma \hat{V}_1^T$.

Also, it can be confirmed that

$$\mathbf{A}\hat{v}'_1 = \sigma_1\hat{u}'_1, \mathbf{A}\hat{v}'_2 = \sigma_2\hat{u}'_2, \mathbf{A}\hat{v}'_3 = \sigma_3\hat{u}'_3$$

so that the matrix pair

$$\hat{U}_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \hat{V}_2 = \begin{bmatrix} 0 & 0 & 1 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}$$

reduces \mathbf{A} to the SVD form

$$\mathbf{A} = \hat{U}_2 \Sigma \hat{V}_2^T$$

However, the corresponding columns of the matrix pair \hat{U}_2, \hat{V}_1 do not satisfy conditions (1.47) and

$$\mathbf{A} \neq \hat{U}_2 \Sigma \hat{V}_1^T$$

To ensure that conditions (1.47) are satisfied it is advisable to select the normalized eigenvectors \hat{v}_i first and then determine the corresponding normalized eigenvectors \hat{u}_i directly from (1.47).

1.8.3 Pseudo inverse

In Section 1.2.5 we considered the solution of the system of simultaneous linear equation

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1.50}$$

where \mathbf{A} is the $n \times n$ square matrix of coefficients and \mathbf{x} is the n vector of unknowns. Here the number of equations is equal to the number of unknowns and a unique solution

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{1.51}$$

exists if and only if the matrix \mathbf{A} is non-singular.

There are situations when the matrix \mathbf{A} is singular or a non-square $m \times n$ matrix. If the matrix \mathbf{A} is a $m \times n$ matrix then:

- if $m > n$ there are more equations than unknowns and this represents the **over determined** case;
- if $m < n$ there are fewer equations than unknowns and this represents the **under determined** case.

Clearly approximate solution vectors \mathbf{x} are desirable in such cases. This can be achieved using the SVD form (1.49) of a $m \times n$ matrix \mathbf{A} . Recognizing the orthogonality of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ the following matrix \mathbf{A}^\dagger is defined

$$\mathbf{A}^\dagger = \hat{\mathbf{V}} \Sigma^* \hat{\mathbf{U}}^T \quad (1.52)$$

where Σ^* is the transpose of Σ in which the singular values σ_i of \mathbf{A} are replaced by their reciprocals. The matrix \mathbf{A}^\dagger is called the **pseudo inverse** (or **generalized inverse**) of the matrix \mathbf{A} . It is also frequently referred to as the **Moore–Penrose pseudo inverse** of \mathbf{A} . It exists for any matrix \mathbf{A} including singular square matrices and non-square matrices. In the particular case when \mathbf{A} is a **square non-singular matrix** $\mathbf{A}^\dagger = \mathbf{A}^{-1}$. Since

$$\mathbf{A}^\dagger \mathbf{A} = \begin{bmatrix} \mathbf{I} & : & 0 \\ : & : & : \\ 0 & : & 0 \end{bmatrix}$$

a solution of (1.50) is $\mathbf{A}^\dagger \mathbf{A} \mathbf{x} = \mathbf{A}^\dagger \mathbf{b}$, that is

$$\mathbf{x} = \mathbf{A}^\dagger \mathbf{b} \quad (1.53)$$

This is the least squares solution of (1.50) in that it minimizes $(\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b})$, the sum of the squares of the errors.

Example 1.37

Determine the pseudo inverse of the matrix

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

and confirm that $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}$.

Solution From Example 1.35 the SVD of \mathbf{A} is

$$\mathbf{A} = \hat{\mathbf{U}} \Sigma \hat{\mathbf{V}}^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{-5}{\sqrt{30}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 \\ 0 & \sqrt{10} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

The matrix Σ^* is obtained by taking the transpose of Σ and inverting the non-zero diagonal elements, giving

$$\Sigma^* = \begin{bmatrix} \frac{1}{\sqrt{12}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{10}} & 0 \end{bmatrix}$$

so from (1.52) the pseudo inverse is

$$\mathbf{A}^\dagger = \hat{\mathbf{V}} \Sigma^* \hat{\mathbf{U}}^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{12}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{10}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{-1}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix} = \frac{1}{60} \begin{bmatrix} 17 & 4 & 5 \\ -7 & 16 & 5 \end{bmatrix}$$

Direct multiplication gives

$$\mathbf{A}^\dagger \mathbf{A} = \frac{1}{60} \begin{bmatrix} 17 & 4 & 5 \\ -7 & 16 & 5 \end{bmatrix} \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix} = \frac{1}{60} \begin{bmatrix} 60 & 0 \\ 0 & 60 \end{bmatrix} = \mathbf{I}$$

so that \mathbf{A}^\dagger is a **left inverse** of \mathbf{A} . However, \mathbf{A}^\dagger cannot be a **right inverse** of \mathbf{A} .

We noted in the solution to Example 1.37 that whilst \mathbf{A}^\dagger was a left inverse of \mathbf{A} it was not a right inverse. Indeed a matrix with more rows than columns cannot have a right inverse, but it will have a left inverse if such an inverse exists. Likewise, a matrix with more columns than rows cannot have a left inverse, but will have a right inverse if such an inverse exists.

There are other ways of computing the pseudo inverse, without having to use SVD. However, most are more restrictive in use and not so generally applicable as the SVD method. It has been shown that \mathbf{A}^\dagger is a unique pseudo inverse of an $m \times n$ matrix \mathbf{A} provided it satisfies the following three conditions:

$$\begin{aligned} \mathbf{AA}^\dagger \text{ and } \mathbf{A}^\dagger \mathbf{A} \text{ are symmetric} \\ \mathbf{AA}^\dagger \mathbf{A} = \mathbf{A} \\ \mathbf{A}^\dagger \mathbf{AA}^\dagger = \mathbf{A}^\dagger \end{aligned} \tag{1.54}$$

For example, if an $m \times n$ matrix \mathbf{A} is of **full rank** then the pseudo inverse may be calculated as follows:

$$\text{if } m > n \text{ then } \mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \tag{1.55a}$$

$$\text{if } m < n \text{ then } \mathbf{A}^\dagger = \mathbf{A}^T (\mathbf{A} \mathbf{A}^T)^{-1} \tag{1.55b}$$

It is left as an exercise to confirm that these two forms satisfy conditions (1.54).

Example 1.38

- (a) Without using SVD determine the pseudo inverse of the matrix

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

- (b) Find the least squares solution of the following systems of simultaneous linear equations

$$\begin{array}{ll} \text{(i)} & 3x - y = 2 \\ & x + 3y = 4 \\ & x + y = 2 \end{array} \quad \begin{array}{ll} \text{(ii)} & 3x - y = 2 \\ & x + 3y = 2 \\ & x + y = 2 \end{array}$$

and comment on the answers.

Solution (a) From the solution to Example 1.33 $\text{rank}(\mathbf{A}) = 2$, so the matrix \mathbf{A} is of full rank. Since in this case $m > n$ we can use (1.55a) to determine the pseudo inverse as

$$\begin{aligned}\mathbf{A}^\dagger &= (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}^{-1} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \\ &= \frac{1}{120} \begin{bmatrix} 11 & -1 \\ -1 & 11 \end{bmatrix} \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix} \\ &= \frac{1}{60} \begin{bmatrix} 17 & 4 & 5 \\ -7 & 16 & 5 \end{bmatrix} = \begin{bmatrix} 0.2833 & 0.0667 & 0.0833 \\ -0.1167 & 0.2667 & 0.0833 \end{bmatrix}\end{aligned}$$

in agreement with the result obtained in Example 1.37.

(b) Both (i) and (ii) are examples of **over determined** (or **over specified**) sets of equations $\mathbf{Ax} = \mathbf{b}$ with \mathbf{A} being an $m \times n$ matrix, $m > n$, \mathbf{b} being an m -vector and \mathbf{x} an n -vector of unknowns. Considering the augmented matrix $(\mathbf{A}: \mathbf{b})$ then:

- if $\text{rank}(\mathbf{A}: \mathbf{b}) > \text{rank}(\mathbf{A})$ the equations are inconsistent and there is no solution (this is the most common situation for over specified sets of equations);
- if $\text{rank}(\mathbf{A}: \mathbf{b}) = \text{rank}(\mathbf{A})$ some of the equations are redundant and there is a solution containing $n - \text{rank}(\mathbf{A})$ free parameters.

(See Section 5.6 of MEM.)

Considering case (i)

$$\mathbf{A} = \begin{bmatrix} 3 & -1 \\ 1 & 3 \\ 1 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\text{rank}(\mathbf{A}: \mathbf{b}) = \text{rank} \begin{bmatrix} 3 & -1 & 2 \\ 1 & 3 & 4 \\ 1 & 1 & 2 \end{bmatrix} = 2 = \text{rank}(\mathbf{A}) \text{ from (a).}$$

Thus the equations are consistent and a unique solution exists. The least squares solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{A}^\dagger \mathbf{b} = \frac{1}{60} \begin{bmatrix} 17 & 4 & 5 \\ -7 & 16 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

which gives the unique solution $x = y = 1$.

Considering case (ii) \mathbf{A} and \mathbf{x} are the same as in (i) and $\mathbf{b} = [2 \ 2 \ 2]^T$

$$\text{rank}(\mathbf{A}: \mathbf{b}) = \text{rank} \begin{bmatrix} 3 & -1 & 2 \\ 1 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix} = 3 > \text{rank}(\mathbf{A}) = 2$$

Thus the equations are inconsistent and there is no unique solution. The least squares solution is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{A}^{\dagger} \mathbf{b} = \frac{1}{60} \begin{bmatrix} 17 & 4 & 5 \\ -7 & 16 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 13 \\ 7 \end{bmatrix}$$

giving $x = \frac{13}{15}$ and $y = \frac{7}{15}$.

As indicated earlier, the least squares solution $\mathbf{x} = \mathbf{A}^{\dagger} \mathbf{b}$ of the system of equations $\mathbf{Ax} = \mathbf{b}$ is the solution that minimizes the square of the error vector $\mathbf{r} = (\mathbf{Ax} - \mathbf{b})$; that is, minimizes $(\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b})$.

In practice, data associated with individual equations within the set may not be equally reliable; so more importance may be attached to some of the errors r_i . To accommodate for this, a **weighting factor** (positive number) w_i is given to the i th equation ($i = 1, 2, \dots, m$) and the least squares solution is the solution that minimizes the square of the vector $\mathbf{W}(\mathbf{Ax} - \mathbf{b})$, where \mathbf{W} is the $n \times n$ diagonal matrix having the square roots $\sqrt{w_i}$ of the weighting factors as its diagonal entries; that is

$$\mathbf{W} = \begin{bmatrix} \sqrt{w_1} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & \sqrt{w_m} \end{bmatrix}$$

The larger w_i the closer the fit of the least squares solution to the i th equation; the smaller w_i the poorer the fit. Care over weighting must be taken when using least squares solution packages. Most times one would notice the heavy weighting, but in automated systems one probably would not notice. Exercise 48 serves to illustrate.



In MATLAB the command

`svd(A)`

returns the singular values of \mathbf{A} in non-decreasing order; whilst the command

`[U, S, V] = svd(A)`

returns the diagonal matrix $\mathbf{S} = \Sigma$ and the two unitary matrices $\mathbf{U} = \hat{\mathbf{U}}$ and $\mathbf{V} = \hat{\mathbf{V}}$ such that $\mathbf{A} = \hat{\mathbf{U}} \hat{\mathbf{S}} \hat{\mathbf{V}}^T$. The commands

`A=sym(A);`

`svd(A)`

return the singular values of the matrix \mathbf{A} in symbolic form. Symbolic singular vectors are not available. The command

`pinv(A)`

returns the pseudo inverse of the matrix \mathbf{A} using the SVD form of \mathbf{A} .

Using the matrix \mathbf{A} of Examples 1.35, 1.36, 1.38 and 1.39 the commands

`A=[3 -1; 1 3; 1 1];`

`[U, S, V] = svd(A)`

```

return
      -0.4082   0.8944  -0.1826
      U= -0.8165  -0.4472  -0.3651
      -0.4082  -0.0000   0.9129
      3.4641      0
      S=    0      3.1623
            0      0
      V=  -0.7071   0.7071
            -0.7071  -0.7071

```

The additional command

```
pinv(A)
```

returns the pseudo inverse of **A** as

```
0.2833  0.0667  0.0833
-0.1167  0.2667  0.0833
```

The commands

```

A=[3 -1;1 3;1 1];
a=sym(A);
S=svd(a)

return

S= 2*3^(1/2)
    10^(1/2)

```

In MAPLE the commands

```

with(LinearAlgebra):
A:=Matrix([[3,-1],[1,3],[1,1]]);
svd:=SingularValues(A,output=['U','S','Vt']);

return

svd= 
$$\begin{bmatrix} -0.4082 & 0.8944 & -0.1826 \\ -0.8165 & -0.4472 & -0.3651 \\ -0.4082 & -1.9429 \times 10^{-16} & 0.9129 \end{bmatrix}, \begin{bmatrix} 3.4641 \\ 3.1623 \\ 0.0000 \end{bmatrix}, \begin{bmatrix} -0.7071 & -0.7071 \\ 0.7071 & -0.07071 \end{bmatrix}$$


```

where the singular values are expressed as a vector. To output the values of **U** and **Vt** separately and to output the singular values as a matrix the following additional commands may be used:

```

U:=svd[1];
Vt:=svd[3];
SS:=matrix(3,2,(i,j) → if i=j then svd[2][i]else 0
fi);#output the singular values into a 3 2 matrix

```

The further command

```
U.SS.Vt;
```

gives the output

$$\begin{bmatrix} 3.0000 & -1.0000 \\ 1.0000 & 3.0000 \\ 1.000 & 1.000 \end{bmatrix}$$

confirming that we reproduce \mathbf{A} .

To obtain the pseudo inverse using MAPLE the normal matrix inverse command is used. Thus the commands

```
with(LinearAlgebra):
A:=Matrix([[3,-1],[1,3],[1,1]]);
MatrixInverse(A);
return
```

$$\begin{bmatrix} \frac{17}{60} & \frac{1}{15} & \frac{1}{12} \\ -\frac{7}{60} & \frac{4}{15} & \frac{1}{12} \end{bmatrix}$$

in agreement with the answer obtained in Example 1.37.

1.8.4 Exercises



Use MATLAB or MAPLE to check your answers.

42 Considering the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 7 & 10 \\ 2 & 1 & 5 & 7 \end{bmatrix}$$

- (a) Determine row rank (\mathbf{A}) and column rank (\mathbf{A}).
- (b) Is the matrix \mathbf{A} of full rank?

43 (a) Find the SVD form of the matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$$

- (b) Use SVD to determine the pseudo inverse \mathbf{A}^\dagger of the matrix \mathbf{A} . Confirm that $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}$.
- (c) Determine the pseudo inverse without using SVD.

44 Show that the matrix

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

is of full rank. Without using SVD determine its pseudo inverse \mathbf{A}^\dagger and confirm that $\mathbf{A}^\dagger \mathbf{A} = \mathbf{I}$.

45 Considering the matrix

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

- (a) What is the rank of \mathbf{A} ?
- (b) Find the SVD of \mathbf{A} .
- (c) Find the pseudo inverse \mathbf{A}^\dagger of \mathbf{A} and confirm that $\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}$ and $\mathbf{A}^\dagger \mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger$.
- (d) Find the least squares solution of the simultaneous equations

$$x - y = 1, -2x + 2y = 2, 2x - 2y = 3$$

- (e) Confirm the answer to (d) by minimizing the square of the error vector $(\mathbf{Ax} - \mathbf{b})$ where $\mathbf{b} = [1 \quad 2 \quad 3]^T$.

46 Considering the matrix

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

- (a) Use the pseudo inverse \mathbf{A}^\dagger determined in Example 1.37 to find the least squares solution for the simultaneous equations

$$3x - y = 1, x + 3y = 2, x + y = 3$$

- (b) Confirm the answer to (a) by minimizing the square of the error vector
 $(Ax - b)$ where $b = [1 \quad 2 \quad 3]^T$.
- (c) By drawing the straight lines represented by the equations illustrate your answer graphically.

47 Considering the matrix

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

- (a) Show that A is of full rank.
(b) Determine the pseudo inverse A^\dagger .
(c) Show that the A^\dagger obtained satisfies the four conditions (1.54).

48 Find the least squares solution of the following pairs of simultaneous linear equations.

(a) (i) $2x + y = 3$ $x + 2y = 3$ $x + y = 2$	(ii) $2x + y = 3$ $x + 2y = 3$ $x + y = 3$
---	--

(b) (i) $2x + y = 3$ $x + 2y = 3$ $10x + 10y = 20$	(ii) $2x + y = 3$ $x + 2y = 3$ $10x + 10y = 30$
--	---

(c) (i) $2x + y = 3$ $x + 2y = 3$ $100x + 100y = 200$	(ii) $2x + y = 3$ $x + 2y = 3$ $100x + 100y = 300$
---	--

Comment on your answers.

By representing the data in the matrix form $Az = y$, where $z = [m \ c]^T$, use the pseudo inverse to find the values of m and c which provide the least squares fit to the linear model $y = mx + c$ for the following data.

k	1	2	3	4	5
x_k	0	1	2	3	4
y_k	1	1	2	2	3

(Compare with Example 2.17 in MEM.)

1.9 State-space representation

In Section 10.11.2 of MEM it was illustrated how the solution of differential equation initial value problems of order n can be reduced to the solution of a set n of first-order differential equations, each with an initial condition. In this section we apply matrix techniques to obtain the solution of such systems.

1.9.1 Single-input-single-output (SISO) systems

First let us consider the **single-input-single-output (SISO) system** characterized by the n th-order linear differential equation

$$a_n \frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \dots + a_1 \frac{dy}{dt} + a_0 y = u(t) \quad (1.56)$$

where the coefficients a_i ($i = 0, 1, \dots, n$) are constants with $a_n \neq 0$ and it is assumed that the initial conditions $y(0), y^{(1)}(0), \dots, y^{(n-1)}(0)$ are known.

We introduce the n variables $x_1(t), x_2(t), \dots, x_n(t)$ defined by

$$x_1(t) = y(t)$$

$$x_2(t) = \frac{dy}{dt} = \dot{x}_1(t)$$

$$x_3(t) = \frac{d^2 y}{dt^2} = \dot{x}_2(t)$$

⋮

$$x_{n-1}(t) = \frac{d^{n-2}y}{dt^{n-2}} = \dot{x}_{n-2}(t)$$

$$x_n(t) = \frac{d^{n-1}y}{dt^{n-1}} = \dot{x}_{n-1}(t)$$

where, as usual, a dot denotes differentiation with respect to time t . Then, by substituting in (1.56), we have

$$a_n \dot{x}_n + a_{n-1} x_n + a_{n-2} x_{n-1} + \cdots + a_1 x_2 + a_0 x_1 = u(t)$$

giving

$$\dot{x}_n = -\frac{a_{n-1}}{a_n} x_n - \frac{a_{n-2}}{a_n} x_{n-1} - \cdots - \frac{a_1}{a_n} x_2 - \frac{a_0}{a_n} x_1 + \frac{1}{a_n} u$$

Thus, we can represent (1.56) as a system of n simultaneous first-order differential equations

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3, \quad \cdots, \quad \dot{x}_{n-1} = x_n$$

$$\dot{x}_n = -\frac{a_0}{a_n} x_1 - \frac{a_1}{a_n} x_2 - \cdots - \frac{a_{n-1}}{a_n} x_n + \frac{1}{a_n} u$$

which may be written as the **vector–matrix differential equation**

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -\frac{a_0}{a_n} & -\frac{a_1}{a_n} & -\frac{a_2}{a_n} & \cdots & -\frac{a_{n-2}}{a_n} & -\frac{a_{n-1}}{a_n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \frac{1}{a_n} \end{bmatrix} u(t) \quad (1.57)$$

(Note: Clearly x_1, x_2, \dots, x_n and u are functions of t and strictly should be written as $x_1(t), x_2(t), \dots, x_n(t)$ and $u(t)$. For the sake of convenience and notational simplicity the argument (t) is frequently omitted when the context is clear.)

Equation (1.57) may be written in the more concise form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u \quad (1.58a)$$

The vector $\mathbf{x}(t)$ is called the system **state vector**, and it contains all the information that one needs to know about the behaviour of the system. Its components are the n **state variables** x_1, x_2, \dots, x_n , which may be considered as representing a set of coordinate axes in the n -dimensional coordinate space over which $\mathbf{x}(t)$ ranges. This is referred to as the **state space**, and as time increases the state vector $\mathbf{x}(t)$ will describe a locus in this space called a **trajectory**. In two dimensions the state space reduces to the **phase plane**. The matrix \mathbf{A} is called the system **matrix** and the particular form adopted in (1.57) is known as the **companion form**, which is widely adopted in practice. Equation (1.58a) is referred to as the system **state equation**.

The output, or response, of the system determined by (1.56) is given by y , which in terms of the state variables is determined by x_1 . Thus

$$y = [1 \ 0 \ \dots \ 0] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

or, more concisely,

$$y = \mathbf{c}^T \mathbf{x} \quad (1.58b)$$

$$\text{where } \mathbf{c} = [1 \ 0 \ \dots \ 0]^T.$$

A distinct advantage of the vector-matrix approach is that it is applicable to multivariable (that is, multi-input–multi-output MIMO) systems, dealt with in Section 1.9.2. In such cases it is particularly important to distinguish between the system state variables and the system outputs, which, in general, are linear combinations of the state variables.

Together the pair of equations (1.58a,b) in the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u \quad (1.59a)$$

$$y = \mathbf{c}^T \mathbf{x} \quad (1.59b)$$

constitute the **dynamic equations** of the system and are commonly referred to as the **state-space model** representation of the system. Such a representation forms the basis of the so-called ‘modern approach’ to the analysis and design of control systems in engineering. An obvious advantage of adopting the vector-matrix representation (1.59) is the compactness of the notation.

More generally the output y could be a linear combination of both the state and input, so that the more general form of the system dynamic equations (1.59) is

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u \quad (1.60a)$$

$$y = \mathbf{c}^T \mathbf{x} + du \quad (1.60b)$$

Comment

It is important to realize that the choice of state variables x_1, x_2, \dots, x_n is not unique. For example, for the system represented by (1.56) we could also take

$$x_1 = \frac{d^{n-1}y}{dt^{n-1}}, \quad x_2 = \frac{d^{n-2}y}{dt^{n-2}}, \quad \dots, \quad x_n = y$$

leading to the state-space model (1.59) with

$$\mathbf{A} = \begin{bmatrix} -\frac{a_{n-1}}{a_n} & -\frac{a_{n-2}}{a_n} & \dots & -\frac{a_1}{a_n} & -\frac{a_0}{a_n} \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \frac{1}{a_n} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \quad (1.61)$$

Example 1.39

Obtain a state-space representation of the system characterized by the third-order differential equation

$$\frac{d^3y}{dt^3} + 3\frac{d^2y}{dt^2} + 2\frac{dy}{dt} - 4y = e^{-t} \quad (1.62)$$

Solution Writing

$$x_1 = y, \quad x_2 = \frac{dy}{dt} = \dot{x}_1, \quad x_3 = \frac{d^2y}{dt^2} = \dot{x}_2$$

we have, from (1.62),

$$\dot{x}_3 = \frac{d^3y}{dt^3} = 4y - 2\frac{dy}{dt} - 3\frac{d^2y}{dt^2} + e^{-t} = 4x_1 - 2x_2 - 3x_3 + e^{-t}$$

Thus the corresponding state equation is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 4 & -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} e^{-t}$$

with the output y being given by

$$y = x_1 = [1 \ 0 \ 0] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

These two equations then constitute the state-space representation of the system.

We now proceed to consider the more general SISO system characterized by the differential equation

$$\frac{d^n y}{dt^n} + a_{n-1} \frac{d^{n-1} y}{dt^{n-1}} + \cdots + a_0 y = b_m \frac{d^m u}{dt^m} + \cdots + b_0 u \quad (m \leq n) \quad (1.63)$$

in which the input involves derivative terms. Again there are various ways of representing (1.63) in the state-space form, depending on the choice of the state variables. As an illustration, we shall consider one possible approach, introducing others in the exercises.

We define \mathbf{A} and \mathbf{b} as in (1.57); that is, we take \mathbf{A} to be the companion matrix of the left-hand side of (1.63), giving

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & & & \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}$$

and we take $\mathbf{b} = [0 \ 0 \ \cdots \ 0 \ 1]^T$. In order to achieve the desired response, the vector \mathbf{c} is then chosen to be

$$\mathbf{c} = [b_0 \ b_1 \ \cdots \ b_m \ 0 \ \cdots \ 0]^T \quad (1.64)$$

It is left as an exercise to confirm that this choice is appropriate (see also Section 5.4.1).

Example 1.40

Obtain the state-space model for the system characterized by the differential equation model

$$\frac{d^3y}{dt^3} + 6\frac{d^2y}{dt^2} + 11\frac{dy}{dt} + 3y = 5\frac{d^2u}{dt^2} + \frac{du}{dt} + u \quad (1.65)$$

Solution Taking \mathbf{A} to be the companion matrix of the left-hand side in (1.65)

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -3 & -11 & -6 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = [0 \ 0 \ 1]^T$$

we have, from (1.64),

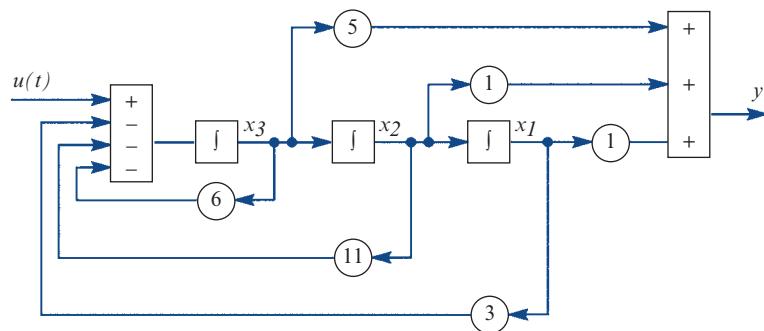
$$\mathbf{c} = [1 \ 1 \ 5]^T$$

Then from (1.59) the state-space model becomes

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{bu}, \quad y = \mathbf{c}^T \mathbf{x}$$

This model structure may be depicted by the block diagram of Figure 1.2. It provides an ideal model for simulation studies, with the state variables being the outputs of the various integrators involved.

Figure 1.2
Block diagram for the state-space model of Example 1.40.



A distinct advantage of this approach to obtaining the state-space model is that \mathbf{A} , \mathbf{b} and \mathbf{c} are readily written down. A possible disadvantage in some applications is that the output y itself is not a state variable. An approach in which y is a state variable is developed in Exercise 44, Section 5.4.2. In practice, it is also fairly common to choose the state variables from a physical consideration.

1.9.2 Multi-input–multi-output (MIMO) systems

Many practical systems are multivariable in nature, being characterized by having more than one input and/or more than one output. In general terms, the state-space model is similar to that in (1.60) for SISO systems, except that the input is now a vector $\mathbf{u}(t)$ as is the output $\mathbf{y}(t)$. Thus the more general form, corresponding to (1.60), of the state-space model representation of an n th-order **multi-input–multi-output (MIMO)** system subject to r inputs and l outputs is

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \quad (1.66a)$$

$$\mathbf{y} = \mathbf{Cx} + \mathbf{Du} \quad (1.66b)$$

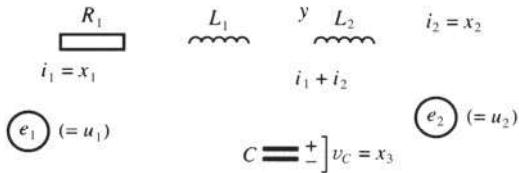
where \mathbf{x} is the n -state vector, \mathbf{u} is the r -input vector, \mathbf{y} is the l -output vector, \mathbf{A} is the $n \times n$ system matrix, \mathbf{B} is the $n \times r$ control (or input) matrix, and \mathbf{C} and \mathbf{D} are respectively $l \times n$ and $l \times r$ output matrices.

Example 1.41

Obtain the state-space model representation characterizing the two-input–one-output parallel network shown in Figure 1.3 in the form

Figure 1.3

Parallel circuit of Example 1.41.



$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}, \quad \mathbf{y} = \mathbf{c}^T \mathbf{x} + \mathbf{d}^T \mathbf{u}$$

where the elements x_1, x_2, x_3 of \mathbf{x} and u_1, u_2 of \mathbf{u} are as indicated in the figure, and the output y is the voltage drop across the inductor L_1 (v_C denotes the voltage drop across the capacitor C).

Solution Applying Kirchhoff's second law (see Chapter 5 and Section 11.4.1 of MEM) to each of the two loops in turn gives

$$R_1 i_1 + L_1 \frac{di_1}{dt} + v_C = e_1 \quad (1.67)$$

$$L_2 \frac{di_2}{dt} + v_C = e_2 \quad (1.68)$$

The voltage drop v_C across the capacitor C is given by

$$\dot{v}_C = \frac{1}{C} (i_1 + i_2) \quad (1.69)$$

The output y , being the voltage drop across the inductor L_1 , is given by

$$y = L_1 \frac{di_1}{dt}$$

which, using (1.67), gives

$$y = -R_1 i_1 - v_C + e_1 \quad (1.70)$$

Writing $x_1 = i_1$, $x_2 = i_2$, $x_3 = v_C$, $u_1 = e_1$ and $u_2 = e_2$, (1.67)–(1.70) give the state-space representation as

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} &= \begin{bmatrix} -\frac{R_1}{L_1} & 0 & -\frac{1}{L_1} \\ 0 & 0 & -\frac{1}{L_2} \\ \frac{1}{C} & \frac{1}{C} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \frac{1}{L_1} & 0 \\ 0 & \frac{1}{L_2} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ y &= [-R_1 \quad 0 \quad -1] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + [1 \quad 0] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \end{aligned}$$

which is of the required form

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$

$$y = \mathbf{c}^T \mathbf{x} + \mathbf{d}^T \mathbf{u}$$

1.9.3 Exercises

- 50 Obtain the state-space forms of the differential equations

$$(a) \frac{d^3y}{dt^3} + 4 \frac{d^2y}{dt^2} + 5 \frac{dy}{dt} + 4y = u(t)$$

$$(b) \frac{d^4y}{dt^4} + 2 \frac{d^2y}{dt^2} + 4 \frac{dy}{dt} = 5u(t)$$

using the companion form of the system matrix in each case.

- 51 Obtain the state-space form of the differential equation models

$$(a) \frac{d^3y}{dt^3} + 6 \frac{d^2y}{dt^2} + 5 \frac{dy}{dt} + 7y = \frac{d^2u}{dt^2} + 3 \frac{du}{dt} + 5u$$

$$(b) \frac{d^3y}{dt^3} + 4 \frac{d^2y}{dt^2} + 3 \frac{dy}{dt} = \frac{d^2u}{dt^2} + 3 \frac{du}{dt} + 2u$$

using the companion form of the system matrix in each case.

- 52 Obtain the state-space model of the single-input–single-output network system of Figure 1.4 in the form $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{bu}$, $y = \mathbf{c}^T \mathbf{x}$, where u , y and the elements x_1 , x_2 , x_3 of \mathbf{x} are as indicated.

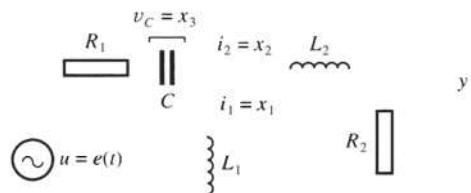


Figure 1.4 Network of Exercise 52.

The mass–spring–damper system of Figure 1.5 models the suspension system of a quarter-car. Obtain a state-space model in which the output represents the body mass vertical movement y and the input represents the tyre vertical movement

$u(t)$ due to the road surface. All displacements are measured from equilibrium positions.

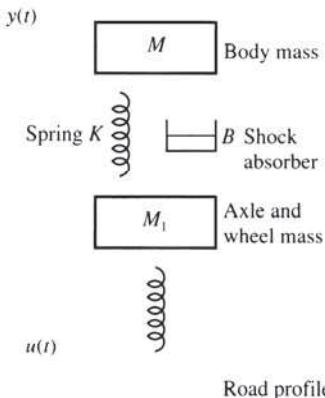


Figure 1.5 Quarter-car suspension model of Exercise 53.

54

Obtain the state-space model, in the form $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{bu}$, $\mathbf{y} = \mathbf{Cx} + \mathbf{d}^T u$ of the one-input–two-output network illustrated in Figure 1.6. The elements x_1, x_2 of the state vector \mathbf{x} and y_1, y_2 of the output vector \mathbf{y} are as indicated. If $R_1 = 1 \text{ k}\Omega$, $R_2 = 5 \text{ k}\Omega$, $R_3 = R_4 = 3 \text{ k}\Omega$, $C_1 = C_2 = 1 \mu\text{F}$ calculate the eigenvalues of the system matrix \mathbf{A} .

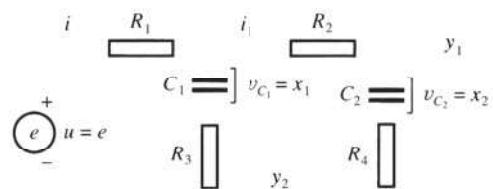


Figure 1.6 Network of Exercise 54.

1.10

Solution of the state equation

In this section we are concerned with seeking the solution of the state equation

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \quad (1.71)$$

given the value of \mathbf{x} at some initial time t_0 to be \mathbf{x}_0 . Having obtained the solution of this state equation, a system response \mathbf{y} may then be readily written down from the linear transformation (1.66b). As mentioned in Section 1.9.1, an obvious advantage of adopting the vector–matrix notation of (1.71) is its compactness. In this section we shall see that another distinct advantage is that (1.71) behaves very much like the corresponding first-order scalar differential equation

$$\frac{dx}{dt} = ax + bu, \quad x(t_0) = x_0 \quad (1.72)$$

1.10.1 Direct form of the solution

Before considering the n th-order system represented by (1.71), let us first briefly review the solution of (1.72). When the input u is zero, (1.72) reduces to the homogeneous equation

$$\frac{dx}{dt} = ax \quad (1.73)$$

which, by separation of variables,

$$\int_{x_0}^x \frac{dx}{x} = \int_{t_0}^t a dt$$

gives

$$\ln x - \ln x_0 = a(t - t_0)$$

leading to the solution

$$x = x_0 e^{a(t-t_0)} \quad (1.74)$$

for the unforced system.

If we consider the nonhomogeneous equation (1.72) directly, a solution can be obtained by first multiplying throughout by the integrating factor e^{-at} to obtain

$$e^{-at} \left(\frac{dx}{dt} - ax \right) = e^{-at} bu(t)$$

or

$$\frac{d}{dt}(e^{-at} x) = e^{-at} bu(t)$$

which on integration gives

$$e^{-at} x - e^{-at_0} x_0 = \int_{t_0}^t e^{-a\tau} bu(\tau) d\tau$$

leading to the solution

$$x(t) = e^{a(t-t_0)} x_0 + \int_{t_0}^t e^{a(t-\tau)} bu(\tau) d\tau \quad (1.75)$$

The first term of the solution, which corresponds to the solution of the unforced system, is a **complementary function**, while the convolution integral constituting the second term, which is dependent on the forcing function $u(t)$, is a **particular integral**.

Returning to (1.71), we first consider the unforced homogeneous system

$$\dot{x} = Ax, \quad x(t_0) = x_0 \quad (1.76)$$

which represents the situation when the system is ‘relaxing’ from an initial state.

The solution is completely analogous to the solution (1.74) of the scalar equation (1.73), and is of the form

$$x = e^{A(t-t_0)} x_0 \quad (1.77)$$

It is readily shown that this is a solution of (1.76). Using (1.33), differentiation of (1.77) gives

$$\dot{x} = A e^{A(t-t_0)} x_0 = Ax$$

so that (1.76) is satisfied. Also, from (1.77),

$$x(t_0) = e^{A(t_0-t_0)} x_0 = I x_0 = x_0$$

using $e^0 = I$. Thus, since (1.77) satisfies the differential equation and the initial conditions, it represents the unique solution of (1.76).

Likewise, the nonhomogeneous equation (1.71) may be solved in an analogous manner to that used for solving (1.72). Premultiplying (1.71) throughout by e^{-At} , we obtain

$$e^{-At} (\dot{x} - Ax) = e^{-At} Bu(t)$$

or using (1.33),

$$\frac{d}{dt}(e^{-At} x) = e^{-At} Bu(t)$$

Integration then gives

$$e^{-\mathbf{A}t}\mathbf{x}(t) - e^{-\mathbf{A}t_0}\mathbf{x}_0 = \int_{t_0}^t e^{-\mathbf{A}\tau} \mathbf{B}\mathbf{u}(\tau) d\tau$$

leading to the solution

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_0)}\mathbf{x}_0 + \int_{t_0}^t e^{\mathbf{A}(t-\tau)} \mathbf{B}\mathbf{u}(\tau) d\tau \quad (1.78)$$

This is analogous to the solution given in (1.75) for the scalar equation (1.72). Again it contains two terms: one dependent on the initial state and corresponding to the solution of the unforced system, and one a convolution integral arising from the input. Having obtained the solution of the state equation, the system output $\mathbf{y}(t)$ is then readily obtained from (1.66b).

1.10.2 The transition matrix

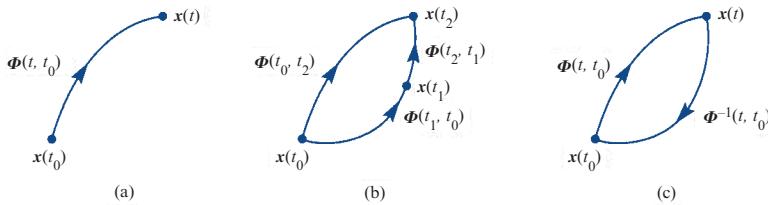
The matrix exponential $e^{\mathbf{A}(t-t_0)}$ is referred to as the **fundamental or transition matrix** and is frequently denoted by $\Phi(t, t_0)$, so that (1.77) is written as

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}_0 \quad (1.79)$$

This is an important matrix, which can be used to characterize a linear system, and in the absence of any input it maps a given state \mathbf{x}_0 at any time t_0 to the state $\mathbf{x}(t)$ at any time t , as illustrated in Figure 1.7(a).

Figure 1.7

- (a) Transition matrix $\Phi(t, t_0)$.
- (b) The transition property.
- (c) The inverse $\Phi^{-1}(t, t_0)$.



Using the properties of the exponential matrix given in Section 1.7, certain properties of the transition matrix may be deduced. From

$$e^{\mathbf{A}(t_1+t_2)} = e^{\mathbf{A}t_1} e^{\mathbf{A}t_2}$$

it follows that $\Phi(t, t_0)$ satisfies the **transition property**

$$\Phi(t_2, t_0) = \Phi(t_2, t_1)\Phi(t_1, t_0) \quad (1.80)$$

for any t_0, t_1 and t_2 , as illustrated in Figure 1.7(b). From

$$e^{\mathbf{A}t} e^{-\mathbf{A}t} = I$$

it follows that the inverse $\Phi^{-1}(t, t_0)$ of the transition matrix is obtained by negating time, so that

$$\Phi^{-1}(t, t_0) = \Phi(-t, -t_0) = \Phi(t_0, t) \quad (1.81)$$

for any t_0 and t , as illustrated in Figure 1.7(c).

1.10.3 Evaluating the transition matrix

Since, when dealing with time-invariant systems, there is no loss of generality in taking $t_0 = 0$, we shall, for convenience, consider the evaluation of the transition matrix

$$\Phi(t) = \Phi(t, 0) = e^{\mathbf{A}t}$$

Clearly, methods of evaluating this are readily applicable to the evaluation of

$$\Phi(t, \tau) = e^{\mathbf{A}(t-\tau)}$$

Indeed, since \mathbf{A} is a constant matrix,

$$\Phi(t, \tau) = \Phi(t - \tau, 0)$$

so, having obtained $\Phi(t)$, we can write down $\Phi(t, \tau)$ by simply replacing t by $t - \tau$.

Since \mathbf{A} is a constant matrix the methods discussed in Section 1.7 are applicable for evaluating the transition matrix. From (1.31a),

$$e^{\mathbf{A}t} = \alpha_0(t)\mathbf{I} + \alpha_1(t)\mathbf{A} + \alpha_2(t)\mathbf{A}^2 + \dots + \alpha_{n-1}(t)\mathbf{A}^{n-1} \quad (1.82a)$$

where, using (1.31b), the $\alpha_i(t)$ ($i = 0, 1, \dots, n - 1$) are obtained by solving simultaneously the n equations

$$e^{\lambda_j t} = \alpha_0(t) + \alpha_1(t)\lambda_j + \alpha_2(t)\lambda_j^2 + \dots + \alpha_{n-1}(t)\lambda_j^{n-1} \quad (1.82b)$$

where λ_j ($j = 1, 2, \dots, n$) are the eigenvalues of \mathbf{A} . As in Section 1.7, if \mathbf{A} has repeated eigenvalues then derivatives of $e^{\lambda t}$, with respect to λ , will have to be used.

Example 1.42

A system is characterized by the state equation

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & -3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t)$$

Given that the input is the unit step function

$$u(t) = H(t) = \begin{cases} 0 & (t < 0) \\ 1 & (t \geq 0) \end{cases}$$

and initially

$$x_1(0) = x_2(0) = 1$$

deduce the state $\mathbf{x}(t) = [x_1(t) \ x_2(t)]^T$ of the system at subsequent time t .

Solution From (1.78), the solution is given by

$$\mathbf{x}(t) = e^{\mathbf{A}t} \mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)} \mathbf{b} u(\tau) d\tau \quad (1.83)$$

where

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 1 & -3 \end{bmatrix}, \quad \mathbf{b} = [1 \ 1]^T$$

Since \mathbf{A} is a 2×2 matrix, it follows from (1.82a) that

$$e^{\mathbf{A}t} = \alpha_0(t)\mathbf{I} + \alpha_1(t)\mathbf{A}$$

The eigenvalues of \mathbf{A} are $\lambda_1 = -1$ and $\lambda_2 = -3$, so, using (1.82b), we have

$$\alpha_0(t) = \frac{1}{2}(3e^{-t} - e^{-3t}), \quad \alpha_1(t) = \frac{1}{2}(e^{-t} - e^{-3t})$$

giving

$$e^{\mathbf{A}t} = \begin{bmatrix} e^{-t} & 0 \\ \frac{1}{2}(e^{-t} - e^{-3t}) & e^{-3t} \end{bmatrix}$$

Thus the first term in (1.83) becomes

$$e^{\mathbf{A}t} \mathbf{x}(0) = \begin{bmatrix} e^{-t} & 0 \\ \frac{1}{2}(e^{-t} - e^{-3t}) & e^{-3t} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} e^{-t} \\ \frac{1}{2}(e^{-t} + e^{-3t}) \end{bmatrix}$$

and the second term is

$$\begin{aligned} \int_0^t e^{\mathbf{A}(t-\tau)} \mathbf{b} u(\tau) d\tau &= \int_0^t \begin{bmatrix} e^{-(t-\tau)} & 0 \\ \frac{1}{2}(e^{-(t-\tau)} - e^{-3(t-\tau)}) & e^{-3(t-\tau)} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} 1 d\tau \\ &= \int_0^t \begin{bmatrix} e^{-(t-\tau)} \\ \frac{1}{2}(e^{-(t-\tau)} + e^{-3(t-\tau)}) \end{bmatrix} d\tau = \begin{bmatrix} e^{-(t-\tau)} \\ \frac{1}{2}(e^{-(t-\tau)} + \frac{1}{3}e^{-3(t-\tau)}) \end{bmatrix}_0^t \\ &= \begin{bmatrix} e^{-0} \\ \frac{1}{2}(e^{-0} + \frac{1}{3}e^{-0}) \end{bmatrix} - \begin{bmatrix} e^{-t} \\ \frac{1}{2}(e^{-t} + \frac{1}{3}e^{-3t}) \end{bmatrix} \\ &= \begin{bmatrix} 1 - e^{-t} \\ \frac{2}{3} - \frac{1}{2}e^{-t} - \frac{1}{6}e^{-3t} \end{bmatrix} \end{aligned}$$

Substituting back in (1.83) gives the required solution

$$\mathbf{x}(t) = \begin{bmatrix} e^{-t} \\ \frac{1}{2}(e^{-t} + e^{-3t}) \end{bmatrix} + \begin{bmatrix} 1 - e^{-t} \\ \frac{2}{3} - \frac{1}{2}e^{-t} - \frac{1}{6}e^{-3t} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{2}{3} + \frac{1}{3}e^{-3t} \end{bmatrix}$$

That is,

$$x_1(t) = 1, \quad x_2(t) = \frac{2}{3} + \frac{1}{3}e^{-3t}$$



Using the Symbolic Math Toolbox in MATLAB the transition matrix e^{At} is generated by the sequence of commands

```
syms t
A=[specify];
A=sym(A);
E=expm(t*A);
pretty(E)
```

Confirm this using the matrix $A = [-1 \ 0; 1 \ -3]$ of Example 1.42.

In MAPLE e^{At} is returned by the commands

```
with(LinearAlgebra):
A:=Matrix([[-1,0], [1,-3]]);
MatrixExponential(A,t);
```

1.10.4 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 55 Obtain the transition matrix $\Phi(t)$ of the system

$$\dot{x} = Ax$$

where

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

Verify that $\Phi(t)$ has the following properties:

- (a) $\Phi(0) = I$;
- (b) $\Phi(t_2) = \Phi(t_2 - t_1)\Phi(t_1)$;
- (c) $\Phi^{-1}(t) = \Phi(-t)$.

- 56 Writing $x_1 = y$ and $x_2 = dy/dt$ express the differential equation

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

in the vector-matrix form $\dot{x} = Ax$, $x = [x_1 \ x_2]^T$. Obtain the transition matrix and hence solve the differential equation given that $y = dy/dt = 1$ when $t = 0$. Confirm your answer by direct solution of the second-order differential equation.

- 57 Solve

$$\dot{x} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

subject to $x(0) = [1 \ 1]^T$.

- 58 Find the solution of

$$\begin{aligned} \dot{x} &= \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ -6 & -5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 6 \end{bmatrix} u(t) \quad (t \geq 0) \end{aligned}$$

where $u(t) = 2$ and $x(0) = [1 \ -1]^T$.

- 59 Using (1.78), find the response for $t \geq 0$ of the system

$$\begin{aligned} \dot{x}_1 &= x_2 + 2u \\ \dot{x}_2 &= -2x_1 - 3x_2 \end{aligned}$$

to an input $u(t) = e^{-t}$ and subject to the initial conditions $x_1(0) = 0$, $x_2(0) = 1$.

- 60 A system is governed by the vector-matrix differential equation

$$\dot{x}(t) = \begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} u(t) \quad (t \geq 0)$$

where $x(t)$ and $u(t)$ are respectively the state and input vectors of the system. Determine the transition matrix of this system, and hence obtain an explicit expression for $x(t)$ for the input $u(t) = [4 \ 3]^T$ and subject to the initial condition $x(0) = [1 \ 2]^T$.

1.10.5 Spectral representation of response

We first consider the unforced system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) \quad (1.84)$$

with the initial state $\mathbf{x}(t_0)$ at time t_0 given, and assume that the matrix \mathbf{A} has as distinct eigenvalues λ_i ($i = 1, 2, \dots, n$) corresponding to n linearly independent eigenvectors \mathbf{e}_i ($i = 1, 2, \dots, n$). Since the n eigenvectors are linearly independent, they may be used as a basis for the n -dimensional state space, so that the system state $\mathbf{x}(t)$ may be written as a linear combination in the form

$$\mathbf{x}(t) = c_1(t)\mathbf{e}_1 + \dots + c_n(t)\mathbf{e}_n \quad (1.85)$$

where, since the eigenvectors are constant, the time-varying nature of $\mathbf{x}(t)$ is reflected in the coefficients $c_i(t)$. Substituting (1.85) into (1.84) gives

$$\dot{c}_1(t)\mathbf{e}_1 + \dots + \dot{c}_n(t)\mathbf{e}_n = \mathbf{A}[c_1(t)\mathbf{e}_1 + \dots + c_n(t)\mathbf{e}_n] \quad (1.86)$$

Since $(\lambda_i, \mathbf{e}_i)$ are **spectral pairs** (that is, eigenvalue–eigenvector pairs) for the matrix \mathbf{A} ,

$$\mathbf{A}\mathbf{e}_i = \lambda_i\mathbf{e}_i \quad (i = 1, 2, \dots, n)$$

(1.86) may be written as

$$[\dot{c}_1(t) - \lambda_1 c_1(t)]\mathbf{e}_1 + \dots + [\dot{c}_n(t) - \lambda_n c_n(t)]\mathbf{e}_n = 0 \quad (1.87)$$

Because the eigenvectors \mathbf{e}_i are linearly independent, it follows from (1.87) that the system (1.84) is completely represented by the set of uncoupled differential equations

$$\dot{c}_i(t) - \lambda_i c_i(t) = 0 \quad (i = 1, 2, \dots, n) \quad (1.88)$$

with solutions of the form

$$c_i(t) = e^{\lambda_i(t-t_0)} c_i(t_0)$$

Then, using (1.85), the system response is

$$\mathbf{x}(t) = \sum_{i=1}^n c_i(t_0) e^{\lambda_i(t-t_0)} \mathbf{e}_i \quad (1.89)$$

Using the given information about the initial state,

$$\mathbf{x}(t_0) = \sum_{i=1}^n c_i(t_0) \mathbf{e}_i \quad (1.90)$$

so that the constants $c_i(t_0)$ may be found from the given initial state using the **reciprocal basis vectors** \mathbf{r}_i ($i = 1, 2, \dots, n$) defined by

$$\mathbf{r}_i^\top \mathbf{e}_j = \delta_{ij}$$

where δ_{ij} is the Kronecker delta. Taking the scalar product of both sides of (1.90) with \mathbf{r}_k , we have

$$\mathbf{r}_k^\top \mathbf{x}(t_0) = \sum_{i=1}^n c_i(t_0) \mathbf{r}_k^\top \mathbf{e}_i = c_k(t_0) \quad (k = 1, 2, \dots, n)$$

which on substituting in (1.89) gives the system response

$$\mathbf{x}(t) = \sum_{i=1}^n \mathbf{r}_i^T \mathbf{x}(t_0) e^{\lambda_i(t-t_0)} \mathbf{e}_i \quad (1.91)$$

which is referred to as the **spectral** or **modal form** of the response. The terms $\mathbf{r}_i^T \mathbf{x}(t_0) e^{\lambda_i(t-t_0)} \mathbf{e}_i$ are called the **modes** of the system. Thus, provided that the system matrix \mathbf{A} has n linearly independent eigenvectors, this approach has the advantage of enabling us to break down the general system response into the sum of its simple modal responses. The amount of excitation of each mode, represented by $\mathbf{r}_i^T \mathbf{x}(t_0)$, is dependent only on the initial conditions, so if, for example, the initial state $\mathbf{x}(t_0)$ is parallel to the i th eigenvector \mathbf{e}_i then only the i th mode will be excited.

It should be noted that if a pair of eigenvalues λ_1, λ_2 are complex conjugates then the modes associated with $e^{\lambda_1(t-t_0)}$ and $e^{\lambda_2(t-t_0)}$ cannot be separated from each other. The combined motion takes place in a plane determined by the corresponding eigenvectors \mathbf{e}_1 and \mathbf{e}_2 and is oscillatory. By retaining only the dominant modes, the spectral representation may be used to approximate high-order systems by lower-order ones.

Example 1.43

Obtain in spectral form the response of the second-order system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

and sketch the trajectory.

Solution The eigenvalues of the matrix

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

are determined by

$$|\mathbf{A} - \lambda \mathbf{I}| = \lambda^2 + 4\lambda + 3 = 0$$

that is,

$$\lambda_1 = -1, \quad \lambda_2 = -3$$

with corresponding eigenvectors

$$\mathbf{e}_1 = [1 \quad 1]^T, \quad \mathbf{e}_2 = [1 \quad -1]^T$$

Denoting the reciprocal basis vectors by

$$\mathbf{r}_1 = [r_{11} \quad r_{12}]^T, \quad \mathbf{r}_2 = [r_{21} \quad r_{22}]^T$$

and using the relationships

$$\mathbf{r}_i^T \mathbf{e}_j = \delta_{ij} \quad (i, j = 1, 2)$$

we have

$$\mathbf{r}_1^T \mathbf{e}_1 = r_{11} + r_{12} = 1, \quad \mathbf{r}_1^T \mathbf{e}_2 = r_{11} - r_{12} = 0$$

giving

$$r_{11} = \frac{1}{2}, \quad r_{12} = \frac{1}{2}, \quad \mathbf{r}_1 = \left[\begin{array}{cc} \frac{1}{2} & \frac{1}{2} \end{array} \right]^T$$

and

$$\mathbf{r}_2^T \mathbf{e}_2 = r_{21} + r_{22} = 0, \quad \mathbf{r}_2^T \mathbf{e}_1 = r_{21} - r_{22} = 1$$

giving

$$r_{21} = \frac{1}{2}, \quad r_{22} = -\frac{1}{2}, \quad \mathbf{r}_2 = \left[\begin{array}{cc} \frac{1}{2} & -\frac{1}{2} \end{array} \right]^T$$

Thus

$$\mathbf{r}_1^T \mathbf{x}(0) = \frac{1}{2} + 1 = \frac{3}{2}, \quad \mathbf{r}_2^T \mathbf{x}(0) = \frac{1}{2} - 1 = -\frac{1}{2}$$

so that, from (1.91), the system response is

$$\mathbf{x}(t) = \sum_{i=1}^2 \mathbf{r}_i^T \mathbf{x}(0) e^{\lambda_i t} \mathbf{e}_i = \mathbf{r}_1^T \mathbf{x}(0) e^{\lambda_1 t} \mathbf{e}_1 + \mathbf{r}_2^T \mathbf{x}(0) e^{\lambda_2 t} \mathbf{e}_2$$

That is,

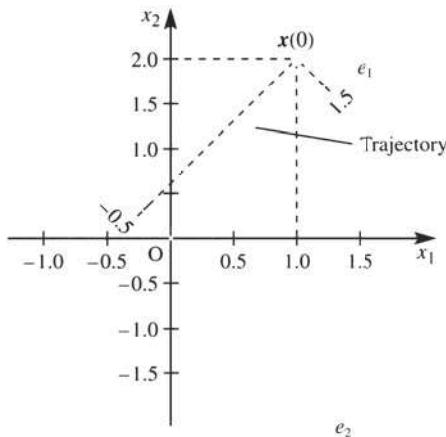
$$\mathbf{x}(t) = \frac{3}{2} e^{-t} \mathbf{e}_1 - \frac{1}{2} e^{-3t} \mathbf{e}_2$$

which is in the required spectral form.

To plot the response, we first draw axes corresponding to the eigenvectors \mathbf{e}_1 and \mathbf{e}_2 , as shown in Figure 1.8. Taking these as coordinate axes, we are at the point $(\frac{3}{2}, -\frac{1}{2})$ at time $t = 0$. As t increases, the movement along the direction of \mathbf{e}_2 is much faster than that in the direction of \mathbf{e}_1 , since e^{-3t} decreases more rapidly than e^{-t} . We can therefore guess the trajectory, without plotting, as sketched in Figure 1.8.

Figure 1.8

Trajectory for Example 1.43.



We can proceed in an analogous manner to obtain the spectral representation of the response to the forced system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$

with $\mathbf{x}(t_0)$ given. Making the same assumption regarding the linear independence of the eigenvectors \mathbf{e}_i ($i = 1, 2, \dots, n$) of the matrix \mathbf{A} , the vector $\mathbf{B}\mathbf{u}(t)$ may also be written as a linear combination of the form

$$\mathbf{B}\mathbf{u}(t) = \sum_{i=1}^n \beta_i(t) \mathbf{e}_i \quad (1.92)$$

so that, corresponding to (1.87), we have

$$[\dot{\mathbf{c}}_1(t) - \lambda_1 c_1(t) - \beta_1(t)] \mathbf{e}_1 + \dots + [\dot{\mathbf{c}}_n(t) - \lambda_n c_n(t) - \beta_n(t)] \mathbf{e}_n = 0$$

As a consequence of the linear independence of the eigenvectors \mathbf{e}_i this leads to the set of uncoupled differential equations

$$\dot{\mathbf{c}}_i(t) - \lambda_i c_i(t) - \beta_i(t) = 0 \quad (i = 1, 2, \dots, n)$$

which, using (1.75), have corresponding solutions

$$c_i(t) = e^{\lambda_i(t-t_0)} c_i(t_0) + \int_{t_0}^t e^{\lambda_i(t-\tau)} \beta_i(\tau) d\tau \quad (1.93)$$

As for $c_i(t_0)$, the reciprocal basis vectors \mathbf{r}_i may be used to obtain the coefficients $\beta_i(\tau)$. Taking the scalar product of both sides of (1.92) with \mathbf{r}_k and using the relationships $\mathbf{r}_i^T \mathbf{e}_j = \delta_{ij}$, we have

$$\mathbf{r}_k^T \mathbf{B}\mathbf{u}(t) = \beta_k(t) \quad (k = 1, 2, \dots, n)$$

Thus, from (1.93),

$$c_i(t) = e^{\lambda_i(t-t_0)} \mathbf{r}_i^T \mathbf{x}(t_0) + \int_{t_0}^t e^{\lambda_i(t-\tau)} \mathbf{r}_i^T \mathbf{B}\mathbf{u}(\tau) d\tau$$

giving the spectral form of the system response as

$$\mathbf{x}(t) = \sum_{i=1}^n c_i(t) \mathbf{e}_i$$

1.10.6 Canonical representation

Consider the state-space representation given in (1.66), namely

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \quad (1.66a)$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u} \quad (1.66b)$$

Applying the transformation

$$\mathbf{x} = \mathbf{T}\mathbf{z}$$

where \mathbf{T} is a non-singular matrix, leads to

$$\dot{\mathbf{T}}\mathbf{z} = \mathbf{A}\mathbf{T}\mathbf{z} + \mathbf{B}\mathbf{u}$$

$$\mathbf{y} = \mathbf{C}\mathbf{T}\mathbf{z} + \mathbf{D}\mathbf{u}$$

which may be written in the form

$$\dot{\mathbf{z}} = \tilde{\mathbf{A}}\mathbf{z} + \tilde{\mathbf{B}}\mathbf{u} \quad (1.94a)$$

$$\mathbf{y} = \tilde{\mathbf{C}}\mathbf{z} + \tilde{\mathbf{D}}\mathbf{u} \quad (1.94b)$$

where \mathbf{z} is now a state vector and

$$\tilde{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \quad \tilde{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B}, \quad \tilde{\mathbf{C}} = \mathbf{C}\mathbf{T}, \quad \tilde{\mathbf{D}} = \mathbf{D}$$

The system input–output relationship is unchanged by the transformation (see Section 5.6.3), and the linear systems (1.66) and (1.94) are said to be **equivalent**. By the transformation the intrinsic properties of the system, such as stability, controllability and observability, which are of interest to the engineer, are preserved, and there is merit in seeking a transformation leading to a system that is more easily analysed.

Since the transformation matrix \mathbf{T} can be arbitrarily chosen, an infinite number of equivalent systems exist. Of particular interest is the case when \mathbf{T} is taken to be the modal matrix \mathbf{M} of the system matrix \mathbf{A} ; that is,

$$\mathbf{T} = \mathbf{M} = [\mathbf{e}_1 \quad \mathbf{e}_2 \quad \dots \quad \mathbf{e}_n]$$

where \mathbf{e}_i ($i = 1, 2, \dots, n$) are the eigenvectors of the matrix \mathbf{A} . Under the assumption that the n eigenvalues are distinct,

$$\tilde{\mathbf{A}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M} = \mathbf{A}, \quad \text{the spectral matrix of } \mathbf{A}$$

$$\tilde{\mathbf{B}} = \mathbf{M}^{-1}\mathbf{B}$$

$$\tilde{\mathbf{C}} = \mathbf{C}\mathbf{M}, \quad \tilde{\mathbf{D}} = \mathbf{D}$$

so that (1.94) becomes

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z} + \mathbf{M}^{-1}\mathbf{B}\mathbf{u} \quad (1.95a)$$

$$\mathbf{y} = \mathbf{C}\mathbf{M}\mathbf{z} + \mathbf{D}\mathbf{u} \quad (1.95b)$$

Equation (1.95a) constitutes a system of uncoupled linear differential equations

$$\dot{z}_i = \lambda_i z_i + b_i^T \mathbf{u} \quad (i = 1, 2, \dots, n) \quad (1.96)$$

where $\mathbf{z} = (z_1, z_2, \dots, z_n)^T$ and \mathbf{b}_i^T is the i th row of the matrix $\mathbf{M}^{-1}\mathbf{B}$. Thus, by reducing (1.66) to the equivalent form (1.95) using the transformation $\mathbf{x} = \mathbf{M}\mathbf{z}$, the modes of the system have been uncoupled, with the new state variables z_i ($i = 1, 2, \dots, n$) being associated with the i th mode only. The representation (1.95) is called the **normal or canonical representation** of the system equations.

From (1.75), the solution of (1.96) is

$$z_i = e^{\lambda_i(t-t_0)} x(t_0) + \int_{t_0}^t e^{\lambda_i(t-\tau)} \mathbf{b}_i^T \mathbf{u}(\tau) d\tau \quad (i = 1, \dots, n)$$

so that the solution of (1.95a) may be written as

$$\mathbf{z}(t) = \mathbf{e}^{\mathbf{A}(t-t_0)} \mathbf{z}(t_0) + \int_{t_0}^t \mathbf{e}^{\mathbf{A}(t-\tau)} \mathbf{M}^{-1} \mathbf{B} \mathbf{u}(\tau) d\tau \quad (1.97)$$

where

$$\mathbf{e}^{\mathbf{A}(t-t_0)} = \begin{bmatrix} e^{\lambda_1(t-t_0)} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n(t-t_0)} \end{bmatrix}$$

In terms of the original state vector $\mathbf{x}(t)$, (1.97) becomes

$$\mathbf{x}(t) = \mathbf{M}\mathbf{z} = \mathbf{M} \mathbf{e}^{\mathbf{A}(t-t_0)} \mathbf{M}^{-1} \mathbf{x}(t_0) + \int_{t_0}^t \mathbf{M} \mathbf{e}^{\mathbf{A}(t-\tau)} \mathbf{M}^{-1} \mathbf{B} \mathbf{u}(\tau) d\tau \quad (1.98)$$

and the system response is then obtained from (1.66b) as

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)$$

By comparing the response (1.98) with that in (1.78), we note that the transition matrix may be written as

$$\Phi(t, t_0) = \mathbf{e}^{\mathbf{A}(t-t_0)} = \mathbf{M} \mathbf{e}^{\mathbf{A}(t-t_0)} \mathbf{M}^{-1}$$

The representation (1.95) may be used to readily infer some system properties. If the system is stable then each mode must be stable, so, from (1.98), each λ_i ($i = 1, 2, \dots, n$) must have a negative real part. If, for example, the j th row of the matrix $\mathbf{M}^{-1}\mathbf{B}$ is zero then, from (1.96), $\dot{z}_j = \lambda_j z_j + 0$, so the input $\mathbf{u}(t)$ has no influence on the j th mode of the system, and the mode is said to be **uncontrollable**. A system is said to be **controllable** if all of its modes are controllable.

If the j th column of the matrix \mathbf{CM} is zero then, from (1.95b), the response y is independent of z_j , so it is not possible to use information about the output to identify z_j . The state z_j is then said to be **unobservable**, and the overall system is not **observable**.

Example 1.44

A third-order system is characterized by the state-space model

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -5 & -6 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ -3 \\ 18 \end{bmatrix} \mathbf{u}, \quad \mathbf{y} = [1 \ 0 \ 0] \mathbf{x}$$

where $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$. Obtain the equivalent canonical representation of the model and then obtain the response of the system to a unit step $u(t) = H(t)$ given that initially $\mathbf{x}(0) = [1 \ 1 \ 0]^T$.

Solution The eigenvalues of the matrix

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

are determined by

$$|\mathbf{A} - \lambda \mathbf{I}| = \begin{vmatrix} -\lambda & 1 & 0 \\ 0 & -\lambda & 1 \\ 0 & -5 & -6-\lambda \end{vmatrix} = 0$$

that is,

$$\lambda(\lambda^2 + 6\lambda + 5) = 0$$

giving $\lambda_1 = 0$, $\lambda_2 = -1$ and $\lambda_3 = -5$, with corresponding eigenvectors

$$\mathbf{e}_1 = [1 \ 0 \ 0]^T, \quad \mathbf{e}_2 = [1 \ -1 \ 1]^T, \quad \mathbf{e}_3 = [1 \ -5 \ 25]^T$$

The corresponding modal and spectral matrices are

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -5 \\ 0 & 1 & 25 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -5 \end{bmatrix}$$

and the inverse modal matrix is determined to be

$$\mathbf{M}^{-1} = \frac{1}{20} \begin{bmatrix} 20 & 25 & 4 \\ 0 & -25 & -5 \\ 0 & 1 & 1 \end{bmatrix}$$

In this case $\mathbf{B} = [1 \ -3 \ 18]^T$, so

$$\mathbf{M}^{-1} \mathbf{B} = \frac{1}{20} \begin{bmatrix} 20 & 25 & 4 \\ 0 & -25 & -5 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -3 \\ 18 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 20 \\ -15 \\ 15 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{3}{4} \\ \frac{3}{4} \end{bmatrix}$$

Likewise, $\mathbf{C} = [1 \ 0 \ 0]$, giving

$$\mathbf{CM} = [1 \ 0 \ 0] \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -5 \\ 0 & 1 & 25 \end{bmatrix} = [1 \ 1 \ 1]$$

Thus, from (1.95), the equivalent canonical state-space representation is

$$\dot{\mathbf{z}} = \begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -5 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} + \begin{bmatrix} 1 \\ -\frac{3}{4} \\ \frac{3}{4} \end{bmatrix} u \quad (1.99a)$$

$$y = [1 \ 1 \ 1] \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \quad (1.99b)$$

When $u(t) = H(t)$, from (1.97) the solution of (1.99a) is

$$z = \begin{bmatrix} e^{0t} & 0 & 0 \\ 0 & e^{-t} & 0 \\ 0 & 0 & e^{-5t} \end{bmatrix} z(0) + \int_0^t \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-(t-\tau)} & 0 \\ 0 & 0 & e^{-5(t-\tau)} \end{bmatrix} \begin{bmatrix} 1 \\ -\frac{3}{4} \\ \frac{3}{4} \end{bmatrix} d\tau$$

where

$$z(0) = M^{-1}x(0) = \frac{1}{20} \begin{bmatrix} 20 & 24 & 4 \\ 0 & -25 & -5 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{44}{20} \\ -\frac{25}{20} \\ \frac{1}{20} \end{bmatrix}$$

leading to

$$\begin{aligned} z &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-t} & 0 \\ 0 & 0 & e^{-5t} \end{bmatrix} \begin{bmatrix} \frac{11}{5} \\ -\frac{5}{4} \\ \frac{1}{20} \end{bmatrix} + \int_0^t \begin{bmatrix} 1 \\ -\frac{3}{4}e^{-(t-\tau)} \\ \frac{3}{4}e^{-5(t-\tau)} \end{bmatrix} d\tau \\ &= \begin{bmatrix} \frac{11}{5} \\ -\frac{5}{4}e^{-t} \\ \frac{1}{20}e^{-5t} \end{bmatrix} + \begin{bmatrix} t \\ -\frac{3}{4} + \frac{3}{4}e^{-t} \\ \frac{3}{20} - \frac{3}{20}e^{-5t} \end{bmatrix} = \begin{bmatrix} t + \frac{11}{5} \\ -\frac{3}{4} - \frac{1}{2}e^{-t} \\ \frac{3}{20} - \frac{1}{10}e^{-5t} \end{bmatrix} \end{aligned}$$

Then, from (1.99b),

$$\begin{aligned} y &= z_1 + z_2 + z_3 = \left(t + \frac{11}{5} \right) + \left(-\frac{3}{4} - \frac{1}{2}e^{-t} \right) + \left(\frac{3}{20} - \frac{1}{10}e^{-5t} \right) \\ &= t + \frac{8}{5} - \frac{1}{2}e^{-t} - \frac{1}{10}e^{-5t} \end{aligned}$$

If we drop the assumption that the eigenvalues of \mathbf{A} are distinct then $\tilde{\mathbf{A}} = \mathbf{M}^{-1}\mathbf{A}\mathbf{M}$ is no longer diagonal, but may be represented by the corresponding Jordan canonical form \mathbf{J} with \mathbf{M} being made up of both eigenvectors and generalized eigenvectors of \mathbf{A} . The equivalent canonical form in this case will be

$$\dot{z} = \mathbf{J}z + \mathbf{M}^{-1}\mathbf{B}u$$

$$y = \mathbf{C}\mathbf{M}z + \mathbf{D}u$$

with the solution corresponding to (1.97) being

$$x(t) = \mathbf{M} e^{\mathbf{J}(t-t_0)} \mathbf{M}^{-1} x(t_0) + \int_{t_0}^t \mathbf{M} e^{\mathbf{J}(t-\tau)} \mathbf{M}^{-1} \mathbf{B} u(\tau) d\tau$$

1.10.7 Exercises

- 61 Obtain in spectral form the response of the unforced second-order system

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -\frac{3}{2} & \frac{3}{4} \\ 1 & -\frac{5}{2} \end{bmatrix} \mathbf{x}(t), \\ \mathbf{x}(0) &= \begin{bmatrix} 2 \\ 4 \end{bmatrix}\end{aligned}$$

Using the eigenvectors as the frame of reference, sketch the trajectory.

- 62 Using the spectral form of the solution given in (1.91), solve the second-order system

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} -2 & 2 \\ 2 & -5 \end{bmatrix} \mathbf{x}(t), \quad \mathbf{x}(0) = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

and sketch the trajectory.

- 63 Repeat Exercise 61 for the system

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & -4 \\ 2 & -4 \end{bmatrix} \mathbf{x}(t), \quad \mathbf{x}(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

- 64 Determine the equivalent canonical representation of the third-order system

$$\begin{aligned}\dot{\mathbf{x}} &= \begin{bmatrix} 1 & 1 & -2 \\ -1 & 2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix} \mathbf{u} \\ \mathbf{y} &= [-2 \ 1 \ 0] \mathbf{x}\end{aligned}$$

- 65 The solution of a third-order linear system is given by

$$\mathbf{x} = \alpha_0 e^{-t} \mathbf{e}_0 + \alpha_1 e^{-2t} \mathbf{e}_1 + \alpha_2 e^{-3t} \mathbf{e}_2$$

where $\mathbf{e}_0, \mathbf{e}_1$ and \mathbf{e}_2 are linearly independent vectors having values

$$\begin{aligned}\mathbf{e}_0 &= [1 \ 1 \ 0]^T, \quad \mathbf{e}_1 = [0 \ 1 \ 1]^T, \\ \mathbf{e}_2 &= [1 \ 2 \ 3]^T\end{aligned}$$

Initially, at time $t = 0$ the system state is $\mathbf{x}(0) = [1 \ 1 \ 1]^T$. Find α_0, α_1 and α_2 using the reciprocal basis method.

- 66 Obtain the eigenvalues and eigenvectors of the matrix

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

Using a suitable transformation $\mathbf{x}(t) = \mathbf{Mz}(t)$, reduce $\dot{\mathbf{x}}(t) = \mathbf{Ax}(t)$ to the canonical form $\dot{\mathbf{z}}(t) = \mathbf{Az}(t)$, where \mathbf{A} is the spectral matrix of \mathbf{A} . Solve the decoupled canonical form for \mathbf{z} , and hence solve for $\mathbf{x}(t)$ given that $\mathbf{x}(0) = [1 \ 4]^T$.

- 67 A second-order system is governed by the state equation

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{u}(t) \quad (t \geq 0)$$

Using a suitable transformation $\mathbf{x}(t) = \mathbf{Mz}(t)$, reduce this to the canonical form

$$\dot{\mathbf{z}}(t) = \mathbf{Az}(t) + \mathbf{Bu}(t)$$

where \mathbf{A} is the spectral matrix of

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

and \mathbf{B} is a suitable 2×2 matrix.

For the input $\mathbf{u}(t) = [4 \ 3]^T$ solve the decoupled canonical form for \mathbf{z} , and hence solve for $\mathbf{x}(t)$ given that $\mathbf{x}(0) = [1 \ 2]^T$. Compare the answer with that for Exercise 59.

In Chapter 5 (in particular Section 5.4) we shall consider the solution of state-space models using the Laplace transform method. If you are unfamiliar with Laplace transforms, see Chapter 11 of MEM. Chapter 6 extends the analysis to discrete-time systems using z transforms.

1.11 Engineering application: Lyapunov stability analysis

The Russian mathematician Aleksandr Mikhailovich Lyapunov (1857–1918) developed an approach to stability analysis which is now referred to as the direct (or second) method of Lyapunov. His approach remained almost unknown in the English-speaking world for around half a century, before it was translated into English in the late 1950s. Publication of Lyapunov's work in English aroused great interest, and it is now widely used for stability analysis of linear and nonlinear systems, both time-invariant and time-varying. Also, the approach has proved to be a useful tool in system design such as, for example, in the design of stable adaptive control systems. The Lyapunov method is in fact a ‘method of approach’ rather than a systematic means of investigating stability and much depends on the ingenuity of the user in obtaining suitable Lyapunov functions. There is no unique Lyapunov function for a given system.

In this section we briefly introduce the Lyapunov approach and will restrict consideration to the unforced (absence of any input) linear time-invariant system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} \quad (1.100)$$

where $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$ is the n -state vector and \mathbf{A} is a constant $n \times n$ matrix. For the linear system (1.100) the origin $\mathbf{x} = \mathbf{0}$ is the only point of equilibrium. If, for any initial state $\mathbf{x}(0)$, the trajectory (solution path) $\mathbf{x}(t)$ of the system approaches zero (the equilibrium point) as $t \rightarrow \infty$ then the system is said to be **asymptotically stable**. In practice the elements of the matrix \mathbf{A} may include system parameters and we are interested in determining what constraints, if any, must be placed on these parameters to ensure system stability. Stability of (1.100) is further discussed in Section 5.6.1, where algebraic criteria for stability are presented. In particular, it is shown that stability of system (1.100) is ensured if and only if all the eigenvalues of the state matrix \mathbf{A} have negative real parts.

To develop the Lyapunov approach we set up a nest of closed surfaces, around the origin (equilibrium point), defined by the scalar function

$$V(\mathbf{x}) = V(x_1, x_2, \dots, x_n) = C \quad (1.101)$$

where C is a **positive constant** (the various surfaces are obtained by increasing the values of C as we move away from the origin). If the function $V(\mathbf{x})$ satisfies the following conditions:

- (a) $V(\mathbf{x}) = 0$ at the origin, that is $V(0) = 0$;
- (b) $V(\mathbf{x}) > 0$ away from the origin;
- (c) $V(\mathbf{x})$ is continuous with continuous partial derivatives;

then it is called a **scalar Lyapunov function**. (Note that conditions (a) and (b) together ensure that $V(\mathbf{x})$ is a **positive-definite function**.) We now consider the rate of change of $V(\mathbf{x})$, called the Eulerian derivative of $V(\mathbf{x})$ and denoted by $\dot{V}(\mathbf{x})$, along the trajectory of the system under investigation; that is,

$$\dot{V}(\mathbf{x}) = \frac{\partial V}{\partial x_1} \frac{dx_1}{dt} + \frac{\partial V}{\partial x_2} \frac{dx_2}{dt} + \dots + \frac{\partial V}{\partial x_n} \frac{dx_n}{dt} \quad (1.102)$$

where the values of $\dot{x}_1, \dot{x}_2, \dots, \dot{x}_n$ are substituted from the given equations representing the system ((1.100) in the case of the linear equations under consideration).

If \dot{V} satisfies the condition

- (d) $\dot{V}(\mathbf{x})$ is negative definite

then it follows that all the trajectories cross the surfaces $V(\mathbf{x}) = C$ in an inward direction and must tend to the origin, the position of equilibrium. Thus asymptotic stability has

been assured without having to solve the differential equations representing the system. The function $V(\mathbf{x})$ which satisfies conditions (a)–(d) is called a **Lyapunov function for the system being considered**.

If we start with a positive-definite $V(\mathbf{x})$ and impose conditions on $\dot{V}(\mathbf{x})$ to be negative-definite, then these conditions will provide sufficient but **not** necessary stability criteria, and in many cases they may be unduly restrictive. However, if we are able to start with a negative-definite $\dot{V}(\mathbf{x})$ and work back to impose conditions on $V(\mathbf{x})$ to be positive-definite, then these conditions **provide necessary and sufficient stability criteria**. This second procedure is far more difficult to apply than the first, although it may be applied in certain cases, and in particular to linear systems.

Of particular importance as Lyapunov functions for linear systems are **quadratic forms** in the variables x_1, x_2, \dots, x_n which were introduced in Section 1.6.4. These may be written in the matrix form $V(\mathbf{x}) = \mathbf{x}^T \mathbf{P} \mathbf{x}$, where \mathbf{P} is a real symmetric matrix. Necessary and sufficient conditions for $V(\mathbf{x})$ to be positive-definite are provided by **Sylvester's criterion**, which states that all the principal minors of \mathbf{P} of order $1, 2, \dots, n$ must be positive; that is

$$p_{11} > 0, \begin{vmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{vmatrix} > 0, \begin{vmatrix} p_{11} & p_{12} & p_{13} \\ p_{12} & p_{22} & p_{23} \\ p_{13} & p_{23} & p_{33} \end{vmatrix} > 0, \dots, |\mathbf{P}| > 0$$

Returning to the linear system (1.100) let us consider as a tentative Lyapunov function the quadratic form

$$V(\mathbf{x}) = \mathbf{x}^T \mathbf{P} \mathbf{x}$$

where \mathbf{P} is an $n \times n$ real symmetric matrix. To obtain the Eulerian derivative of $V(\mathbf{x})$ with respect to system (1.103) we first differentiate $V(\mathbf{x})$ with respect to t

$$\frac{dV}{dt} = \dot{\mathbf{x}}^T \mathbf{P} \mathbf{x} + \mathbf{x}^T \mathbf{P} \dot{\mathbf{x}}$$

and then substitute for $\dot{\mathbf{x}}^T$ and $\dot{\mathbf{x}}$ from (1.100) giving

$$\begin{aligned} \dot{V}(\mathbf{x}) &= (\mathbf{A}\mathbf{x})^T \mathbf{P} \mathbf{x} + \mathbf{x}^T \mathbf{P} (\mathbf{A}\mathbf{x}) \\ \text{that is } \dot{V}(\mathbf{x}) &= \mathbf{x}^T (\mathbf{A}^T \mathbf{P} + \mathbf{P} \mathbf{A}) \mathbf{x} \end{aligned}$$

or alternatively

$$\dot{V}(\mathbf{x}) = -\mathbf{x}^T \mathbf{Q} \mathbf{x} \tag{1.103}$$

$$\text{where } -\mathbf{Q} = \mathbf{A}^T \mathbf{P} + \mathbf{P} \mathbf{A} \tag{1.104}$$

To obtain necessary and sufficient conditions for the stability of the linear system (1.100) we start with any negative definite quadratic form $-\mathbf{x}^T \mathbf{Q} \mathbf{x}$, with an $n \times n$ symmetric matrix \mathbf{Q} , and solve matrix equation (1.104) for the elements of \mathbf{P} . The conditions imposed on \mathbf{P} to ensure that it is positive-definite then provide the required necessary and sufficient stability criteria.

Example 1.45

The vector-matrix differential equation model representing an unforced linear $R-C$ circuit is

$$\dot{\mathbf{x}} = \begin{bmatrix} -4\alpha & 4\alpha \\ 2\alpha & -6\alpha \end{bmatrix} \mathbf{x} \tag{i}$$

Examine its stability using the Lyapunov approach.

Solution Take \mathbf{Q} of equation (1.104) to be the identity matrix \mathbf{I} which is positive-definite (thus $-\mathbf{Q}$ is negative-definite). Then (1.104) may be written

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} -4\alpha & 2\alpha \\ 4\alpha & -6\alpha \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} -4\alpha & 4\alpha \\ 2\alpha & -6\alpha \end{bmatrix} \quad (\text{ii})$$

Equating elements in (ii) gives

$$-8\alpha p_{11} + 4\alpha p_{12} = -1, 4\alpha p_{11} - 10\alpha p_{12} + 2\alpha p_{22} = 0, 8\alpha p_{12} - 12\alpha p_{22} = -1$$

Solving for the elements gives

$$p_{11} = \frac{7}{40\alpha}, p_{12} = \frac{1}{10\alpha}, p_{22} = \frac{3}{20\alpha}$$

so that

$$\mathbf{P} = \frac{1}{40\alpha} \begin{bmatrix} 7 & 4 \\ 4 & 6 \end{bmatrix}$$

The principal minors of $\begin{bmatrix} 7 & 4 \\ 4 & 6 \end{bmatrix}$ are $|7| > 0$ and $|7 \ 4 \ 6| = 26 > 0$.

Thus, by Sylvester's criterion, \mathbf{P} is positive-definite and the system is asymptotically stable provided $\alpha > 0$. Note that the Lyapunov function in this case was

$$V(\mathbf{x}) = \mathbf{x}^T \mathbf{P} \mathbf{x} = \frac{1}{40\alpha} (7x_1^2 + 8x_1x_2 + 6x_2^2)$$

1.11.1 Exercises

- 68 Using the Lyapunov approach investigate the stability of the system described by the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} -4 & 2 \\ 3 & -2 \end{bmatrix} \mathbf{x}$$

Take \mathbf{Q} to be the unit matrix. Confirm your answer by determining the eigenvalues of the state matrix.

- 69 Repeat Exercise 67 for the system described by the state equation

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

- 70 For the system modelled by the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -a & -b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

use the Lyapunov approach to determine the constraints on the parameters a and b that yield necessary and sufficient conditions for asymptotic stability.

Condition (d) in the formulation of a Lyapunov function, requiring $\dot{V}(\mathbf{x})$ to be positive-definite, may be relaxed to $\dot{V}(\mathbf{x})$ being positive-semidefinite provided $\dot{V}(\mathbf{x})$ is not identically zero along any trajectory. A third-order system, in the absence of an input, is modelled by the state equation

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

where $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$ and

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 1 \\ -k & 0 & -1 \end{bmatrix} \text{ with } k \text{ being a constant scalar.}$$

It is required to use the Lyapunov approach to determine the constraints on k to ensure asymptotic stability.

- (a) In (1.103) choose \mathbf{Q} to be the positive-semidefinite matrix

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

so that

$$\dot{V}(\mathbf{x}) = -\mathbf{x}^T \mathbf{Q} \mathbf{x} = -x_3^2$$

Verify that $M(\mathbf{x})$ is identically zero only at the origin (equilibrium point) and is therefore not identically zero along any trajectory.

- (b) Using this matrix \mathbf{Q} solve the matrix equation

$$\mathbf{A}^T \mathbf{P} + \mathbf{P} \mathbf{A} = -\mathbf{Q}$$

to determine the matrix \mathbf{P} .

- (c) Using Sylvester's criterion show that the system is asymptotically stable for $0 < k < 6$.

72

- A feedback control system modelled by the differential equation

$$\ddot{x} + a\dot{x} + kx = 0$$

is known to be asymptotically stable, for $k > 0$, $a > 0$. Set up the state-space form of the equation and show that

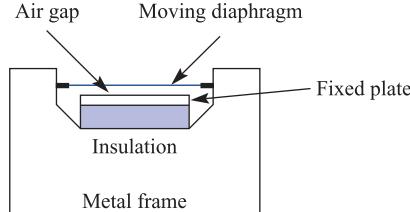
$$V(x_1, x_2) = kx_1^2 + (x_2 + ax_1)^2, x_1 = x, x_2 = \dot{x}$$

is a suitable Lyapunov function for verifying this.

1.12 Engineering application: capacitor microphone

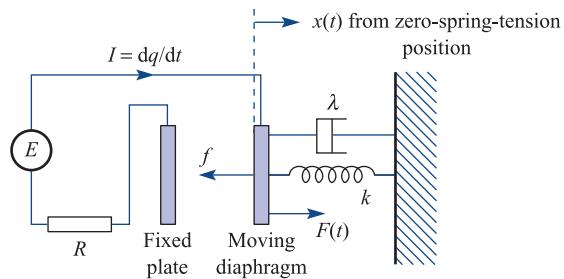
Many smaller portable tape recorders have a capacitor microphone built in, since such a system is simple and robust. It works on the principle that if the distance between the plates of a capacitor changes then the capacitance changes in a known manner, and these changes induce a current in an electric circuit. This current can then be amplified or stored. The basic system is illustrated in Figure 1.9. There is a small air gap (about 0.02 mm) between the moving diaphragm and the fixed plate. Sound waves falling on the diaphragm cause vibrations and small variations in the capacitance C ; these are certainly sufficiently small that the equations can be *linearized*.

Figure 1.9 Capacitor microphone.



We assume that the diaphragm has mass m and moves as a single unit so that its motion is one-dimensional. The housing of the diaphragm is modelled as a spring-and-dashpot system. The plates are connected through a simple circuit containing a resistance and an imposed steady voltage from a battery. Figure 1.10 illustrates the model. The distance $x(t)$ is measured from the position of zero spring tension, F is the imposed force and f is the force required to hold the moving plate in position against the electrical attraction. The mechanical motion is governed by Newton's equation

Figure 1.10 Capacitor microphone model.



$$m\ddot{x} = -kx - \lambda\dot{x} - f + F \quad (1.105)$$

and the electrical circuit equation gives

$$E = RI + \frac{q}{C}, \quad \text{with} \quad \frac{dq}{dt} = I \quad (1.106)$$

The variation of capacitance C with x is given by the standard formula

$$C = \frac{C_0 a}{a + x}$$

where a is the equilibrium distance between the plates. The force f is not so obvious, but the following assumption is standard

$$f = \frac{1}{2} q^2 \frac{d}{dx} \left(\frac{1}{C} \right) = \frac{1}{2} \frac{q^2}{C_0 a}$$

It is convenient to write the equations in the first-order form

$$\dot{x} = v$$

$$m\dot{v} = -kx - \lambda v - \frac{1}{2} \frac{q^2}{C_0 a} + F(t)$$

$$R\dot{q} = -\frac{q(a+x)}{aC_0} + E$$

Furthermore, it is convenient to non-dimensionalize the equations. While it is obvious how to do this for the distance and velocity, for the time and the charge it is less so. There are three natural time scales in the problem: the electrical time $\tau_1 = RC_0$, the spring time $\tau_2^2 = m/k$ and the damping time $\tau_3 = m/\lambda$. Choosing to non-dimensionalize the time with respect to τ_1 , the non-dimensionalization of the charge follows:

$$\tau = \frac{t}{\tau_1}, \quad X = \frac{x}{a}, \quad V = \frac{v}{ka/\lambda}, \quad Q = \frac{q}{\sqrt{(2C_0ka^2)}}$$

Then, denoting differentiation with respect to τ by a prime, the equations are

$$X' = \frac{RC_0 k}{\lambda} V$$

$$\frac{m}{\lambda R C_0} V' = -X - V - Q^2 + \frac{F}{ka}$$

$$Q' = -Q(1 + X) + \frac{EC_0}{\sqrt{(2C_0ka^2)}}$$

There are four non-dimensional parameters: the external force divided by the spring force gives the first, $G = F/ka$; the electrical force divided by the spring force gives the second, $D^2 = (E^2 C_0 / 2a) / ka$; and the remaining two are

$$A = \frac{RC_0 k}{\lambda} = \frac{\tau_1 \tau_3}{\tau_2^2}, \quad B = \frac{m}{\lambda R C_0} = \frac{\tau_3}{\tau_1}$$

The final equations are therefore

$$\left. \begin{aligned} X' &= AV \\ BV' &= -X - V - Q^2 + G \\ Q' &= -Q(1 + X) + D \end{aligned} \right\} \quad (1.107)$$

In equilibrium, with no driving force, $G = 0$ and $V = X' = V' = Q' = 0$, so that

$$\left. \begin{aligned} Q^2 + X &= 0 \\ Q(1 + X) - D &= 0 \end{aligned} \right\} \quad (1.108)$$

or, on eliminating Q ,

$$X(1 + X)^2 = -D^2$$

From Figure 1.11, we see that there is always one solution for $X < -1$, or equivalently $x < -a$. The implication of this solution is that the plates have crossed. This is clearly impossible, so the solution is discarded on physical grounds. There are two other solutions if

$$D^2 < \frac{1}{3} \left(\frac{4}{3}\right)^2 = \frac{4}{27}$$

Figure 1.11 Solutions to equations (1.108).

or

$$\frac{E^2 C_0}{2ka^2} < \frac{4}{27} \quad (1.109)$$

We can interpret this statement as saying that the electrical force must not be too strong, and (1.109) gives a precise meaning to what ‘too strong’ means. There are two physically satisfactory equilibrium solutions $-\frac{1}{3} < X_1 < 0$ and $-1 < X_2 < -\frac{1}{3}$, and the only question left is whether they are stable or unstable.

Stability is determined by small oscillations about the two values X_1 and X_2 , where these values satisfy (1.108). Writing

$$X = X_i + \varepsilon, \quad Q = Q_i + \eta, \quad V = \theta$$

and substituting into (1.107), neglecting terms in $\varepsilon^2, y^2, \theta^2, \varepsilon\theta$ and so on, gives

$$\left. \begin{aligned} \varepsilon' &= A\theta \\ B\theta' &= -\varepsilon - \theta - 2Q_i\eta \\ \eta' &= (-Q_i\varepsilon - (1 + X_i)\eta) \end{aligned} \right\} \quad (1.110)$$

Equations (1.110) are the linearized versions of (1.107) about the equilibrium values. To test for stability, we put $G = 0$ and $\varepsilon = L e^{\alpha\tau}, \theta = M e^{\alpha\tau}, \eta = N e^{\alpha\tau}$ into (1.110):

$$L\alpha = AM$$

$$BM\alpha = -L - M - 2Q_iN$$

$$N\alpha = -Q_iL - (1 + X_i)N$$

which can be written in the matrix form

$$\alpha \begin{bmatrix} L \\ M \\ N \end{bmatrix} = \begin{bmatrix} 0 & A & 0 \\ -1/B & -1/B & -2Q_i/B \\ -Q_i & 0 & -(1+X_i) \end{bmatrix} \begin{bmatrix} L \\ M \\ N \end{bmatrix}$$

Thus the fundamental stability problem is an eigenvalue problem, a result common to all vibrational stability problems. The equations have non-trivial solutions if

$$\begin{aligned} 0 &= \begin{vmatrix} -\alpha & A & 0 \\ -1/B & -(1/B) - \alpha & -2Q_i/B \\ -Q_i & 0 & -(1+X_i) - \alpha \end{vmatrix} \\ &= -[B\alpha^3 + (B(1+X_i) + 1)\alpha^2 + (1+X_i+A)\alpha + A(1+X_i - 2Q_i^2)]/B \end{aligned}$$

For stability, α must have a negative real part, so that the vibrations damp out, and the Routh–Hurwitz criterion (Section 5.3.2) gives the conditions for this to be the case. Each of the coefficients must be positive, and for the first three

$$B > 0, \quad B(1+X_i) + 1 > 0, \quad 1+X_i+A > 0$$

are obviously satisfied since $-1 < X_i < 0$. The next condition is

$$A(1+X_i - 2Q_i^2) > 0$$

which, from (6.118), gives

$$1+3X_i > 0, \quad \text{or} \quad X_i > -\frac{1}{3}$$

Thus the only solution that can possibly be stable is the one for which $X_i > -\frac{1}{3}$; the other solution is unstable. There is one final condition to check,

$$[B(1+X_i) + 1](1+X_i+A) - BA(1+X_i - 2Q_i^2) > 0$$

or

$$B(1+X_i)^2 + 1+X_i+A + 2BAQ_i^2 > 0$$

Since all the terms are positive, the solution $X_i > -\frac{1}{3}$ is indeed a stable solution.

Having established the stability of one of the positions of the capacitor diaphragm, the next step is to look at the response of the microphone to various inputs. The characteristics can most easily be checked by looking at the frequency response, which is the system response to an individual input $G = b e^{j\omega t}$, as the frequency ω varies. This will give information of how the electrical output behaves and for which range of frequencies the response is reasonably flat.

The essential point of this example is to show that a practical vibrational problem gives a stability problem that involves eigenvalues and a response that involves a matrix inversion. The same behaviour is observed for more complicated vibrational problems.

1.13 Review exercises (1–19)



Check your answers using MATLAB or MAPLE whenever possible.

- 1 Obtain the eigenvalues and corresponding eigenvectors of the matrices

$$(a) \begin{bmatrix} -1 & 6 & 12 \\ 0 & -13 & 30 \\ 0 & -9 & 20 \end{bmatrix}$$

$$(b) \begin{bmatrix} 2 & 0 & 1 \\ -1 & 4 & -1 \\ -1 & 2 & 0 \end{bmatrix}$$

$$(c) \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

- 2 Find the principal stress values (eigenvalues) and the corresponding principal stress directions (eigenvectors) for the stress matrix

$$\boldsymbol{T} = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{bmatrix}$$

Verify that the principal stress directions are mutually orthogonal.

- 3 Find the values of b and c for which the matrix

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

has $[1 \ 0 \ 1]^T$ as an eigenvector. For these values of b and c calculate all the eigenvalues and corresponding eigenvectors of the matrix \boldsymbol{A} .

- 4 (a) Using the power method find the dominant eigenvalue and the corresponding eigenvector of the matrix

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2.5 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

starting with an initial vector $[1 \ 1 \ 1]^T$ and working to 3 decimal places.

- (b) Given that another eigenvalue of \boldsymbol{A} is 1.19 correct to 2 decimal places, find the value of the

- third eigenvalue using a property of matrices.
(c) Having determined all the eigenvalues of \boldsymbol{A} , indicate which of these can be obtained by using the power method on the following matrices: (i) \boldsymbol{A}^{-1} ; (ii) $\boldsymbol{A} - 3\boldsymbol{I}$.

- 5 Consider the differential equations

$$\frac{dx}{dt} = 4x + y + z$$

$$\frac{dy}{dt} = 2x + 5y + 4z$$

$$\frac{dz}{dt} = -x - y$$

Show that if it is assumed that there are solutions of the form $x = \alpha e^{\lambda t}$, $y = \beta e^{\lambda t}$ and $z = \gamma e^{\lambda t}$ then the system of equations can be transformed into the eigenvalue problem

$$\begin{bmatrix} 4 & 1 & 1 \\ 2 & 5 & 4 \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \lambda \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}$$

Show that the eigenvalues for this problem are 5, 3 and 1, and find the eigenvectors corresponding to the smallest eigenvalue.

- 6 Find the eigenvalues and corresponding eigenvectors for the matrix

$$\boldsymbol{A} = \begin{bmatrix} 8 & -8 & -2 \\ 4 & -3 & -2 \\ 3 & -4 & 1 \end{bmatrix}$$

Write down the modal matrix \boldsymbol{M} and spectral matrix $\boldsymbol{\Lambda}$ of \boldsymbol{A} , and confirm that

$$\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{M} = \boldsymbol{\Lambda}$$

- 7 Show that the eigenvalues of the symmetric matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -4 \\ 0 & 5 & 4 \\ -4 & 4 & 3 \end{bmatrix}$$

are 9, 3 and -3. Obtain the corresponding eigenvectors in normalized form, and write down the normalized modal matrix $\hat{\boldsymbol{M}}$. Confirm that

$$\hat{\boldsymbol{M}}^T \boldsymbol{A} \hat{\boldsymbol{M}} = \boldsymbol{\Lambda}$$

where $\boldsymbol{\Lambda}$ is the spectral matrix of \boldsymbol{A} .

- 8** In a radioactive series consisting of four different nuclides starting with the parent substance N_1 and ending with the stable product N_4 the amounts of each nuclide present at time t are given by the differential equations model

$$\frac{dN_1}{dt} = -6N_1$$

$$\frac{dN_2}{dt} = 6N_1 - 4N_2$$

$$\frac{dN_3}{dt} = 4N_2 - 2N_3$$

$$\frac{dN_4}{dt} = 2N_3$$

Express these in the vector-matrix form

$$\dot{\mathbf{N}} = \mathbf{A}\mathbf{N}$$

where $\mathbf{N} = [N_1 \ N_2 \ N_3 \ N_4]^T$. Find the eigenvalues and corresponding eigenvectors of \mathbf{A} . Using the spectral form of the solution, determine $N_4(t)$ given that at time $t = 0$, $N_1 = C$ and $N_2 = N_3 = N_4 = 0$.

- 9** (a) Given

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

use the Cayley–Hamilton theorem to find

- (i) $\mathbf{A}^7 - 3\mathbf{A}^6 + \mathbf{A}^4 + 3\mathbf{A}^3 - 2\mathbf{A}^2 + 3\mathbf{I}$
(ii) \mathbf{A}^k , where $k > 0$ is an integer.

- (b) Using the Cayley–Hamilton theorem, find $e^{\mathbf{A}t}$ when

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

- 10** Show that the matrix

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

has an eigenvalue $\lambda = 1$ with algebraic multiplicity 3. By considering the rank of a suitable matrix, show that there is only one corresponding linearly independent eigenvector \mathbf{e}_1 . Obtain the eigenvector \mathbf{e}_1 and two further generalized eigenvectors. Write down the corresponding modal matrix \mathbf{M} and confirm that

$\mathbf{M}^{-1}\mathbf{AM} = \mathbf{J}$, where \mathbf{J} is the appropriate Jordan matrix. (*Hint:* In this example care must be taken in applying the procedure to evaluate the generalized eigenvectors to ensure that the triad of vectors takes the form $\{\mathbf{T}^2\omega, \mathbf{T}\omega, \omega\}$, where $\mathbf{T} = \mathbf{A} - \lambda\mathbf{I}$, with $\mathbf{T}^2\omega = \mathbf{e}_1$.)

- 11**

The equations of motion of three equal masses connected by springs of equal stiffness are

$$\ddot{x} = -2x + y$$

$$\ddot{y} = x - 2y + z$$

$$\ddot{z} = y - 2z$$

Show that for normal modes of oscillation

$$x = X \cos \omega t, \quad y = Y \cos \omega t,$$

$$z = Z \cos \omega t$$

to exist then the condition on $\lambda = \omega^2$ is

$$\begin{vmatrix} \lambda - 2 & 1 & 0 \\ 1 & \lambda - 2 & 1 \\ 0 & 1 & \lambda - 2 \end{vmatrix} = 0$$

Find the three values of λ that satisfy this condition, and find the ratios $X : Y : Z$ in each case.

- 12**

Classify the following quadratic forms:

- (a) $2x^2 + y^2 + 2z^2 - 2xy - 2yz$
(b) $3x^2 + 7y^2 + 2z^2 - 4xy - 4xz$
(c) $16x^2 + 36y^2 + 17z^2 + 32xy + 32xz + 16yz$
(d) $-21x^2 + 30xy - 12xz - 11y^2 + 8yz - 2z^2$
(e) $-x^2 - 3y^2 - 5z^2 + 2xy + 2xz + 2yz$

- 13**

Show that $\mathbf{e}_1 = [1 \ 2 \ 3]^T$ is an eigenvector of the matrix

$$\mathbf{A} = \begin{bmatrix} \frac{7}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 4 & -1 & 0 \\ -\frac{3}{2} & \frac{3}{2} & \frac{1}{2} \end{bmatrix}$$

and find its corresponding eigenvalue. Find the other two eigenvalues and their corresponding eigenvectors.

Write down in spectral form the general solution of the system of differential equations

$$2\frac{dx}{dt} = 7x - y - z$$

$$\frac{dy}{dt} = 4x - y$$

$$2\frac{dz}{dt} = -3x + 3y + z$$

Hence show that if $x = 2$, $y = 4$ and $z = 6$ when $t = 0$ then the solution is

$$x = 2e^t, \quad y = 4e^t, \quad z = 6e^t$$

- 14 (a) Find the SVD form of the matrix

$$\mathbf{A} = \begin{bmatrix} 1.2 & 0.9 & -4 \\ 1.6 & 1.2 & 3 \end{bmatrix}$$

- (b) Use the SVD to determine the pseudo inverse \mathbf{A}^\dagger and confirm it is a right inverse of \mathbf{A} .
(c) Determine the pseudo inverse \mathbf{A}^\dagger without using the SVD.

- 15 From (1.48) the unitary matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ and sigma matrix Σ may be written in the partitioned form:

$$\hat{\mathbf{U}} = [\hat{\mathbf{U}}_r \ \hat{\mathbf{U}}_{m-r}], \hat{\mathbf{V}} = [\hat{\mathbf{V}}_r \ \hat{\mathbf{V}}_{n-r}], \Sigma = \begin{bmatrix} \mathbf{S} & 0 \\ 0 & 0 \end{bmatrix}$$

where \mathbf{S} is $r \times r$ diagonal matrix having the singular values of \mathbf{A} as its diagonal elements and 0 denotes zero matrices having appropriate order.

- (a) Show that the SVD form of \mathbf{A} may be expressed in the form

$$\mathbf{A} = \hat{\mathbf{U}}_r \mathbf{S} \hat{\mathbf{U}}_r^T$$

This is called the reduced singular value decomposition of \mathbf{A} .

- (b) Deduce that the pseudo inverse is given by

$$\mathbf{A}^\dagger = \hat{\mathbf{V}}_r \mathbf{S}^{-1} \hat{\mathbf{U}}_r^T$$

- (c) Use the results of (a) and (b) to determine the SVD form and pseudo inverse of the matrix

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

and check your answers with those obtained in Exercise 45.

16

A linear time-invariant system $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ is modelled by the state-space equations

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t)$$

$$\mathbf{y}(t) = \mathbf{c}^T \mathbf{x}(t)$$

where $\mathbf{x}(t)$ is the n -dimensional state vector, and $u(t)$ and $\mathbf{y}(t)$ are the system input and output respectively. Given that the system matrix \mathbf{A} has n distinct non-zero eigenvalues, show that the system equations may be reduced to the canonical form

$$\dot{\xi}(t) = \Lambda \xi(t) + \mathbf{b}_1 u(t)$$

$$\mathbf{y}(t) = \mathbf{c}_1^T \xi(t)$$

where Λ is a diagonal matrix. What properties of this canonical form determine the controllability and observability of $(\mathbf{A}, \mathbf{b}, \mathbf{c})$?

Reduce to canonical form the system $(\mathbf{A}, \mathbf{b}, \mathbf{c})$ having

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

$$\mathbf{b} = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

and comment on its stability, controllability and observability by considering the ranks of the appropriate Kalman matrices $[\mathbf{b} \ \mathbf{Ab} \ \mathbf{A}^2\mathbf{b}]$ and $[\mathbf{c} \ \mathbf{A}^T\mathbf{c} \ (\mathbf{A}^T)^2\mathbf{c}]$.

17

A third-order system is modelled by the state-space representation

$$\dot{\mathbf{x}} = \begin{bmatrix} -2 & -2 & 0 \\ 0 & 0 & 1 \\ 0 & -3 & -4 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{u}$$

where $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$ and $\mathbf{u} = [u_1 \ u_2]^T$.

Find the transformation $\mathbf{x} = \mathbf{M}\mathbf{z}$ which reduces the model to canonical form and solve for $\mathbf{x}(t)$ given $\mathbf{x}(0) = [10 \ 5 \ 2]^T$ and $\mathbf{u}(t) = [t \ 1]^T$.

- 18** The behaviour of an unforced mechanical system is governed by the differential equation

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 5 & 2 & -1 \\ 3 & 6 & -9 \\ 1 & 1 & 1 \end{bmatrix} \mathbf{x}(t), \quad \mathbf{x}(0) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

- (a) Show that the eigenvalues of the system matrix are 6, 3, 3 and that there is only one linearly independent eigenvector corresponding to the eigenvalue 3. Obtain the eigenvectors corresponding to the eigenvalues 6 and 3 and a further generalized eigenvector for the eigenvalue 3.
- (b) Write down a generalized modal matrix \mathbf{M} and confirm that

$$\mathbf{A}\mathbf{M} = \mathbf{M}\mathbf{J}$$

for an appropriate Jordan matrix \mathbf{J} .

- (c) Using the result

$$\mathbf{x}(t) = \mathbf{M}e^{\mathbf{J}t}\mathbf{M}^{-1}\mathbf{x}(0)$$

obtain the solution to the given differential equation.

- 19** (Extended problem) Many vibrational systems are modelled by the vector–matrix differential equation

$$\ddot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) \quad (1)$$

where \mathbf{A} is a constant $n \times n$ matrix and $\mathbf{x}(t) = [x_1(t) \ x_2(t) \ \dots \ x_n(t)]^T$. By substituting $\mathbf{x} = e^{\lambda t}\mathbf{u}$, show that

$$\lambda^2\mathbf{u} = \mathbf{A}\mathbf{u} \quad (2)$$

and that non-trivial solutions for \mathbf{u} exist provided that

$$|\mathbf{A} - \lambda^2\mathbf{I}| = 0 \quad (3)$$

Let $\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2$ be the solutions of (3) and $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ the corresponding solutions of (2). Define \mathbf{M} to be the matrix having $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ as its columns and \mathbf{S} to be the diagonal matrix having $\lambda_1^2, \lambda_2^2, \dots, \lambda_n^2$ as its diagonal elements. By applying the transformation $\mathbf{x}(t) = \mathbf{M}\mathbf{q}(t)$, where $\mathbf{q}(t) = [q_1(t) \ q_2(t) \ \dots \ q_n(t)]^T$, to (1), show that

$$\ddot{\mathbf{q}} = \mathbf{S}\mathbf{q} \quad (4)$$

and deduce that (4) has solutions of the form

$$q_i = C_i \sin(\omega_i t + \alpha_i) \quad (5)$$

where C_i and α_i are arbitrary constants and $\lambda_i = j\omega_i$, with $j = \sqrt{-1}$.

The solutions λ_i^2 of (3) define the **natural frequencies** ω_i of the system. The corresponding solutions q_i given in (5) are called the **normal modes** of the system. The general solution of (1) is then obtained using $\mathbf{x}(t) = \mathbf{M}\mathbf{q}(t)$.

A mass–spring vibrating system is governed by the differential equations

$$\ddot{x}_1(t) = -3x_1(t) + 2x_2(t)$$

$$\ddot{x}_2(t) = x_1(t) - 2x_2(t)$$

with $x_1(0) = 1$ and $x_2(0) = \dot{x}_1(0) = \dot{x}_2(0) = 2$.

Determine the natural frequencies and the corresponding normal modes of the system. Hence obtain the general displacement $x_1(t)$ and $x_2(t)$ at time $t \geq 0$. Plot graphs of both the normal modes and the general solutions.



2

Numerical Solution of Ordinary Differential Equations

Chapter 2 Contents

2.1	Introduction	114
2.2	Engineering application: motion in a viscous fluid	114
2.3	Numerical solution of first-order ordinary differential equations	115
2.4	Numerical methods for systems of ordinary differential equations and higher-order differential equations	149
2.5	Engineering application: oscillations of a pendulum	162
2.6	Engineering application: heating of an electrical fuse	167
2.7	Review exercises (1–12)	172

2.1

Introduction

Frequently the equations which express mathematical models in both engineering analysis and engineering design involve derivatives and integrals of the models' variables. Equations involving derivatives are called **differential equations** and those which include integrals or both integrals and derivatives are called **integral equations** or **integro-differential equations**. Generally integral and integro-differential equations are more difficult to deal with than purely differential ones.

There are many methods and techniques for the analytical solution of elementary ordinary differential equations. The most common of these are covered in most first-level books on engineering mathematics (e.g. *Modern Engineering Mathematics*). However, many differential equations of interest to engineers are not amenable to analytical solution and in these cases we must resort to numerical solutions. Numerical solutions have many disadvantages (it is, for instance, much less obvious how changes of parameters or coefficients in the equations affect the solutions) so an analytical solution is generally more useful where one is available.

There are many tools available to the engineer which will provide numerical solutions to differential equations. The most versatile of these perhaps are the major computer algebra systems such as MAPLE. These contain functions for both analytical and numerical solution of differential equations. Systems such as MATLAB/Simulink and Mathcad can also provide numerical solutions to differential equations problems. It may sometimes be necessary for the engineer to write a computer program to solve a differential equation numerically, either because suitable software packages are not available or because the packages available provide no method suitable for the particular differential equation under consideration.

Whether the engineer uses a software package or writes a computer program for the specific problem, it is necessary to understand something of how numerical solutions of differential equations are achieved mathematically. The engineer who does not have this understanding cannot critically evaluate the results provided by a software package and may fall into the trap of inadvertently using invalid results. In this chapter we develop the basics of the numerical solution of ordinary differential equations.

2.2

Engineering application: motion in a viscous fluid

The problem of determining the motion of a body falling through a viscous fluid arises in a wide variety of engineering contexts. One obvious example is that of a parachutist, both in free fall and after opening his or her parachute. The dropping of supplies from aircraft provides another example. Many industrial processes involve adding particulate raw materials into process vessels containing fluids, whether gases or liquids, which exert viscous forces on the particles. Often the motion of the raw materials in the process vessel must be understood in order to ensure that the process is effective and efficient. Fluidized bed combustion furnaces involve effectively suspending particles in a moving gas stream through the viscous forces exerted by the gas on the particles. Thus, understanding the mechanics of the motion of a particle through a viscous fluid has important engineering applications.

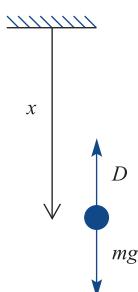


Figure 2.1 A particle falling through a viscous fluid.

When a particle is falling through a viscous fluid it may be modelled simply in the following way. The force of gravity acts downwards and is opposed by a viscous drag force produced by the resistance of the fluid. Figure 2.1 shows a free body diagram of the particle which is assumed to be falling vertically downwards. If the particle's mass is m , the gravitational force is mg , and it is opposed by a drag force, D , acting to oppose motion. The displacement of the particle from its initial position is x .

The equation of motion is

$$m \frac{d^2x}{dt^2} = mg - D \quad (2.1)$$

Before we can solve this equation, the form of the drag term must be determined. For particles moving at a high speed it is often assumed that the drag is proportional to the square of the speed. For slow motion the drag is sometimes assumed to be directly proportional to the speed. In other applications it is more appropriate to assume that drag is proportional to some power of the velocity, so that

$$D = kv^\alpha = k\left(\frac{dx}{dt}\right)^\alpha \quad \text{where, normally, } 1 \leq \alpha \leq 2$$

The differential equation (2.1) then becomes

$$\begin{aligned} m \frac{d^2x}{dt^2} &= mg - k\left(\frac{dx}{dt}\right)^\alpha \\ \text{i.e.} \quad m \frac{d^2x}{dt^2} + k\left(\frac{dx}{dt}\right)^\alpha &= mg \end{aligned} \quad (2.2)$$

This is a second-order, nonlinear, ordinary differential equation for x , the displacement of the particle, as a function of time. In fact, for both $\alpha=1$ and $\alpha=2$, (2.2) can be solved analytically, but for other values of α no such solution exists. If we want to solve the differential equation for such values of α we must resort to numerical techniques.

2.3

Numerical solution of first-order ordinary differential equations

In a book such as this we cannot hope to cover all of the many numerical techniques which have been developed for dealing with ordinary differential equations (ODEs) so we will concentrate on presenting a selection of methods which illustrate the main strands of the theory. In so doing we will meet the main theoretical tools and unifying concepts of the area.

In the last twenty years great advances have been made in the application of computers to the solution of differential equations, particularly using computer algebra packages to assist in the derivation of analytical solutions and the computation of numerical solutions. The MATLAB package is principally oriented towards the solution of numerical problems (although its Symbolic Math Toolbox and the MuPAD version are highly capable) and contains a comprehensive selection of the best modern numerical techniques giving the ability to solve most numerical problems in ODEs. Indeed numerical solutions can be achieved both in native MATLAB and in the Simulink simulation subsystem; which of these paths the user chooses to follow may well be dictated as much by their experience and professional orientation as by theoretical considerations. MAPLE, despite being mainly orientated towards the solution of symbolic problems, also contains a comprehensive suite of numerical solution routines and is, in practice, just as capable as MATLAB in



this area. Moreover, MAPLE gives to the user more control of the solution method used and includes a number of ‘classical’ solution methods. These classical methods include all the methods which are used, in this chapter, to introduce, develop and analyse the main strands of the theory mentioned above. For this reason, MAPLE will be featured rather more frequently than MATLAB, but the practising engineer is as likely to be using MATLAB for the numerical solution of real-world problems as using MAPLE.

Despite the fact that professional engineers are very likely to be using these packages to compute numerical solutions of ODEs, it is still important that they understand the methods which the computer packages use to do their work, for otherwise they are at the mercy of the decisions made by the designers of the packages who have no foreknowledge of the applications to which users may put the package. If the engineering user does not have a sound understanding of the principles being used within the package there is the ever present danger of using results outside their domain of validity. From there it is a short step to engineering failures and human disasters.

2.3.1 A simple solution method: Euler’s method

For a first-order differential equation $dx/dt = f(t, x)$ we can define a **direction field**. The direction field is that two-dimensional vector field in which the vector at any point (t, x) has the gradient dx/dt .

More precisely, we know that the gradient at (t, x) is $f(t, x)$. This means that we can represent the solution of the differential equation in the (t, x) plane by the vector $[1, f(t, x)]$ at each point (t, x) . It is practical to normalize the vectors to give them unit magnitude, thus the direction field is the field

$$\frac{[1, f(t, x)]}{\sqrt{1 + f(t, x)^2}}$$

For instance, Figure 2.2 shows the direction field of the differential equation $dx/dt = x(1 - x)t$.

Since a solution of a differential equation is a function $x(t)$ which has the property $dx/dt = f(t, x)$ at all points (t, x) the solutions of the differential equation are curves in the (t, x) plane to which the direction field lines are tangential at every point. For instance, the curves shown in Figure 2.3 are solutions of the differential equation

Figure 2.2

The direction field for the equation $dx/dt = x(1 - x)t$.

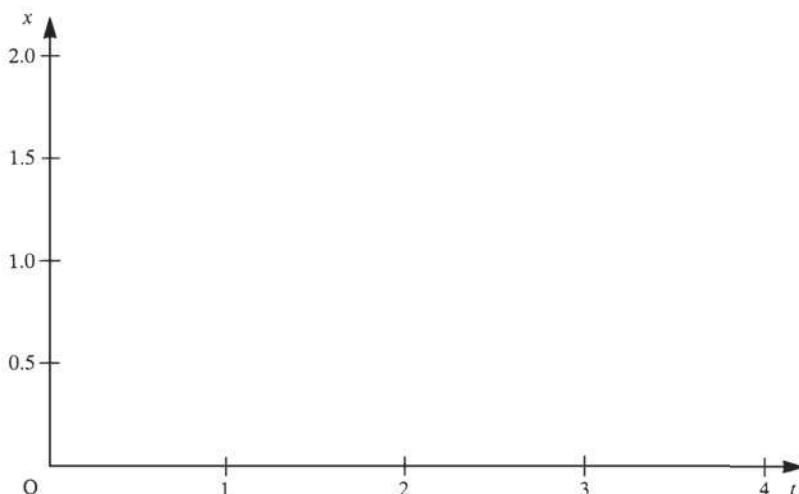
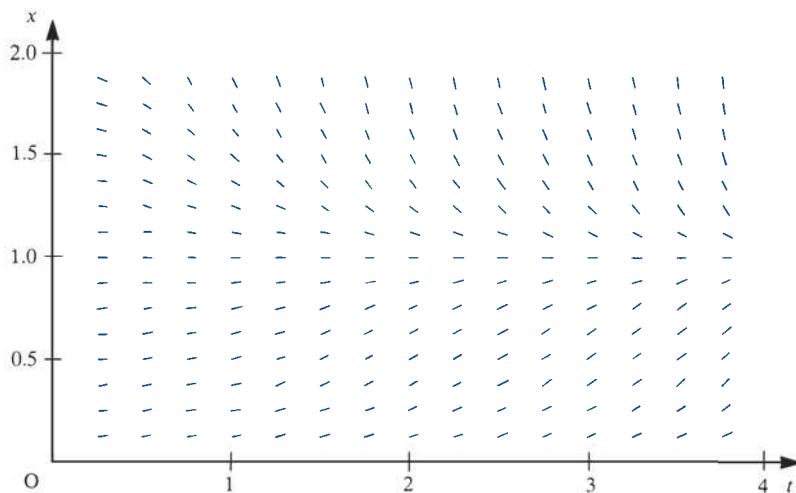


Figure 2.3 Solutions of $dx/dt = x(1 - x)t$ superimposed on its direction field.



$$\frac{dx}{dt} = x(1 - x)t$$

This immediately suggests that a curve representing a solution can be obtained by sketching on the direction field a curve that is always tangential to the lines of the direction field. In Figure 2.4 a way of systematically constructing an approximation to such a curve is shown.

Starting at some point (t_0, x_0) , a straight line parallel to the direction field at that point, $f(t_0, x_0)$, is drawn. This line is followed to a point with abscissa $t_0 + h$. The ordinate at this point is $x_0 + hf(t_0, x_0)$, which we shall call X_1 . The value of the direction field at this new point is calculated, and another straight line from this point with the new gradient is drawn. This line is followed as far as the point with abscissa $t_0 + 2h$. The process can be repeated any number of times, and a curve in the (t, x) plane consisting of a number of short straight-line segments is constructed. The curve is completely defined by the points at which the line segments join, and these can obviously be described by the equations.

Figure 2.4
The construction of a numerical solution of the equation $dx/dt = f(t, x)$.

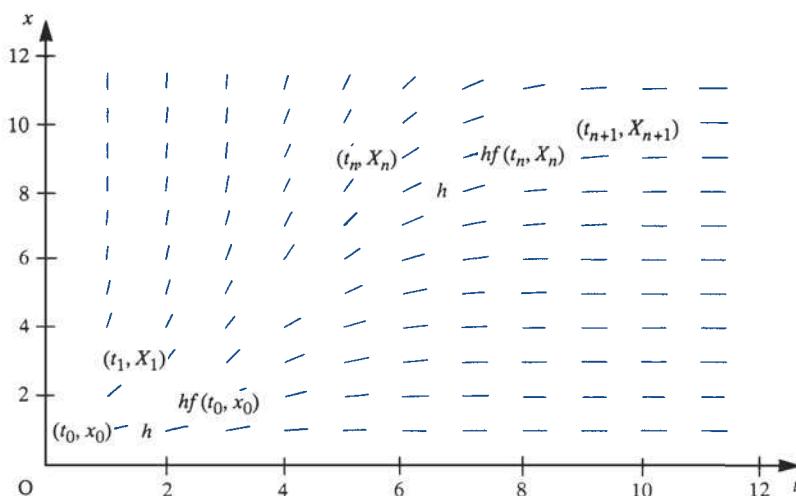
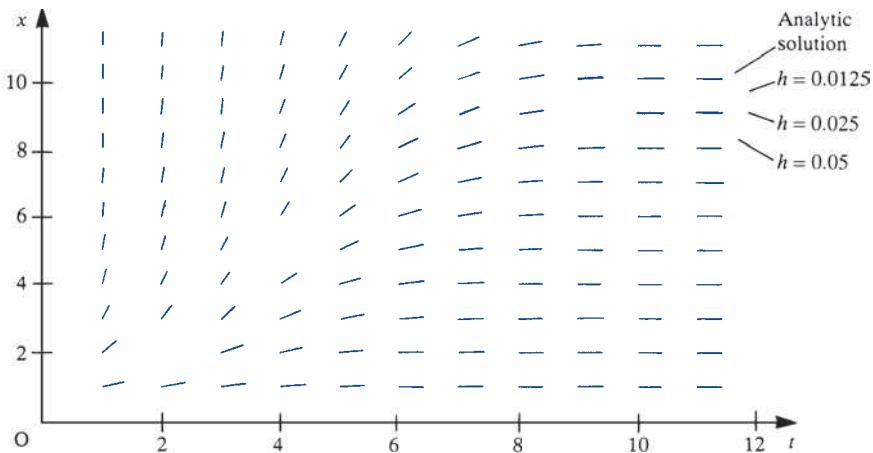


Figure 2.5 The Euler-method solutions of $\frac{dx}{dt} = x^2 t e^{-t}$ for $h = 0.05, 0.025$ and 0.0125 .



$$\begin{aligned}
 t_1 &= t_0 + h, & X_1 &= x_0 + hf(t_0, x_0) \\
 t_2 &= t_1 + h, & X_2 &= X_1 + hf(t_1, X_1) \\
 t_3 &= t_2 + h, & X_3 &= X_2 + hf(t_2, X_2) \\
 \vdots & & \vdots & \\
 t_{n+1} &= t_n + h, & X_{n+1} &= X_n + hf(t_n, X_n)
 \end{aligned}$$

These define, mathematically, the simplest method for integrating first-order differential equations. It is called **Euler's method** (or the forward Euler method). Solutions are constructed step by step, starting from some given starting point (t_0, x_0) . For a given t_0 each different x_0 will give rise to a different solution curve. These curves are all solutions of the differential equation, but each corresponds to a different initial condition.

The solution curves constructed using this method are obviously not exact solutions but only approximations to solutions, because they are only tangential to the direction field at certain points. Between these points, the curves are only approximately tangential to the direction field. Intuitively, we expect that, as the distance for which we follow each straight-line segment is reduced, the curve we are constructing will become a better and better approximation to the exact solution. The increment h in the independent variable t along each straight-line segment is called the **step size** used in the solution. In Figure 2.5 three approximate solutions of the initial-value problem

$$\frac{dx}{dt} = x^2 t e^{-t}, \quad x(0) = 0.91 \quad (2.3)$$

for step sizes $h = 0.05, 0.025$ and 0.0125 are shown. These steps are sufficiently small that the curves, despite being composed of a series of short straight lines, give the illusion of being smooth curves. Equation (2.3) actually has an analytical solution, which can be obtained by separation:

$$x = \frac{1}{(1+t)e^{-t} + C}$$

The analytical solution to the initial-value problem is also shown in Figure 2.5 for comparison. It can be seen that, as we expect intuitively, the smaller the step size the more closely the numerical solution approximates the analytical solution.



MAPLE provides options in the `dsolve` function, the general-purpose ordinary differential equation solver, to return a numerical solution computed using the Euler method. Using this option we can easily generate the solutions plotted on Figure 2.5. In fact we can readily extend the figure to some smaller time steps. The following MAPLE worksheet will produce a figure similar to Figure 2.5 comparing the solutions obtained from the Euler method using time steps of 0.05, 0.025, 0.0125, 0.00625, 0.003125 and the exact solution. The pattern established in Figure 2.5 can be seen to continue with each halving of the time step producing a solution with a yet smaller error when compared with the exact solution.

```

> deq1:=diff(x(t),t)=x(t)^2*t*exp(-t);init1:=x(0)=0.91;
> #solve the differential equation with 5 different
      timesteps
> x1:=dsolve({deq1, init1},
      numeric,method=classical[foreuler],output=listprocedure,
      stepsize=0.05);
> x2:=dsolve({deq1, init1},
      numeric,method=classical[foreuler],output=listprocedure,
      stepsize=0.025);
> x3:=dsolve({deq1, init1},
      numeric,method=classical[foreuler],output=listprocedure,
      stepsize=0.0125);
> x4:=dsolve({deq1, init1},
      numeric,method=classical[foreuler],output=listprocedure,
      stepsize=0.00625);
> x5:=dsolve({deq1, init1},
      numeric,method=classical[foreuler],output=listprocedure,
      stepsize=0.003125);
> #extract the five solutions from the listprocedure
      structures
> for i from 1 to 5 do;solution||i:=op(2,x||i[2]);end do;
> #find the exact solution
> xa:=dsolve({deq1, init1});
> #plot the five numerical solutions and the exact solution
> plot([seq(solution||i(t),i=1..5),op(2,xa)(t)],t=0..12);

```

Example 2.1

The function $x(t)$ satisfies the differential equation

$$\frac{dx}{dt} = \frac{x+t}{xt}$$

and the initial condition $x(1) = 2$. Use Euler's method to obtain an approximation to the value of $x(2)$ using a step size of $h = 0.1$.

Solution

In this example the initial value of t is 1 and $x(1) = 2$. Using the notation above we have $t_0 = 1$, and $x_0 = 2$. The function $f(t, x) = \frac{x+t}{xt}$. So we have

$$t_1 = t_0 + h = 1 + 0.1 = 1.1000$$

Figure 2.6

Computational results
for Example 2.1.

t	X	$X + t$	Xt	$h \frac{X+t}{Xt}$
1.0000	2.0000	3.0000	2.0000	0.1500
1.1000	2.1500	3.2500	2.3650	0.1374
1.2000	2.2874	3.4874	2.7449	0.1271
1.3000	2.4145	3.7145	3.1388	0.1183
1.4000	2.5328	3.9328	3.5459	0.1109
1.5000	2.6437	4.1437	3.9656	0.1045
1.6000	2.7482	4.3482	4.3971	0.0989
1.7000	2.8471	4.5471	4.8400	0.0939
1.8000	2.9410	4.7410	5.2939	0.0896
1.9000	3.0306	4.9306	5.7581	0.0856
2.0000	3.1162			

$$X_1 = x_0 + hf(t_0, x_0) = x_0 + h \frac{x_0 + t_0}{x_0 t_0} = 2 + 0.1 \frac{2 + 1}{2 \cdot 1} = 2.1500$$

$$t_2 = t_1 + h = 1.1000 + 0.1 = 1.2000$$

$$X_2 = x_1 + hf(t_1, x_1) = x_1 + h \frac{x_1 + t_1}{x_1 t_1} = 2.1500 + 0.1 \frac{2.1500 + 1.100}{2.1500 \cdot 1.100} = 2.2874$$

The rest of the solution is obtained step by step as set out in Figure 2.6. The approximation $X(2) = 3.1162$ results.



The solution to this example could easily be obtained using MAPLE as follows:

```
> deq1:=diff(x(t),t)=(x(t)+t)/(x(t)*t);init1:=x(1)=2;
> x1:=dsolve({deq1, init1},
    numeric,method=classical[foreuler],output=listprocedure,
    stepsize=0.1);
> sol:=op(2,x1[2]);sol(2);
```

2.3.2 Analysing Euler's method

We have introduced Euler's method via an intuitive argument from a geometrical understanding of the problem. Euler's method can be seen in another light – as an application of the Taylor series. The Taylor series expansion for a function $x(t)$ gives

$$x(t+h) = x(t) + h \frac{dx}{dt}(t) + \frac{h^2}{2!} \frac{d^2x}{dt^2}(t) + \frac{h^3}{3!} \frac{d^3x}{dt^3}(t) + \dots \quad (2.4)$$

Using this formula, we could, in theory, given the value of $x(t)$ and all the derivatives of x at t , compute the value of $x(t+h)$ for any given h . If we choose a small value for h then the Taylor series truncated after a finite number of terms will provide a good approximation to the value of $x(t+h)$. Euler's method can be interpreted as using the Taylor series truncated after the second term as an approximation to the value of $x(t+h)$.

In order to distinguish between the exact solution of a differential equation and a numerical approximation to the exact solution (and it should be appreciated that all numerical solutions, however accurate, are only approximations to the exact solution), we shall now make explicit the convention that we used in the last section. The exact solution of a differential equation will be denoted by a lower-case letter and a numerical approximation to the exact solution by the corresponding capital letter. Thus, truncating the Taylor series, we write

$$X(t+h) = x(t) + h \frac{dx}{dt}(t) = x(t) + hf(t, x) \quad (2.5)$$

Applying this truncated Taylor series, starting at the point (t_0, x_0) and denoting $t_0 + nh$ by t_n , we obtain

$$X(t_1) = X(t_0 + h) = x(t_0) + hf(t_0, x_0)$$

$$X(t_2) = X(t_1 + h) = X(t_1) + hf(t_1, X_1)$$

$$X(t_3) = X(t_2 + h) = X(t_2) + hf(t_2, X_2)$$

and so on

which is just the Euler-method formula obtained in Section 2.3.1. As an additional abbreviated notation, we shall adopt the convention that $x(t_0 + nh)$ is denoted by x_n , $X(t_0 + nh)$ by X_n , $f(t_n, x_n)$ by f_n , and $f(t_n, X_n)$ by F_n . Hence we may express the Euler method, in general terms, as the recursive rule

$$\begin{aligned} X_0 &= x_0 \\ X_{n+1} &= X_n + hF_n \quad (n \geq 0) \end{aligned}$$

The advantage of viewing Euler's method as an application of Taylor series in this way is that it gives us a clue to obtaining more accurate methods for the numerical solution of differential equations. It also enables us to analyse in more detail how accurate the Euler method may be expected to be. Using the order notation we can abbreviate (2.4) to

$$x(t+h) = x(t) + hf(t, x) + O(h^2)$$

and, combining this with (2.5), we see that

$$X(t+h) = x(t+h) + O(h^2) \quad (2.6)$$

(Note that in obtaining this result we have used the fact that signs are irrelevant in determining the order of terms; that is, $-O(h^p) = O(h^p)$.) Equation (2.6) expresses the fact that at each step of the Euler process the value of $X(t+h)$ obtained has an error of order h^2 , or, to put it another way, the formula used is accurate as far as terms of order h . For this reason Euler's method is known as a **first-order method**. The exact size of the error is, as we intuitively expected, dependent on the size of h , and decreases as h decreases. Since the error is of order h^2 , we expect that halving h , for instance, will reduce the error at each step by a factor of four.

This does not, unfortunately, mean that the error in the solution of the initial value problem is reduced by a factor of four. To understand why this is so, we argue as

follows. Starting from the point (t_0, x_0) and using Euler's method with a step size h to obtain a value of $X(t_0 + 4)$, say, requires $4/h$ steps. At each step an error of order h^2 is incurred. The total error in the value of $X(t_0 + 4)$ will be the sum of the errors incurred at each step, and so will be $4/h$ times the value of a typical step error. Hence the total error is of the order of $(4/h)O(h^2)$; that is, the total error is $O(h)$. From this argument we should expect that if we compare solutions of a differential equation obtained using Euler's method with different step sizes, halving the step size will halve the error in the solution. Examination of Figure 2.5 confirms that this expectation is roughly correct in the case of the solutions presented there.

Example 2.2

Let X_a denote the approximation to the solution of the initial-value problem

$$\frac{dx}{dt} = \frac{x^2}{t+1}, \quad x(0) = 1$$

obtained using Euler's method with a step size $h = 0.1$, and X_b that obtained using a step size of $h = 0.05$. Compute the values of $X_a(t)$ and $X_b(t)$ for $t = 0.1, 0.2, \dots, 1.0$. Compare these values with the values of $x(t)$, the exact solution of the problem. Compute the ratio of the errors in X_a and X_b .

Solution

The exact solution, which may be obtained by separation, is

$$x = \frac{1}{1 - \ln(t+1)}$$

The numerical solutions X_a and X_b and their errors are shown in Figure 2.7. Of course, in this figure the values of X_a are recorded at every step whereas those of X_b are only recorded at alternate steps.

Again, the final column of Figure 2.7 shows that our expectations about the effects of halving the step size when using Euler's method to solve a differential equation are confirmed. The ratio of the errors is not, of course, exactly one-half, because there are some higher-order terms in the errors, which we have ignored.

Figure 2.7

Computational results for Example 2.2.

t	X_a	X_b	$x(t)$	$ x - X_a $	$ x - X_b $	$\frac{ x - X_b }{ x - X_a }$
0.000 00	1.000 00	1.000 00	1.000 00			
0.100 00	1.100 00	1.102 50	1.105 35	0.005 35	0.002 85	0.53
0.200 00	1.210 00	1.216 03	1.222 97	0.012 97	0.006 95	0.54
0.300 00	1.332 01	1.342 94	1.355 68	0.023 67	0.012 75	0.54
0.400 00	1.468 49	1.486 17	1.507 10	0.038 61	0.020 92	0.54
0.500 00	1.622 52	1.649 52	1.681 99	0.059 47	0.032 47	0.55
0.600 00	1.798 03	1.837 91	1.886 81	0.088 78	0.048 90	0.55
0.700 00	2.000 08	2.057 92	2.130 51	0.130 42	0.072 59	0.56
0.800 00	2.235 40	2.318 57	2.425 93	0.190 53	0.107 36	0.56
0.900 00	2.513 01	2.632 51	2.792 16	0.279 15	0.159 65	0.57
1.000 00	2.845 39	3.018 05	3.258 89	0.413 50	0.240 84	0.58

2.3.3 Using numerical methods to solve engineering problems

In Example 2.2 the errors in the values of X_a and X_b are quite large (up to about 14% in the worst case). While carrying out computations with large errors such as these is quite useful for illustrating the mathematical properties of computational methods, in engineering computations we usually need to keep errors very much smaller. Exactly how small they must be is largely a matter of engineering judgement. The engineer must decide how accurately a result is needed for a given engineering purpose. It is then up to that engineer to use the mathematical techniques and knowledge available to carry out the computations to the desired accuracy. The engineering decision about the required accuracy will usually be based on the use that is to be made of the result. If, for instance, a preliminary design study is being carried out then a relatively approximate answer will often suffice, whereas for final design work much more accurate answers will normally be required. It must be appreciated that demanding greater accuracy than is actually needed for the engineering purpose in hand will usually carry a penalty in time, effort or cost.

Let us imagine that, for the problem posed in Example 2.2, we had decided we needed the value of $x(1)$ accurate to 1%. In the cases in which we should normally resort to numerical solution we should not have the analytical solution available, so we must ignore that solution. We shall suppose then that we had obtained the values of $X_a(1)$ and $X_b(1)$ and wanted to predict the step size we should need to use to obtain a better approximation to $x(1)$ accurate to 1%. Knowing that the error in $X_b(1)$ should be approximately one-half the error in $X_a(1)$ suggests that the error in $X_b(1)$ will be roughly the same as the difference between the errors in $X_a(1)$ and $X_b(1)$, which is the same as the difference between $X_a(1)$ and $X_b(1)$; that is, 0.172 66. One per cent of $X_b(1)$ is roughly 0.03, that is roughly one-sixth of the error in $X_b(1)$. Hence we expect that a step size roughly one-sixth of that used to obtain X_b will suffice; that is, a step size $h = 0.008\ 33$. In practice, of course, we shall round to a more convenient non-recurring decimal quantity such as $h = 0.008$. This procedure is closely related to the Aitken extrapolation procedure sometimes used for estimating limits of convergent sequences and series.

Example 2.3

Compute an approximation $X(1)$ to the value of $x(1)$ satisfying the initial-value problem

$$\frac{dx}{dt} = \frac{x^2}{t+1}, \quad x(0) = 1$$

by using Euler's method with a step size $h = 0.008$.

Solution

It is worth commenting here that the calculations performed in Example 2.2 could reasonably be carried out on any hand-held calculator, but this new calculation requires 125 steps. To do this is on the boundaries of what might reasonably be done on a hand-held calculator, and is more suited to computational software such as MAPLE. Repeating the calculation with a step size $h = 0.008$ produces the result $X(1) = 3.21391$.

We had estimated from the evidence available (that is, values of $X(1)$ obtained using step sizes $h = 0.1$ and 0.05) that the step size $h = 0.008$ should provide a value of $X(1)$ accurate to approximately 1%. Comparison of the value we have just computed with the exact solution shows that it is actually in error by approximately 1.4%. This does not quite meet the target of 1% that we set ourselves. This example therefore serves, first, to illustrate how, given two approximations to $x(1)$ derived using Euler's method with different step sizes, we can estimate the step size needed to compute an approximation

within a desired accuracy, and, secondly, to emphasize that the estimate of the appropriate step size is only an *estimate*, and will not *guarantee* an approximate solution to the problem meeting the desired accuracy criterion. If we had been more conservative and rounded the estimated step size down to, say, 0.005, we should have obtained $X(1) = 3.23043$, which is in error by only 0.9% and would have met the required accuracy criterion.



Again the solution to this example could be obtained using MAPLE. The following worksheet computes the numerical solution using a step size of 0.008, then the analytical solution and finally computes the percentage error in the numerical solution.

```
> #set up differential equation
> deq1:=diff(x(t),t)=x(t)^2/(t+1);init1:=x(0)=1;
> #obtain x1, the numerical solution
> x1:=dsolve({deq1, init1},
numeric,method=classical[foreuler],output=listprocedure,
stepsize=0.008);
> #xa is the analytic solution
> xa:=dsolve({deq1, init1});
> #obtain the value of x(t) at t=1
> op(2,x1[2])(1);
> #find the percentage error in the numerical solution
> evalf((op(2,x1[2])(1)-subs(t=1,op(2,xa)))/
subs(t=1,op(2,xa)))*100;
```

Since we have mentioned in Example 2.3 the use of computers to undertake the repetitive calculations involved in the numerical solution of differential equations, it is also worth commenting briefly on the writing of computer programs to implement those numerical solution methods. Whilst it is perfectly possible to write informal, unstructured programs to implement algorithms such as Euler's method, a little attention to planning and structuring a program well will usually be amply rewarded – particularly in terms of the reduced probability of introducing 'bugs'. Another reason for careful structuring is that, in this way, parts of programs can often be written in fairly general terms and can be re-used later for other problems. The two pseudocode algorithms in Figures 2.8 and 2.9 will both produce the table of results in Example 2.2. The pseudocode program of Figure 2.8 is very specific to the problem posed, whereas that of Figure 2.9 is more general, better structured, and more expressive of the structure of mathematical problems. It is generally better to aim at the style of Figure 2.9.

Figure 2.8 A poorly structured algorithm for Example 2.2.

```
x1 ← 1
x2 ← 1
write(vdu, 0, 1, 1, 1)
for i is 1 to 10 do
    x1 ← x1 + 0.1*x1*x1/((i-1)*0.1 + 1)
    x2 ← x2 + 0.05*x2*x2/((i-1)*0.1 + 1)
    x2 ← x2 + 0.05*x2*x2/((i-1)*0.1 + 1.05)
    x ← 1/(1 - ln(i*0.1 + 1))
    write(vdu,0.1*i,x1,x2,x,x - x1,x - x2,(x - x2)/(x - x1))
endfor
```

Figure 2.9 A better structured algorithm for Example 2.2.

```

initial_time ← 0
final_time ← 1
initial_x ← 1
step ← 0.1
t ← initial_time
x1 ← initial_x
x2 ← initial_x
h1 ← step
h2 ← step/2
write(vdu,initial_time,x1,x2,initial_x)
repeat
    euler(t,x1,h1,1 → x1)
    euler(t,x2,h2,2 → x2)
    t ← t + step
    x ← exact_solution(t,initial_time,initial_x)
    write(vdu,t,x1,x2,x,abs(x - x1),abs(x - x2),abs((x - x2)/(x - x1)))
until t ≥ final_time

procedure euler(t_old,x_old,step,number → x_new)
    temp_x ← x_old
    for i is 0 to number -1 do
        temp_x ← temp_x + step*derivative(t_old + step*i,temp_x)
    endfor
    x_new ← temp_x
endprocedure

procedure derivative(t,x → derivative)
    derivative ← x*x/(t + 1)
endprocedure

procedure exact_solution(t,t0,x0 → exact_solution)
    c ← ln(t0 + 1) + 1/x0
    exact_solution ← 1/(c - ln(t + 1))
endprocedure

```

2.3.4 Exercises



All the exercises in this section can be completed using MAPLE in a similar manner to Examples 2.1 and 2.3 above. In particular MAPLE or some other form of computer assistance should be used for Exercises 5, 6 and 7. If you do not have access to MAPLE, you will need to write a program in MATLAB or some other high-level scientific computer programming language (e.g. Fortran, Python or C).

- 1 Find the value of $X(0.3)$ for the initial-value problem

$$\frac{dx}{dt} = -\frac{1}{2}xt, \quad x(0) = 1$$

using Euler's method with step size $h = 0.1$.

- 3 Find the value of $X(1)$ for the initial-value problem

$$\frac{dx}{dt} = \frac{x}{2(t + 1)}, \quad x(0.5) = 1$$

using Euler's method with step size $h = 0.1$.

- 2 Find the value of $X(1.1)$ for the initial-value problem

$$\frac{dx}{dt} = -\frac{1}{2}xt, \quad x(1) = 0.1$$

using Euler's method with step size $h = 0.025$.

- 4 Find the value of $X(0.5)$ for the initial-value problem

$$\frac{dx}{dt} = \frac{4-t}{t+x}, \quad x(0) = 1$$

using Euler's method with step size $h = 0.05$.

- 5 Denote the Euler-method solution of the initial-value problem

$$\frac{dx}{dt} = \frac{xt}{t^2 + 2}, \quad x(1) = 2$$

using step size $h = 0.1$ by $X_a(t)$, and that using $h = 0.05$ by $X_b(t)$. Find the values of $X_a(2)$ and $X_b(2)$. Estimate the error in the value of $X_b(2)$, and suggest a value of step size that would provide a value of $X(2)$ accurate to 0.1%. Find the value of $X(2)$ using this step size. Find the exact solution of the initial-value problem, and determine the actual magnitude of the errors in $X_a(2)$, $X_b(2)$ and your final value of $X(2)$.

- 6 Denote the Euler-method solution of the initial-value problem

$$\frac{dx}{dt} = \frac{1}{xt}, \quad x(1) = 1$$

using step size $h = 0.1$ by $X_a(t)$, and that using $h = 0.05$ by $X_b(t)$. Find the values of $X_a(2)$ and

$X_b(2)$. Estimate the error in the value of $X_b(2)$, and suggest a value of step size that would provide a value of $X(2)$ accurate to 0.2%. Find the value of $X(2)$ using this step size. Find the exact solution of the initial-value problem, and determine the actual magnitude of the errors in $X_a(2)$, $X_b(2)$ and your final value of $X(2)$.

7

Denote the Euler-method solution of the initial-value problem

$$\frac{dx}{dt} = \frac{1}{\ln x}, \quad x(1) = 1.2$$

using step size $h = 0.05$ by $X_a(t)$, and that using $h = 0.025$ by $X_b(t)$. Find the values of $X_a(1.5)$ and $X_b(1.5)$. Estimate the error in the value of $X_b(1.5)$, and suggest a value of step size that would provide a value of $X(1.5)$ accurate to 0.25%. Find the value of $X(1.5)$ using this step size. Find the exact solution of the initial-value problem, and determine the actual magnitude of the errors in $X_a(1.5)$, $X_b(1.5)$ and your final value of $X(1.5)$.

2.3.5 More accurate solution methods: multistep methods

In Section 2.3.2 we discovered that using Euler's method to solve a differential equation is essentially equivalent to using a Taylor series expansion of a function truncated after two terms. Since, by so doing, we are ignoring terms $O(h^2)$, an error of this order is introduced at each step in the solution. Could we not derive a method for calculating approximate solutions of differential equations which, by using more terms of the Taylor series, provides greater accuracy than Euler's method? We can – but there are some disadvantages in so doing, and various methods have to be used to overcome these.

Let us first consider a Taylor series expansion with the first three terms written explicitly. This gives

$$x(t+h) = x(t) + h \frac{dx}{dt}(t) + \frac{h^2}{2!} \frac{d^2x}{dt^2}(t) + O(h^3) \quad (2.7)$$

Substituting $f(t, x)$ for dx/dt , we obtain

$$x(t+h) = x(t) + hf(t, x) + \frac{h^2}{2!} \frac{df}{dt}(t, x) + O(h^3)$$

Dropping the $O(h^3)$ terms provides an approximation

$$X(t+h) = x(t) + hf(t, x) + \frac{h^2}{2!} \frac{df}{dt}(t, x)$$

such that

$$X(t+h) = x(t+h) + O(h^3)$$

in other words, a numerical approximation method which has an error at each step that is not of order h^2 like the Euler method but rather of order h^3 . The corresponding general numerical scheme is

$$X_{n+1} = X_n + hF_n + \frac{h^2}{2} \frac{dF_n}{dt} \quad (2.8)$$

The application of the formula (2.5) in Euler's method was straightforward because an expression for $f(t, x)$ was provided by the differential equation itself. To apply (2.8) as it stands requires an analytical expression for df/dt so that dF_n/dt may be computed. This may be relatively straightforward to provide – or it may be quite complicated. Although, using modern computer algebra systems, it is now often possible to compute analytical expressions for the derivatives of many functions, the need to do so remains a considerable disadvantage when compared with methods which do not require the function's derivatives to be provided.

Fortunately, there are ways to work around this difficulty. One such method hinges on the observation that it is just as valid to write down Taylor series expansions for negative increments as for positive ones. The Taylor series expansion of $x(t - h)$ is

$$x(t - h) = x(t) - h \frac{dx}{dt}(t) + \frac{h^2}{2!} \frac{d^2x}{dt^2}(t) - \frac{h^3}{3!} \frac{d^3x}{dt^3}(t) + \dots$$

If we write only the first three terms explicitly, we have

$$x(t - h) = x(t) - h \frac{dx}{dt}(t) + \frac{h^2}{2!} \frac{d^2x}{dt^2}(t) + O(h^3)$$

or, rearranging the equation,

$$\frac{h^2}{2!} \frac{d^2x}{dt^2}(t) = x(t - h) - x(t) + h \frac{dx}{dt}(t) + O(h^3)$$

Substituting this into (2.7), we obtain

$$x(t + h) = x(t) + h \frac{dx}{dt}(t) + \left[x(t - h) - x(t) + h \frac{dx}{dt}(t) + O(h^3) \right] + O(h^3)$$

That is,

$$x(t + h) = x(t - h) + 2h \frac{dx}{dt}(t) + O(h^3)$$

or, substituting $f(t, x)$ for dx/dt ,

$$x(t + h) = x(t - h) + 2hf(t, x) + O(h^3) \quad (2.9)$$

Alternatively, we could write down the Taylor series expansion of the function dx/dt with an increment of $-h$:

$$\frac{dx}{dt}(t - h) = \frac{dx}{dt}(t) - h \frac{d^2x}{dt^2}(t) + \frac{h^2}{2!} \frac{d^3x}{dt^3}(t) - O(h^3)$$

Writing only the first two terms explicitly and rearranging gives

$$h \frac{d^2x}{dt^2}(t) = \frac{dx}{dt}(t) - \frac{dx}{dt}(t - h) + O(h^2)$$

and substituting this into (2.4) gives

$$x(t+h) = x(t) + h \frac{dx}{dt}(t) + \frac{h}{2} \left[\frac{dx}{dt}(t) - \frac{dx}{dt}(t-h) + O(h^2) \right] + O(h^3)$$

That is,

$$x(t+h) = x(t) + \frac{h}{2} \left[3 \frac{dx}{dt}(t) - \frac{dx}{dt}(t-h) \right] + O(h^3)$$

or, substituting $f(t, x)$ for dx/dt ,

$$x(t+h) = x(t) + \frac{1}{2} h [3f(t, x(t)) - f(t-h, x(t-h))] + O(h^3) \quad (2.10)$$

Equations (2.7), (2.9) and (2.10) each give an expression for $x(t+h)$ in which all terms up to those in h^2 have been made explicit. In the same way as, by ignoring terms of $O(h^3)$ in (2.7), the numerical scheme (2.8) can be obtained, (2.9) and (2.10) give rise to the numerical schemes

$$X_{n+1} = X_{n-1} + 2hF_n \quad (2.11)$$

and

$$X_{n+1} = X_n + \frac{1}{2} h (3F_n - F_{n-1}) \quad (2.12)$$

respectively. Each of these alternative schemes, like (2.8), incurs an error $O(h^3)$ at each step.

The advantage of (2.11) or (2.12) over (2.8) arises because the derivative of $f(t, x)$ in (2.7) has been replaced in (2.9) by the value of the function x at the previous time, $x(t-h)$, and in (2.10) by the value of the function f at time $t-h$. This is reflected in (2.11) and (2.12) by the presence of the terms in X_{n-1} and F_{n-1} respectively and the absence of the term in dF_n/dt . The elimination of the derivative of the function $f(t, x)$ from the numerical scheme is an advantage, but it is not without its penalties. In both (2.11) and (2.12) the value of X_{n+1} depends not only on the values of X_n and F_n but also on the value of one or the other at t_{n-1} . This is chiefly a problem when starting the computation. In the case of the Euler scheme the first step took the form

$$X_1 = X_0 + hF_0$$

In the case of (2.11) and (2.12) the first step would seem to take the forms

$$X_1 = X_{-1} + 2hF_0$$

and

$$X_1 = X_0 + \frac{1}{2} h (3F_0 - F_{-1})$$

respectively. The value of X_{-1} in the first case and F_{-1} in the second is not normally available. The resolution of this difficulty is usually to use some other method to start the computation, and, when the value of X_1 , and therefore also the value of F_1 , is available, change to (2.11) or (2.12). The first step using (2.11) or (2.12) therefore involves

$$X_2 = X_0 + 2hF_1$$

or

$$X_2 = X_1 + \frac{1}{2} h(3F_1 - F_0)$$

Methods like (2.11) and (2.12) that involve the values of the dependent variable or its derivative at more than one value of the independent variable are called **multistep methods**. These all share the problem that we have just noted of difficulties in deciding how to start the computation. We shall return to this problem of starting multistep methods in Section 2.3.7.

Example 2.4

Solve the initial-value problem

$$\frac{dx}{dt} = \frac{x^2}{t+1}, \quad x(0) = 1$$

posed in Example 2.2 using the scheme (2.12) with a step size $h = 0.1$. Compute the values of $X(t)$ for $t = 0.1, 0.2, \dots, 1.0$ and compare them with the values of the exact solution $x(t)$.

Solution

We shall assume that the value of $X(0.1)$ has been computed using some other method and has been found to be 1.105 35. The computation therefore starts with the calculation of the values of F_1 , F_0 and hence X_2 . Using the standard notation we have $t_0 = 0$, and $x_0 = 1$. The function $f(t, x) = x^2/(t+1)$. Using the given value $X(0.1) = 1.105 35$, we have $t_1 = 0.1$, and $X_1 = 1.105 35$. So the first step is

$$t_2 = t_1 + h = 0.100 00 + 0.1 = 0.200 00$$

$$\begin{aligned} X_2 &= X_1 + \frac{1}{2} h(3F_1 - F_0) = X_1 + \frac{1}{2} h[3f(t_1, X_1) - f(t_0, x_0)] \\ &= X_1 + \frac{1}{2} h \left(3 \frac{X_1^2}{t_1 + 1} - \frac{X_0^2}{t_0 + 1} \right) = 1.105 35 + \frac{1}{2} 0.1 \left(3 \frac{1.105 35^2}{0.1 + 1} - \frac{1^2}{0 + 1} \right) = 1.221 96 \end{aligned}$$

The results of the computation are shown in Figure 2.10.

Figure 2.10
Computational results
for Example 2.4.

t	X_n	F_n	$\frac{1}{2} h(3F_n - F_{n-1})$	$x(t)$	$ x - X_n $
0.000 00	1.000 00	1.000 00			
0.100 00	1.105 35	1.110 73	0.116 61	1.105 35	0.000 00
0.200 00	1.221 96	1.244 32	0.131 11	1.222 97	0.001 01
0.300 00	1.353 07	1.408 31	0.149 03	1.355 68	0.002 61
0.400 00	1.502 10	1.611 65	0.171 33	1.507 10	0.004 99
0.500 00	1.673 44	1.866 92	0.199 46	1.681 99	0.008 55
0.600 00	1.872 89	2.192 33	0.235 50	1.886 81	0.013 91
0.700 00	2.108 39	2.614 90	0.282 62	2.130 51	0.022 11
0.800 00	2.391 01	3.176 08	0.345 67	2.425 93	0.034 92
0.900 00	2.736 68	3.941 80	0.432 47	2.792 16	0.055 48
1.000 00	3.169 14			3.258 89	0.089 75

It is instructive to compare the values of X_h computed in Example 2.4 with those computed in Example 2.2. Since the method we are using here is a second-order method, the error at each step should be $O(h^3)$ rather than the $O(h^2)$ error of the Euler method. We are using the same step size as for the solution X_a of Example 2.2, so the errors should be correspondingly smaller. In this example we know the exact solution of the initial value problem and thus can compute the error. Examination of the results shows that they are indeed much smaller than those of the Euler method, and also considerably smaller than the errors in the Euler method solution X_b which used step size $h = 0.05$, half the step size used here.

In fact, some numerical experimentation (which we shall not describe in detail) reveals that to achieve a similarly low level of errors, the Euler method requires a step size $h = 0.016$, and therefore 63 steps are required to find the value of $X(1)$. The second-order method of (2.12) requires only 10 steps to find $X(1)$ to a similar accuracy. Thus the solution of a problem to a given accuracy using a second-order method can be achieved in a much shorter computer processing time than using a first-order method. When very large calculations are involved or simple calculations are repeated very many times, such savings are very important.

How do we choose between methods of equal accuracy such as (2.11) and (2.12)? Numerical methods for the solution of differential equations have other properties apart from accuracy. One important property is **stability**. Some methods have the ability to introduce gross errors into the numerical approximation to the exact solution of a problem. The sources of these gross errors are the so-called **parasitic solutions** of the numerical process, which do not correspond to solutions of the differential equation. The analysis of this behaviour is beyond the scope of this book, but methods that are susceptible to it are intrinsically less useful than those that are not. The method of (2.11) can show unstable behaviour, as demonstrated in Example 2.5.

Further details on the stability of numerical methods can be found in E. Süli and D. Mayers, *An Introduction to Numerical Mathematics* (Cambridge, Cambridge University Press, 2014); R. W. Hamming, *Numerical Methods for Scientists and Engineers* (New York, Dover Publications, 1987); E. Isaacson and H. B. Keller, *Analysis of Numerical Methods* (New York, Dover Publications, 1994).

Example 2.5

Let X_a denote the approximation to the solution of the initial-value problem

$$\frac{dx}{dt} = -3x + 2e^{-t}, \quad x(0) = 2$$

obtained using the method defined by (2.11), and X_b that obtained using the method defined by (2.12), both with step size $h = 0.1$. Compute the values of $X_a(t)$ and $X_b(t)$ for $t = 0.1, 0.2, \dots, 2.0$. Compare these with the values of $x(t)$, the exact solution of the problem. In order to overcome the difficulty of starting the processes, assume that the value $X(0.1) = 1.645\,66$ has been obtained by another method.

Solution

The exact solution of the problem, which is a linear equation and so may be solved by the integrating-factor method, is

$$x = e^{-t} + e^{-3t}$$

Figure 2.11
Computational results
for Example 2.5.

<i>t</i>	X_a	X_b	$x(t)$	$x - X_a$	$x - X_b$
0.000 00	2.000 00	2.000 00	2.000 00	0.000 00	0.000 00
0.100 00	1.645 66	1.645 66	1.645 66	0.000 00	0.000 00
0.200 00	1.374 54	1.376 56	1.367 54	-0.007 00	-0.009 02
0.300 00	1.148 42	1.159 09	1.147 39	-0.001 04	-0.011 70
0.400 00	0.981 82	0.984 36	0.971 51	-0.010 30	-0.012 84
0.500 00	0.827 46	0.842 27	0.829 66	0.002 20	-0.012 61
0.600 00	0.727 95	0.725 83	0.714 11	-0.013 84	-0.011 72
0.700 00	0.610 22	0.629 54	0.619 04	0.008 83	-0.010 50
0.800 00	0.560 45	0.549 22	0.540 05	-0.020 41	-0.009 17
0.900 00	0.453 68	0.481 64	0.473 78	0.020 10	-0.007 86
1.000 00	0.450 88	0.424 32	0.417 67	-0.033 21	-0.006 66
1.100 00	0.330 30	0.375 33	0.369 75	0.039 45	-0.005 58
1.200 00	0.385 84	0.333 15	0.328 52	-0.057 33	-0.004 64
1.300 00	0.219 27	0.296 60	0.292 77	0.073 50	-0.003 83
1.400 00	0.363 29	0.264 75	0.261 59	-0.101 70	-0.003 15
1.500 00	0.099 93	0.236 83	0.234 24	0.134 31	-0.002 59
1.600 00	0.392 59	0.212 25	0.210 13	-0.182 46	-0.002 12
1.700 00	-0.054 86	0.190 52	0.188 78	0.243 64	-0.001 73
1.800 00	0.498 57	0.171 24	0.169 82	-0.328 76	-0.001 42
1.900 00	-0.287 88	0.154 08	0.152 91	0.440 80	-0.001 16
2.000 00	0.731 13	0.138 77	0.137 81	-0.593 32	-0.000 96

The numerical solutions X_a and X_b and their errors are shown in Figure 2.11. It can be seen that X_a exhibits an unexpected oscillatory behaviour, leading to large errors in the solution. This is typical of the type of instability from which the scheme (2.11) and those like it are known to suffer. The scheme defined by (2.11) is not unstable for all differential equations, but only for a certain class. The possibility of instability in numerical schemes is one that should always be borne in mind, and the intelligent user is always critical of the results of numerical work and alert for signs of this type of problem.

In this section we have seen how, starting from the Taylor series for a function, schemes of a higher order of accuracy than Euler's method can be constructed. We have constructed two second-order schemes. The principle of this technique can be extended to produce schemes of yet higher orders. They will obviously introduce more values of X_m or F_m (where $m = n - 2, n - 3, \dots$). The scheme (2.12) is, in fact, a member of a family of schemes known as the **Adams–Bashforth formulae**. The first few members of this family are

$$\begin{aligned}X_{n+1} &= X_n + hF_n \\X_{n+1} &= X_n + \frac{1}{2}h(3F_n - F_{n-1}) \\X_{n+1} &= X_n + \frac{1}{12}h(23F_n - 16F_{n-1} + 5F_{n-2}) \\X_{n+1} &= X_n + \frac{1}{24}h(55F_n - 59F_{n-1} + 37F_{n-2} - 9F_{n-3})\end{aligned}$$

The formulae represent first-, second-, third- and fourth-order methods respectively. The first-order Adams–Bashforth formula is just the Euler method, the second-order

one is the scheme we introduced as (2.12), while the third- and fourth-order formulae are extensions of the principle we have just introduced. Obviously all of these require special methods to start the process in the absence of values of X_{-1} , F_{-1} , X_{-2} , F_{-2} and so on.



Some of the methods used by the standard MATLAB procedures for numerical solution of ODEs are based on more sophisticated versions of the multistep methods which we have just introduced. Multistep methods are particularly suitable for solving equations in which the derivative function, $f(t, x)$, is relatively computationally costly to evaluate. At each step a multistep method can re-use the values of the function already computed at previous steps so the number of evaluations of the derivative function is reduced compared to some other methods.

2.3.6 Local and global truncation errors

In Section 2.3.2 we argued intuitively that, although the Euler method introduces an error $O(h^2)$ at each step, it yields an $O(h)$ error in the value of the dependent variable corresponding to a given value of the independent variable. What is the equivalent result for the second-order methods we have introduced in Section 2.3.5? We shall answer this question with a slightly more general analysis that will also be useful to us in succeeding sections.

First let us define two types of error. The **local error** in a method for integrating a differential equation is the error introduced at each step. Thus if the method is defined by

$$X_{n+1} = g(h, t_n, X_n, t_{n-1}, X_{n-1}, \dots)$$

and analysis shows us that

$$x_{n+1} = g(h, t_n, x_n, t_{n-1}, x_{n-1}, \dots) + O(h^{p+1})$$

then we say that the local error in the method is of order $p + 1$ or that the method is a p th-order method.

The **global error** of an integration method is the error in the value of $X(t_0 + a)$ obtained by using that method to advance the required number of steps from a known value of $x(t_0)$. Using a p th-order method, the first step introduces an error $O(h^{p+1})$. The next step takes the approximation X_1 and derives an estimate X_2 of x_2 that introduces a further error $O(h^{p+1})$. The number of steps needed to calculate the value $X(t_0 + a)$ is a/h . Hence we have

$$X(t_0 + a) = x(t_0 + a) + \frac{a}{h} O(h^{p+1})$$

Dividing a quantity that is $O(h^r)$ by h produces a quantity that is $O(h^{r-1})$, so we must have

$$X(t_0 + a) = x(t_0 + a) + O(h^p)$$

In other words, the global error produced by a method that has a local error $O(h^{p+1})$ is $O(h^p)$. As we saw in Example 2.2, halving the step size for a calculation using Euler's method produces errors that are roughly half as big. This is consistent with the global error being $O(h)$. Since the local error of the Euler method is $O(h^2)$, this is as we should expect. Let us now repeat Example 2.2 using the second-order Adams–Bashforth method, (2.12).

Example 2.6

Let X_a denote the approximation to the solution of the initial-value problem

$$\frac{dx}{dt} = \frac{x^2}{t+1}, \quad x(0) = 1$$

obtained using the second-order Adams–Bashforth method with a step size $h = 0.1$, and X_b that obtained using a step size of $h = 0.05$. Compute the values of $X_a(t)$ and $X_b(t)$ for $t = 0.1, 0.2, \dots, 1.0$. Compare these values with the values of $x(t)$, the exact solution of the problem. Compute the ratio of the errors in X_a and X_b . In order to start the process, assume that the values $X(-0.1) = 0.90468$ and $X(-0.05) = 0.95121$ have already been obtained by another method.

Solution

The exact solution was given in Example 2.2. The numerical solutions X_a and X_b and their errors are shown in Figure 2.12.

Because the method is second-order, we expect the global error to vary like h^2 . Theoretically, then, the error in the solution X_b should be one-quarter that in X_a . We see that this expectation is approximately borne out in practice.

Figure 2.12
Computational results
for Example 2.6.

t	X_a	X_b	$x(t)$	$ x - X_a $	$ x - X_b $	$\frac{ x - X_b }{ x - X_a }$
0.00000	1.00000	1.00000	1.00000	0.00000	0.00000	0.00000
0.10000	1.10453	1.10512	1.10535	0.00082	0.00023	0.28
0.20000	1.22089	1.22239	1.22297	0.00208	0.00058	0.28
0.30000	1.35176	1.35459	1.35568	0.00392	0.00109	0.28
0.40000	1.50049	1.50525	1.50710	0.00661	0.00185	0.28
0.50000	1.67144	1.67903	1.68199	0.01055	0.00296	0.28
0.60000	1.87040	1.88217	1.88681	0.01640	0.00464	0.28
0.70000	2.10525	2.12331	2.13051	0.02525	0.00720	0.29
0.80000	2.38700	2.41470	2.42593	0.03893	0.01123	0.29
0.90000	2.73145	2.77440	2.79216	0.06070	0.01776	0.29
1.00000	3.16220	3.23007	3.25889	0.09670	0.02882	0.30

Just as previously we outlined how, for the Euler method, we could estimate from two solutions of the differential equation the step size that would suffice to compute a solution to any required accuracy, so we can do the same in a more general way. If we use a p th-order method to compute two estimates $X_a(t_0 + a)$ and $X_b(t_0 + a)$ of $x(t_0 + a)$ using step sizes h and $\frac{1}{2}h$ then, because the global error of the process is $O(h^p)$, we expect the error in $X_a(t_0 + a)$ to be roughly 2^p times that in $X_b(t_0 + a)$. Hence the error in $X_b(t_0 + a)$ may be estimated to be

$$\frac{|X_a(t_0 + a) - X_b(t_0 + a)|}{2^p - 1}$$

If the desired error, which may be expressed in absolute terms or may be derived from a desired maximum percentage error, is ε then the factor k , say, by which the error in $X_b(t_0 + a)$ must be reduced is

$$k = \frac{|X_a(t_0 + a) - X_b(t_0 + a)|}{\varepsilon(2^p - 1)}$$

Since reducing the step size by a factor of q will, for a p th-order error, reduce the error by a factor of q^p , the factor by which step size must be reduced in order to meet the error criterion is the p th root of k . The step size used to compute X_b is $\frac{1}{2}h$, so finally we estimate the required step size as

$$h \left(\frac{\epsilon(2^p - 1)}{|X_a(t_0 + a) - X_b(t_0 + a)|} \right)^{1/p} \quad (2.13)$$

This technique of estimating the error in a numerical approximation of an unknown quantity by comparing two approximations of that unknown quantity whose order of accuracy is known is an example of the application of **Richardson extrapolation**.

Example 2.7

Estimate the step size required to compute an estimate of $x(1)$ accurate to 2 decimal places for the initial-value problem in Example 2.6 given the values $X_a(1) = 3.162\ 20$ and $X_b(1) = 3.230\ 07$ obtained using step sizes $h = 0.1$ and 0.05 respectively.

Solution

For the result to be accurate to 2 decimal places the error must be less than 0.005. The estimates $X_a(1)$ and $X_b(1)$ were obtained using a second-order process, so, applying (2.13), with $\epsilon = 0.005$, $\frac{1}{2}h = 0.05$ and $p = 2$, we have

$$h = 0.05 \left(\frac{0.015}{|3.162\ 20 - 3.230\ 07|} \right)^{1/2} = 0.0235$$

In a real engineering problem what we would usually do is round this down to say 0.02 and recompute $X(1)$ using step sizes $h = 0.04$ and 0.02 . These two new estimates of $X(1)$ could then be used to estimate again the error in the value of $X(1)$ and confirm that the desired error criterion had been met.

2.3.7 More accurate solution methods: predictor–corrector methods

In Section 2.3.5 we showed how the third term in the Taylor series expansion

$$x(t+h) = x(t) + h \frac{dx}{dt}(t) + \frac{h^2}{2!} \frac{d^2x}{dt^2}(t) + O(h^3) \quad (2.14)$$

could be replaced by either $x(t-h)$ or $(dx/dt)(t-h)$. These are not the only possibilities. By using appropriate Taylor series expansions, we could replace the term with other values of $x(t)$ or dx/dt . For instance, expanding the function $x(t-2h)$ about $x(t)$ gives rise to

$$x(t-2h) = x(t) - 2h \frac{dx}{dt}(t) + 2h^2 \frac{d^2x}{dt^2}(t) + O(h^3) \quad (2.15)$$

and eliminating the second-derivative term between (2.14) and (2.15) gives

$$x(t+h) = \frac{3}{4}x(t) + \frac{1}{4}x(t-2h) + \frac{3}{2}h \frac{dx}{dt}(t) + O(h^3)$$

which, in turn, would give rise to the integration scheme

$$X_{n+1} = \frac{3}{4}X_n + \frac{1}{4}X_{n-2} + \frac{3}{2}hF_n$$

Such a scheme, however, would not seem to offer any advantages to compensate for the added difficulties caused by a two-step scheme using non-consecutive values of X .

The one alternative possibility that does offer some gains is using the value of $(dx/dt)(t+h)$. Writing the Taylor series expansion of $(dx/dt)(t+h)$ yields

$$\frac{dx}{dt}(t+h) = \frac{dx}{dt}(t) + h \frac{d^2x}{dt^2}(t) + O(h^2)$$

and eliminating the second derivative between this and (2.14) gives

$$x(t+h) = x(t) + \frac{h}{2} \left[\frac{dx}{dt}(t) + \frac{dx}{dt}(t+h) \right] + O(h^3) \quad (2.16)$$

leading to the integration scheme

$$X_{n+1} = X_n + \frac{1}{2}h(F_n + F_{n+1}) \quad (2.17)$$

This, like (2.11) and (2.12), is a second-order scheme. It has the problem that, in order to calculate X_{n+1} , the value of F_{n+1} is needed, which, in its turn, requires that the value of X_{n+1} be known. This seems to be a circular argument!

One way to work around this problem and turn (2.17) into a usable scheme is to start by working out a rough value of X_{n+1} , use that to compute a value of F_{n+1} , and then use (2.17) to compute a more accurate value of X_{n+1} . Such a process can be derived as follows. We know that

$$x(t+h) = x(t) + h \frac{dx}{dt}(t) + O(h^2)$$

Let

$$\hat{x}(t+h) = x(t) + h \frac{dx}{dt}(t) \quad (2.18)$$

so that

$$x(t+h) = \hat{x}(t+h) + O(h^2)$$

or, using the subscript notation defined above,

$$x_{n+1} = \hat{x}_{n+1} + O(h^2)$$

Now

$$\begin{aligned} \frac{dx_{n+1}}{dt} &= f(t_{n+1}, x_{n+1}) \\ &= f(t_{n+1}, \hat{x}_{n+1} + O(h^2)) \\ &= f(t_{n+1}, \hat{x}_{n+1}) + O(h^2) \frac{\partial f}{\partial x}(t_{n+1}, \hat{x}_{n+1}) + O(h^4) \\ &= f(t_{n+1}, \hat{x}_{n+1}) + O(h^2) \end{aligned} \quad (2.19)$$

In the subscript notation (2.16) is

$$x_{n+1} = x_n + \frac{1}{2} h(f(t_n, x_n) + f(t_{n+1}, x_{n+1})) + O(h^3)$$

Substituting (2.19) into this gives

$$x_{n+1} = x_n + \frac{1}{2} h(f(t_n, x_n) + f(t_{n+1}, \hat{x}_{n+1}) + O(h^2)) + O(h^3)$$

That is,

$$x_{n+1} = x_n + \frac{1}{2} h(f(t_n, x_n) + f(t_{n+1}, \hat{x}_{n+1})) + O(h^3) \quad (2.20)$$

Equation (2.20) together with (2.18) forms the basis of what is known as a **predictor–corrector method**, which is defined by the following scheme:

- (1) compute the ‘predicted’ value of X_{n+1} , call it \hat{X}_{n+1} , from

$$\hat{X}_{n+1} = X_n + hf(t_n, X_n) \quad (2.21a)$$

- (2) compute the ‘corrected’ value of X_{n+1} from

$$X_{n+1} = X_n + \frac{1}{2} h(f(t_n, X_n) + f(t_{n+1}, \hat{X}_{n+1})) \quad (2.21b)$$

This predictor–corrector scheme, as demonstrated by (2.20), is a second-order method. It has the advantage over (2.11) and (2.12) of requiring only the value of X_n , not X_{n-1} or F_{n-1} . On the other hand, each step requires two evaluations of the function $f(t, x)$, and so the method is less efficient computationally.

Example 2.8

Solve the initial-value problem

$$\frac{dx}{dt} = \frac{x^2}{t+1}, \quad x(0) = 1$$

posed in Example 2.2 using the second-order predictor–corrector scheme with a step size $h = 0.1$. Compute the values of $X(t)$ for $t = 0.1, 0.2, \dots, 1.0$ and compare them with the values of the exact solution $x(t)$.

Solution

The exact solution was given in Example 2.2. In this example the initial value of t is 0 and $x(0) = 1$. Using the standard notation we have $t_0 = 0$, and $x_0 = x(t_0) = x(0) = 1$. The function $f(t, x) = x^2/(t+1)$. So the first two steps of the computation are thus

$$\hat{X}_1 = x_0 + hf(t_0, x_0) = x_0 + h \frac{x^2}{t+1} = 1 + 0.1 \frac{1^2}{0+1} = 1.100\ 00$$

$$\begin{aligned} X_1 &= x_0 + \frac{1}{2} h[f(t_0, x_0) + f(t_1, \hat{X}_1)] = x_0 + \frac{1}{2} h \left(\frac{x_0^2}{t_0+1} + \frac{\hat{X}_1^2}{t_1+1} \right) \\ &= 1.000\ 00 + \frac{1}{2} 0.1 \left(\frac{1^2}{0+1} + \frac{1.100\ 00^2}{0.100\ 00+1} \right) = 1.105\ 00 \end{aligned}$$

$$\begin{aligned}
 \hat{X}_2 &= X_1 + hf(t_1, X_1) = X_1 + h \frac{X_1^2}{t_1 + 1} \\
 &= 1.105\ 00 + 0.1 \frac{1.105\ 00^2}{0.100\ 00 + 1} = 1.216\ 00 \\
 X_2 &= X_1 + \frac{1}{2} h[f(t_1, X_1) + f(t_2, \hat{X}_2)] \\
 &= X_1 + \frac{1}{2} h \left(\frac{X_1^2}{t_1 + 1} + \frac{\hat{X}_2^2}{t_2 + 1} \right) \\
 &= 1.105\ 00 + \frac{1}{2} 0.1 \left(\frac{1.105\ 00^2}{0.100\ 00 + 1} + \frac{1.216\ 00^2}{0.200\ 00 + 1} \right) = 1.222\ 11
 \end{aligned}$$

The complete computation is set out in Figure 2.13.

Figure 2.13
Computational results
for Example 2.8.

t	X_n	$f(t_n, X_n)$	\hat{X}_{n+1}	$f(t_{n+1}, \hat{X}_{n+1})$	$x(t)$	$ x - X_n $
0.000 00	1.000 00	1.000 00	1.100 00	1.100 00	1.000 00	0.000 00
0.100 00	1.105 00	1.110 02	1.216 00	1.232 22	1.105 35	0.000 35
0.200 00	1.222 11	1.244 63	1.346 58	1.394 82	1.222 97	0.000 86
0.300 00	1.354 08	1.410 42	1.495 13	1.596 72	1.355 68	0.001 60
0.400 00	1.504 44	1.616 67	1.666 11	1.850 61	1.507 10	0.002 65
0.500 00	1.677 81	1.876 69	1.865 47	2.175 00	1.681 99	0.004 18
0.600 00	1.880 39	2.209 92	2.101 38	2.597 53	1.886 81	0.006 42
0.700 00	2.120 76	2.645 67	2.385 33	3.161 00	2.130 51	0.009 75
0.800 00	2.411 10	3.229 66	2.734 06	3.934 26	2.425 93	0.014 83
0.900 00	2.769 29	4.036 30	3.172 92	5.033 72	2.792 16	0.022 87
1.000 00	3.222 79				3.258 89	0.036 10



Again the solution to this example can be obtained using MAPLE. The following worksheet computes the numerical and analytical solutions and compares them at the required points.

```

> #set up differential equation
> deq1:=diff(x(t),t)=x(t)^2/(t+1);init1:=x(0)=1;
> #obtain x1, the numerical solution
> x1:=dsolve({deq1, init1},
numeric,method=classical[heunform],output=listprocedure,
stepsize=0.1);
> #xa is the analytic solution
> xa:=dsolve({deq1, init1});
> #compute values at required solution points
> for i from 1 to 10 do
t:=0.1*i:op(2,x1[2])(t),evalf(op(2,xa)) end do;

```

Comparison of the result of Example 2.8 with those of Examples 2.2 and 2.6 shows that, as we should expect, the predictor–corrector scheme produces results of considerably higher accuracy than the Euler method and of comparable (though slightly better) accuracy to the

second-order Adams–Bashforth scheme. We also expect the scheme to have a global error $O(h^2)$, and, in the spirit of Examples 2.2 and 2.6, we confirm this in Example 2.9.

Example 2.9

Let X_a denote the approximation to the solution of the initial-value problem

$$\frac{dx}{dt} = \frac{x^2}{t+1}, \quad x(0) = 1$$

obtained using the second-order predictor–corrector method with a step size $h = 0.1$, and X_b that obtained using $h = 0.05$. Compute the values of $X_a(t)$ and $X_b(t)$ for $t = 0.1, 0.2, \dots, 1.0$. Compare these with the values of $x(t)$, the exact solution of the problem. Compute the ratio of the errors in X_a and X_b .

Solution

The numerical solutions X_a and X_b and their errors are shown in Figure 2.14. The ratio of the errors confirms that the error behaves roughly as $O(h^2)$.

Figure 2.14

Computational results for Example 2.9.

t	X_a	X_b	$x(t)$	$ x - X_a $	$ x - X_b $	$\frac{ x - X_b }{ x - X_a }$
0.000 00	1.000 00	1.000 00	1.000 00	0.000 00	0.000 00	0.00
0.100 00	1.105 00	1.105 26	1.105 35	0.000 35	0.000 09	0.27
0.200 00	1.222 11	1.222 74	1.222 97	0.000 86	0.000 23	0.27
0.300 00	1.354 08	1.355 25	1.355 68	0.001 60	0.000 43	0.27
0.400 00	1.504 44	1.506 38	1.507 10	0.002 65	0.000 72	0.27
0.500 00	1.677 81	1.680 86	1.681 99	0.004 18	0.001 13	0.27
0.600 00	1.880 39	1.885 07	1.886 81	0.006 42	0.001 73	0.27
0.700 00	2.120 76	2.127 87	2.130 51	0.009 75	0.002 64	0.27
0.800 00	2.411 10	2.421 90	2.425 93	0.014 83	0.004 03	0.27
0.900 00	2.769 29	2.785 92	2.792 16	0.022 87	0.006 24	0.27
1.000 00	3.222 79	3.248 98	3.258 89	0.036 10	0.009 91	0.27

In Section 2.3.5 we mentioned the difficulties that multistep methods introduce with respect to starting the computation. We now have a second-order method that does not need values of X_{n-1} or earlier. Obviously we can use this method just as it stands, but we then pay the penalty, in computer processing time, of the extra evaluation of $f(t, x)$ at each step of the process. An alternative scheme is to use the second-order predictor–corrector for the first step and then, because the appropriate function values are now available, change to the second-order Adams–Bashforth scheme – or even, if the problem is one for which the scheme given by (2.11) (which is called the **central difference scheme**) is stable, to that process. In this way we create a hybrid process that retains the $O(h^2)$ convergence and simultaneously minimizes the computational load.

The principles by which we derive (2.16) and so the integration scheme (2.17) can be extended to produce higher-order schemes. Such schemes are called the **Adams–Moulton formulae** and are as follows:

$$\begin{aligned} X_{n+1} &= X_n + hF_{n+1} \\ X_{n+1} &= X_n + \frac{1}{2}h(F_{n+1} + F_n) \\ X_{n+1} &= X_n + \frac{1}{12}h(5F_{n+1} + 8F_n - F_{n-1}) \\ X_{n+1} &= X_n + \frac{1}{24}h(9F_{n+1} + 19F_n - 5F_{n-1} + F_{n-2}) \end{aligned}$$

These are first-, second-, third- and fourth-order formulae respectively. They are all like the one we derived in this section in that the value of F_{n+1} is required in order to compute the value of X_{n+1} . They are therefore usually used as corrector formulae in predictor–corrector schemes. The most common way to do this is to use the $(p - 1)$ th-order Adams–Bashforth formula as predictor, with the p th-order Adams–Moulton formula as corrector. This combination can be shown to always produce a scheme of p th order. The predictor–corrector scheme we have derived in this section is of this form, with $p = 2$. Of course, for $p > 2$ the predictor–corrector formula produced is no longer self-starting, and other means have to be found to produce the first few values of X . We shall return to this topic in the next section.

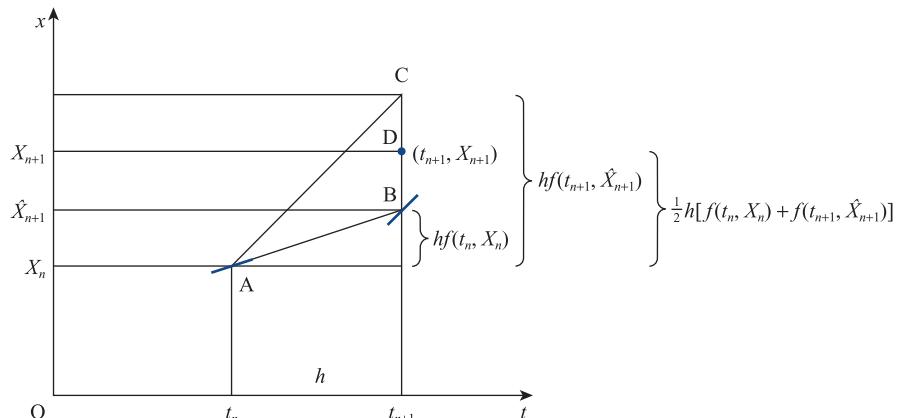


It may be noted that one of the alternative methods offered by MATLAB for the numerical solution of ODEs is based on the families of Adams–Bashforth and Adams–Moulton formulae.

2.3.8 More accurate solution methods: Runge–Kutta methods

Another class of higher-order methods comprises the Runge–Kutta methods. The mathematical derivation of these methods is quite complicated and beyond the scope of this book. However, their general principle can be explained informally by a graphical argument. Mathematical details can be found in the references on page 130 above Example 2.5 and in C. F. Gerald and P. O. Wheatley, *Applied Numerical Analysis* (Upper Saddle River, NJ, Pearson, 2003). Figure 2.15 shows a geometrical interpretation of the second-order predictor–corrector method introduced in the last section. Starting at the point (t_n, X_n) , point A in the diagram, the predicted value \hat{X}_{n+1} is calculated. The line AB has gradient $f(t_n, X_n)$, so the ordinate of the point B is the predicted value \hat{X}_{n+1} . The line AC in the diagram has gradient $f(t_{n+1}, \hat{X}_{n+1})$, the gradient of the direction field of the equation at point B, so point C has ordinate $X_n + hf(t_{n+1}, \hat{X}_{n+1})$. The midpoint of the line BC, point D, has ordinate

Figure 2.15
A geometrical interpretation of the second-order predictor–corrector method.



$X_n + \frac{1}{2}h(f(t_n, X_n) + f(t_{n+1}, \hat{X}_{n+1}))$, which is the value of X_{n+1} given by the corrector formula. Geometrically speaking, the predictor–corrector scheme can be viewed as the process of calculating the gradient of the direction field of the equation at points A and B and then assuming that the average gradient of the solution over the interval (t_n, t_{n+1}) is reasonably well estimated by the average of the gradients at these two points. The Euler method, of course, is equivalent to assuming that the gradient at point A is a good estimate of the average gradient of the solution over the interval (t_n, t_{n+1}) . Given this insight, it is unsurprising that the error performance of the predictor–corrector method is superior to that of the Euler method.

Runge–Kutta methods extend this principle by using the gradient at several points in the interval (t_n, t_{n+1}) to estimate the average gradient of the solution over the interval. The most commonly used Runge–Kutta method is a fourth-order one which can be expressed as follows:

$$c_1 = hf(t_n, X_n) \quad (2.22a)$$

$$c_2 = hf(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_1) \quad (2.22b)$$

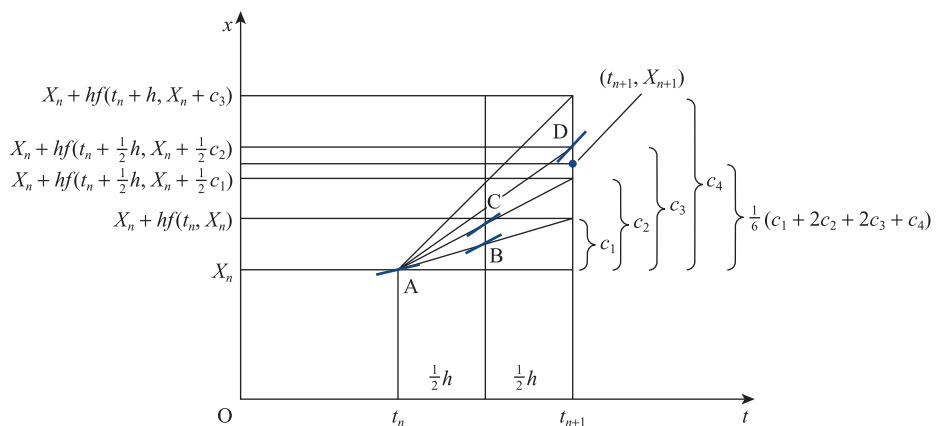
$$c_3 = hf(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_2) \quad (2.22c)$$

$$c_4 = hf(t_n + h, X_n + c_3) \quad (2.22d)$$

$$X_{n+1} = X_n + \frac{1}{6}(c_1 + 2c_2 + 2c_3 + c_4) \quad (2.22e)$$

Geometrically, this can be understood as the process shown in Figure 2.16. The line AB has the same gradient as the equation’s direction field at point A. The ordinate of this line at $t_n + \frac{1}{2}h$ defines point B. The line AC has gradient equal to the direction of the direction field at point B. This line defines point C. Finally, a line AD, with gradient equal to the direction of the direction field at point C, defines point D. The average gradient of the solution over the interval (t_n, t_{n+1}) is then estimated from a weighted average of the gradients at points A, B, C and D. It is intuitively acceptable that such a process is likely to give a highly accurate estimate of the average gradient over the interval.

Figure 2.16
A geometrical interpretation of the fourth-order Runge–Kutta method.



As was said before, the mathematical proof that the process defined by (2.22a–e) is a fourth-order process is beyond the scope of this text. It is interesting to note that the predictor–corrector method defined by (2.21a, b) could also be expressed as

$$\begin{aligned} c_1 &= hf(t_n, X_n) \\ c_2 &= hf(t_n + h, X_n + c_1) \\ X_{n+1} &= X_n + \frac{1}{2}(c_1 + c_2) \end{aligned}$$

This is also of the form of a Runge–Kutta method (the second-order Runge–Kutta method), so we find that the second-order Runge–Kutta method and the second-order Adams–Bashforth/Adams–Moulton predictor–corrector are, in fact, equivalent processes.

Example 2.10

Let X_a denote the approximation to the solution of the initial-value problem

$$\frac{dx}{dt} = \frac{x^2}{t+1}, \quad x(0) = 1$$

obtained using the fourth-order Runge–Kutta method with a step size $h = 0.1$, and X_b that obtained using $h = 0.05$. Compute the values of $X_a(t)$ and $X_b(t)$ for $t = 0.1, 0.2, \dots, 1.0$. Compare these with the values of $x(t)$, the exact solution of the problem. Compute the ratio of the errors in X_a and X_b .

Solution

The exact solution was given in Example 2.2. The numerical solutions X_a and X_b and their errors are presented in Figure 2.17.

This example shows, first, that the Runge–Kutta scheme, being a fourth-order scheme, has considerably smaller errors, in absolute terms, than any of the other methods we have met so far (note that Figure 2.17 does not give raw errors but errors times 1000!) and, second, that the expectation we have that the global error should be $O(h^4)$ is roughly borne out in practice (the ratio of $|x - X_a|$ to $|x - X_b|$ is roughly 16 : 1).

Figure 2.17
Computational results
for Example 2.10.

t	X_a	X_b	$x(t)$	$ x - X_a \times 10^3$	$ x - X_b \times 10^3$	$\frac{ x - X_b }{ x - X_a }$
0.00000	1.000 000 0	1.000 000 0	1.000 000 0	0.000 55	0.000 04	0.0682
0.10000	1.105 350 7	1.105 351 2	1.105 351 2	0.000 55	0.000 09	0.0680
0.20000	1.222 973 3	1.222 974 5	1.222 974 6	0.001 33	0.000 17	0.0679
0.30000	1.355 680 2	1.355 682 5	1.355 682 7	0.002 46	0.000 28	0.0678
0.40000	1.507 091 8	1.507 095 7	1.507 095 9	0.004 10	0.000 44	0.0678
0.50000	1.681 980 5	1.681 986 6	1.681 987 1	0.006 53	0.000 69	0.0677
0.60000	1.886 795 2	1.886 804 7	1.886 805 4	0.010 20	0.001 08	0.0677
0.70000	2.130 491 5	2.130 506 4	2.130 507 4	0.015 92	0.001 71	0.0677
0.80000	2.425 903 1	2.425 926 6	2.425 928 3	0.025 19	0.002 78	0.0677
0.90000	2.792 115 5	2.792 153 7	2.792 156 5	0.041 03	0.004 74	0.0678
1.00000	3.258 821 4	3.258 886 6	3.258 891 4	0.069 94	0.004 74	0.0678



The table of values in Figure 2.17 can be obtained using MAPLE with the appropriate setting of the numerical method. The following worksheet computes the solutions specified and composes the required table.

```

> #set up differential equation
> deq1:=diff(x(t),t)=x(t)^2/(t+1);init1:=x(0)=1;
> #obtain x1 and x2, the numerical solutions
> x1:=dsolve({deq1, init1}, numeric,method=classical[rk4],
           output=listprocedure,stepsize=0.1);
> x2:=dsolve({deq1, init1},numeric,method=classical[rk4],
           output=listprocedure,stepsize=0.05);
> #xa is the analytic solution
> xa:=dsolve({deq1, init1});
> printlevel:=0:
fmtstr:="%5.1f,%12.7f,%12.7f,%12.7f,%10.5f,%10.5f,
           %10.4f,\n":
for i from 1 to 10 do
  t:=0.1*i:
  xx1:=op(2,x1[2])(t):
  xx2:=op(2,x2[2])(t):
  xxa:=evalf(subs(t=1,op(2,xa))):
  printf(fmtstr,t,xx1,xx2,xxa,abs(xx1-xxa)*1e3,
         abs(xx2-xxa)*1e3,(xx2-xxa)/(xx1-xxa));
end do;
```

It is interesting to note that the MAPLE results in the right-hand column, the ratio of the errors in the two numerical solutions, vary slightly from those in Figure 2.17. The results in Figure 2.17 were computed using the high-level programming language Pascal which uses a different representation of floating point numbers from that used by MAPLE. The variation in the results is an effect of the differing levels of precision in the two languages. The differences are, of course, small and do not change the overall message obtained from the figure.

Runge–Kutta schemes are single-step methods in the sense that they only require the value of X_n , not the value of X at any steps prior to that. Therefore, they are entirely self-starting, unlike the predictor–corrector and other multistep methods. On the other hand, Runge–Kutta methods proceed by effectively creating substeps within each step. Thus they require more evaluations of the function $f(t, x)$ at each step than multistep methods of equivalent order of accuracy. For this reason, they are computationally less efficient. Because they are self-starting, however, Runge–Kutta methods can be used to start the process for multistep methods. An example of an efficient scheme that consistently has a fourth-order local error is as follows. Start by taking two steps using the fourth-order Runge–Kutta method. At this point values of X_0 , X_1 and X_2 are available, so, to achieve computational efficiency, change to the three-step fourth-order predictor–corrector consisting of the third-order Adams–Bashforth/fourth-order Adams–Moulton pair.

2.3.9 Exercises

(Note that Questions 8–15 may be attempted using a hand-held calculator, particularly if it is of the programmable variety. The arithmetic will, however, be found to be tedious, and the use of computer assistance is recommended if the maximum benefit is to be obtained from completing these questions.)

- 8** Using the second-order Adams–Bashforth method (start the process with a single step using the second-order predictor–corrector method),
- compute an estimate of $x(0.5)$ for the initial-value problem

$$\frac{dx}{dt} = x^2 \sin t - x, \quad x(0) = 0.2$$

using step size $h = 0.1$;

- compute an estimate of $x(1.2)$ for the initial-value problem

$$\frac{dx}{dt} = x^2 e^x, \quad x(0.5) = 0.5$$

using step size $h = 0.1$.

- 9** Using the third-order Adams–Bashforth method (start the process with two second-order predictor–corrector method steps) compute an estimate of $x(0.5)$ for the initial-value problem

$$\frac{dx}{dt} = \sqrt{(x^2 + 2t)}, \quad x(0) = 1$$

using step size $h = 0.1$.

- 10** Using the second-order predictor–corrector method,
- compute an estimate of $x(0.5)$ for the initial-value problem

$$\frac{dx}{dt} = (2t + x) \sin 2t, \quad x(0) = 0.5$$

using step size $h = 0.05$;

- compute an estimate of $x(1)$ for the initial-value problem

$$\frac{dx}{dt} = -\frac{1+x}{\sin(t+1)}, \quad x(0) = -2$$

using step size $h = 0.1$.

11

Write down the first three terms of the Taylor series expansions of the functions

$$\frac{dx}{dt}(t-h) \quad \text{and} \quad \frac{dx}{dt}(t-2h)$$

about $x(t)$. Use these two equations to eliminate

$$\frac{d^2x}{dt^2}(t) \quad \text{and} \quad \frac{d^3x}{dt^3}(t)$$

from the Taylor series expansion of the function $x(t+h)$ about $x(t)$. Show that the resulting formula for $x(t+h)$ is the third member of the Adams–Bashforth family, and hence confirm that this Adams–Bashforth method is a third-order method.

12

Write down the first three terms of the Taylor series expansions of the functions

$$\frac{dx}{dt}(t+h) \quad \text{and} \quad \frac{dx}{dt}(t-h)$$

about $x(t)$. Use these two equations to eliminate

$$\frac{d^2x}{dt^2}(t) \quad \text{and} \quad \frac{d^3x}{dt^3}(t)$$

from the Taylor series expansion of the function $x(t+h)$ about $x(t)$. Show that the resulting formula for $x(t+h)$ is the third member of the Adams–Moulton family, and hence confirm that this Adams–Moulton method is a third-order method.

13

Write down the first four terms of the Taylor series expansion of the function $x(t-h)$ about $x(t)$, and the first three terms of the expansion of the function

$$\frac{dx}{dt}(t-h)$$

about $x(t)$. Use these two equations to eliminate

$$\frac{d^2x}{dt^2}(t) \quad \text{and} \quad \frac{d^3x}{dt^3}(t)$$

from the Taylor series expansion of the function $x(t+h)$ about $x(t)$. Show that the resulting formula is

$$X_{n+1} = -4X_n + 5X_{n-1} + h(4F_n + 2F_{n-1}) + O(h^4)$$

Show that this method is a linear combination of the second-order Adams–Bashforth method and the central difference method (that is, the scheme based on (2.9)). What do you think, in view of this, might be its disadvantages?

14

-  Using the third-order Adams–Bashforth–Moulton predictor–corrector method (that is, the second-order Adams–Bashforth formula as predictor and the third-order Adams–Moulton formula as corrector), compute an estimate of $x(0.5)$ for the initial-value problem

$$\frac{dx}{dt} = x^2 + t^2, \quad x(0.3) = 0.1$$

using step size $h = 0.05$. (You will need to employ another method for the first step to start this scheme – use the fourth-order Runge–Kutta method.)

15

-  Using the fourth-order Runge–Kutta method,

- (a) compute an estimate of $x(0.75)$ for the initial-value problem

$$\frac{dx}{dt} = x + t + xt, \quad x(0) = 1$$

using step size $h = 0.15$;

- (b) compute an estimate of $x(2)$ for the initial-value problem

$$\frac{dx}{dt} = \frac{1}{x+t}, \quad x(1) = 2$$

using step size $h = 0.1$.

16

- Consider the initial-value problem

$$\frac{dx}{dt} = x^2 + t^{3/2}, \quad x(0) = -1$$

- (a) Compute estimates of $x(2)$ using the second-order Adams–Bashforth scheme (using the second-order predictor–corrector to start the computation) with step sizes $h = 0.2$ and 0.1 . From these two estimates of $x(2)$ estimate what step size would be needed to compute an estimate of $x(2)$ accurate to 3 decimal places. Compute $X(2)$, first using your estimated step size and second using half your estimated step size. Does the required accuracy appear to have been achieved?
- (b) Compute estimates of $x(2)$ using the second-order predictor–corrector scheme with step sizes $h = 0.2$ and 0.1 . From these two estimates of $x(2)$

estimate what step size would be needed with this scheme to compute an estimate of $x(2)$ accurate to 3 decimal places. Compute $X(2)$, first using your estimated step size and second using half your estimated step size. Does the required accuracy appear to have been achieved?

- (c) Compute estimates of $x(2)$ using the fourth-order Runge–Kutta scheme with step sizes $h = 0.4$ and 0.2 . From these two estimates of $x(2)$ estimate what step size would be needed to compute an estimate of $x(2)$ accurate to 5 dp. Compute $X(3)$, first using your estimated step size and second using half your estimated step size. Does the required accuracy appear to have been achieved?

17

- For the initial-value problem

$$\frac{dx}{dt} = x^2 e^{-t}, \quad x(1) = 1$$

find, by any method, an estimate, accurate to 5dp, of the value of $x(3)$.



Note: All of the exercises in this section can be completed by programming the algorithms in a high-level computer language such as Pascal, C and Java. Programming in a similar high-level style can be achieved using the language constructs embedded within the MATLAB and MAPLE packages. MAPLE, as we have already seen, and MATLAB also allow a higher-level style of programming using their built-in procedures for numerical solution of ODEs. Both MATLAB and MAPLE have very sophisticated built-in procedures, but MAPLE also allows the user to specify that it should use simpler algorithms (which it calls ‘classic’ algorithms). Amongst these simpler algorithms are many of the algorithms we discuss in this chapter. In the preceding exercise set, those which specify the Runge–Kutta method and the second-order predictor–corrector could be completed using MAPLE’s `dsolve` procedure specifying the relevant ‘classic’ solution methods.

2.3.10 Stiff equations

There is a class of differential equations, known as **stiff differential equations**, that are apt to be somewhat troublesome to solve numerically. It is beyond the scope of this text to explore the topic of stiff equations in any great detail. It is, however, important to be aware of the possibility of difficulties from this source and to be able to recognize the sort of equations that are likely to be stiff. In that spirit we shall present a very informal treatment of stiff equations and the sort of troubles that they cause. Example 2.11 shows the sort of behaviour that is typical of stiff differential equations.

Example 2.11

The equation

$$\frac{dx}{dt} = 1 - x, \quad x(0) = 2 \quad (2.23)$$

has analytical solution $x = 1 + e^{-t}$. The equation

$$\frac{dx}{dt} = 50(1 - x) + 50e^{-t}, \quad x(0) = 2 \quad (2.24)$$

has analytical solution $x = 1 + \frac{1}{49}(50e^{-t} - e^{-50t})$. The two solutions are shown in Figure 2.18.

Suppose that it were not possible to solve the two equations analytically and that numerical solutions must be sought. The form of the two solutions shown in Figure 2.18 is not very different, and it might be supposed (at least naively) that the numerical solution of the two equations would present similar problems. This, however, is far from the case.

Figure 2.19 shows the results of solving the two equations using the second-order predictor–corrector method with step size $h = 0.01$. The numerical and exact solutions of (2.23) are denoted by X_a and x_a respectively, and those of (2.24) by X_b and x_b . The third and fifth columns give the errors in the numerical solutions (compared with the exact solutions), and the last column gives the ratio of the errors. The solution X_a is seen to be considerably more accurate than X_b using the same step size.

Figure 2.18
The analytical solutions of (2.23) and (2.24).

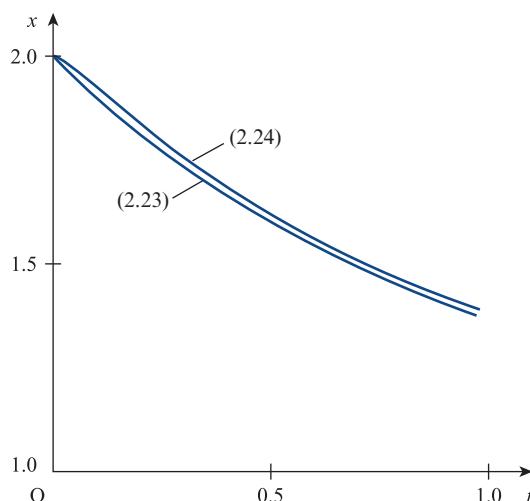


Figure 2.19

Computational results for Example 2.11;
 $h = 0.01$.

t	X_a	$ X_a - x_a $	X_b	$ X_b - x_b $	Ratio of errors
0.000 00	2.000 00	0.000 000	2.000 00	0.000 000	
0.100 00	1.904 84	0.000 002	1.923 15	0.000 017	11.264 68
0.200 00	1.818 73	0.000 003	1.835 47	0.000 028	10.022 19
0.300 00	1.740 82	0.000 004	1.755 96	0.000 026	6.864 34
0.400 00	1.670 32	0.000 005	1.684 02	0.000 023	5.150 07
0.500 00	1.606 54	0.000 005	1.618 93	0.000 021	4.120 06
0.600 00	1.548 82	0.000 006	1.560 03	0.000 019	3.433 38
0.700 00	1.496 59	0.000 006	1.506 74	0.000 017	2.942 90
0.800 00	1.449 34	0.000 006	1.458 51	0.000 016	2.575 03
0.900 00	1.406 58	0.000 006	1.414 88	0.000 014	2.288 92
1.000 00	1.367 89	0.000 006	1.375 40	0.000 013	2.060 02

Figure 2.20

Computational results for Example 2.11;
 $h = 0.025$.

t	X_a	$ X_a - x_a $	X_b	$ X_b - x_b $	Ratio of errors
0.000 00	2.000 00	0.000 000	2.000 00	0.000 000	
0.10000	1.904 85	0.000 010	1.922 04	0.001 123	116.951 24
0.20000	1.818 75	0.000 017	1.835 67	0.000 231	13.270 10
0.30000	1.740 84	0.000 024	1.756 25	0.000 317	13.438 84
0.40000	1.670 35	0.000 028	1.684 30	0.000 296	10.384 39
0.50000	1.606 56	0.000 032	1.619 18	0.000 268	8.328 98
0.60000	1.548 85	0.000 035	1.560 25	0.000 243	6.942 36
0.70000	1.496 62	0.000 037	1.506 94	0.000 220	5.950 68
0.80000	1.449 37	0.000 038	1.458 70	0.000 199	5.206 82
0.90000	1.406 61	0.000 039	1.415 05	0.000 180	4.628 26
1.00000	1.367 92	0.000 039	1.375 55	0.000 163	4.165 42

Figure 2.21

Computational results for Example 2.11;
 $h = 0.05$.

t	X_a	$ X_a - x_a $	X_b	$ X_b - x_b $
0.000 00	2.000 00	0.000 000	2.000 00	0.000 000
0.100 00	1.904 88	0.000 039	1.873 43	0.049 740
0.200 00	1.818 80	0.000 071	1.707 36	0.128 075
0.300 00	1.740 91	0.000 096	1.421 02	0.334 914
0.400 00	1.670 44	0.000 116	0.802 59	0.881 408
0.500 00	1.606 66	0.000 131	-0.705 87	2.324 778
0.600 00	1.548 95	0.000 142	-4.576 42	6.136 434
0.700 00	1.496 74	0.000 150	-14.695 10	16.201 818
0.800 00	1.449 48	0.000 156	-41.322 43	42.780 932
0.900 00	1.406 73	0.000 158	-111.551 73	112.966 595
1.000 00	1.368 04	0.000 159	-296.925 40	298.300 783

Figure 2.20 is similar to Figure 2.19, but with a step size $h = 0.025$. As we might expect, the error in the solution X_a is larger by a factor of roughly six (the global error of the second-order predictor–corrector method is $O(h^2)$). The errors in X_b , however, are larger by more than the expected factor, as is evidenced by the increase in the ratio of the error in X_b to that in X_a .

Figure 2.21 shows the results obtained using a step size $h = 0.05$. The errors in X_a are again larger by about the factor expected (25 when compared with Figure 2.19). The

solution X_b , however, shows little relationship to the exact solution x_b – so little that the error at $t = 1$ is over 20 000% of the exact solution. Obviously a numerical method that causes such large errors to accumulate is not at all satisfactory.

In Section 2.3.5 we met the idea that some numerical methods can, when applied to some classes of differential equation, show instability. What has happened here is, of course, that the predictor–corrector method is showing instability when used to solve (2.24) with a step size larger than some critical limit. Unfortunately the same behaviour is also manifest by the other methods that we have already come across – the problem lies with the equation (2.24), which is an example of a stiff differential equation.

The typical pattern with stiff differential equations is that, in order to avoid instability, the step size used to solve the equation using normal numerical methods must be very small when compared with the interval over which the equation is to be solved. In other words, the number of steps to be taken is very large and the solution is costly in time and computing resources. Essentially, stiff equations are equations whose solution contains terms involving widely varying time scales. That (2.24) is of this type is evidenced by the presence of terms in both e^{-t} and e^{-50t} in the analytical solution. In order to solve such equations accurately, a step must be chosen that is small enough to cope with the shortest time scale. If the solution is required for times comparable to the long time scales, this can mean that very large numbers of steps are needed and the computer processing time needed to solve the problem becomes prohibitive. In Example 2.11 the time scale of the rapidly varying and the more slowly varying components of the solution differed by only a factor of 50. It is not unusual, in the physical problems arising from engineering investigations, to find time scales differing by three or more orders of magnitude; that is, factors of 1000 or more. In these cases the problems caused are proportionately amplified. Fortunately a number of numerical methods that are particularly efficient at solving stiff differential equations have been developed. It is beyond the scope of this text to treat these in any detail.

From the engineering point of view, the implication of the existence of stiff equations is that engineers must be aware of the possibility of meeting such equations and also of the nature of the difficulty for the numerical methods – the widely varying time scales inherent in the problem. It is probably easier to recognize that an engineering problem is likely to give rise to a stiff equation or equations because of the physical nature of the problem than it is to recognize a stiff equation in its abstract form isolated from the engineering context from which it arose. As is often the case, a judicious combination of mathematical reasoning and engineering intuition is more powerful than either approach in isolation.



Both MAPLE and MATLAB feature procedures for the numerical solution of ODEs which are designed to deal efficiently with stiff equations. The user may be tempted to think that a simple way to negotiate the problem of stiff equations is to use the stiff equation solvers for all ODEs. However, the stiff equation methods are less computationally efficient for non-stiff equations so it is worth trying to identify which type of equation one is facing and using the most appropriate methods.

2.3.11 Computer software libraries

In the last few sections we have built up some basic methods for the integration of first-order ODEs. These methods, particularly the more sophisticated ones – the fourth-order

Runge–Kutta and the predictor–corrector methods – suffice for many of the problems arising in engineering practice. However, for more demanding problems – demanding in terms of the scale of the problem or because the problem is characterized by ill behaviour of some form – there exist more sophisticated methods than those we are able to present in this book.

All the methods that we have presented in the last few sections use a fixed step size. Among the more sophisticated methods to which we have just alluded are some that use a variable step size. In Section 2.3.6 we showed how Richardson extrapolation can be used to estimate the size of the error in a numerical solution and, furthermore, to estimate the step size that should be used in order to compute a solution of a differential equation to some desired accuracy. The principle of the variable-step methods is that a running check is kept of the estimated error in the solution being computed. The error may be estimated by a formula derived along principles similar to that of Richardson extrapolation. This running estimate of the error is used to predict, at any point in the computation, how large a step can be taken while still computing a solution within any given error bound specified by the user. The step size used in the solution can be altered accordingly. If the error is approaching the limits of what is acceptable then the step size can be reduced; if it is very much smaller than that which can be tolerated then the step size may be increased in the interests of speedy and efficient computing. For multistep methods the change of step size can lead to quite complicated formulae or procedures. As an alternative, or in addition, to a change of step size, changes can be made in the order of the integration formula used. When increased accuracy is required, instead of reducing the step size, the order of the integration method can be increased, and vice versa. Implementations of the best of these more sophisticated schemes are readily available in software packages, such as MAPLE and MATLAB, and software libraries such as the NAG library.

The availability of complex and sophisticated ‘state of the art’ methods is not the only argument for the use of software packages and libraries. It is a good engineering principle that, if an engineer wishes to design and construct a reliable engineering artefact, tried and proven components of known reliability and performance characteristics should be used. This principle can also be extended to engineering software. It is almost always both more efficient, in terms of expenditure of time and intellectual energy, and more reliable, in terms of elimination of bugs and unwanted side-effects, to use software from a known and proven source than to write programs from scratch.

For both of the foregoing reasons, when reliable mathematical packages, such as MAPLE and MATLAB, and software libraries are available, their use is strongly recommended. MAPLE offers both symbolic manipulation (computer algebra) and numerical problem solving across the whole span of mathematics. Amongst these, as we have already noted, MAPLE includes routines for the numerical solution of systems of ODEs. These routines are highly sophisticated, offering alternative methods suitable for stiff and non-stiff problems, using fixed time steps or variable time steps and optimized either for speed or for accuracy. The MATLAB package, with its Simulink continuous system modelling add-on, also offers sophisticated facilities for solving differential equations numerically. Again the package offers the choice of both fixed and variable time step methods, methods suitable for stiff problems as well as non-stiff ones, and a choice of optimizations aimed at either best speed or highest accuracy. Amongst the best known, and probably the most widely used, library of software procedures today is the NAG library. This library has a long history and has been compiled by leading experts in the field of numerical mathematics. Routines are available in a variety of programming languages. The routines provided for the solution of ODEs again

encompass a variety of methods chosen to deal with stiff and non-stiff problems and to offer the user considerable flexibility in choice of method to suit every possible engineering requirement. By choosing an appropriate, high-quality software package or library the engineer can be assured that the implementation will be, as far as possible, bug free, that the methods used will be efficient and reliable, and that the algorithms will have been chosen from the best ‘state of the art’ methods.

It is tempting to believe that the use of software libraries solves all the problems of numerical analysis that an engineering user is likely to meet. Faced with a problem for which analytical methods fail, the engineer simply needs to thumb through the index to some numerical analysis software library until a method for solving the type of problem currently faced is found. Unfortunately such undiscerning use of packaged software will almost certainly, sooner or later, lead to a gross failure of some sort. If the user is fortunate, the software will be sophisticated enough to detect that the problem posed is outside its capabilities and to return an error message to that effect. If the user is less fortunate, the writer of the software will not have been able to foresee all the possible uses and misuses to which the software might be subjected and the software will not be proof against such use outside its range of applicability. In that case the software may produce seemingly valid answers while giving no indication of any potential problem. Under such circumstances the undiscerning user of engineering software is on the verge of committing a major engineering blunder. From such circumstances result failed bridges and crashed aircraft! It has been the objective of these sections on the numerical solution of differential equations both to equip readers with numerical methods suitable for the less demanding problems that will arise in their engineering careers and to give them sufficient understanding of the basics of this branch of numerical analysis that they may become discriminating, intelligent and wary users of packaged software and other aids to numerical computing.

2.4

Numerical methods for systems of ordinary differential equations and higher-order differential equations

Obviously, the classes of second- and higher-order differential equations that can be solved analytically, while representing an important subset of the totality of such equations, are relatively restricted. Just as for first-order equations, those for which no analytical solution exists can still be solved by numerical means. The numerical solution of second- and higher-order equations does not, in fact, need any significant new mathematical theory or technique.

2.4.1 Numerical solution of coupled first-order equations

In Section 2.3 we met various methods for the numerical solution of equations of the form

$$\frac{dx}{dt} = f(t, x)$$

that is, first-order differential equations involving a single dependent variable and a single independent variable. However it is possible to have sets of coupled first-order

equations, each involving the same independent variable but with more than one dependent variable. An example of these types of equation is

$$\frac{dx}{dt} = x - y^2 + xt \quad (2.25a)$$

$$\frac{dy}{dt} = 2x^2 + xy - t \quad (2.25b)$$

This is a pair of differential equations in the dependent variables x and y with the independent variable t . The derivative of each of the dependent variables depends not only on itself and on the independent variable t , but also on the other dependent variable. Neither of the equations can be solved in isolation or independently of the other – both must be solved simultaneously, or side by side. A pair of coupled differential equations such as (2.25) may be characterized as

$$\frac{dx}{dt} = f_1(t, x, y) \quad (2.26a)$$

$$\frac{dy}{dt} = f_2(t, x, y) \quad (2.26b)$$

For a set of p such equations it is convenient to denote the dependent variables not by x, y, z, \dots but by $x_1, x_2, x_3, \dots, x_p$ and the set of equations by

$$\frac{dx_i}{dt} = f_i(t, x_1, x_2, \dots, x_p) \quad (i = 1, 2, \dots, p)$$

or equivalently, using vector notation,

$$\frac{d}{dt}[\mathbf{x}] = \mathbf{f}(t, \mathbf{x})$$

where $\mathbf{x}(t)$ is a vector function of t given by

$$\mathbf{x}(t) = [x_1(t) \quad x_2(t) \quad \cdots \quad x_p(t)]^T$$

$\mathbf{f}(t, \mathbf{x})$ is a vector-valued function of the scalar variable t and the vector variable \mathbf{x} .

The Euler method for the solution of a single differential equation takes the form

$$X_{n+1} = X_n + h f(t_n, X_n)$$

If we were to try to apply this method to (2.26a), we should obtain

$$X_{n+1} = X_n + h f_1(t_n, X_n, Y_n)$$

In other words, the value of X_{n+1} depends not only on t_n and X_n but also on Y_n . In the same way, we would obtain

$$Y_{n+1} = Y_n + h f_2(t_n, X_n, Y_n)$$

for Y_{n+1} . In practice, this means that to solve two simultaneous differential equations, we must advance the solution of both equations simultaneously in the manner shown in Example 2.12.

Example 2.12

Find the value of $X(1.4)$ satisfying the following initial-value problem:

$$\frac{dx}{dt} = x - y^2 + xt, \quad x(1) = 0.5$$

$$\frac{dy}{dt} = 2x^2 + xy - t, \quad y(1) = 1.2$$

using the Euler method with time step $h = 0.1$.

Solution The right-hand sides of the two equations will be denoted by $f_1(t, x, y)$ and $f_2(t, x, y)$ respectively, so

$$f_1(t, x, y) = x - y^2 + xt \quad \text{and} \quad f_2(t, x, y) = 2x^2 + xy - t$$

The initial condition is imposed at $t = 1$, so t_n will denote $1 + nh$, X_n will denote $X(1 + nh)$, and Y_n will denote $Y(1 + nh)$. Then we have

$$\begin{aligned} X_1 &= x_0 + hf_1(t_0, x_0, y_0) & Y_1 &= y_0 + hf_2(t_0, x_0, y_0) \\ &= 0.5 + 0.1f_1(1, 0.5, 1.2) & &= 1.2 + 0.1f_2(1, 0.5, 1.2) \\ &= 0.4560 & &= 1.2100 \end{aligned}$$

for the first step. The next step is therefore

$$\begin{aligned} X_2 &= X_1 + hf_1(t_1, X_1, Y_1) & Y_2 &= Y_1 + hf_2(t_1, X_1, Y_1) \\ &= 0.4560 & &= 1.2100 \\ &+ 0.1f_1(1.1, 0.4560, 1.2100) & &+ 0.1f_2(1.1, 0.4560, 1.2100) \\ &= 0.4054 & &= 1.1968 \end{aligned}$$

and the third step is

$$\begin{aligned} X_3 &= 0.4054 & Y_3 &= 1.1968 \\ &+ 0.1f_1(1.2, 0.4054, 1.1968) & &+ 0.1f_2(1.2, 0.4054, 1.1968) \\ &= 0.3513 & &= 1.1581 \end{aligned}$$

Finally, we obtain

$$\begin{aligned} X_4 &= 0.3513 + 0.1f_1(1.3, 0.3513, 1.1581) \\ &= 0.2980 \end{aligned}$$

Hence we have $X(1.4) = 0.2980$.



MAPLE's `dsolve` procedure can find the numerical solution of sets of coupled ordinary differential equations as readily as for a single differential equation. The following worksheet finds the solution required in the example above.

```
> #set up the two differential equations
> deq1:=diff(x(t),t)=x(t)*(1+t)-y(t)^2;
deq2:=diff(y(t),t)=2*x(t)^2+x(t)*y(t)-t;
deqsystem:=deq1,deq2;
> #set up the initial conditions
> inits:=x(1)=0.5,y(1)=1.2;
> #procedure "dsolve" used to solve s system of two coupled
      differential equations
> sol:=dsolve({deqsystem, inits}, numeric,
method=classical[foreuler],output=listprocedure,
stepsize=0.1);
> #obtain numerical solution required
> xx:=op(2,sol[2]);xx(1.4);
```

The principle of solving the two equations side by side extends in exactly the same way to the solution of more than two simultaneous equations and to the solution of simultaneous differential equations by methods other than the Euler method.

Example 2.13

Find the value of $X(1.4)$ satisfying the following initial-value problem:

$$\frac{dx}{dt} = x - y^2 + xt, \quad x(1) = 0.5$$

$$\frac{dy}{dt} = 2x^2 + xy - t, \quad y(1) = 1.2$$

using the second-order predictor–corrector method with time step $h = 0.1$.

Solution

First step:

predictor

$$\begin{aligned}\hat{X}_1 &= x_0 + hf_1(t_0, x_0, y_0) \\ &= 0.4560\end{aligned}$$

$$\begin{aligned}\hat{Y}_1 &= y_0 + hf_2(t_0, x_0, y_0) \\ &= 1.2100\end{aligned}$$

corrector

$$\begin{aligned}X_1 &= x_0 + \frac{1}{2}h[f_1(t_0, x_0, y_0) \\ &\quad + f_1(t_1, \hat{X}_1, \hat{Y}_1)] \\ &= 0.5 + 0.05[f_1(1, 0.5, 1.2) \\ &\quad + f_1(1.1, 0.456, 1.21)] \\ &= 0.4527\end{aligned}$$

$$\begin{aligned}Y_1 &= y_0 + \frac{1}{2}h[f_2(t_0, x_0, y_0) \\ &\quad + f_2(t_1, \hat{X}_1, \hat{Y}_1)] \\ &= 1.2 + 0.05[f_2(1, 0.5, 1.2) \\ &\quad + f_2(1.1, 0.456, 1.21)] \\ &= 1.1984\end{aligned}$$

Second step:

predictor

$$\begin{aligned}\hat{X}_2 &= X_1 + hf_1(t_1, X_1, Y_1) \\ &= 0.4042\end{aligned}$$

$$\begin{aligned}\hat{Y}_2 &= Y_1 + hf_2(t_1, X_1, Y_1) \\ &= 1.1836\end{aligned}$$

corrector

$$\begin{aligned}X_2 &= X_1 + \frac{1}{2}h[f_1(t_1, X_1, Y_1) \\ &\quad + f_1(t_2, \hat{X}_2, \hat{Y}_2)] \\ &= 0.4527 \\ &\quad + 0.05[f_1(1.1, 0.4527, 1.1984) \\ &\quad + f_1(1.2, 0.4042, 1.1836)] \\ &= 0.4028\end{aligned}$$

$$\begin{aligned}Y_2 &= Y_1 + \frac{1}{2}h[f_2(t_1, X_1, Y_1) \\ &\quad + f_2(t_2, \hat{X}_2, \hat{Y}_2)] \\ &= 1.1984 \\ &\quad + 0.05[f_2(1.1, 0.4527, 1.1984) \\ &\quad + f_2(1.2, 0.4042, 1.1836)] \\ &= 1.1713\end{aligned}$$

Third step:

predictor

$$\begin{aligned}\hat{X}_3 &= X_2 + hf_1(t_2, X_2, Y_2) \\ &= 0.3542\end{aligned}$$

$$\begin{aligned}\hat{Y}_3 &= Y_2 + hf_2(t_2, X_2, Y_2) \\ &= 1.1309\end{aligned}$$

corrector

$$\begin{aligned}X_3 &= X_2 + \frac{1}{2}h[f_1(t_2, X_2, Y_2) \\ &\quad + f_1(t_3, \hat{X}_3, \hat{Y}_3)] \\ &= 0.4028 \\ &\quad + 0.05[f_1(1.2, 0.4028, 1.1713) \\ &\quad + f_1(1.3, 0.3542, 1.1309)] \\ &= 0.3553\end{aligned}$$

$$\begin{aligned}Y_3 &= Y_2 + \frac{1}{2}h[f_2(t_2, X_2, Y_2) \\ &\quad + f_2(t_3, \hat{X}_3, \hat{Y}_3)] \\ &= 1.1713 \\ &\quad + 0.05[f_2(1.2, 0.4028, 1.1713) \\ &\quad + f_2(1.3, 0.3542, 1.1309)] \\ &= 1.1186\end{aligned}$$

Fourth step:

predictor

$$\begin{aligned}\hat{X}_4 &= X_3 + hf_1(t_3, X_3, Y_3) \\ &= 0.3119\end{aligned}$$

$$\begin{aligned}\hat{Y}_4 &= Y_3 + hf_2(t_3, X_3, Y_3) \\ &= 1.0536\end{aligned}$$

corrector

$$\begin{aligned}X_4 &= X_3 + \frac{1}{2}h[f_1(t_3, X_3, Y_3) + f_1(t_4, \hat{X}_4, \hat{Y}_4)] \\ &= 0.3553 + 0.05[f_1(1.3, 0.3553, 1.1186) + f_1(1.4, 0.3119, 1.0536)]\end{aligned}$$

Hence finally we have $X(1.4) = 0.3155$.

The MAPLE worksheet at the end of Example 2.12 can be easily modified to reproduce the solution of Example 2.13 by changing the name of the required numerical method from `foreuler` to `heunform`.

It should be obvious from Example 2.13 that the main drawback of extending the methods we already have at our disposal to sets of differential equations is the additional labour and tedium of the computations. Intrinsically, the computations are no more difficult, merely much more laborious – a prime example of a problem ripe for computerization.

2.4.2 State-space representation of higher-order systems

The solution of differential equation initial-value problems of order greater than one can be reduced to the solution of a set of first-order differential equations using the state-space representation introduced in Section 1.9. This is achieved by a simple transformation, illustrated by Example 2.14.

Example 2.14

The initial-value problem

$$\frac{d^2x}{dt^2} + x^2t \frac{dx}{dt} - xt^2 = \frac{1}{2}t^2, \quad x(0) = 1.2, \quad \frac{dx}{dt}(0) = 0.8$$

can be transformed into two coupled first-order differential equations by introducing an additional variable

$$y = \frac{dx}{dt}$$

With this definition, we have

$$\frac{d^2x}{dt^2} = \frac{dy}{dt}$$

and so the differential equation becomes

$$\frac{dy}{dt} + x^2ty - xt^2 = \frac{1}{2}t^2$$

Thus the original differential equation can be replaced by a pair of coupled first-order differential equations, together with initial conditions:

$$\frac{dx}{dt} = y, \quad x(0) = 1.2$$

$$\frac{dy}{dt} = -x^2ty + xt^2 + \frac{1}{2}t^2, \quad y(0) = 0.8$$

This process can be extended to transform a p th-order initial-value problem into a set of p first-order equations, each with an initial condition. Once the original equation has been transformed in this way, its solution by numerical methods is just the same as if it had been a set of coupled equations in the first place.

Example 2.15

Find the value of $X(0.2)$ satisfying the initial-value problem

$$\frac{d^3x}{dt^3} + xt \frac{d^2x}{dt^2} + t \frac{dx}{dt} - t^2x = 0, \quad x(0) = 1, \quad \frac{dx}{dt}(0) = 0.5, \quad \frac{d^2x}{dt^2}(0) = -0.2$$

using the fourth-order Runge–Kutta scheme with step size $h = 0.05$.

Solution Since this is a third-order equation, we need to introduce two new variables:

$$y = \frac{dx}{dt} \quad \text{and} \quad z = \frac{dy}{dt} = \frac{d^2x}{dt^2}$$

Then the equation is transformed into a set of three first-order differential equations

$$\frac{dx}{dt} = y \quad x(0) = 1$$

$$\frac{dy}{dt} = z \quad y(0) = 0.5$$

$$\frac{dz}{dt} = -xtz - ty + t^2x \quad z(0) = -0.2$$

Applied to the set of differential equations

$$\frac{dx}{dt} = f_1(t, x, y, z)$$

$$\frac{dy}{dt} = f_2(t, x, y, z)$$

$$\frac{dz}{dt} = f_3(t, x, y, z)$$

the Runge–Kutta scheme is of the form

$$c_{11} = hf_1(t_n, X_n, Y_n, Z_n)$$

$$c_{21} = hf_2(t_n, X_n, Y_n, Z_n)$$

$$c_{31} = hf_3(t_n, X_n, Y_n, Z_n)$$

$$c_{12} = hf_1(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_{11}, Y_n + \frac{1}{2}c_{21}, Z_n + \frac{1}{2}c_{31})$$

$$c_{22} = hf_2(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_{11}, Y_n + \frac{1}{2}c_{21}, Z_n + \frac{1}{2}c_{31})$$

$$c_{32} = hf_3(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_{11}, Y_n + \frac{1}{2}c_{21}, Z_n + \frac{1}{2}c_{31})$$

$$c_{13} = hf_1(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_{12}, Y_n + \frac{1}{2}c_{22}, Z_n + \frac{1}{2}c_{32})$$

$$c_{23} = hf_2(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_{12}, Y_n + \frac{1}{2}c_{22}, Z_n + \frac{1}{2}c_{32})$$

$$c_{33} = hf_3(t_n + \frac{1}{2}h, X_n + \frac{1}{2}c_{12}, Y_n + \frac{1}{2}c_{22}, Z_n + \frac{1}{2}c_{32})$$

$$c_{14} = hf_1(t_n + h, X_n + c_{13}, Y_n + c_{23}, Z_n + c_{33})$$

$$c_{24} = hf_2(t_n + h, X_n + c_{13}, Y_n + c_{23}, Z_n + c_{33})$$

$$c_{34} = hf_3(t_n + h, X_n + c_{13}, Y_n + c_{23}, Z_n + c_{33})$$

$$\begin{aligned}X_{n+1} &= X_n + \frac{1}{6}(c_{11} + 2c_{12} + 2c_{13} + c_{14}) \\Y_{n+1} &= Y_n + \frac{1}{6}(c_{21} + 2c_{22} + 2c_{23} + c_{24}) \\Z_{n+1} &= Z_n + \frac{1}{6}(c_{31} + 2c_{32} + 2c_{33} + c_{34})\end{aligned}$$

Note that each of the four substeps of the Runge–Kutta scheme must be carried out in parallel on each of the equations, since the intermediate values for all the independent variables are needed in the next substep for each variable; for instance, the computation of c_{13} requires not only the value of c_{12} but also the values of c_{22} and c_{32} . The first step of the computation in this case proceeds thus:

$$\begin{aligned}X_0 &= x_0 = 1 & Y_0 &= y_0 = 0.5 & Z_0 &= z_0 = -0.2 \\c_{11} &= hf_1(t_0, X_0, Y_0, Z_0) \\&= hY_0 \\&= 0.025\ 000 & c_{21} &= hf_2(t_0, X_0, Y_0, Z_0) \\&= hZ_0 \\&= -0.010\ 000 & c_{31} &= hf_3(t_0, X_0, Y_0, Z_0) \\&= h(-X_0 t_0 Z_0 - t_0 Y_0 + t_0^2 X_0) \\&= 0.000\ 000 \\c_{12} &= hf_1(t_0 + \frac{1}{2}h, X_0 + \frac{1}{2}c_{11}, Y_0 + \frac{1}{2}c_{21}, Z_0 + \frac{1}{2}c_{31}) \\&= h(Y_0 + \frac{1}{2}c_{21}) \\&= 0.024\ 750 \\c_{22} &= hf_2(t_0 + \frac{1}{2}h, X_0 + \frac{1}{2}c_{11}, Y_0 + \frac{1}{2}c_{21}, Z_0 + \frac{1}{2}c_{31}) \\&= h(Z_0 + \frac{1}{2}c_{31}) \\&= -0.010\ 000 \\c_{32} &= hf_3(t_0 + \frac{1}{2}h, X_0 + \frac{1}{2}c_{11}, Y_0 + \frac{1}{2}c_{21}, Z_0 + \frac{1}{2}c_{31}) \\&= h(-(X_0 + \frac{1}{2}c_{11})(t_0 + \frac{1}{2}h)(Z_0 + \frac{1}{2}c_{31}) \\&\quad - (t_0 + \frac{1}{2}h)(Y_0 + \frac{1}{2}c_{21}) + (t_0 + \frac{1}{2}h)^2(X_0 + \frac{1}{2}c_{11})) \\&= -0.000\ 334 \\c_{13} &= hf_1(t_0 + \frac{1}{2}h, X_0 + \frac{1}{2}c_{12}, Y_0 + \frac{1}{2}c_{22}, Z_0 + \frac{1}{2}c_{32}) \\&= h(Y_0 + \frac{1}{2}c_{22}) \\&= 0.024\ 750 \\c_{23} &= hf_2(t_0 + \frac{1}{2}h, X_0 + \frac{1}{2}c_{12}, Y_0 + \frac{1}{2}c_{22}, Z_0 + \frac{1}{2}c_{32}) \\&= h(Z_0 + \frac{1}{2}c_{32}) \\&= -0.010\ 008 \\c_{33} &= hf_3(t_0 + \frac{1}{2}h, X_0 + \frac{1}{2}c_{12}, Y_0 + \frac{1}{2}c_{22}, Z_0 + \frac{1}{2}c_{32}) \\&= h(-(X_0 + \frac{1}{2}c_{12})(t_0 + \frac{1}{2}h)(Z_0 + \frac{1}{2}c_{32}) \\&\quad - (t_0 + \frac{1}{2}h)(Y_0 + \frac{1}{2}c_{22}) + (t_0 + \frac{1}{2}h)^2(X_0 + \frac{1}{2}c_{12})) \\&= -0.000\ 334\end{aligned}$$

$$\begin{aligned}c_{14} &= hf_1(t_0 + h, X_0 + c_{13}, Y_0 + c_{23}, Z_0 + c_{33}) \\&= h(Y_0 + c_{23}) \\&= 0.024\,499\end{aligned}$$

$$\begin{aligned}c_{24} &= hf_2(t_0 + h, X_0 + c_{13}, Y_0 + c_{23}, Z_0 + c_{33}) \\&= h(Z_0 + c_{33}) \\&= -0.010\,016\end{aligned}$$

$$\begin{aligned}c_{34} &= hf_3(t_0 + h, X_0 + c_{13}, Y_0 + c_{23}, Z_0 + c_{33}) \\&= h(-(X_0 + c_{13})(t_0 + h)(Z_0 + c_{33}) \\&\quad - (t_0 + h)(Y_0 + c_{23}) + (t_0 + h)^2(X_0 + c_{13})) \\&= -0.000\,584\end{aligned}$$

$$X_1 = 1.024\,750, \quad Y_1 = 0.489\,994, \quad Z_1 = -0.200\,320$$

The second and subsequent steps are similar – we shall not present the details of the computations. It should be obvious by now that computations like these are sufficiently tedious to justify the effort of writing a computer program to carry out the actual arithmetic. The essential point for the reader to grasp is not the mechanics, but rather the principle whereby methods for the solution of first-order differential equations can be extended to the solution of sets of equations and hence to higher-order equations.



Again MAPLE could be used to find the numerical solution of this set of coupled ordinary differential equations. However, the MAPLE `dsolve` procedure is also able to do the conversion of the higher-order equation into a set of first-order equations internally so the numerical solution of the example above using the fourth-order Runge–Kutta algorithm could be achieved with the following worksheet.

```
> #set up the differential equation
> deq:=diff(x(t),t,t,t)+x(t)*t*diff(x(t),t,t)
                  +t*diff(x(t),t)-t^2*x(t)=0;
> #set up the initial conditions
> inits:=x(0)=1,D(x)(0)=0.5,D(D(x))(0)=-0.2;
> #procedure "dsolve" used to solve third order
                  differential equations
> sol:=dsolve({deq, inits}, numeric,method=classical[rk4],
                  output=listprocedure,stepsize=0.05);
> #obtain the numerical solution required
> xx:=op(2,sol[2]);xx(0.05);xx(0.2);
```

2.4.3 Exercises

- 18 Transform the following initial-value problems into sets of first-order differential equations with appropriate initial conditions:

$$(a) \frac{d^2x}{dt^2} + 6(x^2 - t)\frac{dx}{dt} - 4xt = 0$$

$$x(0) = 1, \quad \frac{dx}{dt}(0) = 2$$

$$(b) \frac{d^2x}{dt^2} + 4(x^2 - t^2)^{1/2} = 0$$

$$x(1) = 2, \quad \frac{dx}{dt}(1) = 0.5$$

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

$$x(0) = 0, \quad \frac{dx}{dt}(0) = 0$$

$$(d) \frac{d^3x}{dt^3} + t \frac{d^2x}{dt^2} + 6e^t \frac{dx}{dt} - x^2 t = e^{2t}$$

$$x(0) = 1, \quad \frac{dx}{dt}(0) = 2, \quad \frac{d^2x}{dt^2}(0) = 0$$

$$(e) \frac{d^3x}{dt^3} + t \frac{d^2x}{dt^2} + x^2 = \sin t$$

$$x(1) = 1, \quad \frac{dx}{dt}(1) = 0, \quad \frac{d^2x}{dt^2}(1) = -2$$

$$(f) \left(\frac{d^3x}{dt^3}\right)^{1/2} + t \frac{d^2x}{dt^2} + x^2 t^2 = 0$$

$$x(2) = 0, \quad \frac{dx}{dt}(2) = 0, \quad \frac{d^2x}{dt^2}(2) = 2$$

$$(g) \frac{d^4x}{dt^4} + x \frac{d^2x}{dt^2} + x^2 = \ln t, \quad x(0) = 0, \quad \frac{dx}{dt}(0) = 0,$$

$$\frac{d^2x}{dt^2}(0) = 4, \quad \frac{d^3x}{dt^3}(0) = -3$$

$$(h) \frac{d^4x}{dt^4} + \left(\frac{dx}{dt} - 1\right)t \frac{d^3x}{dt^3} + \frac{dx}{dt} - (xt)^{1/2} \\ = t^2 + 4t - 5$$

$$x(0) = a, \quad \frac{dx}{dt}(0) = 0, \quad \frac{d^2x}{dt^2}(0) = b, \quad \frac{d^3x}{dt^3}(0) = 0$$

- 19 Find the value of $X(0.3)$ for the initial-value problem

$$\frac{d^2x}{dt^2} + x^2 \frac{dx}{dt} + x = \sin t, \quad x(0) = 0, \quad \frac{dx}{dt}(0) = 1$$

using the Euler method with step size $h = 0.1$.

- 20 The second-order Adams–Bashforth method for the integration of a single first-order differential equation

$$\frac{dx}{dt} = f(t, x)$$

is

$$X_{n+1} = X_n + \frac{1}{2} h [3f(t_n, X_n) - f(t_{n-1}, X_{n-1})]$$

Write down the appropriate equations for applying the same method to the solution of the pair of differential equations

$$\frac{dx}{dt} = f_1(t, x, y), \quad \frac{dy}{dt} = f_2(t, x, y)$$

Hence find the value of $X(0.3)$ for the initial-value problem

$$\frac{d^2x}{dt^2} + x^2 \frac{dx}{dt} + x = \sin t, \quad x(0) = 0, \quad \frac{dx}{dt}(0) = 1$$

using this Adams–Bashforth method with step size $h = 0.1$. Use the second-order predictor–corrector method for the first step to start the computation.

- 21 Use the second-order predictor–corrector method (that is, the first-order Adams–Bashforth formula as predictor and the second-order Adams–Moulton formula as corrector) to compute an approximation $X(0.65)$ to the solution $x(0.65)$ of the initial-value problem

$$\frac{d^3x}{dt^3} + (x-t) \frac{d^2x}{dt^2} + \left(\frac{dx}{dt}\right)^2 - x^2 = 0$$

$$x(0.5) = -1, \quad \frac{dx}{dt}(0.5) = 1, \quad \frac{d^2x}{dt^2}(0.5) = 2$$

using a step size $h = 0.05$.

- 22 Write a computer program to solve the initial-value problem

$$\frac{d^2x}{dt^2} + x^2 \frac{dx}{dt} + x = \sin t, \quad x(0) = 0, \quad \frac{dx}{dt}(0) = 1$$

using the fourth-order Runge–Kutta method. Use your program to find the value of $X(1.6)$ using step sizes $h = 0.4$ and 0.2 . Estimate the accuracy of your value of $X(1.6)$ and estimate the step size that would be necessary to obtain a value of $X(1.6)$ accurate to 6dp.

23

Write a computer program to solve the initial-value problem



$$\frac{d^3x}{dt^3} + (x-t)\frac{d^2x}{dt^2} + \left(\frac{dx}{dt}\right)^2 - x^2 = 0$$

$$x(0.5) = -1, \quad \frac{dx}{dt}(0.5) = 1, \quad \frac{d^2x}{dt^2}(0.5) = 2$$



using the third-order predictor–corrector method (that is, the second-order Adams–Bashforth formula as predictor with the third-order Adams–Moulton as corrector). Use the fourth-order Runge–Kutta method to overcome the starting problem with this process. Use your program to find the value of $X(2.2)$ using step sizes $h = 0.1$ and 0.05 . Estimate the accuracy of your value of $X(2.2)$ and estimate the step size that would be necessary to obtain a value of $X(2.2)$ accurate to 6dp.

Note: The comment on the use of high-level computer language and the MATLAB and MAPLE packages at the end of Section 2.3.9 is equally applicable to the immediately preceding exercises in this section.

2.4.4 Boundary-value problems

Because first-order ODEs only have one boundary condition, that condition can always be treated as an initial condition. Once we turn to second- and higher-order differential equations, there are, at least for fully determined problems, two or more boundary conditions. If the boundary conditions are all imposed at the same point then the problem is an initial-value problem and can be solved by the methods we have already described. The problems that have been used as illustrations in Sections 2.4.1 and 2.4.2 were all initial-value problems. Boundary-value problems are somewhat more difficult to solve than initial-value problems.

To illustrate the difficulties of boundary-value problems, let us consider second-order differential equations. These have two boundary conditions. If they are both imposed at the same point (and so are initial conditions), the conditions will usually be a value of the dependent variable and of its derivative, for instance a problem like

$$L[x(t)] = f(t), \quad x(a) = p, \quad \frac{dx}{dt}(a) = q$$

where L is some differential operator. Occasionally, a mixed boundary condition such as

$$Cx(a) + D\frac{dx}{dt}(a) = p$$

will arise. Provided that a second boundary condition on x or dx/dt is imposed at the same point, this causes no difficulty, since the boundary conditions can be decoupled, that is solved to give values of $x(a)$ and $(dx/dt)(a)$, before the problem is solved.

If the two boundary conditions are imposed at different points then they could consist of two values of the dependent variable, the value of the dependent variable at one boundary and its derivative at the other, or even linear combinations of the values of the dependent variable and its derivative. For instance, we may have

$$L[x(t)] = f(t), \quad x(a) = p, \quad x(b) = q$$

or

$$L[x(t)] = f(t), \quad \frac{dx}{dt}(a) = p, \quad x(b) = q$$

or

$$L[x(t)] = f(t), \quad x(a) = p, \quad \frac{dx}{dt}(b) = q$$

or even such systems as

$$L[x(t)] = f(t), \quad x(a) = p, \quad Ax(b) + B\frac{dx}{dt}(b) = q$$

The increased range of possibilities introduced by boundary-value problems almost inevitably increases the problems which may arise in their solution. For instance, it may at first sight seem that it should also be possible to solve problems with boundary conditions consisting of the derivative at both boundaries, such as

$$L[x(t)] = f(t), \quad \frac{dx}{dt}(a) = p, \quad \frac{dx}{dt}(b) = q$$

Things are unfortunately not that simple – as Example 2.16 shows.

Example 2.16

Solve the boundary-value problem

$$\frac{d^2x}{dt^2} = 4, \quad \frac{dx}{dt}(0) = p, \quad \frac{dx}{dt}(1) = q$$

Solution

Integrating twice easily yields the general solution

$$x = 2t^2 + At + B$$

The boundary conditions then impose

$$A = p \quad \text{and} \quad 4 + A = q$$

It is obviously not possible to find a value of A satisfying both these equations unless $q = p + 4$. In any event, whether or not p and q satisfy this relation, it is not possible to determine the constant B .

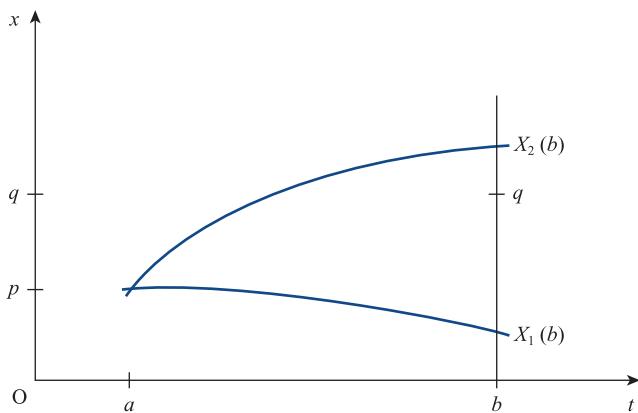
Example 2.16 illustrates the fact that if derivative boundary conditions are to be applied, a supplementary compatibility condition is needed. In addition, there may be a residual uncertainty in the solution. The complete analysis of what types of boundary conditions are allowable for two-point boundary-value problems is beyond the scope of this book. Differential equations of orders higher than two increase the range of possibilities even further and introduce further complexities into the determination of what boundary conditions are allowable and valid.

2.4.5 The method of shooting

One obvious way of solving two-point boundary-value problems is a form of systematic trial and error in which the boundary-value problem is replaced by an initial-value

Figure 2.22

The solution of a differential equation by the method of shooting: initial trials.



problem with initial values given at one of the two boundary points. The initial-value problem can be solved by an appropriate numerical technique and the value of whatever function is involved in the boundary condition at the second boundary point determined. The initial values are then adjusted and another initial-value problem solved. This process is repeated until a solution is found with the appropriate value at the second boundary point.

As an illustration, we shall consider a second-order boundary-value problem of the form

$$L[x] = f(t), \quad x(a) = p, \quad x(b) = q \quad (2.27)$$

The related initial-value problem

$$L[x] = f(t), \quad x(a) = p, \quad \frac{dx}{dt}(a) = 0 \quad (2.28)$$

could be solved as described in Section 2.4.2. Suppose that doing this results in an approximate solution of (2.28) denoted by X_1 . In the same way, denote the solution of the problem

$$L[x] = f(t), \quad x(a) = p, \quad \frac{dx}{dt}(a) = 1 \quad (2.29)$$

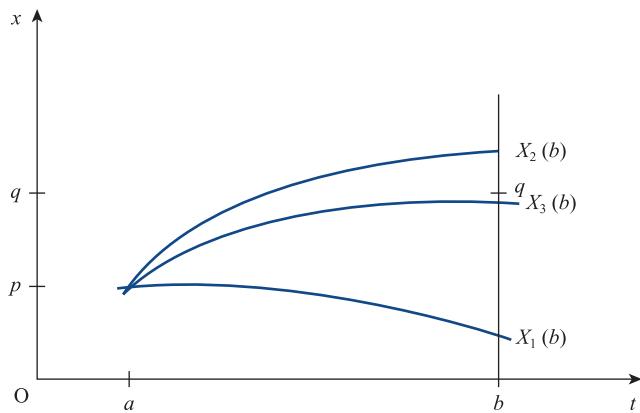
by X_2 . We now have a situation as shown in Figure 2.22. The values of the two solutions at the point $t = b$ are $X_1(b)$ and $X_2(b)$. The original boundary-value problem (2.27) requires a value q at b . Since q is roughly three-quarters of the way between $X_1(b)$ and $X_2(b)$, we should intuitively expect that solving the initial-value problem

$$L[x] = f(t), \quad x(a) = p, \quad \frac{dx}{dt}(a) = 0.75 \quad (2.30)$$

will produce a solution with $X(b)$ much closer to q . What we have done, of course, is to assume that $X(b)$ varies continuously and roughly in proportion to $(dx/dt)(a)$ and then to use linear interpolation to estimate a better value of $(dx/dt)(a)$. It is unlikely, of course, that $X(b)$ will vary exactly linearly with $(dx/dt)(a)$ so the solution of (2.30), call it X_3 , will be something like that shown in Figure 2.23. The process of linear

Figure 2.23

The solution of a differential equation by the method of shooting: first refinement.



interpolation to estimate a value of $(dx/dt)(a)$ and the subsequent solution of the resulting initial-value problem can be repeated until a solution is found with a value of $X(b)$ as close to q as may be required. This method of solution is known, by an obvious analogy with the bracketing method employed by artillerymen to find their targets, as the **method of shooting**. Shooting is not restricted to solving two-point boundary-value problems in which the two boundary values are values of the dependent variable. Problems involving boundary values on the derivatives can be solved in an analogous manner.

The solution of a two-point boundary-value problem by the method of shooting involves repeatedly solving a similar initial-value problem. It is therefore obvious that the amount of computation required to obtain a solution to a two-point boundary-value problem by this method is certain to be an order of magnitude or more greater than that required to solve an initial-value problem of the same order to the same accuracy. The method for finding the solution that satisfies the boundary condition at the second boundary point which we have just described used linear interpolation. It is possible to reduce the computation required by using more sophisticated interpolation methods. For instance, a version of the method of shooting that utilizes Newton–Raphson iteration is described in R. D. Milne, *Applied Functional Analysis, An Introductory Treatment* (London, Pitman, 1979).

2.5 Engineering application: oscillations of a pendulum

The simple pendulum has been used for hundreds of years as a timing device. A pendulum clock, using either a falling weight or a clockwork spring device to provide motive power, relies on the natural periodic oscillations of a pendulum to ensure good timekeeping. Generally we assume that the period of a pendulum is constant regardless of its amplitude. But this is only true for infinitesimally small amplitude oscillations. In reality the period of a pendulum's oscillations depends on its amplitude. In this section we will use our knowledge of numerical analysis to assist in an investigation of this relationship.

Figure 2.24 shows a simple rigid pendulum mounted on a frictionless pivot swinging in a single plane. By resolving forces in the tangential direction we have, following the classical analysis of such pendulums,

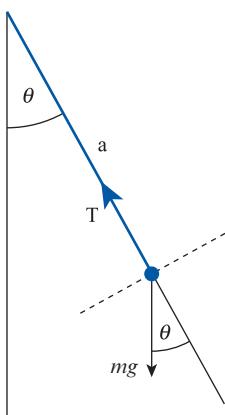


Figure 2.24 A simple pendulum.

$$ma \frac{d^2\theta}{dt^2} = -mg \sin \theta$$

that is,

$$\frac{d^2\theta}{dt^2} + \frac{g}{a} \sin \theta = 0 \quad (2.31)$$

For small oscillations of the pendulum we can use the approximation $\sin \theta \approx \theta$ so the equation becomes

$$\frac{d^2\theta}{dt^2} + \frac{g}{a} \theta = 0 \quad (2.32)$$

which is, of course, the simple harmonic motion equation with solutions

$$\theta = A \cos\left(\sqrt{\frac{g}{a}} t\right) + B \sin\left(\sqrt{\frac{g}{a}} t\right)$$

Hence the period of the oscillations is $2\pi\sqrt{(a/g)}$ and is independent of the amplitude of the oscillations.

In reality, of course, the amplitude of the oscillations may not be small enough for the linear approximation $\sin \theta \approx \theta$ to be valid, so it would be useful to be able to solve (2.31). Equation (2.31) is nonlinear so its solution is rather more problematical than (2.32). We will solve the equation numerically. In order to make the solution a little more transparent we will scale it so that the period of the oscillations of the linear approximation (2.32) is unity. This is achieved by setting $t = 2\pi\sqrt{(a/g)}\tau$. Equation (2.31) then becomes

$$\frac{d^2\theta}{d\tau^2} + 4\pi^2 \sin \theta = 0 \quad (2.33)$$

For an initial amplitude of 30° , the pseudocode algorithm shown in Figure 2.25, which implements the fourth-order Runge–Kutta method described in Section 2.3.8, produces the results $\Theta(6.0) = 23.965\,834$ using a time step of 0.05 and $\Theta(6.0) = 24.018\,659$ with a step of 0.025. Using Richardson extrapolation (see Section 2.3.6) we can predict that the time step needed to achieve 5 decimal places of accuracy (i.e. an error less than 5×10^{-6}) with this fourth-order method is

$$\left[\frac{0.000\,005 \times (2^4 - 1)}{|23.965\,834 - 24.018\,659|} \right]^{1/4} \times 0.025 = 0.0049$$

repeating the calculation with time steps 0.01 and 0.005 gives $\Theta(6.0) = 24.021\ 872\ 7$ and $\Theta(6.0) = 24.021\ 948\ 1$ for which Richardson extrapolation implies an error of 5×10^{-6} as predicted.



These results could also have been obtained using MAPLE as shown by the following worksheet:

```
> deqsys:=diff(x(t),t$2)+4*Pi^2*sin(x(t))=0;
> inits:=x(0)=60/180*Pi,D(x)(0)=0;
> sol:=dsolve({deqsys, inits}, numeric,method=classical
               [rk4],output=listprocedure,stepsize=0.005);
> xx:=op(2,sol[2]);xx(6);evalf(xx(6)*180/Pi);
```

As a check we can draw the graph of $|\Theta_{0.01}(\tau) - \Theta_{0.005}(\tau)|$, shown in Figure 2.26. This confirms that the error grows as the solution advances and that the maximum error is around 7.5×10^{-6} .

What we actually wanted is an estimate of the period of the oscillations. The most satisfactory way to determine this is to find the interval between the times of successive zero crossings. The time of a zero crossing can be estimated by linear interpolation between the data points produced in numerical solution of the differential equation. At a zero crossing the successive values of Θ have the opposite sign. Figure 2.27 shows a modified version of the main part of the algorithm of Figure 2.25. This version determines the times of successive positive to negative zero crossings and the differences between them.

Figure 2.28 shows some results from a program based on the algorithm of Figure 2.27; it is evident that the period has been determined to 6 sf accuracy. Figure 2.29 has been compiled from similar results for other amplitudes of oscillation.

Some spring-powered pendulum clocks are observed to behave in a counter-intuitive way – as the spring winds down the clock gains time where most people intuitively expect it to run more slowly and hence lose time. Figure 2.29 explains this phenomenon. The reason is that, in a spring-powered clock, the spring, acting through the escapement mechanism, exerts forces on the pendulum which, over each cycle of oscillation of the pendulum, result in the application of a tiny net impulse. The result is that just sufficient work is done on the pendulum to overcome the effects of bearing friction, air resistance and any other dissipative effects, and to keep the pendulum swinging with constant amplitude. But, as the spring unwinds the force available is reduced and the impulse gets smaller. The result is that, as the

Figure 2.25

A pseudocode algorithm for solving the nonlinear pendulum (2.33).

```

tol ← 0.00001
t_start ← 0
t_end ← 6
write(vdu,'Enter amplitude => ')
read(keyb, x0)
x_start ← pi*x0/180
v_start ← 0
write(vdu,'Enter stepsize => ')
read(keyb,h)
write(vdu,t_start,' ',deg(x_start))
t ← t_start
x ← x_start
v ← v_start
repeat
    rk4(x,v,h → xn,vn)
    x ← xn
    v ← vn
    t ← t+h
until abs(t - t_end) < tol
write(vdu,t, ' ',deg(x))

procedure rk4(x,v,h → xn,vn)
    c11 ← h*f1(x,v)
    c21 ← h*f2(x,v)
    c12 ← h*f1(x + c11/2,v + c21/2)
    c22 ← h*f2(x + c11/2,v + c21/2)
    c13 ← h*f1(x + c12/2,v + c22/2)
    c23 ← h*f2(x + c12/2,v + c22/2)
    c14 ← h*f1(x + c13,v + c23)
    c24 ← h*f2(x + c13,v + c23)
    xn ← x + (c11 + 2*(c12 + c13) + c14)/6
    vn ← v + (c21 + 2*(c22 + c23) + c24)/6
endprocedure

procedure f1(x,v → f1)
    f1 ← v
endprocedure

procedure f2(x,v → f2)
    f2 ← -4*pi*pi*sin (x)
endprocedure

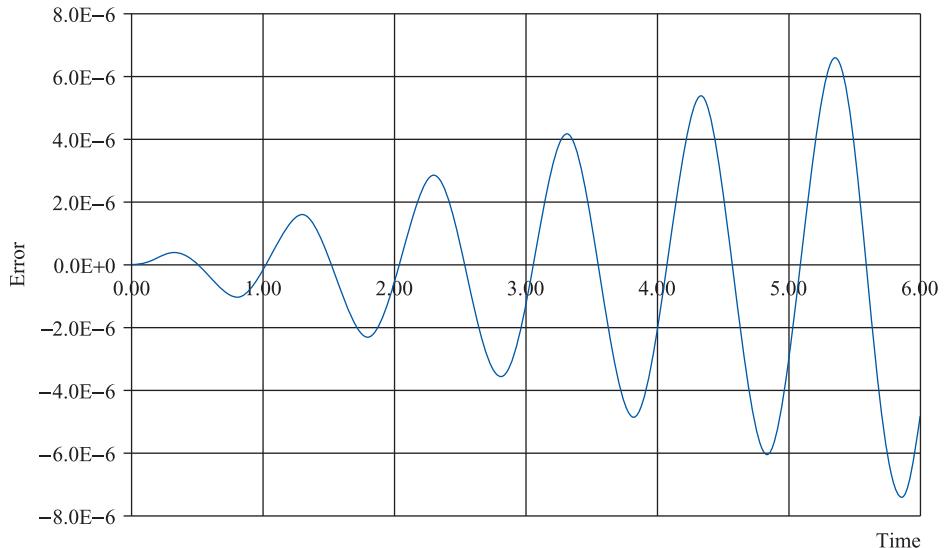
procedure deg(x → deg)
    deg ← 180*x/pi
endprocedure

```

clock winds down, the amplitude of oscillation of the pendulum decreases slightly. Figure 2.29 shows that as the amplitude decreases the period also decreases. Since the period of the pendulum controls the speed of the clock, the clock runs faster as the period decreases! Of course, as the clock winds down even further, the spring

Figure 2.26

Error in solution of (2.33) using algorithm (2.30) with $h = 0.005$.

**Figure 2.27**

Modification of pseudocode algorithm to find the period of oscillations of (2.33).

```

tol ← 0.000 01
t_start ← 0
t_end ← 6
write(vdu,'Enter amplitude => ')
read(keyb,x0)
x_start ← pi*x0/180
v_start ← 0
write(vdu,'Enter stepsize => ')
read(keyb,h)
write(vdu,t_start,' ',deg(x_start))
t ← t_start
x ← x_start
v ← v_start
t_previous_cross ← t_start
repeat
    rk4(x,v,h → xn,vn)
    if(xn*x < 0) and (x > 0) then
        t_cross ← (t*xn - (t + h)*x)/(xn-x)
        write(vdu,t_cross,' ',t_cross - t_previous_cross)
        t_previous_cross ← t_cross
    endif
    x ← xn
    v ← vn
    t ← t+h
until abs(t - t_end) < tol

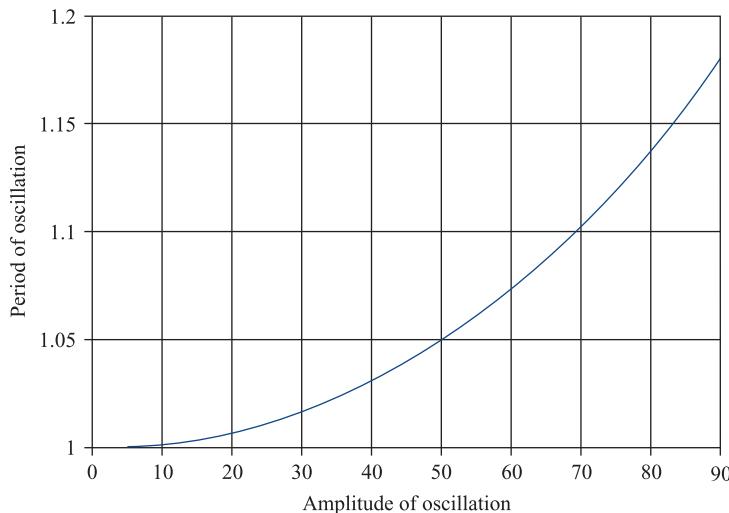
```

reaches a point where it is no longer capable of applying a sufficient impulse to overcome the dissipative forces, the pendulum ceases swinging and the clock finally stops.

Figure 2.28 Periods of successive oscillations of (2.33), $\Theta_0 = 30^\circ$, $h = 0.005$.

Time of crossing	Period of last cycle
0.254 352 13	
1.271 761 06	1.017 408 93
2.289 169 73	1.017 408 67
3.306 578 68	1.017 408 95
4.323 987 34	1.017 408 66
5.341 396 30	1.017 408 96

Figure 2.29 Variation of period of oscillations of (2.33) with amplitude.



The periods of the oscillations can also be measured using MAPLE. The procedure `fsolve` finds numerically the roots of a function. The output of the procedure `dsolve` is a function so we can use `fsolve` to find the zeros of that function, as in the following MAPLE worksheet. Note that the period of successive cycles is found more accurately and consistently using MAPLE. This is because the procedure `fsolve` uses a higher-order method to locate the zeros of the function rather than the linear interpolation method outlined in the algorithm in Figure 2.27.

```
> printlevel:=0:
> for i from 1 to 6 do;
    t1:=fsolve(xx(t)=0,t,(i-1)..(i-1+0.99)):
    t2:=fsolve(xx(t)=0,t,i..(i+0.99)):
    printf("%12.7f,%12.7f,%12.7f,\n",t1,t2,t2-t1);
end do;
```

2.6 Engineering application: heating of an electrical fuse

The electrical fuse is a simple device for protecting an electrical apparatus or circuit from overload and possible damage after the failure of one or more components in the apparatus. A fuse is usually a short length of thin wire through which the electrical current

powering the apparatus flows. If the apparatus fails in such a way as to draw a dangerously increased current, the fuse wire heats up and eventually melts thus disconnecting the apparatus from the power source. In order to design fuses which will not fail during normal use but which will operate reliably and rapidly in abnormal circumstances we must understand the heating of a thin wire carrying an electrical current.

The equation governing the heat generation and dissipation in a wire carrying an electrical current can be formulated as

$$-k\pi r^2 \frac{d^2 T}{dx^2} + 2\pi rh(T - T_e)^\alpha = I^2 \frac{\rho}{\pi r^2} \quad (2.34)$$

where T is the temperature of the fuse wire, x is the distance along the wire, k is the thermal conductivity of the material of which the wire is composed, r is the radius of the wire, h is the convective heat transfer coefficient from the surface of the wire, T_e is the ambient temperature of the fuse's surroundings, α is an empirical constant with a value around 1.25, I is the current in the wire and ρ is the resistivity of the wire. Equation (2.34) expresses the balance, in the steady state, between heat generation and heat loss. The first term of the equation represents the transfer of heat along the wire by conduction, the second term is the loss of heat from the surface of the wire by convection and the third term is the generation of heat in the wire by the electrical current.

Taking $\theta = (T - T_e)$ and dividing by $k\pi r^2$, (2.34) can be expressed as

$$\frac{d^2 \theta}{dx^2} - \frac{2h}{kr} \theta^\alpha = -\frac{\rho I^2}{k\pi^2 r^4} \quad (2.35)$$

Letting the length of the fuse be $2a$ and scaling the space variable, x , by setting $x = 2aX$, (2.35) becomes

$$\frac{d^2 \theta}{dX^2} - \frac{8a^2 h}{kr} \theta^\alpha = -\frac{4a^2 \rho I^2}{k\pi^2 r^4}$$

The boundary conditions are that the two ends of the wire, which are in contact with the electrical terminals in the fuse unit, are kept at some fixed temperature (we will assume that this temperature is the same as T_e). In addition, the fuse has symmetry about its midpoint $x = a$. Hence we may express the complete differential equation problem as

$$\frac{d^2 \theta}{dX^2} - \frac{8a^2 h}{kr} \theta^\alpha = -\frac{4a^2 \rho I^2}{k\pi^2 r^4}, \quad \theta(0) = 0, \quad \frac{d\theta}{dX}(1) = 0 \quad (2.36)$$

Equation (2.36) is a nonlinear second-order ODE. There is no straightforward analytical technique for tackling it so we must use numerical means. The problem is a boundary-value problem so we could use the method of shooting or some function approximation method. Figure 2.30 shows a pseudocode algorithm for this problem and Figure 2.31 gives the supporting procedures. The procedure `desolve` assumes initial conditions of the form $\theta(0) = 0$, $d\theta/dX(0) = \theta'_0$ and solves the differential equation using the third-order predictor–corrector method (with a single fourth-order Runge–Kutta step to start the multistep process). The main program uses the method of *regula falsa* to iterate from two starting values of θ'_0 which bracket that value of θ'_0 corresponding to $d\theta/dX(1) = 0$ which we seek.

Figure 2.32 shows the result of computations using a program based on the algorithm in Figure 2.30. Taking the values of the physical constants as $h = 100 \text{ W m}^{-2} \text{ K}^{-1}$, $a = 0.01 \text{ m}$, $k = 63 \text{ W m}^{-1} \text{ K}^{-1}$, $\rho = 16 \times 10^{-8} \Omega \text{ m}$ and $r = 5 \times 10^{-4} \text{ m}$, and taking I as 20 amps and 40 amps, gives the lower and upper curves in Figure 2.32 respectively.

Figure 2.30
Pseudocode algorithm
for solving (2.36).

```

rho ← 16e-8
kappa ← 63
r ← 5e-4
a ← 1e-2
hh ← 1e2
i ← 20
pconst ← 8*hh*a*a/(kappa*r)
qconst ← 4*a*a*rho*i*i/(kappa*pi*pi*r*r*r*r)
tol ← 1e-5
x_start ← 0.0
x_end ← 1.0
theta_start ← 0.0
write(vdu,'Enter stepsize -->')
read(keyb,h)
write(vdu,'Enter lower limit -->')
read(keyb,theta_dash_low)
write(vdu,'Enter upper limit -->')
read(keyb,theta_dash_high)
desolve(x_start,x_end,h,theta_start,theta_dash_low → th,ql)
desolve(x_start,x_end,h,theta_start,theta_dash_high → th,qh)
repeat
    theta_dash_new ← (qh*theta_dash_low - ql*theta_dash_high)/(qh - ql)
    desolve (x_start,x_end,h,theta_start,theta_dash_new → th,qn)
    if ql*qn>0 then
        ql ← qn
        theta_dash_low ← theta_dash_new
    else
        qh ← qn
        theta_dash_high ← theta_dash_new
    endif
until abs(qn) < tol
write(vdu,th,qn)

procedure desolve(x_0,x_end,h,v1_0,v2_0 → v1_f,v2_f)
    x ← x_0
    v1_o ← v1_0
    v2_o ← v2_0
    rk4(x,v1_o,v2_o,h → v1,v2)
    x ← x+h
    repeat
        pc3(x,v1_o,v2_o,v1,v2,h, → v1_n,v2_n)
        v1_o ← v1
        v2_o ← v2
        v1 ← v1_n
        v2 ← v2_n
        x ← x+h
    until abs(x - x_end) < tol
    v1_f ← v1
    v2_f ← v2
endprocedure

```

Evidently at 20 amps the operating temperature of the middle part of the wire is about 77° above the ambient temperature. If the current increases to 40 amps the temperature increases to about 245° above ambient – just above the melting point of tin! The procedure could obviously be used to design and validate appropriate dimensions (length and diameter) for fuses made from a variety of metals for a variety of applications and rated currents.

Figure 2.31

Subsidiary procedures for pseudocode algorithm for solving (2.36).

```

procedure rk4 (x,v1,v2,h → v1n,v2n)
    c11 ← h*f1(x,v1,v2)
    c21 ← h*f2(x,v1,v2)
    c12 ← h*f1(x + h/2,v1 + c11/2,v2 + c21/2)
    c22 ← h*f2(x + h/2,v1 + c11/2,v2 + c21/2)
    c13 ← h*f1(x + h/2,v1 + c12/2,v2 + c22/2)
    c23 ← h*f2(x + h/2,v1 + c12/2,v2 + c22/2)
    c14 ← h*f1(x + h,v1 + c13,v2 + c23)
    c24 ← h*f2(x + h,v1 + c13,v2 + c23)
    v1n ← v1 + (c11 + 2*(c12 + c13) + c14)/6
    v2n ← v2 + (c21 + 2*(c22 + c23) + c24)/6
endprocedure

procedure pc3(x, v1_o,v2_o,v1,v2,h → v1_n,v2_n)
    v1_p ← v1 + h*(3*f1(x,v1,v2) - f1(x - h,v1_o,v2_o))/2
    v2_p ← v2 + h*(3*f2(x,v1,v2) - f2(x - h,v1_o,v2_o))/2
    v1_n ← v1 + h*(5*f1(x + h,v1_p,v2_p)
        + 8*f1(x,v1,v2) - f1(x - h,v1_o,v2_o))/12
    v2_n ← v2 + h*(5*f2(x + h,v1_p, v2_p)
        + 8*f2(x,v1,v2) - f2(x - h,v1_o,v2_o))/12
endprocedure

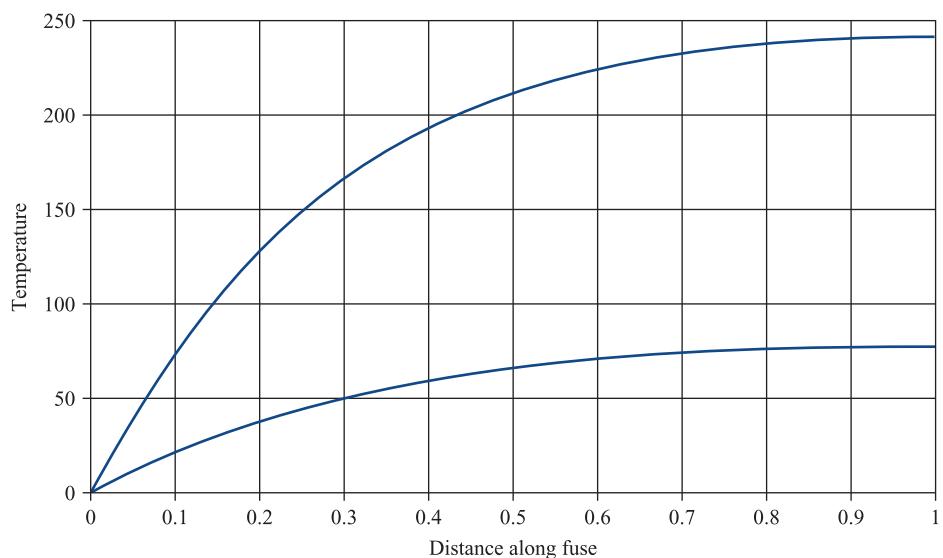
procedure f1(x, theta,theta_dash → f1)
    f1 ← theta_dash;
endprocedure

procedure f2(x,theta,theta_dash → f2)
    if theta < tol then
        f2 ← -qconst
    else
        f2 ← pconst*exp(ln (theta)*1.25) - qconst
    endif
endprocedure

```

Figure 2.32

Comparison of temperatures in a fuse wire carrying 20 amps and 40 amps.





The differential equation problem to be solved in this application is a boundary-value problem rather than an initial-value problem. MAPLE's `dsolve` procedure can readily handle this type of problem. The following MAPLE worksheet reproduces the temperature profiles shown in Figure 2.32.

```
> deqsys:=diff(theta(x),x,x)-8*a^2*h/
(k*r)*theta(x)^alpha=-4*a^2*ro*i^2/(k*pi^2*r^4);
> inits:=theta(0)=0,D(theta)(1)=0;
> alpha:=1.25;h:=100;a:=0.01;k:=63;ro:=16e-8;r:=5e-4;
                                         i:=20;
> sol1:=dsolve({deqsys, inits},
               numeric,output=listprocedure,maxmesh=512);
> i:=40;
> sol2:=dsolve({deqsys, inits},
               numeric,output=listprocedure,maxmesh=512);
> op(2,sol1[2])(1);op(2,sol2[2])(1);
> plot([op(2,sol1[2]),op(2,sol2[2])],0..1);
```

To find a numerical solution of a second-order differential equation using MATLAB, the user must first carry out the transformation to a set of two first-order equations; MATLAB, unlike MAPLE, cannot complete this stage internally. Then the following MATLAB M-file solves the differential equation and reproduce the temperature profiles shown in Figure 2.32.

```
function engineering_app2
a=0.01;h=100;k=63;r=5e-4;alpha=1.25;ro=16e-8;i=20;
solinit = bvpinit(linspace(0,1,10),[40 0.5]);
sol1 = bvp4c(@odefun,@bcfun,solinit);
i=40;
sol2 = bvp4c(@odefun,@bcfun,solinit);
x = linspace(0,1);
y1 = deval(sol1,x);
y2 = deval(sol2,x);
plot(x,y1(1,:),x,y2(1,:));
y1(1,100)
y2(1,100)

function dydx = odefun(x,y)
dydx = [ y(2)
          8*a^2*h/(k*r)*y(1)^alpha-4*a^2*ro*i^2/(k*pi^2*r^4) ];
end
function res = bcfun(ya,yb)
res = [ ya(1)
        yb(2)];
end
end
```

2.7 Review exercises (1–12)

- 1 Find the value of $X(0.5)$ for the initial-value problem

$$\frac{dx}{dt} = \sqrt{x}, \quad x(0) = 1$$

using Euler's method with step size $h = 0.1$.

- 2 Find the value of $X(1.2)$ for the initial-value problem

$$\frac{dx}{dt} = -e^{xt}, \quad x(1) = 1$$

using Euler's method with step size $h = 0.05$.

- 3 Solve the differential equation

$$\frac{dx}{dt} = \sqrt{\frac{xt}{x^2 + t^2}}, \quad x(0) = 1$$

to find the value of $X(0.4)$ using the Euler method with steps of size 0.1 and 0.05. By comparing the two estimates of $x(0.4)$ estimate the accuracy of the better of the two values which you have obtained and also the step size you would need to use in order to calculate an estimate of $x(0.4)$ accurate to 2 decimal places.

- 4 Solve the differential equation

$$\frac{dx}{dt} = \sin(t^2), \quad x(0) = 2$$

to find the value of $X(0.25)$ using the Euler method with steps of size 0.05 and 0.025. By comparing the two estimates of $x(0.25)$ estimate the accuracy of the better of the two values which you have obtained and also the step size you would need to use in order to calculate an estimate of $x(0.25)$ accurate to 3 decimal places.

- 5 Let X_1 , X_2 and X_3 denote the estimates of the function $x(t)$ satisfying the differential equation

$$\frac{dx}{dt} = \sqrt{(xt + t)}, \quad x(1) = 2$$

which are calculated using the second-order predictor–corrector method with steps of 0.1, 0.05 and 0.025 respectively. Compute $X_1(1.2)$, $X_2(1.2)$ and $X_3(1.2)$. Show that the ratio of $|X_2 - X_1|$ and

$|X_3 - X_2|$ should tend to 4 : 1 as the step size tends to zero. Do your computations bear out this expectation?

- 6 Compute the solution of the differential equation

$$\frac{dx}{dt} = e^{-xt}, \quad x(0) = 5$$

for $x = 0$ to 2 using the fourth-order Runge–Kutta method with step sizes of 0.2, 0.1 and 0.05. Estimate the accuracy of the most accurate of your three solutions.

- 7 In a thick cylinder subjected to internal pressure the radial pressure $p(r)$ at distance r from the axis of the cylinder is given by

$$p + r \frac{dp}{dr} = 2a - p$$

where a is a constant (which depends on the geometry of the cylinder).

If the stress has magnitude p_0 at the inner wall, $r = r_0$, and may be neglected at the outer wall, $r = r_1$, show that

$$p(r) = \frac{p_0 r_0^2}{r_1^2 - r_0^2} \left(\frac{r_1^2}{r^2} - 1 \right)$$

If $r_0 = 1$, $r_1 = 2$ and $p_0 = 1$, compare the value of $p(1.5)$ obtained from this analytic solution with the numerical value obtained using the fourth-order Runge–Kutta method with step size $h = 0.5$. (Note: With these values of r_0 , r_1 and p_0 , $a = -1/3$.)

- 8 Find the values of $X(t)$ for t up to 2 where $X(t)$ is the solution of the differential equation problem

$$\frac{d^3x}{dt^3} + \left(\frac{d^2x}{dt^2} \right)^2 + 4 \left(\frac{dx}{dt} \right)^2 - tx = \sin t,$$

$$x(1) = 0.2, \quad \frac{dx}{dt}(1) = 1, \quad \frac{d^2x}{dt^2}(1) = 0$$

using the Euler method with steps of 0.025. Repeat the computation with a step size of 0.0125. Hence estimate the accuracy of the value of $X(2)$ given by your solution.

- 9 Find the solution of the differential equation problem



$$\frac{d^2x}{dt^2} + (x^2 - 1)\frac{dx}{dt} + 40x = 0,$$

$$x(0) = 0.02, \quad \frac{dx}{dt}(0) = 0$$

using the second-order predictor–corrector method. Hence find an estimate of the value of $x(4)$ accurate to 4 decimal places.

- 10 Find the solution of the differential equation problem



$$\frac{d^3x}{dt^3} + \left| \frac{d^2x}{dt^2} \right|^{\frac{1}{2}} + 4\left(\frac{dx}{dt} \right)^3 - tx = \sin t,$$

$$x(1) = -1, \quad \frac{dx}{dt}(1) = 1, \quad \frac{d^2x}{dt^2}(1) = 2$$

using the fourth-order Runge–Kutta method. Hence find an estimate of the value of $x(2.5)$ accurate to 4 decimal places.

- 11 (Extended, open-ended problem.) The second-order, nonlinear, ODEs



$$\frac{d^2x}{dt^2} + \mu(x^2 - 1)\frac{dx}{dt} + \lambda^2x = 0$$

governs the oscillations of the Van der Pol oscillator. By scaling the time variable the equation can be reduced to

$$\frac{d^2x}{dt^2} + \mu(x^2 - 1)\frac{dx}{dt} + (2\pi)^2x = 0$$

Investigate the properties of the Van der Pol oscillator. In particular show that the oscillator shows limit cycle behaviour (that is, the oscillations tend to a form which is independent of the initial conditions and depends only on the parameter μ). Determine the dependence of the limit cycle period on μ .

- 12

(Extended, open-ended problem.) The equation of simple harmonic motion



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

is generally used to model the undamped oscillations of a mass supported on the end of a linear spring (that is, a spring whose tension is strictly proportional to its extension). Most real springs are actually nonlinear because as their extension or compression increases their stiffness changes. This can be modelled by the equation

$$\frac{d^2x}{dt^2} + 4\pi^2(1 + \beta x^2)x = 0$$

For a ‘hard’ spring stiffness increases with displacement ($\beta > 0$) and a soft spring’s stiffness decreases ($\beta < 0$). Investigate the oscillations of a mass supported by a hard or soft spring. In particular determine the connection between the frequency of the oscillations and their amplitude.



3

Vector Calculus

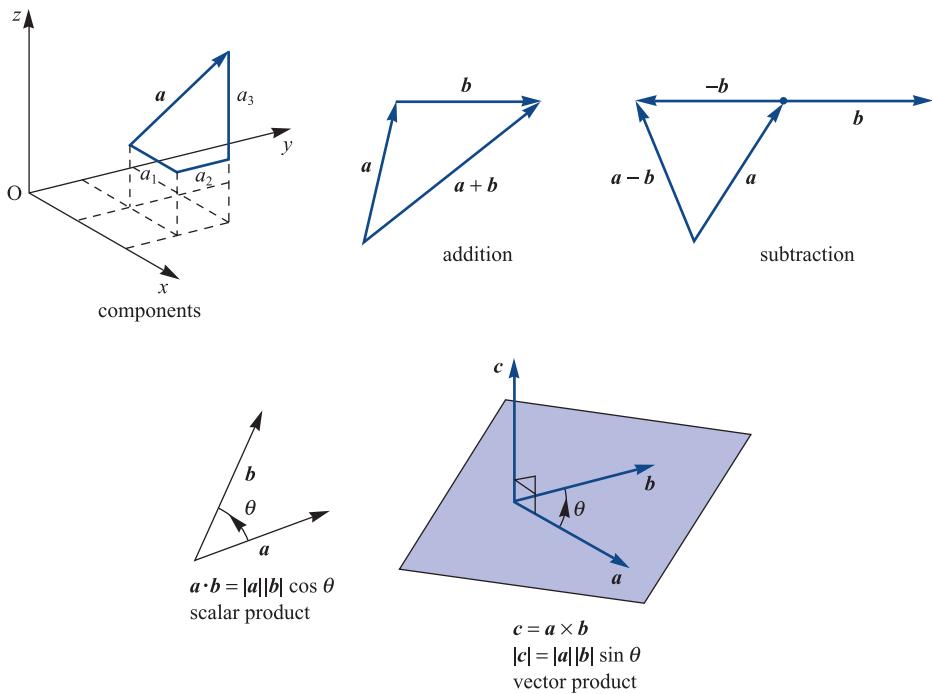
Chapter 3 Contents

3.1	Introduction	176
3.2	Derivatives of a scalar point function	192
3.3	Derivatives of a vector point function	196
3.4	Topics in integration	206
3.5	Engineering application: streamlines in fluid dynamics	240
3.6	Engineering application: heat transfer	242
3.7	Review exercises (1–21)	246

3.1 Introduction

In many applications we use functions of the space variable $\mathbf{r} = xi + yj + zk$ as models for quantities that vary from point to point in three-dimensional space. There are two types of such functions. There are **scalar point functions**, which model scalar quantities like the temperature at a point in a body, and **vector point functions**, which model vector quantities like the velocity of the flow at a point in a liquid. We can express this more formally in the following way. For each scalar point function f we have a **rule**, $u = f(\mathbf{r})$, which assigns to each point with coordinate \mathbf{r} in the **domain** of the function a unique real number u . For vector point functions the rule $\mathbf{v} = \mathbf{F}(\mathbf{r})$ assigns to each \mathbf{r} a unique vector \mathbf{v} in the range of the function. Vector calculus was designed to measure the variation of such functions with respect to the space variable \mathbf{r} . That development made use of the ideas about vectors (components, addition, subtraction, scalar and vector products) described in Chapter 4 of *Modern Engineering Mathematics* (MEM) and summarized here in Figure 3.1.

Figure 3.1
Elementary
vector algebra.



In component form if $\mathbf{a} = (a_1, a_2, a_3)$ and $\mathbf{b} = (b_1, b_2, b_3)$ then

$$\mathbf{a} \pm \mathbf{b} = (a_1 \pm b_1, a_2 \pm b_2, a_3 \pm b_3)$$

$$\mathbf{a} \cdot \mathbf{b} = (a_1 b_1 + a_2 b_2 + a_3 b_3) = \mathbf{b} \cdot \mathbf{a}$$

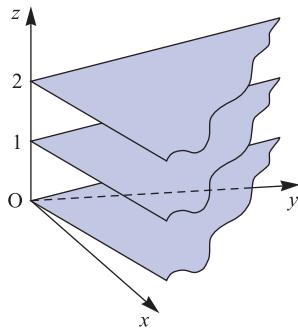
$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = -\mathbf{b} \times \mathbf{a}$$

$$= (a_2 b_3 - b_2 a_3, b_1 a_3 - a_1 b_3, a_1 b_2 - b_1 a_2).$$

The recent development of computer packages for the modelling of engineering problems involving vector quantities has relieved designers of much tedious analysis and computation. To be able to use those packages effectively, however, designers need a good understanding of the mathematical tools they bring to their tasks. It is on that basic understanding that this chapter focuses.

3.1.1 Basic concepts

Figure 3.2
Level surfaces
of $f(\mathbf{r}) = (2, 2, -1) \cdot \mathbf{r} = 2x + 2y - z$.



We can picture a scalar point function $f(\mathbf{r})$ by means of its level surfaces $f(\mathbf{r}) = \text{constant}$. For example, the level surfaces of $f(\mathbf{r}) = 2x + 2y - z$ are planes parallel to the plane $z = 2x + 2y$, as shown in Figure 3.2. On the level surface the function value does not change, so the rate of change of the function will be zero along any line drawn on the level surface. An alternative name for a scalar point function is **scalar field**. This is in contrast to the vector point function (or **vector field**). We picture a vector field by its field (or flow) lines. A field line is a curve in space represented by the position vector $\mathbf{r}(t)$ such that at each point of the curve its tangent is parallel to the vector field. Thus the field lines of $\mathbf{F}(\mathbf{r})$ are given by the differential equation

$$\frac{d\mathbf{r}}{dt} = \mathbf{F}(\mathbf{r}), \quad \text{where } \mathbf{r}(t_0) = \mathbf{r}_0$$

and \mathbf{r}_0 is the point on the line corresponding to $t = t_0$. This vector equation represents the three simultaneous ordinary differential equations

$$\frac{dx}{dt} = P(x, y, z),$$

$$\frac{dy}{dt} = Q(x, y, z),$$

$$\frac{dz}{dt} = R(x, y, z)$$

where $\mathbf{F} = (P, Q, R)$.



Modern computer algebra packages make it easier to draw both the level surfaces of scalar functions and the field lines of vector functions, but to underline the basic ideas we shall consider two simple examples.

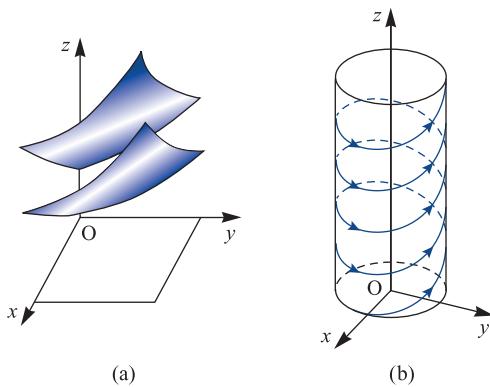
Example 3.1

Sketch

- (a) the level surfaces of the scalar point function $f(\mathbf{r}) = z e^{-xy}$;
 (b) the field lines of the vector point function $\mathbf{F}(\mathbf{r}) = (-y, x, 1)$.

Solution (a) Consider the level surface given by $f(\mathbf{r}) = c$, where c is a number. Then $z e^{-xy} = c$ and so $z = c e^{xy}$. For c, x and y all positive we can easily sketch part of the surface as shown in Figure 3.3(a), from which we can deduce the appearance of the whole family of level surfaces.

Figure 3.3 (a) Level surfaces of $f(\mathbf{r}) = z e^{-xy}$; and (b) field lines of $\mathbf{F}(\mathbf{r}) = (-y, x, 1)$ of Example 3.1.



- (b) For the function $\mathbf{F}(\mathbf{r}) = (-y, x, 1)$ the field lines are given by

$$\frac{d\mathbf{r}}{dt} = (-y, x, 1)$$

that is, by the simultaneous differential equations

$$\frac{dx}{dt} = -y, \quad \frac{dy}{dt} = x, \quad \frac{dz}{dt} = 1$$

The general solution of these simultaneous equations is

$$x(t) = A \cos t + B \sin t, \quad y(t) = -B \cos t + A \sin t, \quad z(t) = t + C$$

where A, B and C are arbitrary constants. Considering, in particular, the field line that passes through $(1, 0, 0)$, we determine the parametric equation

$$(x(t), y(t), z(t)) = (\cos t, \sin t, t)$$

This represents a circular helix as shown in Figure 3.3(b), from which we can deduce the appearance of the whole family of flow lines.



In MATLAB a level surface may be drawn using the `ezsrf` function. Using the Symbolic Math Toolbox the commands

```
syms x y z c
for c = [1, 2, 3]
    fsurf(a(x,y)c*exp(-x*y), [0, 2, 0, 2]);
    hold on
    xlabel('x')
    ylabel('y')
    title('c exp(-xy)')
end
```

will produce three of the level surfaces of $z = e^{-xy}$ on the same set of axes. The surfaces may also be produced in MAPLE using the `ezsrf` function. The field lines may be plotted in MATLAB using the `streamline` function.

To investigate the properties of scalar and vector fields further we need to use the calculus of several variables. Here we shall describe the basic ideas and definitions needed for vector calculus. A fuller treatment is given in Chapter 9 of MEM.

Given a function $f(x)$ of a single variable x , we measure its rate of change (or gradient) by its derivative with respect to x . This is

$$\frac{df}{dx} = f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

However, a function $f(x, y, z)$ of three independent variables x, y and z does not have a unique rate of change. The value of the latter depends on the direction in which it is measured. The rate of change of the function $f(x, y, z)$ in the x direction is given by its **partial derivative** with respect to x , namely

$$\frac{\partial f}{\partial x} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y, z) - f(x, y, z)}{\Delta x}$$

This measures the rate of change of $f(x, y, z)$ with respect to x when y and z are held constant. We can calculate such partial derivatives by differentiating $f(x, y, z)$ with respect to x , treating y and z as constants. Similarly,

$$\frac{\partial f}{\partial y} = \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y, z) - f(x, y, z)}{\Delta y}$$

and

$$\frac{\partial f}{\partial z} = \lim_{\Delta z \rightarrow 0} \frac{f(x, y, z + \Delta z) - f(x, y, z)}{\Delta z}$$

define the partial derivatives of $f(x, y, z)$ with respect to y and z respectively.

For conciseness we sometimes use a suffix notation to denote partial derivatives, for example writing f_x for $\partial f/\partial x$. The rules for partial differentiation are essentially the same as for ordinary differentiation, but it must always be remembered which variables are being held constant.

Higher-order partial derivatives may be defined in a similar manner, with, for example,

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x} \left(\frac{\partial f}{\partial x} \right) = f_{xx}$$

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial x} \right) = f_{xy}$$

$$\frac{\partial^3 f}{\partial z \partial y \partial x} = \frac{\partial}{\partial z} \left(\frac{\partial^2 f}{\partial y \partial x} \right) = f_{xyz}$$

Example 3.2

Find the first partial derivatives of the functions $f(x, y, z)$ with formula (a) $x + 2y + z^3$, (b) $x^2(y + 2z)$ and (c) $(x + y)/(z^3 + x)$.

Solution

- (a) $f(x, y, z) = x + 2y + z^3$. To obtain f_x , we differentiate $f(x, y, z)$ with respect to x , keeping y and z constant. Thus $f_x = 1$, since the derivative of a constant ($2y + z^3$) with respect to x is zero. Similarly, $f_y = 2$ and $f_z = 3z^2$.
- (b) $f(x, y, z) = x^2(y + 2z)$. Here we use the same idea: when we differentiate with respect to one variable, we treat the other two as constants. Thus

$$\frac{\partial}{\partial x} [x^2(y + 2z)] = (y + 2z) \frac{\partial}{\partial x} (x^2) = 2x(y + 2z)$$

$$\frac{\partial}{\partial y} [x^2(y + 2z)] = x^2 \frac{\partial}{\partial y} (y + 2z) = x^2(1) = x^2$$

$$\frac{\partial}{\partial z} [x^2(y + 2z)] = x^2 \frac{\partial}{\partial z} (y + 2z) = x^2(2) = 2x^2$$

- (c) $f(x, y, z) = (x + y)/(z^3 + x)$. Here we use the same idea, together with basic rules from ordinary differentiation:

$$\frac{\partial f}{\partial x} = \frac{(1)(z^3 + x) - (x + y)(1)}{(z^3 + x)^2} \quad (\text{quotient rule})$$

$$= \frac{z^3 - y}{(z^3 + x)^2}$$

$$\frac{\partial f}{\partial y} = \frac{1}{z^3 + x}$$

$$\frac{\partial f}{\partial z} = \frac{-3z^2(x + y)}{(z^3 + x)^2} \quad (\text{chain rule})$$

In Example 3.2 we used the **chain** (or **composite-function**) **rule** of ordinary differentiation

$$\frac{df}{dx} = \frac{df}{du} \frac{du}{dx}$$

to obtain the partial derivative $\partial f / \partial z$. The multivariable calculus form of the chain rule is a little more complicated. If the variables u , v and w are defined in terms of x , y and z then the partial derivative of $f(u, v, w)$ with respect to x is

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial u} \frac{\partial u}{\partial x} + \frac{\partial f}{\partial v} \frac{\partial v}{\partial x} + \frac{\partial f}{\partial w} \frac{\partial w}{\partial x}$$

with similar expressions for $\partial f / \partial y$ and $\partial f / \partial z$.

Example 3.3

Find $\partial T / \partial r$ and $\partial T / \partial \theta$ when

$$T(x, y) = x^3 - xy + y^3$$

and

$$x = r \cos \theta \quad \text{and} \quad y = r \sin \theta$$

Solution By the chain rule,

$$\frac{\partial T}{\partial r} = \frac{\partial T}{\partial x} \frac{\partial x}{\partial r} + \frac{\partial T}{\partial y} \frac{\partial y}{\partial r}$$

In this example

$$\frac{\partial T}{\partial x} = 3x^2 - y \quad \text{and} \quad \frac{\partial T}{\partial y} = -x + 3y^2$$

and

$$\frac{\partial x}{\partial r} = \cos \theta \quad \text{and} \quad \frac{\partial y}{\partial r} = \sin \theta$$

so that

$$\frac{\partial T}{\partial r} = (3x^2 - y)\cos \theta + (-x + 3y^2)\sin \theta$$

Substituting for x and y in terms of r and θ gives

$$\frac{\partial T}{\partial r} = 3r^2(\cos^3 \theta + \sin^3 \theta) - 2r \cos \theta \sin \theta$$

Similarly,

$$\begin{aligned} \frac{\partial T}{\partial \theta} &= (3x^2 - y)(-r \sin \theta) + (-x + 3y^2)r \cos \theta \\ &= 3r^3(\sin \theta - \cos \theta)\cos \theta \sin \theta + r^2(\sin^2 \theta - \cos^2 \theta) \end{aligned}$$

Example 3.4

Find dH/dt when

$$H(t) = \sin(3x - y)$$

and

$$x = 2t^2 - 3 \quad \text{and} \quad y = \frac{1}{2}t^2 - 5t + 1$$

Solution We note that x and y are functions of t only, so that the chain rule becomes

$$\frac{dH}{dt} = \frac{\partial H}{\partial x} \frac{dx}{dt} + \frac{\partial H}{\partial y} \frac{dy}{dt}$$

Note the mixture of partial and ordinary derivatives. H is a function of the one variable t , but its dependence is expressed through the two variables x and y .

Substituting for the derivatives involved, we have

$$\begin{aligned} \frac{dH}{dt} &= 3[\cos(3x - y)]4t - [\cos(3x - y)](t - 5) \\ &= (11t + 5)\cos(3x - y) \\ &= (11t + 5)\cos(\frac{11}{2}t^2 + 5t - 10) \end{aligned}$$

Example 3.5

A scalar point function $f(r)$ can be expressed in terms of rectangular cartesian coordinates (x, y, z) or in terms of spherical polar coordinates (r, θ, ϕ) , where

$$x = r \sin \theta \cos \phi, \quad y = r \sin \theta \sin \phi, \quad z = r \cos \theta$$

as shown in Figure 3.4. Find $\partial f / \partial x$ in terms of the partial derivatives of the function with respect to r, θ and ϕ .

Solution

Using the chain rule, we have

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial r} \frac{\partial r}{\partial x} + \frac{\partial f}{\partial \theta} \frac{\partial \theta}{\partial x} + \frac{\partial f}{\partial \phi} \frac{\partial \phi}{\partial x}$$

From Figure 3.4, $r^2 = x^2 + y^2 + z^2$, $\tan \phi = y/x$ and $\tan \theta = (x^2 + y^2)^{1/2}/z$, so that

$$\frac{\partial r}{\partial x} = \frac{x}{r} = \sin \theta \cos \phi$$

$$\frac{\partial \phi}{\partial x} = \frac{\partial}{\partial x} \left(\tan^{-1} \frac{y}{x} \right) = -\frac{y}{x^2 + y^2} = -\frac{\sin \phi}{r \sin \theta}$$

$$\frac{\partial \theta}{\partial x} = \frac{\partial}{\partial x} \left\{ \tan^{-1} \left[\frac{(x^2 + y^2)^{1/2}}{z} \right] \right\} = \frac{xz}{(x^2 + y^2 + z^2)(x^2 + y^2)^{1/2}}$$

$$= \frac{\cos \phi \cos \theta}{r}$$

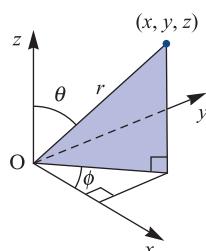


Figure 3.4 Spherical polar coordinates of Example 3.5.

Thus

$$\frac{\partial f}{\partial x} = \sin \theta \cos \phi \frac{\partial f}{\partial r} - \frac{\sin \phi}{r \sin \theta} \frac{\partial f}{\partial \phi} + \frac{\cos \phi \cos \theta}{r} \frac{\partial f}{\partial \theta}$$

Example 3.6

The Laplace equation in two dimensions is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

where x and y are rectangular cartesian coordinates. Show that expressed in polar coordinates (r, θ) , where $x = r \cos \theta$ and $y = r \sin \theta$, the Laplace equation may be written

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = 0$$

Solution Using the chain rule, we have

$$\begin{aligned} \frac{\partial u}{\partial r} &= \frac{\partial u}{\partial x} \frac{\partial x}{\partial r} + \frac{\partial u}{\partial y} \frac{\partial y}{\partial r} \\ &= \frac{\partial u}{\partial x} \cos \theta + \frac{\partial u}{\partial y} \sin \theta \end{aligned}$$

and

$$\frac{\partial^2 u}{\partial r^2} = \frac{\partial^2 u}{\partial x^2} \cos^2 \theta + \frac{\partial^2 u}{\partial y^2} \sin^2 \theta + 2 \frac{\partial^2 u}{\partial x \partial y} \sin \theta \cos \theta$$

Similarly

$$\frac{\partial u}{\partial \theta} = \frac{\partial u}{\partial x} (-r \sin \theta) + \frac{\partial u}{\partial y} (r \cos \theta)$$

and

$$\begin{aligned} \frac{\partial^2 u}{\partial \theta^2} &= \frac{\partial^2 u}{\partial x^2} (-r \sin \theta)^2 + \frac{\partial^2 u}{\partial y^2} (r \cos \theta)^2 - 2 \frac{\partial^2 u}{\partial x \partial y} r^2 \sin \theta \cos \theta \\ &\quad - \frac{\partial u}{\partial x} (r \cos \theta) - \frac{\partial u}{\partial y} (r \sin \theta) \end{aligned}$$

so that

$$\begin{aligned} \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} &= \frac{\partial^2 u}{\partial x^2} \sin^2 \theta + \frac{\partial^2 u}{\partial y^2} \cos^2 \theta - 2 \frac{\partial^2 u}{\partial x \partial y} r^2 \sin \theta \cos \theta \\ &\quad - \frac{1}{r} \left(\frac{\partial u}{\partial x} \cos \theta + \frac{\partial u}{\partial y} \sin \theta \right) \end{aligned}$$

Hence

$$\frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{1}{r} \frac{\partial u}{\partial r} = \frac{\partial^2 u}{\partial x^2} \sin^2 \theta + \frac{\partial^2 u}{\partial y^2} \cos^2 \theta - 2 \frac{\partial^2 u}{\partial x \partial y} r^2 \sin \theta \cos \theta$$

and

$$\frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{1}{r} \frac{\partial u}{\partial r} + \frac{\partial^2 u}{\partial r^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

Since

$$\frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) = r \frac{\partial^2 u}{\partial r^2} + \frac{\partial u}{\partial r}$$

we obtain the polar form of the Laplace equation in two dimensions

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} = 0$$

3.1.2 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 1 Sketch the contours (in two dimensions) of the scalar functions
- (a) $f(x, y) = \ln(x^2 + y^2 - 1)$
- (b) $f(x, y) = \tan^{-1}[y/(1+x)]$
- 2 Sketch the flow lines (in two dimensions) of the vector functions
- (a) $\mathbf{F}(x, y) = y\mathbf{i} + (6x^2 - 4x)\mathbf{j}$
- (b) $\mathbf{F}(x, y) = y\mathbf{i} + (\frac{1}{6}x^3 - x)\mathbf{j}$
- where \mathbf{i} and \mathbf{j} are unit vectors in the direction of the x and y axes respectively.
- 3 Sketch the level surfaces of the functions
- (a) $f(\mathbf{r}) = z - xy$
- (b) $f(\mathbf{r}) = z - e^{-xy}$
- 4 Sketch the field lines of the functions
- (a) $\mathbf{F}(\mathbf{r}) = (xy, y^2 + 1, z)$
- (b) $\mathbf{F}(\mathbf{r}) = (yz, zx, xy)$
- 5 Find all the first and second partial derivatives of the functions
- (a) $f(\mathbf{r}) = xyz - x^2 + y - z$
- (b) $f(\mathbf{r}) = x^2yz^3$
- (c) $f(\mathbf{r}) = z \tan^{-1}(y/x)$
- 6 Find df/dt , where
- (a) $f(\mathbf{r}) = x^2 + y^2 - z$, and $x = t^3 - 1$, $y = 2t$, $z = 1/(t-1)$
- 7 Find $\partial f/\partial y$ and $\partial f/\partial z$ in terms of the partial derivatives of f with respect to spherical polar coordinates (r, θ, ϕ) (see Example 3.5).
- 8 Show that if $u(\mathbf{r}) = f(r)$, where $r^2 = x^2 + y^2 + z^2$, as usual, and
- $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$
- then
- $$\frac{d^2 f}{dr^2} + \frac{2}{r} \frac{df}{dr} = 0$$
- Hence find the general form for $f(r)$.
- 9 Show that
- $$V(x, y, z) = \frac{1}{z} \exp\left(-\frac{x^2 + y^2}{4z}\right)$$
- satisfies the differential equation
- $$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = \frac{\partial V}{\partial z}$$
- 10 Verify that $V(x, y, z) = \sin 3x \cos 4y \cosh 5z$ satisfies the differential equation
- $$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} + \frac{\partial^2 V}{\partial z^2} = 0$$

3.1.3 Transformations

Example 3.3 may be viewed as an example of *transformation of coordinates*. For example, consider the transformation or mapping from the (x, y) plane to the (s, t) plane defined by

$$s = s(x, y), \quad t = t(x, y) \quad (3.1)$$

Then a function $u = f(x, y)$ of x and y becomes a function $u = F(s, t)$ of s and t under the transformation, and the partial derivatives are related by

$$\begin{aligned} \frac{\partial u}{\partial x} &= \frac{\partial u}{\partial s} \frac{\partial s}{\partial x} + \frac{\partial u}{\partial t} \frac{\partial t}{\partial x} \\ \frac{\partial u}{\partial y} &= \frac{\partial u}{\partial s} \frac{\partial s}{\partial y} + \frac{\partial u}{\partial t} \frac{\partial t}{\partial y} \end{aligned} \quad (3.2)$$

In matrix notation this becomes

$$\begin{bmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{\partial s}{\partial x} & \frac{\partial t}{\partial x} \\ \frac{\partial s}{\partial y} & \frac{\partial t}{\partial y} \end{bmatrix} \begin{bmatrix} \frac{\partial u}{\partial s} \\ \frac{\partial u}{\partial t} \end{bmatrix} \quad (3.3)$$

The determinant of the matrix of the transformation is called the **Jacobian** of the transformation defined by (3.1) and is abbreviated to

$$\frac{\partial(s, t)}{\partial(x, y)} \quad \text{or simply to} \quad J$$

so that

$$J = \frac{\partial(s, t)}{\partial(x, y)} = \begin{vmatrix} \frac{\partial s}{\partial x} & \frac{\partial t}{\partial x} \\ \frac{\partial s}{\partial y} & \frac{\partial t}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \\ \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \end{vmatrix} \quad (3.4)$$

The matrix itself is referred to as the **Jacobian matrix** and is generally expressed in

the form $\begin{bmatrix} \frac{\partial s}{\partial x} & \frac{\partial s}{\partial y} \\ \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \end{bmatrix}$. The Jacobian plays an important role in various applications of

mathematics in engineering, particularly in implementing changes in variables in multiple integrals, as considered later in this chapter.

As indicated earlier, (3.1) define a transformation of the (x, y) plane to the (s, t) plane and give the coordinates of a point in the (s, t) plane corresponding to a point in the (x, y) plane. If we solve (3.1) for x and y , we obtain

$$x = X(s, t), \quad y = Y(s, t) \quad (3.5)$$

which represent a transformation of the (s, t) plane into the (x, y) plane. This is called the inverse transformation of the transformation defined by (3.1), and, analogously to (3.2), we can relate the partial derivatives by

$$\left. \begin{aligned} \frac{\partial u}{\partial s} &= \frac{\partial u}{\partial x} \frac{\partial x}{\partial s} + \frac{\partial u}{\partial y} \frac{\partial y}{\partial s} \\ \frac{\partial u}{\partial t} &= \frac{\partial u}{\partial x} \frac{\partial x}{\partial t} + \frac{\partial u}{\partial y} \frac{\partial y}{\partial t} \end{aligned} \right\} \quad (3.6)$$

The Jacobian of the inverse transformation (3.5) is

$$J_1 = \frac{\partial(x, y)}{\partial(s, t)} = \begin{vmatrix} x_s & y_s \\ x_t & y_t \end{vmatrix}$$

where the suffix notation has been used to denote the partial derivatives. Provided $J \neq 0$, it is always true that $J_1 = J^{-1}$ or

$$\frac{\partial(x, y)}{\partial(s, t)} \frac{\partial(s, t)}{\partial(x, y)} = 1$$

If $J = 0$ then the variables s and t defined by (3.1) are functionally dependent; that is, a relationship of the form $f(s, t) = 0$ exists. This implies a non-unique correspondence between points in the (x, y) and (s, t) planes.

Example 3.7

- (a) Obtain the Jacobian J of the transformation

$$s = 2x + y, \quad t = x - 2y$$

- (b) Determine the inverse transformation of the above transformation and obtain its Jacobian J_1 . Confirm that $J_1 = J^{-1}$.

Solution (a) Using (3.4), the Jacobian of the transformation is

$$J = \frac{\partial(s, t)}{\partial(x, y)} = \begin{vmatrix} 2 & 1 \\ 1 & -2 \end{vmatrix} = -5$$

- (b) Solving the pair of equations in the transformation for x and y gives the inverse transformation as

$$x = \frac{1}{5}(2s + t), \quad y = \frac{1}{5}(s - 2t)$$

The Jacobian of this inverse transformation is

$$J_1 = \frac{\partial(x, y)}{\partial(s, t)} = \begin{vmatrix} \frac{2}{5} & \frac{1}{5} \\ \frac{1}{5} & -\frac{2}{5} \end{vmatrix} = -\frac{1}{5}$$

confirming that $J_1 = J^{-1}$.

Example 3.8

Show that the variables x and y given by

$$x = \frac{s+t}{s}, \quad y = \frac{s+t}{t} \quad (3.7)$$

are functionally dependent, and obtain the relationship $f(x, y) = 0$.

Solution

The Jacobian of the transformation (3.7) is

$$J = \frac{\partial(x, y)}{\partial(s, t)} = \begin{vmatrix} x_s & y_s \\ x_t & y_t \end{vmatrix} = \begin{vmatrix} -\frac{t}{s^2} & \frac{1}{t} \\ \frac{1}{s} & -\frac{s}{t^2} \end{vmatrix} = \frac{1}{st} - \frac{1}{st} = 0$$

Since $J = 0$, the variables x and y are functionally related.

Rearranging (3.7), we have

$$x = 1 + \frac{t}{s}, \quad y = \frac{s}{t} + 1$$

so that

$$(x-1)(y-1) = \frac{t}{s} \cdot \frac{s}{t} = 1$$

giving the functional relationship as

$$xy - (x+y) = 0$$

The definition of a Jacobian is not restricted to functions of two variables, and it is readily extendable to functions of many variables. For example, for functions of three variables, if

$$u = U(x, y, z), \quad v = V(x, y, z), \quad w = W(x, y, z) \quad (3.8)$$

represents a transformation in three dimensions from the variables x, y, z to the variables u, v, w then the corresponding Jacobian is

$$J = \frac{\partial(u, v, w)}{\partial(x, y, z)} = \begin{vmatrix} u_x & v_x & w_x \\ u_y & v_y & w_y \\ u_z & v_z & w_z \end{vmatrix} = \begin{vmatrix} u_x & u_y & u_z \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{vmatrix}$$

Again, if $J = 0$, it follows that there exists a functional relationship $f(u, v, w) = 0$ between the variables u, v and w defined by (3.8).

3.1.4 Exercises

11 Show that if $x + y = u$ and $y = uv$, then

$$\frac{\partial(x, y)}{\partial(u, v)} = u$$

12 Show that, if $x + y + z = u$, $y + z = uv$ and $z = uvw$, then

$$\frac{\partial(x, y, z)}{\partial(u, v, w)} = u^2v$$

13 If $x = e^u \cos v$ and $y = e^u \sin v$, obtain the two Jacobians

$$\frac{\partial(x, y)}{\partial(u, v)} \quad \text{and} \quad \frac{\partial(u, v)}{\partial(x, y)}$$

and verify that they are mutual inverses.

14 Find the values of the constant parameter λ for which the functions

$$u = \cos x \cos y - \lambda \sin x \sin y$$

$$v = \sin x \cos y + \lambda \cos x \sin y$$

are functionally dependent.

15 Find the value of the constant K for which

$$u = Kx^2 + 4y^2 + z^2$$

$$v = 3x + 2y + z$$

$$w = 2yz + 3zx + 6xy$$

are functionally related, and obtain the corresponding relation.

16 Show that, if $u = g(x, y)$ and $v = h(x, y)$, then

$$\begin{aligned}\frac{\partial x}{\partial u} &= \frac{\partial v}{\partial y} \Big| J & \frac{\partial x}{\partial v} &= -\frac{\partial u}{\partial y} \Big| J \\ \frac{\partial y}{\partial u} &= -\frac{\partial v}{\partial x} \Big| J & \frac{\partial y}{\partial v} &= \frac{\partial u}{\partial x} \Big| J\end{aligned}$$

where in each case

$$J = \frac{\partial(u, v)}{\partial(x, y)}$$

17 Use the results of Exercise 16 to obtain the partial derivatives

$$\frac{\partial x}{\partial u}, \frac{\partial x}{\partial v}, \frac{\partial y}{\partial u}, \frac{\partial y}{\partial v}$$

where

$$u = e^x \cos y \quad \text{and} \quad v = e^{-x} \sin y$$

3.1.5 The total differential

Consider a function $u = f(x, y)$ of two variables x and y . Let Δx and Δy be increments in the values of x and y . Then the corresponding increment in u is given by

$$\Delta u = f(x + \Delta x, y + \Delta y) - f(x, y)$$

We rewrite this as two terms: one showing the change in u due to the change in x , and the other showing the change in u due to the change in y . Thus

$$\Delta u = [f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)] + [f(x, y + \Delta y) - f(x, y)]$$

Dividing the first bracketed term by Δx and the second by Δy gives

$$\Delta u = \frac{f(x + \Delta x, y + \Delta y) - f(x, y + \Delta y)}{\Delta x} \Delta x + \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} \Delta y$$

From the definition of the partial derivative, we may approximate this expression by

$$\Delta u \approx \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y$$

We define the **differential** du by the equation

$$du = \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y \quad (3.9)$$

By setting $f(x, y) = f_1(x, y) = x$ and $f(x, y) = f_2(x, y) = y$ in turn in (3.9), we see that

$$dx = \frac{\partial f_1}{\partial x} \Delta x + \frac{\partial f_1}{\partial y} \Delta y = \Delta x \quad \text{and} \quad dy = \Delta y$$

so that for the independent variables increments and differentials are equal. For the dependent variable we have

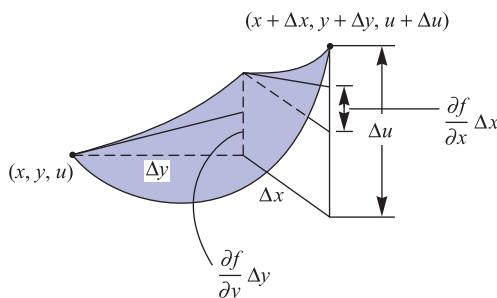
$$du = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy \quad (3.10)$$

We see that the differential du is an approximation to the change Δu in $u = f(x, y)$ resulting from small changes Δx and Δy in the independent variables x and y ; that is,

$$\Delta u \approx du = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y \quad (3.11)$$

a result illustrated in Figure 3.5.

Figure 3.5
Illustration of result
(3.11).



This extends to functions of as many variables as we please, provided that the partial derivatives exist. For example, for a function of three variables (x, y, z) defined by $u = f(x, y, z)$ we have

$$\begin{aligned} \Delta u \approx du &= \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz \\ &= \frac{\partial f}{\partial x} \Delta x + \frac{\partial f}{\partial y} \Delta y + \frac{\partial f}{\partial z} \Delta z \end{aligned}$$

The differential of a function of several variables is often called a **total differential**, emphasizing that it shows the variation of the function with respect to small changes in *all* the independent variables.

Example 3.9 Find the total differential of $u(x, y) = x^y$.

Solution Taking partial derivatives we have

$$\frac{\partial u}{\partial x} = yx^{y-1} \quad \text{and} \quad \frac{\partial u}{\partial y} = x^y \ln x$$

Hence, using (3.10),

$$du = yx^{y-1} dx + x^y \ln x dy$$

Differentials sometimes arise naturally when modelling practical problems. When this occurs, it is often possible to analyse the problem further by testing to see if the expression in which the differentials occur is a total differential. Consider the equation

$$P(x, y) dx + Q(x, y) dy = 0$$

connecting x, y and their differentials. The left-hand side of this equation is said to be an **exact differential** if there is a function $f(x, y)$ such that

$$df = P(x, y) dx + Q(x, y) dy$$

Now we know that

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy$$

so if $f(x, y)$ exists then

$$P(x, y) = \frac{\partial f}{\partial x} \quad \text{and} \quad Q(x, y) = \frac{\partial f}{\partial y}$$

For functions with continuous second derivatives we have

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$$

Thus if $f(x, y)$ exists then

$$\frac{\partial P}{\partial y} = \frac{\partial Q}{\partial x} \tag{3.12}$$

This gives us a test for the existence of $f(x, y)$, but does not tell us how to find it! The technique for finding $f(x, y)$ is shown in Example 3.10.

Example 3.10

Show that

$$(6x + 9y + 11)dx + (9x - 4y + 3)dy$$

is an exact differential and find the relationship between y and x given

$$\frac{dy}{dx} = -\frac{6x + 9y + 11}{9x - 4y + 3}$$

and the condition $y = 1$ when $x = 0$.

Solution In this example

$$P(x, y) = 6x + 9y + 11 \quad \text{and} \quad Q(x, y) = 9x - 4y + 3$$

First we test whether the expression is an exact differential. In this example

$$\frac{\partial P}{\partial y} = 9 \quad \text{and} \quad \frac{\partial Q}{\partial x} = 9$$

so from (3.12), we have an exact differential. Thus we know that there is a function $f(x, y)$ such that

$$\frac{\partial f}{\partial x} = 6x + 9y + 11 \quad \text{and} \quad \frac{\partial f}{\partial y} = 9x - 4y + 3 \quad (3.13a, b)$$

Integrating (3.13a) with respect to x , keeping y constant (that is, reversing the partial differentiation process), we have

$$f(x, y) = 3x^2 + 9xy + 11x + g(y) \quad (3.14)$$

Note that the ‘constant’ of integration is a function of y . You can check that this expression for $f(x, y)$ is correct by differentiating it partially with respect to x . But we also know from (3.13b) the partial derivative of $f(x, y)$ with respect to y , and this enables us to find $g'(y)$. Differentiating (3.14) partially with respect to y and equating it to (3.13b), we have

$$\frac{\partial f}{\partial y} = 9x + \frac{dg}{dy} = 9x - 4y + 3$$

(Note that since g is a function of y only we use dg/dy rather than $\partial g/\partial y$.) Thus

$$\frac{dg}{dy} = -4y + 3$$

so, on integrating,

$$g(y) = -2y^2 + 3y + C$$

Substituting back into (3.13b) gives

$$f(x, y) = 3x^2 + 9xy + 11x - 2y^2 + 3y + C$$

Now we are given that

$$\frac{dy}{dx} = -\frac{6x + 9y + 11}{9x - 4y + 3}$$

which implies that

$$(6x + 9y + 11)dx + (9x - 4y + 3)dy = 0$$

which in turn implies that

$$3x^2 + 9xy + 11x - 2y^2 + 3y + C = 0$$

The arbitrary constant C is fixed by applying the given condition $y = 1$ when $x = 0$, giving $C = -1$. Thus x and y satisfy the equation

$$3x^2 + 9xy + 11x - 2y^2 + 3y = 1$$

3.1.6 Exercises

- 18 Determine which of the following are exact differentials of a function, and find, where appropriate, the corresponding function.

- (a) $(y^2 + 2xy + 1)dx + (2xy + x^2)dy$
- (b) $(2xy^2 + 3y \cos 3x)dx + (2x^2y + \sin 3x)dy$
- (c) $(6xy - y^2)dx + (2xe^y - x^2)dy$
- (d) $(z^3 - 3y)dx + (12y^2 - 3x)dy + 3xz^2dz$

- 19 Find the value of the constant λ such that
- $$(y \cos x + \lambda \cos y)dx + (x \sin y + \sin x + y)dy$$

is the exact differential of a function $f(x, y)$. Find the corresponding function $f(x, y)$ that also satisfies the condition $f(0, 1) = 0$.

- 20 Show that the differential

$$g(x, y) = (10x^2 + 6xy + 6y^2)dx + (9x^2 + 4xy + 15y^2)dy$$

is not exact, but that a constant m can be chosen so that

$$(2x + 3y)^m g(x, y)$$

is equal to dz , the exact differential of a function $z = f(x, y)$. Find $f(x, y)$.

3.2

Derivatives of a scalar point function

In many practical problems it is necessary to measure the rate of change of a scalar point function. For example, in heat transfer problems we need to know the rate of change of temperature from point to point, because that determines the rate at which heat flows. Similarly, if we are investigating the electric field due to static charges, we need to know the variation of the electric potential from point to point. To determine such information, the ideas of calculus were extended to vector quantities. The first development of this was the concept of the gradient of a scalar point function.

3.2.1 The gradient of a scalar point function

We described in Section 3.1.1 how the gradient of a scalar field depended on the direction along which its rate of change was measured. We now explore this idea further. Consider the rate of change of the function $f(\mathbf{r})$ at the point (x, y, z) in the direction of the unit vector (l, m, n) . To find this, we need to evaluate the limit

$$\lim_{\Delta r \rightarrow 0} \frac{f(\mathbf{r} + \Delta \mathbf{r}) - f(\mathbf{r})}{\Delta r}$$

where $\Delta \mathbf{r}$ is in the direction of (l, m, n) . In terms of coordinates, this means

$$\begin{aligned} \mathbf{r} + \Delta \mathbf{r} &= \mathbf{r} + \Delta r(l, m, n) \\ &= (x + \Delta x, y + \Delta y, z + \Delta z) \end{aligned}$$

so that

$$\Delta x = l\Delta r, \quad \Delta y = m\Delta r, \quad \Delta z = n\Delta r$$

Thus we have to consider the limit

$$\lim_{\Delta r \rightarrow 0} \frac{f(x + l\Delta r, y + m\Delta r, z + n\Delta r) - f(x, y, z)}{\Delta r}$$

We can rewrite this as

$$\begin{aligned} & \lim_{\Delta r \rightarrow 0} \left[\frac{f(x + l\Delta r, y + m\Delta r, z + n\Delta r) - f(x, y + m\Delta r, z + n\Delta r)}{l\Delta r} \right] l \\ & + \lim_{\Delta r \rightarrow 0} \left[\frac{f(x, y + m\Delta r, z + n\Delta r) - f(x, y, z + n\Delta r)}{m\Delta r} \right] m \\ & + \lim_{\Delta r \rightarrow 0} \left[\frac{f(x, y, z + n\Delta r) - f(x, y, z)}{n\Delta r} \right] n \end{aligned}$$

Evaluating the limits, remembering that $\Delta x = l\Delta r$ and so on, we find that the rate of change of $f(\mathbf{r})$ in the direction of the unit vector (l, m, n) is

$$\frac{\partial f}{\partial x}l + \frac{\partial f}{\partial y}m + \frac{\partial f}{\partial z}n = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right) \cdot (l, m, n)$$

The vector

$$\left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right)$$

is called the **gradient** of the scalar point function $f(x, y, z)$, and is denoted by $\text{grad } f$ or by ∇f , where ∇ is the vector operator

$$\nabla = i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z}$$

where i, j and k are the usual triad of unit vectors.

The symbol ∇ is called ‘del’ or sometimes ‘nabla’. Then

$$\text{grad } f = \nabla f = \frac{\partial f}{\partial x}i + \frac{\partial f}{\partial y}j + \frac{\partial f}{\partial z}k \equiv \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right) \quad (3.15)$$

Thus we can calculate the rate of change of $f(x, y, z)$ along any direction we please. If $\hat{\mathbf{u}}$ is the unit vector in that direction then

$$(\text{grad } f) \cdot \hat{\mathbf{u}}$$

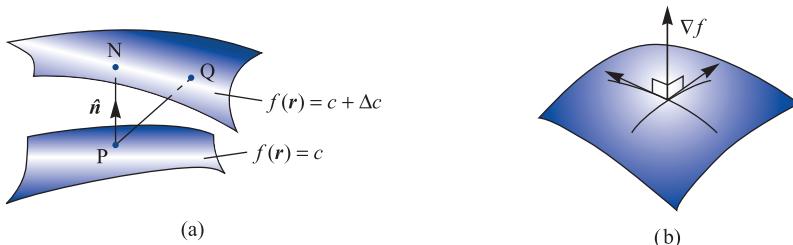
gives the required **directional derivative**, that is the rate of change of $f(x, y, z)$ in the direction of $\hat{\mathbf{u}}$. Remembering that $\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta$, where θ is the angle between the two vectors, it follows that the rate of change of $f(x, y, z)$ is zero along directions perpendicular to $\text{grad } f$ and is maximum along the direction parallel to $\text{grad } f$. Furthermore, $\text{grad } f$ acts along the normal direction to the level surface of $f(x, y, z)$. We can see this by considering the level surfaces of the function corresponding to c and $c + \Delta c$, as shown in Figure 3.6(a). In going from P on the surface $f(\mathbf{r}) = c$ to any point Q on $f(\mathbf{r}) = c + \Delta c$, the increase in f is the same whatever point Q is chosen, but the distance PQ will be smallest, and hence the rate of change of $f(x, y, z)$ greatest, when Q lies on the normal $\hat{\mathbf{n}}$ to the surface at P. Thus $\text{grad } f$ at P is in the direction of the outward normal $\hat{\mathbf{n}}$ to the surface $f(\mathbf{r}) = c$, and represents in magnitude and direction the greatest rate of increase of $f(x, y, z)$ with distance (Figure 3.6(b)). It is frequently written as

$$\text{grad } f = \frac{\partial f}{\partial n} \hat{\mathbf{n}}$$

where $\partial f / \partial n$ is referred to as the normal derivative to the surface $f(\mathbf{r}) = c$.

Figure 3.6

- (a) Adjacent level surfaces of $f(\mathbf{r})$; (b) $\text{grad } f$ acts normally to the surface $f(\mathbf{r}) = c$.



Example 3.11

Find $\text{grad } f$ for $f(\mathbf{r}) = 3x^2 + 2y^2 + z^2$ at the point $(1, 2, 3)$. Hence calculate

- the directional derivative of $f(\mathbf{r})$ at $(1, 2, 3)$ in the direction of the unit vector $\frac{1}{3}(2, 2, 1)$;
- the maximum rate of change of the function at $(1, 2, 3)$ and its direction.

Solution (a) Since $\partial f / \partial x = 6x$, $\partial f / \partial y = 4y$ and $\partial f / \partial z = 2z$, we have from (3.15) that

$$\text{grad } f = \nabla f = 6x\mathbf{i} + 4y\mathbf{j} + 2z\mathbf{k}$$

At the point $(1, 2, 3)$

$$\text{grad } f = 6\mathbf{i} + 8\mathbf{j} + 6\mathbf{k}$$

Thus the directional derivative of $f(\mathbf{r})$ at $(1, 2, 3)$ in the direction of the unit vector $(\frac{2}{3}, \frac{2}{3}, \frac{1}{3})$ is

$$(6\mathbf{i} + 8\mathbf{j} + 6\mathbf{k}) \cdot (\frac{2}{3}\mathbf{i} + \frac{2}{3}\mathbf{j} + \frac{1}{3}\mathbf{k}) = \frac{34}{3}$$

- The maximum rate of change of $f(\mathbf{r})$ at $(1, 2, 3)$ occurs along the direction parallel to $\text{grad } f$ at $(1, 2, 3)$; that is, parallel to $(6, 8, 6)$. The unit vector in that direction is $(3, 4, 3)/\sqrt{34}$ and the maximum rate of change of $f(\mathbf{r})$ is $|\text{grad } f| = 2\sqrt{34}$.

If a surface in three dimensions is specified by the equation $f(x, y, z) = c$, or equivalently $f(\mathbf{r}) = c$, then $\text{grad } f$ is a vector perpendicular to that surface. This enables us to calculate the normal vector at any point on the surface, and consequently to find the equation of the tangent plane at that point.

Example 3.12

A paraboloid of revolution has equation $2z = x^2 + y^2$. Find the unit normal vector to the surface at the point $(1, 3, 5)$. Hence obtain the equation of the normal and the tangent plane to the surface at that point.

Solution A vector normal to the surface $2z = x^2 + y^2$ is given by

$$\text{grad}(x^2 + y^2 - 2z) = 2x\mathbf{i} + 2y\mathbf{j} - 2\mathbf{k}$$

At the point $(1, 3, 5)$ the vector has the value $2\mathbf{i} + 6\mathbf{j} - 2\mathbf{k}$. Thus the normal unit vector at the point $(1, 3, 5)$ is $(\mathbf{i} + 3\mathbf{j} - \mathbf{k})/\sqrt{11}$. The equation of the line through $(1, 3, 5)$ in the direction of this normal is

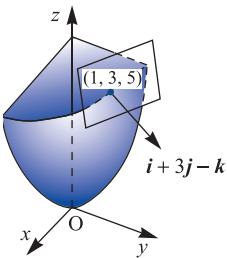


Figure 3.7 Tangent plane at $(1, 3, 5)$ to the paraboloid $2z = x^2 + y^2$ of Example 3.12.

$$\frac{x-1}{1} = \frac{y-3}{3} = \frac{z-5}{-1}$$

and the equation of the tangent plane is

$$(1)(x-1) + (3)(y-3) + (-1)(z-5) = 0$$

which simplifies to $x + 3y - z = 5$ (see Figure 3.7).

The concept of the gradient of a scalar field occurs in many applications. The simplest, perhaps, is when $f(\mathbf{r})$ represents the potential in an electric field due to static charges. Then the electric force is in the direction of the greatest decrease of the potential. Its magnitude is equal to that rate of decrease, so that the force is given by $-\nabla f$.

3.2.2 Exercises

- 21 Find ∇f for $f(\mathbf{r}) = x^2yz^2$ at the point $(1, 2, 3)$. Hence calculate

- (a) the directional derivative of $f(\mathbf{r})$ at $(1, 2, 3)$ in the direction of the vector $(-2, 3, -6)$;
- (b) the maximum rate of change of the function at $(1, 2, 3)$ and its direction.

- 22 Find ∇f where $f(\mathbf{r})$ is

- (a) $x^2 + y^2 - z$
- (b) $z \tan^{-1}(y/x)$
- (c) $e^{-x-y+z}/\sqrt{(x^3 + y^2)}$
- (d) $xyz \sin(\pi(x+y+z))$

- 23 Find the directional derivative of $f(\mathbf{r}) = x^2 + y^2 - z$ at the point $(1, 1, 2)$ in the direction of the vector $(4, 4, -2)$.

- 24 Find a unit normal to the surface $xy^2 - 3xz = -5$ at the point $(1, -2, 3)$.

- 25 If \mathbf{r} is the usual position vector $\mathbf{r} = xi + yj + zk$, with $|\mathbf{r}| = r$, evaluate

- (a) ∇r
- (b) $\nabla\left(\frac{1}{r}\right)$

- 26 If $\nabla\phi = (2xy + z^2)\mathbf{i} + (x^2 + z)\mathbf{j} + (y + 2xz)\mathbf{k}$, find a possible value for ϕ .

- 27 Given the scalar function of position

$$\phi(x, y, z) = x^2y - 3xyz + z^3$$

find the value of $\nabla\phi$ at the point $(3, 1, 2)$. Also find the directional derivative of ϕ at this point in the direction of the vector $(3, -2, 6)$; that is, in the direction $3\mathbf{i} - 2\mathbf{j} + 6\mathbf{k}$.

- 28 Find the angle between the surfaces $x^2 + y^2 + z^2 = 9$ and $z = x^2 + y^2 - 3$ at the point $(2, -1, 2)$.

- 29 Find the equations of the tangent plane and normal line to the surfaces

- (a) $x^2 + 2y^2 + 3z^2 = 6$ at $(1, 1, 1)$
- (b) $2x^2 + y^2 - z^2 = -3$ at $(1, 2, 3)$
- (c) $x^2 + y^2 - z = 1$ at $(1, 2, 4)$.

- 30 (Spherical polar coordinates) When a function $f(\mathbf{r})$ is specified in polar coordinates, it is usual to express ∇f in terms of the partial derivatives of f with respect to r , θ and ϕ and the unit vectors \mathbf{u}_r , \mathbf{u}_θ and \mathbf{u}_ϕ in the directions of increasing r , θ and ϕ as shown in Figure 3.8. Working from first principles, show that

$$\nabla f = \nabla f = \frac{\partial f}{\partial r} \mathbf{u}_r + \frac{1}{r} \frac{\partial f}{\partial \theta} \mathbf{u}_\theta + \frac{1}{r \sin \theta} \frac{\partial f}{\partial \phi} \mathbf{u}_\phi$$

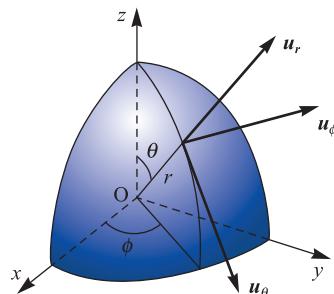


Figure 3.8 Unit vectors associated with spherical polar coordinates of Exercise 30.

3.3

Derivatives of a vector point function

When we come to consider the rate of change of a vector point function $\mathbf{F}(\mathbf{r})$, we see that there are two ways of combining the vector operator ∇ with the vector \mathbf{F} . Thus we have two cases to consider, namely

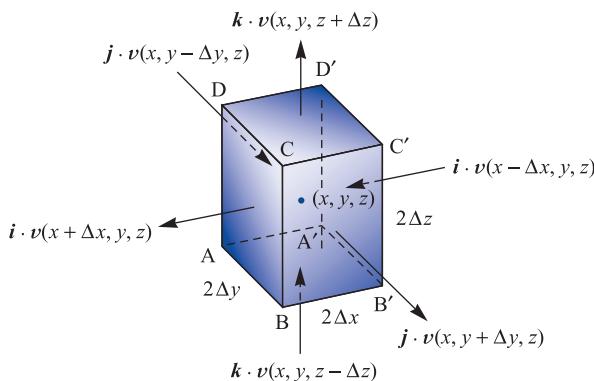
$$\nabla \cdot \mathbf{F} \quad \text{and} \quad \nabla \times \mathbf{F}$$

that is, the scalar product and vector product respectively. Both of these ‘derivatives’ have physical meanings, as we shall discover in the following sections. Roughly, if we picture a vector field as a fluid flow then at every point in the flow we need to measure the rate at which the field is flowing away from that point and also the amount of spin possessed by the particles of the fluid at that point. The two ‘derivatives’ given formally above provide these measures.

3.3.1 Divergence of a vector field

Consider the steady motion of a fluid in a region R such that a particle of fluid instantaneously at the point \mathbf{r} with coordinates (x, y, z) has a velocity $\mathbf{v}(\mathbf{r})$ that is independent of time. To measure the flow away from this point in the fluid, we surround the point by an ‘elementary’ cuboid of side $(2\Delta x) \times (2\Delta y) \times (2\Delta z)$, as shown in Figure 3.9, and calculate the average flow out of the cuboid per unit volume.

Figure 3.9 Flow out of a cuboid.



The flow out of the cuboid is the sum of the flows across each of its six faces. Representing the velocity of the fluid at (x, y, z) by \mathbf{v} , the flow out of the face ABCD is given approximately by

$$i \cdot \mathbf{v}(x + \Delta x, y, z)(4\Delta y \Delta z)$$

The flow out of the face A'B'C'D' is given approximately by

$$-i \cdot \mathbf{v}(x - \Delta x, y, z)(4\Delta y \Delta z)$$

There are similar expressions for the remaining four faces of the cuboid, so that the total flow out of the latter is

$$\begin{aligned} & \mathbf{i} \cdot [\mathbf{v}(x + \Delta x, y, z) - \mathbf{v}(x - \Delta x, y, z)](4\Delta y \Delta z) \\ & + \mathbf{j} \cdot [\mathbf{v}(x, y + \Delta y, z) - \mathbf{v}(x, y - \Delta y, z)](4\Delta x \Delta z) \\ & + \mathbf{k} \cdot [\mathbf{v}(x, y, z + \Delta z) - \mathbf{v}(x, y, z - \Delta z)](4\Delta x \Delta y) \end{aligned}$$

Dividing by the volume $8\Delta x \Delta y \Delta z$, and proceeding to the limit as $\Delta x, \Delta y, \Delta z \rightarrow 0$, we see that the flow away from the point (x, y, z) per unit time is given by

$$\mathbf{i} \cdot \frac{\partial \mathbf{v}}{\partial x} + \mathbf{j} \cdot \frac{\partial \mathbf{v}}{\partial y} + \mathbf{k} \cdot \frac{\partial \mathbf{v}}{\partial z}$$

This may be rewritten as

$$\left(\mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z} \right) \cdot \mathbf{v}$$

or simply as $\nabla \cdot \mathbf{v}$. Thus we see that the flow away from this point is given by the scalar product of the vector operator ∇ with the velocity vector \mathbf{v} . This is called the **divergence** of the vector \mathbf{v} , and is written as $\operatorname{div} \mathbf{v}$. In terms of components,

$$\begin{aligned} \operatorname{div} \mathbf{v} = \nabla \cdot \mathbf{v} &= \left(\mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z} \right) \cdot (\mathbf{i} v_1 + \mathbf{j} v_2 + \mathbf{k} v_3) \\ &= \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z} \end{aligned} \tag{3.16}$$

When \mathbf{v} is specified in this way, it is easy to compute its divergence. Note that the divergence of a vector field is a scalar quantity.

Example 3.13

Find the divergence of the vector $\mathbf{v} = (2x - y^2, 3z + x^2, 4y - z^2)$ at the point $(1, 2, 3)$.

Solution Here $v_1 = 2x - y^2$, $v_2 = 3z + x^2$ and $v_3 = 4y - z^2$, so that

$$\frac{\partial v_1}{\partial x} = 2, \quad \frac{\partial v_2}{\partial y} = 0, \quad \frac{\partial v_3}{\partial z} = -2z$$

Thus from (3.16), at a general point (x, y, z) ,

$$\operatorname{div} \mathbf{v} = \nabla \cdot \mathbf{v} = 2 - 2z$$

so that at the point $(1, 2, 3)$

$$\nabla \cdot \mathbf{v} = -4$$

A more general way of defining the divergence of a vector field $\mathbf{F}(\mathbf{r})$ at the point \mathbf{r} is to enclose the point in an elementary volume ΔV and find the flow or flux out of ΔV per unit volume. Thus

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \lim_{\Delta V \rightarrow 0} \frac{\text{flow out of } \Delta V}{\Delta V}$$

A non-zero divergence at a point in a fluid measures the rate, per unit volume, at which the fluid is flowing away from or towards that point. That implies that either the density of the fluid is changing at the point or there is a source or sink of fluid there. In the case of a non-material vector field, for example temperature gradient in heat transfer, a non-zero divergence indicates a point of generation or absorption. When the divergence is everywhere zero, the flow entering any element of the space is exactly balanced by the outflow. This implies that the lines of flow of the field $\mathbf{F}(\mathbf{r})$ where $\operatorname{div} \mathbf{F} = 0$ must either form closed curves or finish at boundaries or extend to infinity. Vectors satisfying this condition are sometimes termed **solenoidal**.



Using MuPAD in MATLAB the divergence of a vector field is given by the divergence function. For example, the divergence of the vector

$$\mathbf{v} = (2x - y^2, 3z + x^2, 4y - z^2)$$

considered in Example 3.13, is given by the commands

```
delete x, y, z;
linalg :: divergence([2*x -y^2, 3*z + x^2, 4*y - x^2],
[x, y, z])
```

which return the answer

$$2 - 2z$$

In MAPLE the answer is returned using the commands

```
with(VectorCalculus):
SetCoordinates('cartesian' [ x, y, z]);
F:= VectorField(<2*x -y^2, 3*z + x^2, 4*y - x^2>);
Divergence(F); or Del.F;
```

3.3.2 Exercises

- 31 Find $\operatorname{div} \mathbf{v}$ where

(a) $\mathbf{v}(\mathbf{r}) = 3x^2\mathbf{i} + \mathbf{j} + x^2\mathbf{k}$

(b) $\mathbf{v}(\mathbf{r}) = (3x + y)\mathbf{i} + (2z + x)\mathbf{j} + (z - 2y)\mathbf{k}$

- 32 If $\mathbf{F} = (2xy^2 + z^2)\mathbf{i} + (3x^2z^2 - y^2z^3)\mathbf{j} + (yz^2 - xz^3)\mathbf{k}$, calculate $\operatorname{div} \mathbf{f}$ at the point $(-1, 2, 3)$.

- 33 Find $\nabla(\mathbf{a} \cdot \mathbf{r})$, $(\mathbf{a} \cdot \nabla)\mathbf{r}$ and $\mathbf{a}(\nabla \cdot \mathbf{r})$, where \mathbf{a} is a constant vector and, as usual, \mathbf{r} is the position vector $\mathbf{r} = (x, y, z)$.

- 34 The vector \mathbf{v} is defined by $\mathbf{v} = \mathbf{r}\mathbf{r}^{-1}$, where $\mathbf{r} = (x, y, z)$ and $r = |\mathbf{r}|$. Show that

$$\nabla(\nabla \cdot \mathbf{v}) \equiv \operatorname{grad} \operatorname{div} \mathbf{v} = -\frac{2}{r^3} \mathbf{r}$$

- 35 Find the value of the constant λ such that the vector field defined by

$$\mathbf{F} = (2x^2y^2 + z^2)\mathbf{i} + (3xy^3 - x^2z)\mathbf{j} + (\lambda xy^2z + xy)\mathbf{k}$$

is solenoidal.

- 36

(Spherical polar coordinates) Using the notation introduced in Exercise 30, show, working from first principles, that

$$\begin{aligned} \nabla \cdot \mathbf{v} = \operatorname{div} \mathbf{v} &= \frac{1}{r^2} \frac{\partial}{\partial r} (r^2 v_r) + \frac{1}{r \sin \theta} \frac{\partial}{\partial \theta} (v_\theta \sin \theta) \\ &\quad + \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} (v_\phi) \end{aligned}$$

$$\text{where } \mathbf{v} = v_r \mathbf{u}_r + v_\theta \mathbf{u}_\theta + v_\phi \mathbf{u}_\phi.$$

- 37

A force field \mathbf{F} , defined by the inverse square law, is given by

$$\mathbf{F} = \mathbf{r}/r^3$$

Show that $\nabla \cdot \mathbf{F} = 0$.

3.3.3 Curl of a vector field

It is clear from observations (for example, by watching the movements of marked corks on water) that many fluid flows involve rotational motion of the fluid particles. Complete determination of this motion requires knowledge of the axis of rotation, the rate of rotation and its sense (clockwise or anticlockwise). The measure of rotation is thus a vector quantity, which we shall find by calculating its x , y and z components separately. Consider the vector field $\mathbf{v}(r)$. To find the flow around an axis in the x direction at the point r , we take an elementary rectangle surrounding r perpendicular to the x direction, as shown in Figure 3.10.

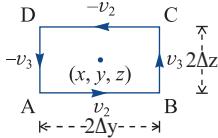


Figure 3.10 Flow around a rectangle.

To measure the circulation around the point r about an axis parallel to the x direction, we calculate the flow around the elementary rectangle ABCD and divide by its area, giving

$$[v_2(x, y^*, z - \Delta z)(2\Delta y) + v_3(x, y + \Delta y, z^*)(2\Delta z) \\ - v_2(x, \tilde{y}, z + \Delta z)(2\Delta y) - v_3(x, y - \Delta y, T)(2\Delta z)]/(4\Delta y \Delta z)$$

where $y^*, \tilde{y} \in (y - \Delta y, y + \Delta y)$, $z^*, \tilde{z} \in (z - \Delta z, z + \Delta z)$ and $\mathbf{v} = v_1 \mathbf{i} + v_2 \mathbf{j} + v_3 \mathbf{k}$. Rearranging, we obtain

$$-[v_2(x, \tilde{y}, z + \Delta z) - v_2(x, y^*, z - \Delta z)]/(2\Delta z) \\ + [v_3(x, y + \Delta y, z^*) - v_3(x, y - \Delta y, \tilde{z})]/(2\Delta y)$$

Proceeding to the limit as $\Delta y \Delta z \rightarrow 0$, we obtain the x component of this vector as

$$\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z}$$

By similar arguments, we obtain the y and z components as

$$\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x}, \quad \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}$$

respectively.

The vector measuring the rotation about a point in the fluid is called the **curl** of \mathbf{v} :

$$\text{curl } \mathbf{v} = \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z} \right) \mathbf{i} + \left(\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x} \right) \mathbf{j} + \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right) \mathbf{k} \\ = \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z}, \frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x}, \frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right) \quad (3.17)$$

It may be written formally as

$$\text{curl } \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix} \quad (3.18)$$

or more compactly as

$$\text{curl } \mathbf{v} = \nabla \times \mathbf{v}$$

Example 3.14 Find the curl of the vector $\mathbf{v} = (2x - y^2, 3z + x^2, 4y - z^2)$ at the point $(1, 2, 3)$.

Solution Here $v_1 = 2x - y^2$, $v_2 = 3z + x^2$, $v_3 = 4y - z^2$, so that

$$\begin{aligned}\operatorname{curl} \mathbf{v} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 2x - y^2 & 3z + x^2 & 4y - z^2 \end{vmatrix} \\ &= \mathbf{i} \left[\frac{\partial}{\partial y} (4y - z^2) - \frac{\partial}{\partial z} (3z + x^2) \right] \\ &\quad - \mathbf{j} \left[\frac{\partial}{\partial x} (4y - z^2) - \frac{\partial}{\partial z} (2x - y^2) \right] \\ &\quad + \mathbf{k} \left[\frac{\partial}{\partial x} (3z + x^2) - \frac{\partial}{\partial y} (2x - y^2) \right] \\ &= \mathbf{i}(4 - 3) - \mathbf{j}(0 - 0) + \mathbf{k}(2x + 2y) = \mathbf{i} + 2(x + y)\mathbf{k}\end{aligned}$$

Thus, at the point $(1, 2, 3)$, $\nabla \times \mathbf{v} = (1, 0, 6)$.

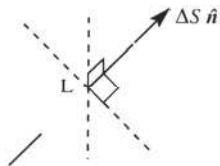


Figure 3.11
Circulation around
the element ΔS .

More generally, the component of the curl of a vector field $\mathbf{F}(\mathbf{r})$ in the direction of the unit vector $\hat{\mathbf{n}}$ at a point L is found by enclosing L by an elementary area ΔS that is perpendicular to $\hat{\mathbf{n}}$, as in Figure 3.11, and calculating the flow around ΔS per unit area. Thus

$$(\operatorname{curl} \mathbf{F}) \cdot \hat{\mathbf{n}} = \lim_{\Delta s \rightarrow 0} \frac{\text{flow round } \Delta S}{\Delta S}$$

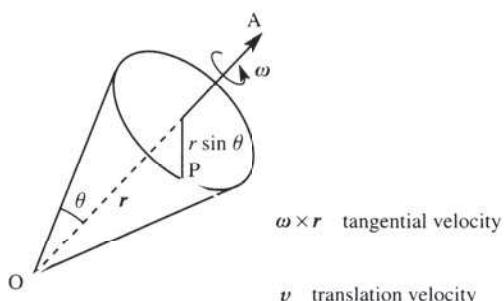
Another way of visualizing the meaning of the curl of a vector is to consider the motion of a rigid body. We can describe such motion by specifying the angular velocity ω of the body about an axis OA, where O is a fixed point in the body, together with the translational (linear) velocity \mathbf{v} of O itself. Then at any point P in the body the velocity \mathbf{u} is given by

$$\mathbf{u} = \mathbf{v} + \omega \times \mathbf{r}$$

as shown in Figure 3.12. Here \mathbf{v} and ω are independent of (x, y, z) . Thus

$$\operatorname{curl} \mathbf{u} = \operatorname{curl} \mathbf{v} + \operatorname{curl} (\omega \times \mathbf{r}) = 0 + \operatorname{curl} (\omega \times \mathbf{r})$$

Figure 3.12
Rotation of a
rigid body.



The vector $\omega \times r$ is given by

$$\begin{aligned}\omega \times r &= (\omega_1, \omega_2, \omega_3) \times (x, y, z) \\ &= (\omega_2 z - \omega_3 y) \mathbf{i} + (\omega_3 x - \omega_1 z) \mathbf{j} + (\omega_1 y - \omega_2 x) \mathbf{k}\end{aligned}$$

and

$$\begin{aligned}\text{curl}(\omega \times r) &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ \omega_2 z - \omega_3 y & \omega_3 x - \omega_1 z & \omega_1 y - \omega_2 x \end{vmatrix} \\ &= 2\omega_1 \mathbf{i} + 2\omega_2 \mathbf{j} + 2\omega_3 \mathbf{k} = 2\omega\end{aligned}$$

Thus

$$\text{curl } \mathbf{u} = 2\omega$$

that is,

$$\omega = \frac{1}{2} \text{curl } \mathbf{u}$$

Hence when any rigid body is in motion, the curl of its linear velocity at any point is twice its angular velocity in magnitude and has the same direction.

Applying this result to the motion of a fluid, we can see by regarding particles of the fluid as miniature bodies that when the curl of the velocity is zero there is no rotation of the particle, and the motion is said to be **curl-free** or **irrotational**. When the curl is non-zero, the motion is **rotational**.



Using MuPAD in MATLAB the command `linalg :: curl(v, x)` computes the curl of the three-dimensional vector field v with respect to the three-dimensional vector x in cartesian coordinates. For example, the curl of the vector

$$\mathbf{v} = (2x - y^2, 3z + x^2, 4y - z^2)$$

considered in Example 3.14, is given by the commands

```
delete x, y, z;
linalg :: curl([2*x -y^2, 3*z + x^2, 4*y - z^2],
[x, y, z])
```

which return the answer $\begin{pmatrix} 1 \\ 0 \\ 2x + 2y \end{pmatrix}$.

In MAPLE the answer is returned using the commands

```
with(VectorCalculus):
SetCoordinates('cartesian' [ x, y, z]);
F:= VectorField(<2*x -y^2, 3*z + x^2, 4*y - z^2>);
Curl(F); or Del & x F;
```

3.3.4 Exercises

38 Find $\mathbf{u} = \operatorname{curl} \mathbf{v}$ when $\mathbf{v} = (3xz^2, -yz, x + 2z)$.

39 A vector field is defined by $\mathbf{v} = (yz, xz, xy)$. Show that $\operatorname{curl} \mathbf{v} = 0$.

40 Show that if $\mathbf{v} = (2x + yz, 2y + zx, 2z + xy)$ then $\operatorname{curl} \mathbf{v} = 0$, and find $f(\mathbf{r})$ such that $\mathbf{v} = \operatorname{grad} f$.

41 By evaluating each term separately, verify the identity

$$\nabla \times (f\mathbf{v}) = f(\nabla \times \mathbf{v}) + (\nabla f) \times \mathbf{v}$$

for $f(\mathbf{r}) = x^3 - y$ and $\mathbf{v}(\mathbf{r}) = (z, 0, -x)$.

42 Find constants a, b and c such that the vector field defined by

$$\mathbf{F} = (4xy + az^3)\mathbf{i} + (bx^2 + 3z)\mathbf{j} + (6xz^2 + cy)\mathbf{k}$$

is irrotational. With these values of a, b and c , determine a scalar function $\phi(x, y, z)$ such that $\mathbf{F} = \nabla\phi$.

43 If $\mathbf{v} = -yi + xj + xyzk$ is the velocity vector of a fluid, find the local value of the angular velocity at the point $(1, 3, 2)$.

44 If the velocity of a fluid at the point (x, y, z) is given by

$$\mathbf{v} = (ax + by)\mathbf{i} + (cx + dy)\mathbf{j}$$

find the conditions on the constants a, b, c and d in order that

$$\operatorname{div} \mathbf{v} = 0, \quad \operatorname{curl} \mathbf{v} = \mathbf{0}$$

Verify that in this case

$$\mathbf{v} = \frac{1}{2} \operatorname{grad}(ax^2 + 2bxy - ay^2)$$

45 (Spherical polar coordinates) Using the notation introduced in Exercise 30, show that

$$\nabla \times \mathbf{v} = \operatorname{curl} \mathbf{v}$$

$$= \frac{1}{r^2 \sin \theta} \begin{vmatrix} \mathbf{u}_r & r\mathbf{u}_\theta & r \sin \mathbf{u}_\phi \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \theta} & \frac{\partial}{\partial \phi} \\ v_r & rv_\theta & r \sin v_\phi \end{vmatrix}$$

3.3.5 Further properties of the vector operator ∇

So far we have used the vector operator in three ways:

$$\nabla f = \operatorname{grad} f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}, \quad f(\mathbf{r}) \text{ a scalar field}$$

$$\nabla \cdot \mathbf{F} = \operatorname{div} \mathbf{F} = \frac{\partial f_1}{\partial x} + \frac{\partial f_2}{\partial y} + \frac{\partial f_3}{\partial z}, \quad \mathbf{F}(\mathbf{r}) \text{ a vector field}$$

$$\nabla \times \mathbf{F} = \operatorname{curl} \mathbf{F}$$

$$= \left(\frac{\partial f_3}{\partial y} - \frac{\partial f_2}{\partial z} \right) \mathbf{i} + \left(\frac{\partial f_1}{\partial z} - \frac{\partial f_3}{\partial x} \right) \mathbf{j} + \left(\frac{\partial f_2}{\partial x} - \frac{\partial f_1}{\partial y} \right) \mathbf{k}, \quad \mathbf{F}(\mathbf{r}) \text{ a vector field}$$

A further application is in determining the directional derivative of a vector field:

$$\begin{aligned} \mathbf{a} \cdot \nabla \mathbf{F} &= \left(a_1 \frac{\partial}{\partial x} + a_2 \frac{\partial}{\partial y} + a_3 \frac{\partial}{\partial z} \right) \mathbf{F} \\ &= \left(a_1 \frac{\partial f_1}{\partial x} + a_2 \frac{\partial f_1}{\partial y} + a_3 \frac{\partial f_1}{\partial z} \right) \mathbf{i} + \left(a_1 \frac{\partial f_2}{\partial x} + a_2 \frac{\partial f_2}{\partial y} + a_3 \frac{\partial f_2}{\partial z} \right) \mathbf{j} \\ &\quad + \left(a_1 \frac{\partial f_3}{\partial x} + a_2 \frac{\partial f_3}{\partial y} + a_3 \frac{\partial f_3}{\partial z} \right) \mathbf{k} \end{aligned}$$

The ordinary rules of differentiation carry over to this vector differential operator, but they have to be applied with care, using the rules of vector algebra. For non-orthogonal coordinate systems a specialist textbook should be consulted. Thus for scalar fields $f(\mathbf{r})$, $g(\mathbf{r})$ and vector fields $\mathbf{u}(\mathbf{r})$, $\mathbf{v}(\mathbf{r})$ we have

$$\nabla[f(g(\mathbf{r}))] = \frac{df}{dg} \nabla g \quad (3.19a)$$

$$\nabla[f(\mathbf{r})g(\mathbf{r})] = g(\mathbf{r})\nabla f(\mathbf{r}) + f(\mathbf{r})\nabla g(\mathbf{r}) \quad (3.19b)$$

$$\nabla[\mathbf{u}(\mathbf{r}) \cdot \mathbf{v}(\mathbf{r})] = \mathbf{v} \times (\nabla \times \mathbf{u}) + \mathbf{u} \times (\nabla \times \mathbf{v}) + (\mathbf{v} \cdot \nabla)\mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{v} \quad (3.19c)$$

$$\nabla \cdot [\mathbf{f}(\mathbf{r})\mathbf{u}(\mathbf{r})] = \mathbf{u} \cdot \nabla f + f \nabla \cdot \mathbf{u} \quad (3.19d)$$

$$\nabla \times [\mathbf{f}(\mathbf{r})\mathbf{u}(\mathbf{r})] = (\nabla f) \times \mathbf{u} + f \nabla \times \mathbf{u} \quad (3.19e)$$

$$\nabla \cdot [\mathbf{u}(\mathbf{r}) \times \mathbf{v}(\mathbf{r})] = \mathbf{v} \cdot (\nabla \times \mathbf{u}) - \mathbf{u} \cdot (\nabla \times \mathbf{v}) \quad (3.19f)$$

$$\nabla \times [\mathbf{u}(\mathbf{r}) \times \mathbf{v}(\mathbf{r})] = (\mathbf{v} \cdot \nabla)\mathbf{u} - \mathbf{v}(\nabla \cdot \mathbf{u}) - (\mathbf{u} \cdot \nabla)\mathbf{v} + \mathbf{u}(\nabla \cdot \mathbf{v}) \quad (3.19g)$$

Higher-order derivatives can also be formed, giving the following:

$$\operatorname{div}[\operatorname{grad} f(\mathbf{r})] = \nabla \cdot \nabla f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \nabla^2 f \quad (3.20)$$

where ∇^2 is called the **Laplacian operator** (sometimes denoted by Δ);

$$\operatorname{curl}[\operatorname{grad} f(\mathbf{r})] = \nabla \times \nabla f(\mathbf{r}) \equiv 0 \quad (3.21)$$

since

$$\begin{aligned} \nabla \times \nabla f &= \left(\frac{\partial^2 f}{\partial y \partial z} - \frac{\partial^2 f}{\partial z \partial y} \right) \mathbf{i} + \left(\frac{\partial^2 f}{\partial z \partial x} - \frac{\partial^2 f}{\partial x \partial z} \right) \mathbf{j} + \left(\frac{\partial^2 f}{\partial x \partial y} - \frac{\partial^2 f}{\partial y \partial x} \right) \mathbf{k} \\ &= 0 \end{aligned}$$

when all second-order derivatives of $f(\mathbf{r})$ are continuous;

$$\operatorname{div}[\operatorname{curl} \mathbf{v}(\mathbf{r})] = \nabla \cdot (\nabla \times \mathbf{v}) \equiv 0 \quad (3.22)$$

since

$$\frac{\partial}{\partial x} \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z} \right) + \frac{\partial}{\partial y} \left(\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x} \right) + \frac{\partial}{\partial z} \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y} \right) = 0$$

$$\operatorname{grad}(\operatorname{div} \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) = \left(\mathbf{i} \frac{\partial}{\partial x} + \mathbf{j} \frac{\partial}{\partial y} + \mathbf{k} \frac{\partial}{\partial z} \right) \left(\frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z} \right) \quad (3.23)$$

$$\nabla^2 \mathbf{v} = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) (v_1 \mathbf{i} + v_2 \mathbf{j} + v_3 \mathbf{k}) \quad (3.24)$$

$$\operatorname{curl}[\operatorname{curl} \mathbf{v}(\mathbf{r})] = \nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v} \quad (3.25)$$

Example 3.15

Verify that $\nabla \times (\nabla \times \mathbf{v}) = \nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v}$ for the vector field $\mathbf{v} = (3xz^2, -yz, x + 2z)$.

Solution

$$\begin{aligned}\nabla \times \mathbf{v} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 3xz^2 & -yz & x + 2z \end{vmatrix} = (y, 6xz - 1, 0) \\ \nabla \times (\nabla \times \mathbf{v}) &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ y & 6xz - 1 & 0 \end{vmatrix} = (-6x, 0, 6z - 1) \\ \nabla \cdot \mathbf{v} &= \frac{\partial}{\partial x}(3xz^2) + \frac{\partial}{\partial y}(-yz) + \frac{\partial}{\partial z}(x + 2z) = 3z^2 - z + 2 \\ \nabla(\nabla \cdot \mathbf{v}) &= (0, 0, 6z - 1) \\ \nabla^2 \mathbf{v} &= (\nabla^2(3xz^2), \nabla^2(-yz), \nabla^2(x + 2z)) = (6x, 0, 0)\end{aligned}$$

Thus

$$\nabla(\nabla \cdot \mathbf{v}) - \nabla^2 \mathbf{v} = (-6x, 0, 6z - 1) = \nabla \times (\nabla \times \mathbf{v})$$

Similar verifications for other identities are suggested in Exercises 3.3.6.

Example 3.16

Maxwell's equations in free space may be written, in Gaussian units, as

$$(a) \operatorname{div} \mathbf{H} = 0, \quad (b) \operatorname{div} \mathbf{E} = 0$$

$$(c) \operatorname{curl} \mathbf{H} = \nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t}, \quad (d) \operatorname{curl} \mathbf{E} = \nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t}$$

where c is the velocity of light (assumed constant). Show that these equations are satisfied by

$$\mathbf{H} = \frac{1}{c} \frac{\partial}{\partial t} \operatorname{grad} \phi \times \mathbf{k}, \quad \mathbf{E} = -\mathbf{k} \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} + \frac{\partial}{\partial z} \operatorname{grad} \phi$$

where ϕ satisfies

$$\nabla^2 \phi = \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2}$$

and \mathbf{k} is a unit vector along the z axis.

$$\text{Solution (a)} \quad \mathbf{H} = \frac{1}{c} \frac{\partial}{\partial t} \operatorname{grad} \phi \times \mathbf{k}$$

gives

$$\begin{aligned}\operatorname{div} \mathbf{H} &= \frac{1}{c} \frac{\partial}{\partial t} \operatorname{div} (\operatorname{grad} \phi \times \mathbf{k}) \\ &= \frac{1}{c} \frac{\partial}{\partial t} [\mathbf{k} \cdot \operatorname{curl} (\operatorname{grad} \phi) - (\operatorname{grad} \phi) \cdot \operatorname{curl} \mathbf{k}], \quad \text{from (3.19f)}$$

By (3.21), $\operatorname{curl} (\operatorname{grad} \phi) = 0$, and since \mathbf{k} is a constant vector, $\operatorname{curl} \mathbf{k} = 0$, so that

$$\operatorname{div} \mathbf{H} = 0$$

$$(b) \quad \mathbf{E} = -\frac{\mathbf{k}}{c^2} \frac{\partial^2 \phi}{\partial t^2} + \frac{\partial}{\partial z} \operatorname{grad} \phi$$

gives

$$\begin{aligned} \operatorname{div} \mathbf{E} &= -\frac{1}{c^2} \operatorname{div} \left(\mathbf{k} \frac{\partial^2 \phi}{\partial t^2} \right) + \frac{\partial}{\partial z} \operatorname{div} \operatorname{grad} \phi \\ &= -\frac{1}{c^2} \frac{\partial}{\partial z} \left(\frac{\partial^2 \phi}{\partial t^2} \right) + \frac{\partial}{\partial z} (\nabla^2 \phi), \quad \text{by (3.20)} \\ &= \frac{\partial}{\partial z} \left(\nabla^2 \phi - \frac{1}{c^2} \frac{\partial^2 \phi}{\partial t^2} \right) \end{aligned}$$

and since $\nabla^2 \phi = (1/c^2) \partial^2 \phi / \partial t^2$, we have

$$\operatorname{div} \mathbf{E} = 0$$

$$\begin{aligned} (c) \quad \operatorname{curl} \mathbf{H} &= \frac{1}{c} \frac{\partial}{\partial t} \operatorname{curl} (\operatorname{grad} \phi \times \mathbf{k}) \\ &= \frac{1}{c} \frac{\partial}{\partial t} [(\mathbf{k} \cdot \nabla) \operatorname{grad} \phi \\ &\quad - \mathbf{k} (\operatorname{div} \operatorname{grad} \phi) - (\operatorname{grad} \phi \cdot \nabla) \mathbf{k} + \operatorname{grad} \phi (\nabla \cdot \mathbf{k})], \quad \text{from (3.19g)} \\ &= \frac{1}{c} \frac{\partial}{\partial t} \left(\frac{\partial}{\partial z} \operatorname{grad} \phi - \mathbf{k} \nabla^2 \phi \right), \quad \text{since } \mathbf{k} \text{ is a constant vector} \\ &= \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t} \end{aligned}$$

$$\begin{aligned} (d) \quad \operatorname{curl} \mathbf{E} &= -\frac{1}{c^2} \operatorname{curl} \left(\mathbf{k} \frac{\partial^2 \phi}{\partial t^2} \right) + \frac{\partial}{\partial z} \operatorname{curl} \operatorname{grad} \phi \\ &= -\frac{1}{c^2} \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 0 & 0 & \frac{\partial^2 \phi}{\partial t^2} \end{bmatrix}, \quad \text{since } \operatorname{curl} \operatorname{grad} \phi = 0 \text{ by (3.21)} \\ &= -\frac{1}{c^2} \left(\mathbf{i} \frac{\partial^3 \phi}{\partial y \partial t^2} - \mathbf{j} \frac{\partial^3 \phi}{\partial x \partial t^2} \right) \end{aligned}$$

Also,

$$\begin{aligned} \frac{\partial \mathbf{H}}{\partial t} &= \frac{1}{c} \frac{\partial^2}{\partial t^2} \operatorname{grad} \phi \times \mathbf{k} \\ &= \frac{1}{c} \frac{\partial^2}{\partial t^2} (\operatorname{grad} \phi \times \mathbf{k}), \quad \text{since } \mathbf{k} \text{ is a constant vector} \\ &= \frac{1}{c} \frac{\partial^2}{\partial t^2} \left[\left(\frac{\partial \phi}{\partial x} \mathbf{i} + \frac{\partial \phi}{\partial y} \mathbf{j} + \frac{\partial \phi}{\partial z} \mathbf{k} \right) \times \mathbf{k} \right] = \frac{1}{c} \frac{\partial^2}{\partial t^2} \left(\mathbf{i} \frac{\partial \phi}{\partial y} - \mathbf{j} \frac{\partial \phi}{\partial x} \right) = \frac{1}{c} \left(\mathbf{i} \frac{\partial^3 \phi}{\partial y \partial t^2} - \mathbf{j} \frac{\partial^3 \phi}{\partial x \partial t^2} \right) \end{aligned}$$

so that we have

$$\nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t}$$

3.3.6 Exercises

46 Show that if g is a function of $\mathbf{r} = (x, y, z)$ then

$$\operatorname{grad} g = \frac{1}{r} \frac{dg}{dr} \mathbf{r}$$

Deduce that if \mathbf{u} is a vector field then

$$\operatorname{div}[(\mathbf{u} \times \mathbf{r})g] = (\mathbf{r} \cdot \operatorname{curl} \mathbf{u})g$$

- 47 For $\phi(x, y, z) = x^2y^2z^3$ and $\mathbf{F}(x, y, z) = x^2\mathbf{i} + xy^2\mathbf{j} - yz^2\mathbf{k}$ determine
 (a) $\nabla^2\phi$ (b) $\operatorname{grad} \operatorname{div} \mathbf{F}$ (c) $\operatorname{curl} \operatorname{curl} \mathbf{F}$

48 Show that if \mathbf{a} is a constant vector and \mathbf{r} is the position vector $\mathbf{r} = (x, y, z)$ then

$$\operatorname{div}\{\operatorname{grad}[(\mathbf{r} \cdot \mathbf{r})(\mathbf{r} \cdot \mathbf{a})]\} = 10(\mathbf{r} \cdot \mathbf{a})$$

49 Verify the identity

$$\nabla^2\mathbf{v} = \operatorname{grad} \operatorname{div} \mathbf{v} - \operatorname{curl} \operatorname{curl} \mathbf{v}$$

for the vector field $\mathbf{v} = x^2y(x\mathbf{i} + y\mathbf{j} + z\mathbf{k})$.

50 Verify, by calculating each term separately, the identities

$$\operatorname{div}(\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot \operatorname{curl} \mathbf{u} - \mathbf{u} \cdot \operatorname{curl} \mathbf{v}$$

$$\begin{aligned} \operatorname{curl}(\mathbf{u} \times \mathbf{v}) &= \mathbf{u} \operatorname{div} \mathbf{v} - \mathbf{v} \operatorname{div} \mathbf{u} + (\mathbf{v} \cdot \nabla) \mathbf{u} \\ &\quad - (\mathbf{u} \cdot \nabla) \mathbf{v} \end{aligned}$$

when $\mathbf{u} = xy\mathbf{j} + xz\mathbf{k}$ and $\mathbf{v} = xy\mathbf{i} + yz\mathbf{k}$.

51 If \mathbf{r} is the usual position vector $\mathbf{r} = (x, y, z)$, show that

$$(a) \operatorname{div} \operatorname{grad}\left(\frac{1}{r}\right) = 0$$

$$(b) \operatorname{curl}\left[\mathbf{k} \times \operatorname{grad}\left(\frac{1}{r}\right)\right] + \operatorname{grad}\left[\mathbf{k} \cdot \operatorname{grad}\left(\frac{1}{r}\right)\right] = 0$$

52 If \mathbf{A} is a constant vector and \mathbf{r} is the position vector $\mathbf{r} = (x, y, z)$, show that

$$(a) \operatorname{grad}\left(\frac{\mathbf{A} \cdot \mathbf{r}}{r^3}\right) = \frac{\mathbf{A}}{r^3} - 3\frac{(\mathbf{A} \cdot \mathbf{r})}{r^5} \mathbf{r}$$

$$(b) \operatorname{curl}\left(\frac{\mathbf{A} \times \mathbf{r}}{r^3}\right) = \frac{2\mathbf{A}}{r^3} + \frac{3}{r^5}(\mathbf{A} \times \mathbf{r}) \times \mathbf{r}$$

53 If \mathbf{r} is the position vector $\mathbf{r} = (x, y, z)$, and \mathbf{a} and \mathbf{b} are constant vectors, show that

$$(a) \nabla \times \mathbf{r} = 0$$

$$(b) (\mathbf{a} \cdot \nabla) \mathbf{r} = \mathbf{a}$$

$$(c) \nabla \times [(\mathbf{a} \cdot \mathbf{r})\mathbf{b} - (\mathbf{b} \cdot \mathbf{r})\mathbf{a}] = 2(\mathbf{a} \times \mathbf{b})$$

$$(d) \nabla \cdot [(\mathbf{a} \cdot \mathbf{r})\mathbf{b} - (\mathbf{b} \cdot \mathbf{r})\mathbf{a}] = 0$$

54 By evaluating $\nabla \cdot (\nabla f)$, show that the Laplacian in spherical polar coordinates (see Exercise 30) is given by

$$\begin{aligned} \nabla^2 f &= \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial f}{\partial \theta} \right) \\ &\quad + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 f}{\partial \phi^2} \end{aligned}$$

55 Show that Maxwell's equations in free space, namely

$$\operatorname{div} \mathbf{H} = 0, \quad \operatorname{div} \mathbf{E} = 0$$

$$\nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t}, \quad \nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t}$$

are satisfied by

$$\mathbf{H} = \frac{1}{c} \operatorname{curl} \frac{\partial \mathbf{Z}}{\partial t}$$

$$\mathbf{E} = \operatorname{curl} \operatorname{curl} \mathbf{Z}$$

where the Hertzian vector \mathbf{Z} satisfies

$$\nabla^2 \mathbf{Z} = \frac{1}{c} \frac{\partial^2 \mathbf{Z}}{\partial t^2}$$

3.4 Topics in integration

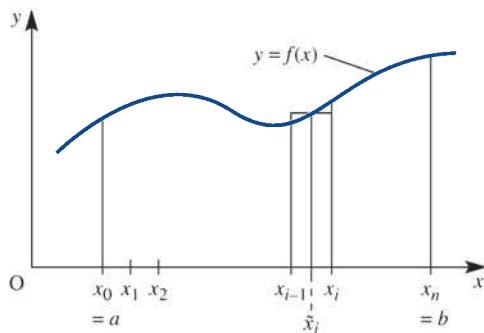
In the previous sections we saw how the idea of the differentiation of a function of a single variable is generalized to include scalar and vector point functions. We now turn to the inverse process of integration. The fundamental idea of an integral is that of

summing all the constituent parts that make a whole. More formally, we define the integral of a function $f(x)$ by

$$\int_a^b f(x) dx = \lim_{\substack{n \rightarrow \infty \\ \text{all } \Delta x_i \rightarrow 0}} \sum_{i=1}^n f(\tilde{x}_i) \Delta x_i$$

where $a = x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n = b$, $\Delta x_i = x_i - x_{i-1}$ and $x_{i-1} \leq \tilde{x}_i \leq x_i$. Geometrically, we can interpret this integral as the area between the graph $y = f(x)$, the x axis and the lines $x = a$ and $x = b$, as illustrated in Figure 3.13.

Figure 3.13 Definite integral as an area.



3.4.1 Line integrals

Consider the integral

$$\int_b^a f(x, y) dx, \quad \text{where } y = g(x)$$

This can be evaluated in the usual way by first substituting for y in terms of x in the integrand and then performing the integration

$$\int_b^a f(x, g(x)) dx$$

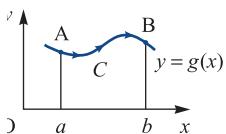


Figure 3.14 Integral along a curve.

Clearly the value of the integral will, in general, depend on the function $y = g(x)$. It may be interpreted as evaluating the integral $\int_a^b f(x, y) dx$ along the curve $y = g(x)$, as shown in Figure 3.14. Note, however, that the integral is *not* represented in this case by the area under the curve. This type of integral is called a **line integral**.

There are many different types of such integrals, for example

$$\int_C^B f(x, y) dx, \quad \int_C^B f(x, y) ds, \quad \int_{t_1}^{t_2} f(x, y) dt, \quad \int_C^B [f_1(x, y) dx + f_2(x, y) dy]$$

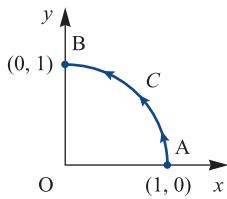
Here the letter under the integral sign indicates that the integral is evaluated along the curve (or **path**) C . This path is not restricted to two dimensions, and may be in as many dimensions as we please. It is normal to omit the points A and B , since they are usually implicit in the specification of C .

Example 3.17

Evaluate $\int_C xy \, dx$ from A(1, 0) to B(0, 1) along the curve C that is the portion of $x^2 + y^2 = 1$ in the first quadrant.

Figure 3.15

Portion of circle of Example 3.17.

**Solution**

The curve C is the first quadrant of the unit circle as shown in Figure 3.15. On the curve, $y = \sqrt{1 - x^2}$, so that

$$\int_C xy \, dx = \int_1^0 x \sqrt{1 - x^2} \, dx = \left[-\frac{1}{2} \cdot \frac{2}{3} (1 - x^2)^{3/2} \right]_1^0 = -\frac{1}{3}$$

Example 3.18

Evaluate the integral

$$I = \int_C [(x^2 + 2y) \, dx + (x + y^2) \, dy]$$

from A(0, 1) to B(2, 3) along the curve C defined by $y = x + 1$.

Solution

The curve C is the straight line $y = x + 1$ from the point A(0, 1) to the point B(2, 3). In this case we can eliminate either x or y . Using

$$y = x + 1 \quad \text{and} \quad dy = dx$$

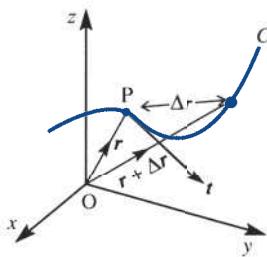
we have, on eliminating y ,

$$\begin{aligned} I &= \int_{x=0}^{x=2} \{ [x^2 + 2(x+1)] \, dx + [x + (x+1)^2] \, dy \} \\ &= \int_0^2 (2x^2 + 5x + 3) \, dx = \left[\frac{2}{3}x^3 + \frac{5}{2}x^2 + 3x \right]_0^2 = \frac{64}{3} \end{aligned}$$

In many practical problems line integrals involving vectors occur. Let $P(\mathbf{r})$ be a point on a curve C in three dimensions, and let \mathbf{t} be the unit tangent vector at P in the sense of the integration (that is, in the sense of increasing arclength s), as indicated in Figure 3.16. Then $\mathbf{t} \, ds$ is the vector element of arc at P, and

$$\mathbf{t} \, ds = \left[\frac{dx}{ds} \mathbf{i} + \frac{dy}{ds} \mathbf{j} + \frac{dz}{ds} \mathbf{k} \right] ds = dx \mathbf{i} + dy \mathbf{j} + dz \mathbf{k} = d\mathbf{r}$$

Figure 3.16
Element of arclength.



If $f_1(x, y, z), f_2(x, y, z)$ and $f_3(x, y, z)$ are the scalar components of a vector field $\mathbf{F}(r)$ then

$$\begin{aligned} & \int_C [f_1(x, y, z) dx + f_2(x, y, z) dy + f_3(x, y, z) dz] \\ &= \int_C \left[f_1(x, y, z) \frac{dx}{ds} ds + f_2(x, y, z) \frac{dy}{ds} ds + f_3(x, y, z) \frac{dz}{ds} ds \right] \\ &= \int_C \mathbf{F} \cdot \mathbf{t} ds = \int_C \mathbf{F} \cdot d\mathbf{r} \end{aligned}$$

Thus, given a vector field $\mathbf{F}(r)$, we can evaluate line integrals of the form $\int_C \mathbf{F} \cdot d\mathbf{r}$. In order to make it clear that we are integrating along a curve, the line integral is sometimes written as $\int_C \mathbf{F} \cdot ds$, where $ds = d\mathbf{r}$ (some authors use $d\mathbf{l}$ instead of ds in order to avoid confusion with dS , the element of surface area). In a similar manner we can evaluate line integrals of the form $\int_C \mathbf{F} \times d\mathbf{r}$.

Example 3.19

Calculate (a) $\int_C \mathbf{F} \cdot d\mathbf{r}$ and (b) $\int_C \mathbf{F} \times d\mathbf{r}$, where C is the part of the spiral $\mathbf{r} = (a \cos \theta, a \sin \theta, a\theta)$ corresponding to $0 \leq \theta \leq \frac{1}{2}\pi$, and $\mathbf{F} = r^2 \mathbf{i}$.

Solution

The curve C is illustrated in Figure 3.17.

(a) Since $\mathbf{r} = a \cos \theta \mathbf{i} + a \sin \theta \mathbf{j} + a\theta \mathbf{k}$,

$$d\mathbf{r} = -a \sin \theta d\theta \mathbf{i} + a \cos \theta d\theta \mathbf{j} + a d\theta \mathbf{k}$$

so that

$$\begin{aligned} \mathbf{F} \cdot d\mathbf{r} &= r^2 \mathbf{i} \cdot (-a \sin \theta d\theta \mathbf{i} + a \cos \theta d\theta \mathbf{j} + a d\theta \mathbf{k}) \\ &= -ar^2 \sin \theta d\theta \\ &= -a^3(\cos^2 \theta + \sin^2 \theta + \theta^2) \sin \theta d\theta = -a^3(1 + \theta^2) \sin \theta d\theta \end{aligned}$$

since $r = |\mathbf{r}| = \sqrt{(a^2 \cos^2 \theta + a^2 \sin^2 \theta + a^2 \theta^2)}$. Thus,

$$\begin{aligned} \int_C \mathbf{F} \cdot d\mathbf{r} &= -a^3 \int_0^{\pi/2} (1 + \theta^2) \sin \theta d\theta \\ &= -a^3 [\cos \theta + 2\theta \sin \theta - \theta^2 \cos \theta]_0^{\pi/2}, \text{ using integration by parts} \\ &= -a^3(\pi - 1) \end{aligned}$$

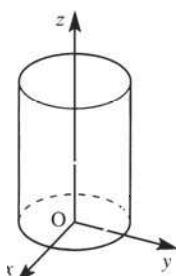


Figure 3.17
The spiral
 $\mathbf{r} = (a \cos \theta, a \sin \theta, a\theta)$ of
Example 3.19.

$$\begin{aligned}
 \text{(b)} \quad \mathbf{F} \times d\mathbf{r} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ r^2 & 0 & 0 \\ -a \sin \theta d\theta & a \cos \theta d\theta & a d\theta \end{vmatrix} \\
 &= -ar^2 d\theta \mathbf{j} + ar^2 \cos \theta d\theta \mathbf{k} \\
 &= -a^3(1 + \theta^2) d\theta \mathbf{j} + a^3(1 + \theta^2) \cos \theta d\theta \mathbf{k}
 \end{aligned}$$

so that

$$\begin{aligned}
 \int_C \mathbf{F} \times d\mathbf{r} &= -ja^3 \int_0^{\pi/2} (1 + \theta^2) d\theta + ka^3 \int_0^{\pi/2} (1 + \theta^2) \cos \theta d\theta \\
 &= -\frac{\pi a^3}{24} (12 + \pi^2) \mathbf{j} + \frac{a^3}{4} (\pi^2 - 4) \mathbf{k}
 \end{aligned}$$

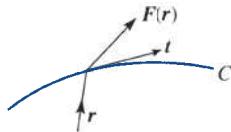


Figure 3.18 Work done by a force \mathbf{F} .

The work done as the point of application of a force \mathbf{F} moves along a given path C as illustrated in Figure 3.18 can be expressed as a line integral. The work done as the point of application moves from $P(\mathbf{r})$ to $P'(\mathbf{r} + d\mathbf{r})$, where $\overrightarrow{PP'} = d\mathbf{r}$, is $dW = |d\mathbf{r}| |\mathbf{F}| \cos \theta = \mathbf{F} \cdot d\mathbf{r}$. Hence the total work done as P goes from A to B is

$$W = \int_C \mathbf{F} \cdot d\mathbf{r}$$

In general, W depends on the path chosen. If, however, $\mathbf{F}(\mathbf{r})$ is such that $\mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$ is an exact differential, say $-dU$, then $W = \int_C -dU = U_A - U_B$, which depends only on A and B and is the same for all paths C joining A and B. Such a force is a **conservative** force, and $U(\mathbf{r})$ is its potential energy, with $\mathbf{F}(\mathbf{r}) = -\text{grad } U$. Forces that do not have this property are said to be **dissipative** or **non-conservative**.

Similarly, if $\mathbf{v}(\mathbf{r})$ represents the velocity field of a fluid then $\oint_C \mathbf{v} \cdot d\mathbf{r}$ is the flow around the closed curve C in unit time. This is sometimes termed the **net circulation integral** of \mathbf{v} . If $\oint_C \mathbf{v} \cdot d\mathbf{r} = 0$ then the fluid is curl-free or irrotational, and in this case \mathbf{v} has a potential function $\phi(\mathbf{r})$ such that $\mathbf{v} = -\text{grad } \phi$.

3.4.2 Exercises

- 56 Evaluate $\int_C y ds$ along the parabola $y = 2\sqrt{x}$ from

$A(3, 2\sqrt{3})$ to $B(24, 4\sqrt{6})$.

Recall: $\left(\frac{ds}{dy} \right)^2 = 1 + \left(\frac{dx}{dy} \right)^2$.

- 57 Evaluate $\int_A^B [2xy dx + (x^2 - y^2) dy]$ along the arc of the circle $x^2 + y^2 = 1$ in the first quadrant from $A(1, 0)$ to $B(0, 1)$.

- 58 Evaluate the integral $\int_C V \cdot d\mathbf{r}$, where $\mathbf{V} = (2yz + 3x^2, y^2 + 4xz, 2z^2 + 6xy)$, and C is the curve with parametric equations $x = t^3$, $y = t^2$, $z = t$ joining the points $(0, 0, 0)$ and $(1, 1, 1)$.

- 59 If $\mathbf{A} = (2y + 3)\mathbf{i} + xz\mathbf{j} + (yz - x)\mathbf{k}$, evaluate $\int_C \mathbf{A} \cdot d\mathbf{r}$ along the following paths C :

- (a) $x = 2t^2$, $y = t$, $z = t^3$ from $t = 0$ to $t = 1$;
- (b) the straight lines from $(0, 0, 0)$ to $(0, 0, 1)$, then to $(0, 1, 1)$ and then to $(2, 1, 1)$;
- (c) the straight line joining $(0, 0, 0)$ to $(2, 1, 1)$.

- 60 Prove that $\mathbf{F} = (y^2 \cos x + z^3)\mathbf{i} + (2y \sin x - 4)\mathbf{j} + (3xz^2 + z)\mathbf{k}$ is a conservative force field. Hence find the work done in moving an object in this field from $(0, 1, -1)$ to $(\pi/2, -1, 2)$.

- 61 Find the work done in moving a particle in the force field $\mathbf{F} = 3x^2\mathbf{i} + (2xz - y)\mathbf{j} + z\mathbf{k}$ along
- the curve defined by $x^2 = 4y$, $3x^3 = 8z$ from $x = 0$ to $x = 2$;
 - the straight line from $(0, 0, 0)$ to $(2, 1, 3)$.
- (c) Does this mean that \mathbf{F} is a conservative force? Give reasons for your answer.
- 62 Prove that the vector field $\mathbf{F} = (3x^2 - y, 2yz^2 - x, 2y^2z)$ is conservative, but not solenoidal. Hence evaluate the scalar line integral $\int_C \mathbf{F} \cdot d\mathbf{r}$ along
- any curve C joining the point $(0, 0, 0)$ to the point $(1, 2, 3)$.
- 63 If $\mathbf{F} = xy\mathbf{i} - z\mathbf{j} + x^2\mathbf{k}$ and C is the curve $x = t^2$, $y = 2t$, $z = t^3$ from $t = 0$ to $t = 1$, evaluate the vector line integral $\int_C \mathbf{F} \times d\mathbf{r}$.
- 64 If $\mathbf{A} = (3x + y, -x, y - z)$ and $\mathbf{B} = (2, -3, 1)$ evaluate the line integral $\oint_C (\mathbf{A} \times \mathbf{B}) \times d\mathbf{r}$ around the circle in the (x, y) plane having centre at the origin and radius 2, traversed in the positive direction.

3.4.3 Double integrals

In the introduction to Section 3.4 we defined the definite integral of a function $f(x)$ of one variable by the limit

$$\int_a^b f(x) dx = \lim_{\substack{n \rightarrow \infty \\ \text{all } \Delta x_i \rightarrow 0}} \sum_{i=1}^n f(\tilde{x}_i) \Delta x_i$$

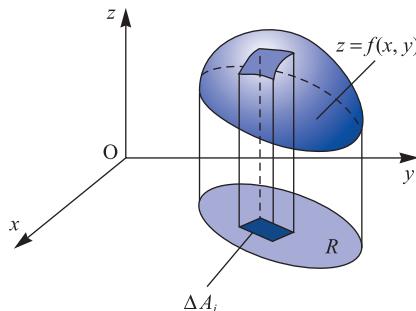
where $a = x_0 < x_1 < x_2 < \dots < x_n = b$, $\Delta x_i = x_i - x_{i-1}$ and $x_{i-1} \leq \tilde{x}_i \leq x_i$. This integral is represented by the area between the curve $y = f(x)$ and the x axis and between $x = a$ and $x = b$, as shown in Figure 3.13.

Now consider $z = f(x, y)$ and a region R of the (x, y) plane, as shown in Figure 3.19. Define the integral of $f(x, y)$ over the region R by the limit

$$\iint_R f(x, y) dA = \lim_{\substack{n \rightarrow \infty \\ \text{all } \Delta A_i \rightarrow 0}} \sum_{i=1}^n f(\tilde{x}_i, \tilde{y}_i) \Delta A_i$$

where ΔA_i ($i = 1, \dots, n$) is a partition of R into n elements of area ΔA_i and $(\tilde{x}_i, \tilde{y}_i)$ is a point in ΔA_i . Now $z = f(x, y)$ represents a surface, and so $f(\tilde{x}_i, \tilde{y}_i) \Delta A_i = \tilde{z}_i \Delta A_i$ is the volume between $z = 0$ and $z = \tilde{z}_i$ on the base ΔA_i . The integral $\iint_R f(x, y) dA$ is the limit of the sum of all such volumes, and so it is the volume under the surface $z = f(x, y)$ above the region R .

Figure 3.19 Volume as an integral.



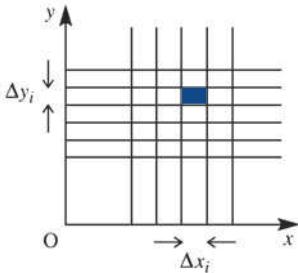


Figure 3.20 A possible grid for the partition of R (rectangular cartesian).

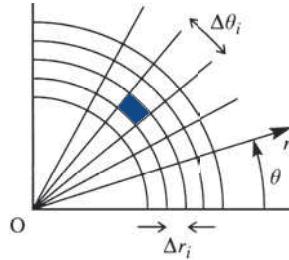


Figure 3.21 Another possible grid for the partition of R (polar).

The partition of R into elementary areas can be achieved using grid lines parallel to the x and y axes as shown in Figure 3.20. Then $\Delta A_i = \Delta x_i \Delta y_i$, and we can write

$$\iint_R f(x, y) dA = \iint_R f(x, y) dx dy = \lim_{n \rightarrow \infty} \sum_{i=1}^n f(\tilde{x}_i, \tilde{y}_i) \Delta x_i \Delta y_i$$

Other partitions may be chosen, for example a polar grid as in Figure 3.21. Then the element of area is $(r_i \Delta \theta_i) \Delta r_i = \Delta A_i$ and

$$\iint_R f(x, y) dA = \iint_R f(r \cos \theta, r \sin \theta) r dr d\theta \quad (3.26)$$

The expression for ΔA is more complicated when the grid lines do not intersect at right angles; we shall discuss this case in Section 3.4.5.

We can evaluate integrals of the type $\iint_R f(x, y) dx dy$ as repeated single integrals in x and y . Consequently, they are usually called **double integrals**.

Consider the region R shown in Figure 3.22, with boundary ACBD. Let the curve ACB be given by $y = g_1(x)$ and the curve ADB by $y = g_2(x)$. Then we can evaluate $\iint_R f(x, y) dx dy$ by summing for y first over the Δy_i , holding x constant ($x = \tilde{x}_i$, say), from $y = g_1(x_i)$ to $y = g_2(x_i)$, and then summing all such strips from A to B; that is, from $x = a$ to $x = b$. Thus we may write

$$\begin{aligned} \iint_R f(x, y) dA &= \lim_{\substack{n_1 \rightarrow \infty \\ \text{all } \Delta x_i, \Delta y_i \rightarrow 0}} \sum_{i=1}^{n_1} \left[\sum_{j=1}^{n_2} f(\tilde{x}_i, y_j) \Delta y_j \right] \Delta x_i \quad (n = \min(n_1, n_2)) \\ &= \int_a^b \left[\int_{y=g_1(x)}^{y=g_2(x)} f(x, y) dy \right] dx \end{aligned}$$

Here the integral inside the brackets is evaluated first, integrating with respect to y , keeping the value of x fixed, and then the result of this integration is integrated with respect to x .

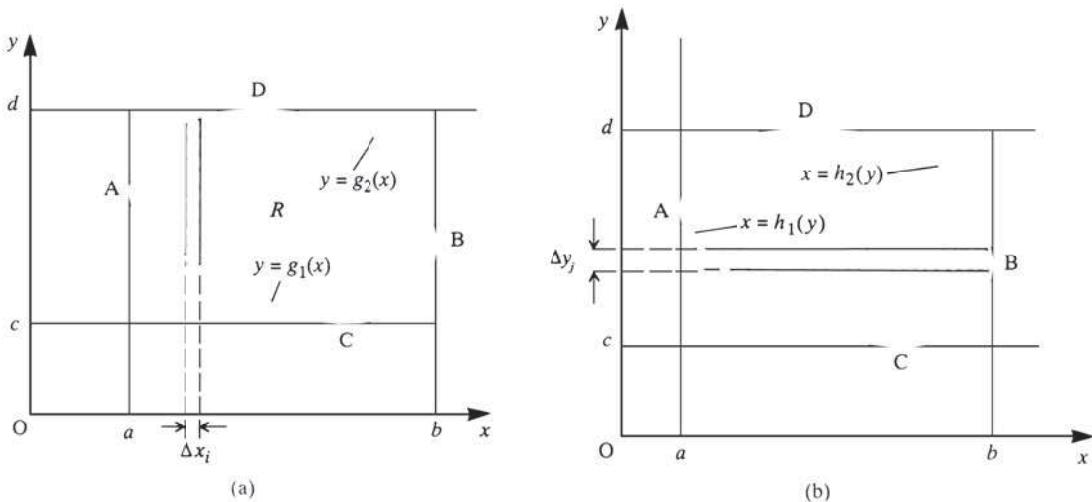


Figure 3.22 The region R .

Alternatively, we can sum for x first and then y . If the curve CAD is represented by $x = h_1(y)$ and the curve CBD by $x = h_2(y)$, we can write the integral as

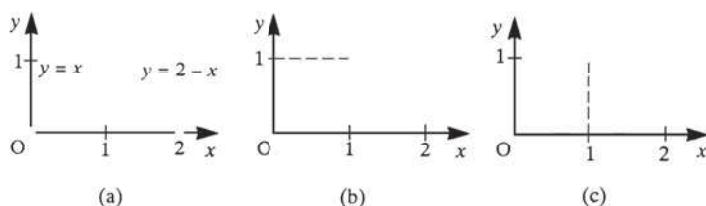
$$\begin{aligned} \iint_S f(x, y) dA &= \lim_{\substack{n \rightarrow \infty \\ \text{all } \Delta y_j, \Delta x_i \rightarrow 0}} \sum_{j=1}^{n_1} \left[\sum_{i=1}^{n_2} f(x_i, \tilde{y}_j) \Delta x_i \right] \Delta y_j \quad (n = \min(n_1, n_2)) \\ &= \int_c^d \left[\int_{x=h_1(y)}^{x=h_2(y)} f(x, y) dx \right] dy \end{aligned}$$

If the double integral exists then these two results are equal, and in going from one to the other we have changed the order of integration. Notice that the limits of integration are also changed in the process. Often, when evaluating an integral analytically, it is easier to perform the evaluation one way rather than the other.

Example 3.20

Evaluate $\iint_R (x^2 + y^2) dA$ over the triangle with vertices at $(0, 0)$, $(2, 0)$ and $(1, 1)$.

Figure 3.23 Domain of integration for Example 3.20.



Solution The domain of integration is shown in Figure 3.23(a). The triangle is bounded by the lines $y = 0$, $y = x$ and $y = 2 - x$.

- (a) Integrating with respect to x first, as indicated in Figure 3.23(b), gives

$$\begin{aligned}\iint_R (x^2 + y^2) dA &= \int_0^1 \int_{x=y}^{x=2-y} (x^2 + y^2) dx dy \\ &= \int_0^1 [\frac{1}{3}x^3 + y^2 x]_{x=y}^{x=2-y} dy \\ &= \int_0^1 [\frac{8}{3} - 4y + 4y^2 - \frac{8}{3}y^3] dy = \frac{4}{3}\end{aligned}$$

- (b) Integrating with respect to y first, as indicated in Figure 3.23(c), gives

$$\iint_R (x^2 + y^2) dA = \int_0^1 \int_{y=0}^{y=x} (x^2 + y^2) dy dx + \int_1^2 \int_{y=0}^{y=2-x} (x^2 + y^2) dy dx$$

Note that because the upper boundary of the region R has different equations for it along different parts, the integral has to be split up into convenient subintegrals. Evaluating the integrals we have

$$\begin{aligned}\int_0^1 \int_{y=0}^{y=x} (x^2 + y^2) dy dx &= \int_0^1 [x^2 y + \frac{1}{3}y^3]_{y=0}^{y=x} dx = \int_0^1 \frac{4}{3}x^3 dx = \frac{1}{3} \\ \int_0^2 \int_{y=0}^{y=2-x} (x^2 + y^2) dy dx &= \int_1^2 [x^2 y + \frac{1}{3}y^3]_{y=0}^{y=2-x} dx \\ &= \int_1^2 (\frac{8}{3} - 4x + 4x^2 - \frac{4}{3}x^3) dx = 1\end{aligned}$$

Thus

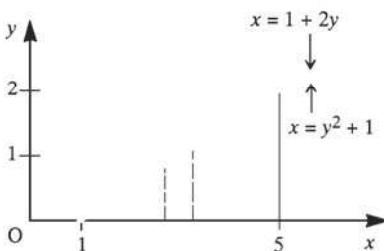
$$\iint_R (x^2 + y^2) dA = \frac{1}{3} + 1 = \frac{4}{3}, \quad \text{as before}$$

Clearly, in this example it is easier to integrate with respect to x first.

Example 3.21

Evaluate $\iint_R (x + 2y)^{-1/2} dA$ over the region $x - 2y \leq 1$ and $x \geq y^2 + 1$.

Figure 3.24 Domain of integration for Example 3.21.



Solution The bounding curves intersect where $2y + 1 = y^2 + 1$, which gives $y = 0$ (with $x = 1$) and $y = 2$ (with $x = 5$). The region R is shown in Figure 3.24. In this example we choose to take x first because the formula for the boundary is easier to deal with: $x = y^2 + 1$ rather than $y = (x - 1)^{1/2}$. Thus we obtain

$$\begin{aligned} \iint_R (x+2y)^{-1/2} dA &= \int_0^2 \int_{y^2+1}^{2y+1} (x+2y)^{-1/2} dx dy \\ &= \int_0^2 [2(x+2y)^{1/2}]_{x=y^2+1}^{x=2y+1} dy \\ &= \int_0^2 [2(4y+1)^{1/2} - 2(y+1)] dy \\ &= [\frac{1}{3}(4y+1)^{3/2} - y^2 - 2y]_0^2 = \frac{2}{3} \end{aligned}$$

As indicated earlier, the evaluation of integrals over a domain R is not restricted to the use of rectangular cartesian coordinates (x, y) . Example 3.22 shows how polar coordinates can be used in some cases to simplify the analytical process.

Example 3.22

Evaluate $\iint_R x^2 y dA$, where R is the region $x^2 + y^2 \leq 1$.

Solution

The fact that the domain of integration is a circle suggests that polar coordinates are a natural choice for the integration process. Then, from (3.26), $x = r \cos \theta$, $y = r \sin \theta$ and $dA = r d\theta dr$, and the integral becomes

$$\begin{aligned} \iint_R x^2 y dA &= \int_{r=0}^1 \int_{\theta=0}^{2\pi} r^2 \cos^2 \theta r \sin \theta r d\theta dr \\ &= \int_{r=0}^1 \int_{\theta=0}^{2\pi} r^4 \cos^2 \theta \sin \theta d\theta dr \end{aligned}$$

Note that in this example the integration is such that we can separate the variables r and θ and write

$$\iint_R x^2 y dA = \int_{r=0}^1 r^4 \int_{\theta=0}^{2\pi} \cos^2 \theta \sin \theta d\theta dr$$

Furthermore, since the limits of integration with respect to θ do not involve r , we can write

$$\iint_R x^2 y dA = \int_{r=0}^1 r^4 dr \int_{\theta=0}^{2\pi} \cos^2 \theta \sin \theta d\theta$$

and the double integral in this case reduces to a product of integrals. Thus we obtain

$$\iint_R x^2 y \, dA = \left[\frac{1}{5} r^5 \right]_0^1 \left[-\frac{1}{3} \cos^3 \theta \right]_0^{2\pi} = 0$$

Reflecting on the nature of the integrand and the domain of integration, this is the result one would anticipate.



There are several ways of evaluating double integrals using MATLAB. The simplest uses the command `dblquad (f, x0, x1, y0, y1)`. For example, consider

$$\int_1^2 \int_0^3 (x^2 + y^2) \, dx \, dy$$

Here we define the integrand as an `inline` function

```
f = inline ('x.^2 + y.^2', 'x', 'y');
```

(Note that `x` is taken as a vector argument.)

```
I = dblquad (f, 1, 2, 0, 3)
```

returns the answer

```
I = 16
```

For non-rectangular domains, the same command is used but the integrand is modified as shown below. Consider

$$\int_0^1 \int_0^x (x^2 + y^2) \, dx \, dy$$

from Example 3.20 (b). Here we define the integrand as the `inline` function

```
f = inline ('(x.^2 + y.^2).* (y-x <= 0)', 'x', 'y');
```

where the logical expression `(y - x <= 0)` returns 1 if the expression is true and 0 otherwise, so that the command

```
I = dblquad (f, 0, 1, 0, 1)
```

returns the required answer

```
I = 0.3333
```

despite integrating over a rectangular domain.

3.4.4 Exercises

65 Evaluate the following:

(a) $\int_0^3 \int_1^2 xy(x+y) \, dy \, dx$

(c) $\int_{-1}^1 \int_{-2}^2 (2x^2 + y^2) \, dy \, dx$

66 Evaluate

$$\iint \frac{x^2}{y} \, dx \, dy$$

over the rectangle bounded by the lines $x = 0$, $x = 2$, $y = 1$ and $y = 2$.

- 67 Evaluate $\iint (x^2 + y^2) \, dx \, dy$ over the region for which $x \geq 0, y \geq 0$ and $x + y \leq 1$.

- 68 Sketch the domain of integration and evaluate

$$(a) \int_1^2 dx \int_x^{2x} \frac{dy}{x^2 + y^2} \quad (b) \int_0^1 dx \int_0^{1-x} (x+y) \, dy$$

$$(c) \int_0^1 dx \int_{\sqrt{x^2 - 1}}^{\sqrt{1-x^2}} \frac{1}{\sqrt{(1-x^2-y^2)}} \, dy$$

- 69 Evaluate $\iint \sin \frac{1}{2}\pi(x+y) \, dx \, dy$ over the triangle whose vertices are $(0, 0), (2, 1), (1, 2)$.

- 70 Sketch the domains of integration of the double integrals

$$(a) \int_0^1 dx \int_x^1 \frac{xy \, dy}{\sqrt{1+y^4}}$$

$$(b) \int_0^{\pi/2} dy \int_0^y (\cos 2y) \sqrt{1-k^2 \sin^2 x} \, dx$$

Change the order of integration, and hence evaluate the integrals.

- 71 Evaluate

$$\int_0^1 dy \int_{\sqrt{y}}^1 \frac{dx}{\sqrt{[y(1+x^2)]}}$$

- 72 Sketch the domain of integration of the double integral

$$\int_0^1 \int_0^{\sqrt{x-x^2}} \frac{x}{\sqrt{(x^2+y^2)}} \, dy \, dx$$

Express the integral in polar coordinates, and hence show that its value is $\frac{1}{3}$.

Sketch the domain of integration of the double integral

$$\int_0^1 dx \int_0^{\sqrt{1-x^2}} \frac{x+y}{\sqrt{(x^2+y^2)}} \, dy$$

and evaluate the integral.

- 74 Evaluate

$$\iint \frac{x+y}{x^2+y^2+a^2} \, dx \, dy$$

over the portion of the first quadrant lying inside the circle $x^2 + y^2 = a^2$.

By using polar coordinates, evaluate the double integral

$$\iint \frac{x^2-y^2}{x^2+y^2} \, dx \, dy$$

over the region in the first quadrant bounded by the arc of the parabola $y^2 = 4(1-x)$ and the coordinate axes.

75 By transforming to polar coordinates, show that the double integral

$$\iint \frac{(x^2+y^2)^2}{(xy)^2} \, dx \, dy$$

taken over the area common to the two circles $x^2 + y^2 = ax$ and $x^2 + y^2 = by$ is ab .

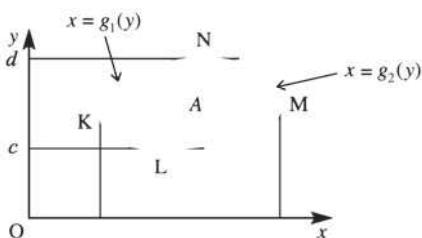
3.4.5 Green's theorem in a plane

This theorem shows the relationship between line integrals and double integrals, and will also provide a justification for the general change of variables in a double integral.

Consider a simple closed curve, C , enclosing the region A as shown in Figure 3.25. If $P(x, y)$ and $Q(x, y)$ are continuous functions with continuous partial derivatives then

$$\oint_C (P \, dx + Q \, dy) = \iint_A \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) \, dx \, dy \quad (3.27)$$

Figure 3.25 Green's theorem.



where C is traversed in the positive sense (that is, so that the bounded area is always on the left). This result is called **Green's theorem** in a plane.

The proof of this result is straightforward. Consider the first term on the right-hand side. Then, with reference to Figure 3.25,

$$\begin{aligned} \iint_R \frac{\partial Q}{\partial x} dx dy &= \int_c^d \left[\int_{g_1(y)}^{g_2(y)} \frac{\partial Q}{\partial x} dx \right] dy \\ &= \int_c^d [Q(g_2(y), y) - Q(g_1(y), y)] dy \\ &= \underset{\text{LMN}}{\int} Q(x, y) dy - \underset{\text{LKN}}{\int} Q(x, y) dy \\ &= \underset{\text{LMNKL}}{\int} Q(x, y) dy = \oint_C Q(x, y) dy \end{aligned}$$

Similarly,

$$-\iint_A \frac{\partial P}{\partial y} dx dy = \oint_C P(x, y) dx$$

and hence

$$\iint_A \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dx dy = \oint_C [P(x, y) dx + Q(x, y) dy]$$

An elementary application is shown in Example 3.23.

Example 3.23

Evaluate $\oint [2x(x + y) dx + (x^2 + xy + y^2) dy]$ around the square with vertices at $(0, 0)$, $(1, 0)$, $(1, 1)$ and $(0, 1)$ illustrated in Figure 3.26.

Solution Here $P(x, y) = 2x(x + y)$ and $Q(x, y) = x^2 + xy + y^2$, so that $\partial P/\partial y = 2x$, $\partial Q/\partial x = 2x + y$ and $\partial Q/\partial x - \partial P/\partial y = y$. Thus the line integral transforms into an easy double integral

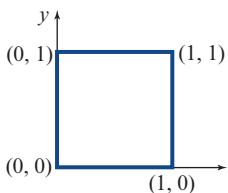


Figure 3.26 Path of integration for Example 3.23.

$$\oint_C [2x(x+y) \, dx + (x^2 + xy + y^2) \, dy] = \iint_A y \, dx \, dy$$

$$= \int_0^1 \int_0^1 y \, dx \, dy$$

$$= \int_0^1 y \, dy \int_0^1 dx = \frac{1}{2}$$

It follows immediately from Green's theorem (3.27) that the area A enclosed by the closed curve C is given by

$$A = \iint_A 1 \, dx \, dy = \oint_C x \, dy = - \oint_C y \, dx = \frac{1}{2} \oint_C (-y \, dx + x \, dy)$$

Suppose that under a transformation of coordinates $x = x(u, v)$ and $y = y(u, v)$, the curve becomes C' , enclosing an area A' . Then

$$\begin{aligned} A' &= \iint_{A'} du \, dv = \oint_{C'} u \, dv = \oint_C u \left(\frac{\partial v}{\partial x} \, dx + \frac{\partial v}{\partial y} \, dy \right) \\ &= \iint_A \left[\frac{\partial}{\partial x} \left(u \frac{\partial v}{\partial y} \right) - \frac{\partial}{\partial y} \left(u \frac{\partial v}{\partial x} \right) \right] \, dx \, dy \\ &= \iint_A \left\{ \left[\frac{\partial u}{\partial x} \frac{\partial v}{\partial y} + u \frac{\partial^2 v}{\partial x \partial y} \right] - \left[\frac{\partial u}{\partial y} \frac{\partial v}{\partial x} + u \frac{\partial^2 v}{\partial y \partial x} \right] \right\} \, dx \, dy \\ &= \iint_A \left(\frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \right) \, dx \, dy \end{aligned}$$

This implies that the element of area $du \, dv$ is equivalent to the element

$$\left| \left(\frac{\partial u}{\partial x} \frac{\partial v}{\partial y} - \frac{\partial u}{\partial y} \frac{\partial v}{\partial x} \right) \right| \, dx \, dy$$

Here the modulus sign is introduced to preserve the orientation of the curve under the mapping. Similarly, we may prove that

$$dx \, dy = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| \, du \, dv \quad (3.28)$$

where $\partial(x, y)/\partial(u, v)$ is the **Jacobian**

$$\frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u} = J(x, y)$$

This enables us to make a general change of coordinates in a double integral:

$$\iint_A f(x, y) dx dy = \iint_{A'} f(x(u, v), y(u, v)) |J| du dv \quad (3.29)$$

where A' is the region in the (u, v) plane corresponding to A in the (x, y) plane.

Note that the above discussion confirms the result

$$\frac{\partial(u, v)}{\partial(x, y)} = \left[\frac{\partial(x, y)}{\partial(u, v)} \right]^{-1}$$

as shown in Section 3.1.3. Using (3.29), the result (3.26) when using polar coordinates is readily confirmed.

Example 3.24

Evaluate $\iint xy dx dy$ over the region in $x \geq 0, y \geq 0$ bounded by $y = x^2 + 4$, $y = x^2$, $y = 6 - x^2$ and $y = 12 - x^2$.

Solution

The domain of integration is shown in Figure 3.27(a). The bounding curves can be rewritten as $y - x^2 = 4$, $y - x^2 = 0$, $y + x^2 = 6$ and $y + x^2 = 12$, so that a natural change of coordinates is to set

$$u = y + x^2, \quad v = y - x^2$$

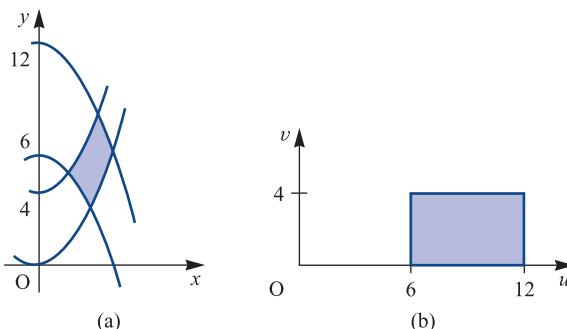
Under this transformation, the region of integration becomes the rectangle $6 \leq u \leq 12$, $0 \leq v \leq 4$, as shown in Figure 3.27(b). Thus since

$$J(x, y) = \frac{\partial(x, y)}{\partial(u, v)} = \left[\frac{\partial(u, v)}{\partial(x, y)} \right]^{-1} = \frac{1}{4x}$$

the integral simplifies to

$$\iint_A xy dx dy = \iint_{A'} xy \frac{1}{4x} du dv$$

Figure 3.27
Domain of integration for Example 3.24:
(a) in the (x, y) plane;
(b) in the (u, v) plane.



Hence

$$\begin{aligned} \iint_A xy \, dx \, dy &= \frac{1}{4} \iint_A y \, du \, dv = \frac{1}{8} \iint_A (u+v) \, du \, dv, \quad \text{since } y = (u+v)/2 \\ &= \frac{1}{8} \int_0^4 dv \int_6^{12} (u+v) \, du = 33 \end{aligned}$$

We remark in passing that Green's theorem in a plane may be generalized to three dimensions. Note that the result (3.27) may be written as

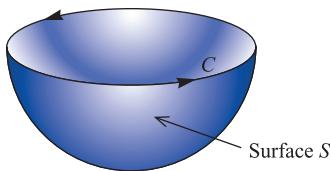
$$\oint_C (P, Q, 0) \cdot d\mathbf{r} = \iint_A \operatorname{curl} [(P, Q, 0)] \cdot \mathbf{k} \, dx \, dy$$

For a general surface S with bounding curve C as shown in Figure 3.28 this identity becomes

$$\oint_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \iint_S \operatorname{curl} \mathbf{F}(\mathbf{r}) \cdot d\mathbf{S}$$

where $d\mathbf{S} = \hat{\mathbf{n}} dS$ is the vector element of surface area and $\hat{\mathbf{n}}$ is a unit vector along the normal. This generalization is called Stokes' theorem, and will be discussed in Section 3.4.12 after we have formally introduced the concept of a surface integral.

Figure 3.28
Three-dimensional
generalization of
Green's theorem.



3.4.6 Exercises

77 Evaluate the line integral

$$\oint_C [\sin y \, dx + (x - \cos y) \, dy]$$

taken in the anticlockwise sense, where C is the perimeter of the triangle formed by the lines

$$y = \frac{1}{2}\pi x, \quad y = \frac{1}{2}\pi, \quad x = 0$$

Verify your answer using Green's theorem in a plane.

78 Use Green's theorem in a plane to evaluate

$$\oint_C [(xy^2 - y) \, dx + (x + y^2) \, dy]$$

as a double integral, where C is the triangle with vertices at $(0, 0)$, $(2, 0)$ and $(2, 2)$ and is traversed in the anticlockwise direction.

79 Evaluate the line integral

$$I = \oint_C (xy \, dx + x \, dy)$$

where C is the closed curve consisting of $y = x^2$ from $x = 0$ to $x = 1$ and $y = \sqrt{x}$ from $x = 1$ to $x = 0$. Confirm your answer by applying Green's theorem in the plane and evaluating I as a double integral.

80 Use Green's theorem in a plane to evaluate the line integral

$$\oint_C [(e^x - 3y^2) \, dx + (e^y + 4x^2) \, dy]$$

where C is the circle $x^2 + y^2 = 4$. (Hint: Use polar coordinates to evaluate the double integral.)

81 Evaluate

$$\int_0^a dx \int_x^{2a-x} \frac{y-x}{4a^2 + (y+x)^2} \, dy$$

using the transformation of coordinates $u = x + y$, $v = x - y$.

82 Using the transformation

$$x + y = u, \quad \frac{y}{x} = v$$

show that

$$\int_0^1 dy \int_y^{2-y} \frac{x+y}{x^2} e^{x+y} \, dx = \int_0^2 du \int_0^1 e^u \, dv = e^2 - 1$$

3.4.7 Surface Integrals

The extensions of the idea of an integral to line and double integrals are not the only generalizations that can be made. We can also extend the idea to integration over a general surface S . Two types of such integrals occur:

$$(a) \iint_S f(x, y, z) \, dS$$

$$(b) \iint_S \mathbf{F}(\mathbf{r}) \cdot \hat{\mathbf{n}} \, dS = \iint_S \mathbf{F}(\mathbf{r}) \cdot d\mathbf{S}$$

In case (a) we have a scalar field $f(\mathbf{r})$ and in case (b) a vector field $\mathbf{F}(\mathbf{r})$. Note that $d\mathbf{S} = \hat{\mathbf{n}} \, dS$ is the vector element of area, where $\hat{\mathbf{n}}$ is the unit outward-drawn normal vector to the element dS .

In general, the surface S can be described in terms of two parameters, u and v say, so that on S

$$\mathbf{r} = \mathbf{r}(u, v) = (x(u, v), y(u, v), z(u, v))$$

The surface S can be specified by a scalar point function $C(\mathbf{r}) = c$, where c is a constant. Curves may be drawn on that surface, and in particular if we fix the value of one of the two parameters u and v then we obtain two families of curves. On one, $C_u(\mathbf{r}(u, v_0))$, the value of u varies while v is fixed, and on the other, $C_v(\mathbf{r}(u_0, v))$, the value of v varies while u is fixed, as shown in Figure 3.29. Then as indicated in Figure 3.29, the vector element of area $d\mathbf{S}$ is given by

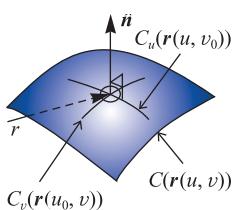


Figure 3.29 Parametric curves on a surface.

$$\begin{aligned} dS &= \frac{\partial \mathbf{r}}{\partial u} du \times \frac{\partial \mathbf{r}}{\partial v} dv = \frac{\partial \mathbf{r}}{\partial u} \times \frac{\partial \mathbf{r}}{\partial v} du dv \\ &= \left(\frac{\partial x}{\partial u}, \frac{\partial y}{\partial u}, \frac{\partial z}{\partial u} \right) \times \left(\frac{\partial x}{\partial v}, \frac{\partial y}{\partial v}, \frac{\partial z}{\partial v} \right) du dv = (J_1 \mathbf{i} + J_2 \mathbf{j} + J_3 \mathbf{k}) du dv \end{aligned}$$

where

$$J_1 = \frac{\partial y}{\partial u} \frac{\partial z}{\partial v} - \frac{\partial y}{\partial v} \frac{\partial z}{\partial u}, \quad J_2 = \frac{\partial z}{\partial u} \frac{\partial x}{\partial v} - \frac{\partial z}{\partial v} \frac{\partial x}{\partial u}, \quad J_3 = \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u} \quad (3.30)$$

Hence

$$\begin{aligned} \iint_S \mathbf{F}(\mathbf{r}) \cdot d\mathbf{S} &= \iint_A (PJ_1 + QJ_2 + RJ_3) du dv \\ \iint_S f(x, y, z) dS &= \iint_A f(u, v) \sqrt{(J_1^2 + J_2^2 + J_3^2)} du dv \end{aligned}$$

where $\mathbf{F}(\mathbf{r}) = (P, Q, R)$ and A is the region of the (u, v) plane corresponding to S . Here, of course, the terms in the integrands have to be expressed in terms of u and v .

In particular, u and v can be chosen as any two of x, y and z . For example, if $z = z(x, y)$ describes a surface as in Figure 3.30 then

$$\mathbf{r} = (x, y, z(x, y))$$

with x and y as independent variables. This gives

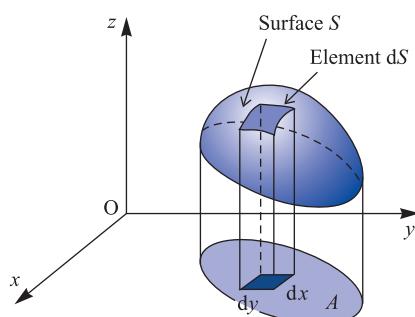
$$J_1 = -\frac{\partial z}{\partial x}, \quad J_2 = -\frac{\partial z}{\partial y}, \quad J_3 = 1$$

and

$$\iint_S \mathbf{F}(\mathbf{r}) \cdot d\mathbf{S} = \iint_A \left(-P \frac{\partial z}{\partial x} - Q \frac{\partial z}{\partial y} + R \right) dx dy \quad (3.31a)$$

$$\iint_S f(x, y, z) dS = \iint_A f(x, y, z(x, y)) \sqrt{1 + \left(\frac{\partial z}{\partial x} \right)^2 + \left(\frac{\partial z}{\partial y} \right)^2} dx dy \quad (3.31b)$$

Figure 3.30 A surface described by $z = z(x, y)$.



Example 3.25

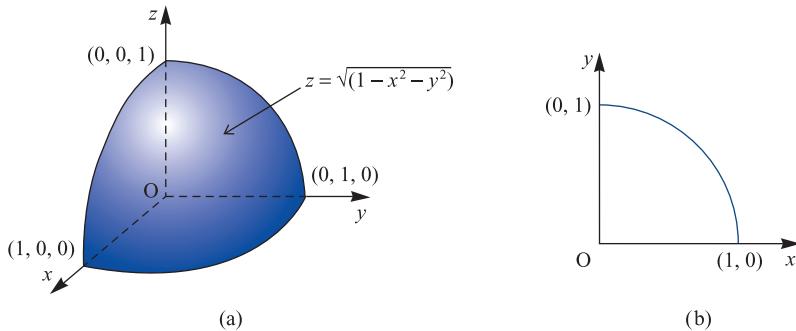
Evaluate the surface integral

$$\iint_S (x + y + z) \, dS$$

where S is the portion of the sphere $x^2 + y^2 + z^2 = 1$ that lies in the first quadrant.

Figure 3.31

(a) Surface S for Example 3.25;
 (b) quadrant of a circle in the (x, y) plane.



Solution The surface S is illustrated in Figure 3.31(a). Taking

$$z = \sqrt{1 - x^2 - y^2}$$

we have

$$\frac{\partial z}{\partial x} = \frac{-x}{\sqrt{1 - x^2 - y^2}}, \quad \frac{\partial z}{\partial y} = \frac{-y}{\sqrt{1 - x^2 - y^2}}$$

giving

$$\begin{aligned} \sqrt{1 + \left(\frac{\partial z}{\partial x}\right)^2 + \left(\frac{\partial z}{\partial y}\right)^2} &= \sqrt{\frac{x^2 + y^2 + (1 - x^2 - y^2)}{(1 - x^2 - y^2)}} \\ &= \frac{1}{\sqrt{1 - x^2 - y^2}} \end{aligned}$$

Using (3.31) then gives

$$\iint_S (x + y + z) \, dS = \iint_A [x + y + \sqrt{1 - x^2 - y^2}] \frac{1}{\sqrt{1 - x^2 - y^2}} \, dx \, dy$$

where A is the quadrant of a circle in the (x, y) plane illustrated in Figure 3.31(b).

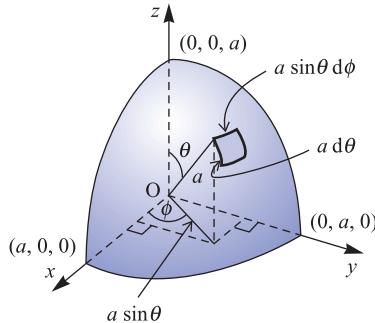
Thus

$$\begin{aligned}
 \iint_S (x + y + z) dS &= \int_0^1 dx \int_0^{\sqrt{1-x^2}} \left[\frac{x}{\sqrt{1-x^2-y^2}} + \frac{y}{\sqrt{1-x^2-y^2}} + 1 \right] dy \\
 &= \int_0^1 \left[x \sin^{-1} \left(\frac{y}{\sqrt{1-x^2}} \right) - \sqrt{1-x^2-y^2} + y \right]_0^{\sqrt{1-x^2}} dx \\
 &= \int_0^1 \left[\frac{\pi}{2}x + 2\sqrt{1-x^2} \right] dx \\
 &= \left[\frac{\pi}{4}x^2 + x\sqrt{1-x^2} + \sin^{-1}x \right]_0^1 \\
 &= \frac{3}{4}\pi
 \end{aligned}$$

An alternative approach to evaluating the surface integral in Example 3.25 is to evaluate it directly over the surface of the sphere using spherical polar coordinates. As illustrated in Figure 3.32, on the surface of a sphere of radius a we have

$$\begin{aligned}
 x &= a \sin \theta \cos \phi, & y &= a \sin \theta \sin \phi \\
 z &= a \cos \theta, & dS &= a^2 \sin \theta d\theta d\phi
 \end{aligned}$$

Figure 3.32 Surface element in spherical polar coordinates.



In the sphere of Example 3.25 the radius $a = 1$, so that

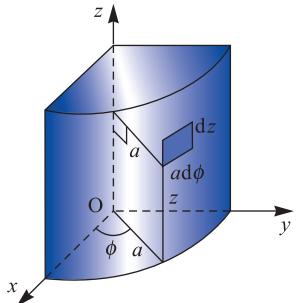
$$\begin{aligned}
 \iint_S (x + y + z) dS &= \int_0^{\pi/2} \int_0^{\pi/2} (\sin \theta \cos \phi + \sin \theta \sin \phi + \cos \theta) \sin \theta d\theta d\phi \\
 &= \int_0^{\pi/2} \left[\frac{1}{4}\pi \cos \phi + \frac{1}{4}\pi \sin \phi + \frac{1}{2} \right] d\phi = \frac{3}{4}\pi
 \end{aligned}$$

as determined in Example 3.25.

In a similar manner, when evaluating surface integrals over the surface of a cylinder of radius a , we have, as illustrated in Figure 3.33,

$$x = a \cos \phi, \quad y = a \sin \phi, \quad z = z, \quad dS = a dz d\phi$$

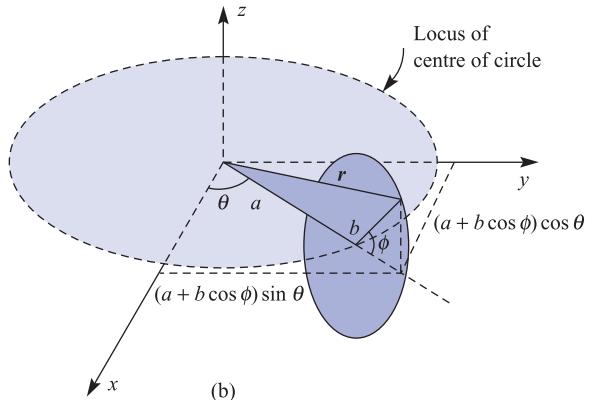
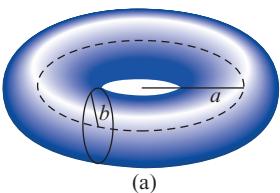
Figure 3.33
Surface element in cylindrical polar coordinates.



Example 3.26

Find the surface area of the torus shown in Figure 3.34(a) formed by rotating a circle of radius b about an axis distance a from its centre.

Figure 3.34 (a) Torus of Example 3.26;
(b) position vector of a point on the surface of the torus.



Solution From Figure 3.34(b), the position vector \mathbf{r} of a point on the surface is given by

$$\mathbf{r} = (a + b \cos \phi) \cos \theta \mathbf{i} + (a + b \cos \phi) \sin \theta \mathbf{j} + b \sin \phi \mathbf{k}$$

(Notice that θ and ϕ are not the angles used for spherical polar coordinates.) Thus using (3.30),

$$J_1 = (a + b \cos \phi) \cos \theta (b \cos \phi) - (-b \sin \phi \sin \theta)(0)$$

$$J_2 = (0)(-b \sin \phi \cos \theta) - (b \cos \phi)(a + b \cos \phi)(-\sin \theta)$$

$$J_3 = -(a + b \cos \phi) \sin \theta (-b \sin \phi \sin \theta) - (-b \sin \phi \cos \theta)(a + b \cos \phi) \cos \theta$$

Simplifying, we obtain

$$J_1 = b(a + b \cos \phi) \cos \theta \cos \phi$$

$$J_2 = b(a + b \cos \phi) \sin \theta \cos \phi$$

$$J_3 = b(a + b \cos \phi) \sin \phi$$

and the surface area is given by

$$\begin{aligned} S &= \int_0^{2\pi} \int_0^{2\pi} \sqrt{(J_1^2 + J_2^2 + J_3^2)} d\theta d\phi \\ &= \int_0^{2\pi} \int_0^{2\pi} b(a + b \cos \phi) d\theta d\phi \\ &= 4\pi^2 ab \end{aligned}$$

Thus the surface area of the torus is the product of the circumferences of the two circles that generate it.

Example 3.27

Evaluate $\iint_S V \cdot dS$, where $V = z\mathbf{i} + x\mathbf{j} - 3y^2z\mathbf{k}$ and S is the surface of the cylinder $x^2 + y^2 = 16$ in the first octant between $z = 0$ and $z = 5$.

Solution

The surface S is illustrated in Figure 3.35. From Section 3.2.1, the outward normal to the surface is in the direction of the vector

$$\mathbf{n} = \text{grad}(x^2 + y^2 - 16) = 2x\mathbf{i} + 2y\mathbf{j}$$

so that the unit outward normal $\hat{\mathbf{n}}$ is given by

$$\hat{\mathbf{n}} = \frac{2x\mathbf{i} + 2y\mathbf{j}}{2\sqrt{x^2 + y^2}}$$

Hence on the surface $x^2 + y^2 = 16$,

$$\hat{\mathbf{n}} = \frac{1}{4}(x\mathbf{i} + y\mathbf{j})$$

giving

$$dS = dS \hat{\mathbf{n}} = \frac{1}{4} dS(x\mathbf{i} + y\mathbf{j})$$

Projecting the element of surface dS onto the (x, z) plane as illustrated in Figure 3.35, the area $dx dz$ of the projected element is given by

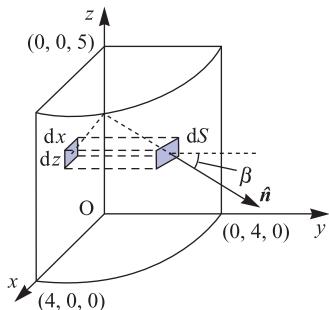
$$dx dz = dS \cos \beta$$

where β is the angle between the normal $\hat{\mathbf{n}}$ to the surface element and the normal \mathbf{j} to the (x, z) plane. Thus

$$dx dz = dS |\hat{\mathbf{n}} \cdot \mathbf{j}| = \frac{1}{4} dS |(x\mathbf{i} + y\mathbf{j}) \cdot \mathbf{j}| = \frac{1}{4} dS y$$

Figure 3.35

Surface S for Example 3.26.



giving

$$dS = \frac{4}{y} dx dz$$

Also,

$$\mathbf{V} \cdot d\mathbf{S} = \mathbf{V} \cdot \hat{\mathbf{n}} dS = (zi + xj - 3y^2zk) \cdot \left(\frac{x\mathbf{i} + y\mathbf{j}}{4}\right) \frac{4}{y} dx dz = \frac{xz + xy}{y} dx dz$$

so that

$$\iint_S \mathbf{V} \cdot d\mathbf{S} = \iint_A \frac{xz + xy}{y} dx dz$$

where A is the rectangular region in the (x, z) plane bounded by $0 \leq x \leq 4$, $0 \leq z \leq 5$. Noting that the integrand is still evaluated on the surface, we can write $y = \sqrt{(16 - x^2)}$, so that

$$\begin{aligned} \iint_S \mathbf{V} \cdot d\mathbf{S} &= \int_0^4 \int_0^5 \left[x + \frac{xz}{\sqrt{(16 - x^2)}} \right] dz dx \\ &= \int_0^4 \left[xz + \frac{xz^2}{2\sqrt{(16 - x^2)}} \right]_0^5 dx \\ &= \int_0^4 \left[5x + \frac{25x}{2\sqrt{(16 - x^2)}} \right] dx \\ &= \left[\frac{5}{2}x^2 - \frac{25}{2}\sqrt{(16 - x^2)} \right]_0^4 \\ &= 90 \end{aligned}$$

An alternative approach in this case is to evaluate $\frac{1}{4} \iint_s (xz + xy) dS$ directly over the surface using cylindrical polar coordinates. This is left as Exercise 90, in Exercises 3.4.8.

3.4.8 Exercises

- 83 Evaluate the area of the surface $z = 2 - x^2 - y^2$ lying above the (x, y) plane. (Hint: Use polar coordinates to evaluate the double integral.)
- 84 Evaluate
- $\iint_S (x^2 + y^2) dS$, where S is the surface area of the plane $2x + y + 2z = 6$ cut off by the planes $z = 0, z = 2, y = 0, y = 3$;
 - $\iint_S z dS$, where S is the surface area of the hemisphere $x^2 + y^2 + z^2 = 1 (z > 0)$ cut off by the cylinder $x^2 + y^2 = 0$.
- 85 Evaluate $\iint_S \mathbf{v} \cdot d\mathbf{S}$, where
- $\mathbf{v} = (xy, -x^2, x + z)$ and S is the part of the plane $2x + 2y + z = 6$ included in the first octant;
 - $\mathbf{v} = (3y, 2x^2, z^3)$ and S is the surface of the cylinder $x^2 + y^2 = 1, 0 < z < 1$.
- 86 Show that $\iint_S z^2 dS = \frac{2}{3}\pi$, where S is the surface of the sphere $x^2 + y^2 + z^2 = 1, z \geq 0$.
- 87 Evaluate the surface integral $\iint_S U(x, y, z) dS$, where S is the surface of the paraboloid $z = 2 - (x^2 + y^2)$ above the (x, y) plane and $U(x, y, z)$ is given by
- 1
 - $x^2 + y^2$
 - z
- Give a physical interpretation in each case.
- 88 Determine the surface area of the plane $2x + y + 2z = 16$ cut off by $x = 0, y = 0$ and $x^2 + y^2 = 64$.
- 89 Show that the area of that portion of the surface of the paraboloid $x^2 + y^2 = 4z$ included between the planes $z = 1$ and $z = 3$ is $\frac{16}{3}\pi(4 - \sqrt{2})$.
- 90 Evaluate the surface integral in Example 3.27 using cylindrical polar coordinates.
- 91 If $\mathbf{F} = yi + (x - 2xz)\mathbf{j} - xy\mathbf{k}$, evaluate the surface integral $\iint_S (\operatorname{curl} \mathbf{F}) \cdot d\mathbf{S}$, where S is the surface of the sphere $x^2 + y^2 + z^2 = a^2, z \geq 0$.

3.4.9 Volume integrals

In Section 3.4.7 we defined the integral of a function over a curved surface in three dimensions. This idea can be extended to define the integral of a function of three variables through a region T of three-dimensional space by the limit

$$\iiint_T f(x, y, z) dV = \lim_{\substack{n \rightarrow \infty \\ \text{all } \Delta V_i \rightarrow 0}} \sum_{i=1}^n f(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i) \Delta V_i$$

where $\Delta V_i (i = 1, \dots, n)$ is a partition of T into n elements of volume, and $(\tilde{x}_i, \tilde{y}_i, \tilde{z}_i)$ is a point in ΔV_i as illustrated in Figure 3.36.

In terms of rectangular cartesian coordinates the triple integral can, as illustrated in Figure 3.37, be written as

$$\iiint_T f(x, y, z) dV = \int_a^b dx \int_{g_1(x)}^{g_2(x)} dy \int_{h_1(x, y)}^{h_2(x, y)} f(x, y, z) dz \quad (3.32)$$

Note that there are six different orders in which the integration in (3.32) can be carried out.

As we saw for double integrals in (3.28), the expression for the element of volume $dV = dx dy dz$ under the transformation $x = x(u, v, w), y = y(u, v, w), z = z(u, v, w)$ may be obtained using the Jacobian

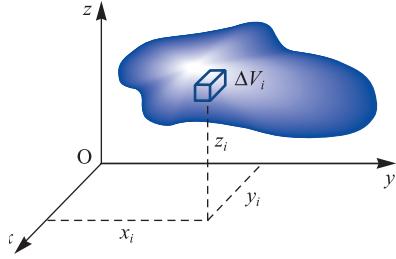


Figure 3.36 Partition of region T into volume elements ΔV_i .

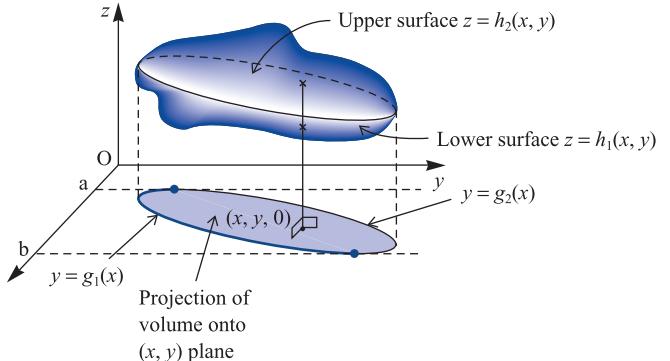


Figure 3.37 The volume integral in terms of rectangular cartesian coordinates.

$$J = \frac{\partial(x, y, z)}{\partial(u, v, w)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} & \frac{\partial z}{\partial v} \\ \frac{\partial x}{\partial w} & \frac{\partial y}{\partial w} & \frac{\partial z}{\partial w} \end{vmatrix}$$

as

$$dV = dx dy dz = |J| du dv dw \quad (3.33)$$

For example, in the case of cylindrical polar coordinates

$$x = \rho \cos \phi, \quad y = \rho \sin \phi, \quad z = z$$

$$J = \rho \begin{vmatrix} \cos \phi & \sin \phi & 0 \\ -\sin \phi & \cos \phi & 0 \\ 0 & 0 & 1 \end{vmatrix} = \rho$$

so that

$$dV = \rho d\rho d\phi dz \quad (3.34)$$

a result illustrated in Figure 3.38.

Similarly, for spherical polar coordinates (r, θ, ϕ)

$$x = r \sin \theta \cos \phi, \quad y = r \sin \theta \sin \phi, \quad z = r \cos \theta$$

$$J = \begin{vmatrix} \sin \theta \cos \phi & \sin \theta \sin \phi & \cos \theta \\ r \cos \theta \cos \phi & r \cos \theta \sin \phi & -r \sin \theta \\ -r \sin \theta \sin \phi & r \sin \theta \cos \phi & 0 \end{vmatrix} = r^2 \sin \theta$$

so that

$$dV = r^2 \sin \theta dr d\theta d\phi \quad (3.35)$$

a result illustrated in Figure 3.39.

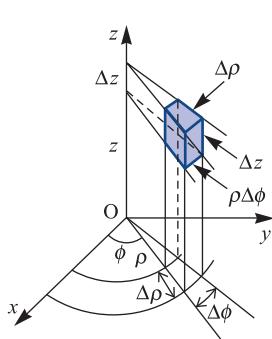


Figure 3.38 Volume element in cylindrical polar coordinates.

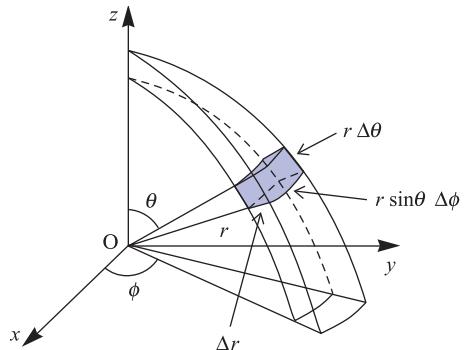


Figure 3.39 Volume element in spherical polar coordinates.

Example 3.28

Find the volume and the coordinates of the centroid of the tetrahedron defined by $x \geq 0$, $y \geq 0$, $z \geq 0$ and $x + y + z \leq 1$.

Solution

The tetrahedron is shown in Figure 3.40. Its volume is

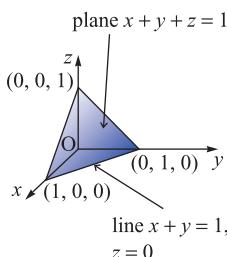


Figure 3.40
Tetrahedron for Example 3.28.

$$\begin{aligned} V &= \iiint_{\text{tetrahedron}} dx dy dz = \int_{x=0}^{x=1} dx \int_{y=0}^{y=1-x} dy \int_{z=0}^{z=1-x-y} dz \\ &= \int_0^1 dx \int_0^{1-x} (1-x-y) dy = \int_0^1 \frac{1}{2}(1-x)^2 dx = \frac{1}{6} \end{aligned}$$

Let the coordinates of the centroid be $(\bar{x}, \bar{y}, \bar{z})$; then, taking moments about the line $x = 0$, $z = \bar{z}$,

$$\begin{aligned} \bar{x}V &= \iiint_{\text{tetrahedron}} x dV = \iiint_{\text{tetrahedron}} x dx dy dz \\ &= \int_0^1 dx \int_0^{1-x} dy \int_0^{1-x-y} x dz = \int_0^1 \frac{1}{2}x(1-x)^2 dx = \frac{1}{24} \end{aligned}$$

Hence $\bar{x} = \frac{1}{4}$, and by symmetry $\bar{y} = \bar{z} = \frac{1}{4}$.

Example 3.29

Find the moment of inertia of a uniform sphere of mass M and radius a about a diameter.

Solution

A sphere of radius a has volume $4\pi a^3/3$, so that its density is $3M/4\pi a^3$. Then the moment of inertia of the sphere about the z axis is

$$I = \frac{3M}{4\pi a^3} \iiint_{\text{sphere}} (x^2 + y^2) dx dy dz$$

In this example it is natural to use spherical polar coordinates, so that

$$\begin{aligned} I &= \frac{3M}{4\pi a^3} \iiint_{\text{sphere}} (r^2 \sin^2 \theta) r^2 \sin \theta \, dr \, d\theta \, d\phi \\ &= \frac{3M}{4\pi a^3} \int_0^a r^4 \, dr \int_0^\pi \sin^3 \theta \, d\theta \int_0^{2\pi} \, d\phi = \frac{3M}{4\pi a^3} (\frac{1}{5}a^5)(\frac{4}{3})(2\pi) \\ &= \frac{2}{5} Ma^2 \end{aligned}$$



Evaluating triple integrals using MATLAB uses the command `triplequad`. For example, consider (see Example 3.28):

$$\int_0^1 \int_0^{1-x} \int_0^{1-x-y} x \, dx \, dy \, dz$$

Here we write the integrand as the `inline` function

```
F = inline ('x.* ( x + y + z <=1 ) ', 'x', 'y', 'z');
```

so that the command

```
I = triplequad (f , 0 , 1 , 0 , 1, 0 , 1)
```

returns the answer

```
I = 0.0416
```

This procedure could be slow because of the large number of points at which the integrand is evaluated.

3.4.10 Exercises

- 92 Evaluate the triple integrals

(a) $\int_0^1 dx \int_0^2 dy \int_0^3 x^2 yz \, dz$

(b) $\int_0^2 \int_1^3 \int_2^4 xyz^2 \, dz \, dy \, dx$

- 93 Show that

$$\int_{-1}^1 dz \int_0^z dx \int_{x-z}^{x+z} (x + y + z) \, dy = 0$$

- 94 Evaluate $\iiint_V \sin(x + y + z) \, dx \, dy \, dz$ over the portion of the positive octant cut off by the plane $x + y + z = \pi$.

- 95 Evaluate $\iiint_V xyz \, dx \, dy \, dz$, where V is the region bounded by the planes $x = 0$, $y = 0$, $z = 0$ and $x + y + z = 1$.

- 96 Sketch the region contained between the parabolic cylinders $y = x^2$ and $x = y^2$ and the planes $z = 0$ and $x + y + z = 2$. Show that the volume of the region may be expressed as the triple integral

$$\int_0^1 \int_{x^2}^{\sqrt{x}} \int_0^{2-x-y} dz \, dy \, dx$$

and evaluate it.

- 97 Use spherical polar coordinates to evaluate

$$\iiint_V x(x^2 + y^2 + z^2) \, dx \, dy \, dz$$

where V is the region in the first octant lying within the sphere $x^2 + y^2 + z^2 = 1$.

- 98 Evaluate $\iiint_V x^2y^2z^2(x+y+z) \, dx \, dy \, dz$ throughout the region defined by $x+y+z \leq 1$, $x \geq 0$, $y \geq 0$, $z \geq 0$.

- 99 Show that if $x+y+z = u$, $y+z = uv$ and $z = uvw$ then

$$\frac{\partial(x, y, z)}{\partial(u, v, w)} = u^2v$$

Hence evaluate the triple integral

$$\iiint_V \exp[-(x+y+z)^3] \, dx \, dy \, dz$$

where V is the volume of the tetrahedron bounded by the planes $x = 0$, $y = 0$, $z = 0$ and $x + y + z = 1$.

- 100 Evaluate $\iiint_V yz \, dx \, dy \, dz$ taken throughout the prism with sides parallel to the z axis, whose base is the triangle with vertices at $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$ and whose top is the triangle with vertices at $(0, 0, 2)$, $(1, 0, 1)$, $(0, 1, 1)$. Find also the position of the centroid of this prism.

- 101 Evaluate $\iiint_V z \, dx \, dy \, dz$ throughout the region defined by $x^2 + y^2 \leq z^2$, $x^2 + y^2 + z^2 \leq 1$, $z > 0$.

- 102 Using spherical polar coordinates, evaluate $\iiint_V x \, dx \, dy \, dz$ throughout the positive octant of the sphere $x^2 + y^2 + z^2 = a^2$.

3.4.11 Gauss's divergence theorem

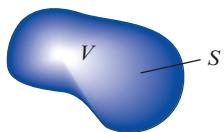


Figure 3.41
Closed volume V with surface S .

In the same way that Green's theorem relates surface and line integrals, Gauss's theorem relates surface and volume integrals.

Consider the closed volume V with surface area S shown in Figure 3.41. The surface integral $\iint_S \mathbf{F} \cdot d\mathbf{S}$ may be interpreted as the flow of a liquid with velocity field $\mathbf{F}(r)$ out of the volume V . In Section 3.3.1 we saw that the divergence of \mathbf{F} could be expressed as

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \lim_{\Delta V \rightarrow 0} \frac{\text{flow out of } \Delta V}{\Delta V}$$

In terms of differentials, this may be written

$$\operatorname{div} \mathbf{F} \, dV = \text{flow out of } dV$$

Consider now a partition of the volume V given by ΔV_i ($i = 1, \dots, n$). Then the total flow out of V is the sum of the flows out of each ΔV_i . That is,

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \lim_{n \rightarrow \infty} \sum_{i=1}^n (\text{flow out of } \Delta V_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^n (\operatorname{div} \mathbf{F} \Delta V_i)$$

giving

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \iiint_V \operatorname{div} \mathbf{F} \, dV \quad (3.36)$$

This result is known as the **divergence theorem** or **Gauss's theorem**. It enables us to convert surface integrals into volume integrals, and often simplifies their evaluation.

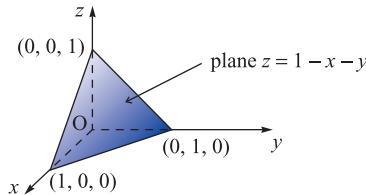
Example 3.30

A vector field $\mathbf{F}(r)$ is given by

$$\mathbf{F}(r) = x^3y\mathbf{i} + x^2y^2\mathbf{j} + x^2yz\mathbf{k}$$

Find $\iint_S \mathbf{F} \cdot d\mathbf{S}$, where S is the surface of the region in the first octant for which $x + y + z \leq 1$.

Figure 3.42 Region V and surface S for Example 3.30.

**Solution**

We begin by sketching the region V enclosed by S , as shown in Figure 3.42. It is clear that evaluating the surface integral directly will be rather clumsy, involving four separate integrals (one over each of the four surfaces). It is simpler in this case to transform it into a volume integral using the divergence theorem (3.36):

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \iiint_V \operatorname{div} \mathbf{F} dV$$

Here

$$\operatorname{div} \mathbf{F} = 3x^2y + 2x^2y + x^2y = 6x^2y$$

and we obtain

$$\begin{aligned} \iint_S \mathbf{F} \cdot d\mathbf{S} &= \int_0^1 dx \int_0^{1-x} dy \int_0^{1-x-y} 6x^2y dz \\ &= 6 \int_0^1 x^2 dx \int_0^{1-x} y dy \int_0^{1-x-y} dz \\ &= 6 \int_0^1 x^2 dx \int_0^{1-x} [(1-x)y - y^2] dy \\ &= \int_0^1 x^2(1-x)^3 dx = \frac{1}{60} \end{aligned} \quad (\text{see Example 3.28})$$

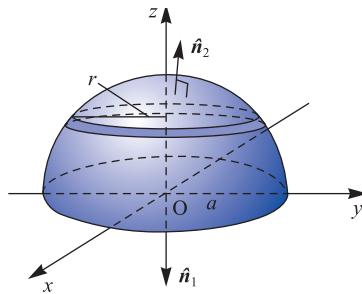
Example 3.31

Verify the divergence theorem

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \iiint_V \operatorname{div} \mathbf{F} dV$$

when $\mathbf{F} = 2xz\mathbf{i} + yz\mathbf{j} + z^2\mathbf{k}$ and V is the volume enclosed by the upper hemisphere $x^2 + y^2 + z^2 = a^2, z \geq 0$.

Figure 3.43
Hemisphere for Example 3.31.



Solution The volume V and surface S of the hemisphere are illustrated in Figure 3.43. Note that since the theorem relates to a closed volume, the surface S consists of the flat circular base in the (x, y) plane as well as the hemispherical surface. In this case

$$\operatorname{div} \mathbf{F} = 2z + z + 2z = 5z$$

so that the volume integral is readily evaluated as

$$\iiint_V 5z \, dx \, dy \, dz = \int_0^a 5z \pi r^2 \, dz = \int_0^a 5\pi z(a^2 - z^2) \, dz = \frac{5}{4}\pi a^4$$

Considering the surface integral

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \iint_{\text{circular base}} \mathbf{F} \cdot \hat{\mathbf{n}}_1 \, dS + \iint_{\text{hemisphere}} \mathbf{F} \cdot \hat{\mathbf{n}}_2 \, dS$$

The unit normal to the base is clearly $\hat{\mathbf{n}}_1 = -\mathbf{k}$, so

$$\mathbf{F} \cdot \hat{\mathbf{n}}_1 = -z^2$$

giving

$$\iint_{\text{circular base}} \mathbf{F} \cdot \hat{\mathbf{n}}_1 \, dS = 0$$

since $z = 0$ on this surface.

The hemispherical surface is given by

$$f(x, y, z) = x^2 + y^2 + z^2 - a^2 = 0$$

so the outward unit normal $\hat{\mathbf{n}}_2$ is

$$\hat{\mathbf{n}}_2 = \frac{\nabla f}{|\nabla f|} = \frac{2xi + 2yj + 2zk}{2\sqrt{x^2 + y^2 + z^2}}$$

Since $x^2 + y^2 + z^2 = a^2$ on the surface,

$$\hat{\mathbf{n}}_2 = \frac{x}{a}\mathbf{i} + \frac{y}{a}\mathbf{j} + \frac{z}{a}\mathbf{k}$$

giving

$$\mathbf{F} \cdot \hat{\mathbf{n}}_2 = \frac{2x^2z}{a} + \frac{y^2z}{a} + \frac{z^3}{a} = \frac{x^2z}{a} + \frac{z}{a}(x^2 + y^2 + z^2)$$

Hence

$$\iint_{\text{hemisphere}} \mathbf{F} \cdot \hat{\mathbf{n}}_2 dS = \iint_{\text{hemisphere}} \frac{z}{a}(x^2 + a^2) dS$$

since $x^2 + y^2 + z^2 = a^2$ on the surface. Transforming to spherical polar coordinates,

$$x = a \sin \theta \cos \phi, \quad z = a \cos \theta, \quad dS = a^2 \sin \theta d\theta d\phi$$

the surface integral becomes

$$\begin{aligned} \iint_{\text{hemisphere}} \mathbf{F} \cdot \hat{\mathbf{n}}_2 dS &= a^4 \int_0^{2\pi} \int_0^{\pi/2} (\sin \theta \cos \theta + \sin^3 \theta \cos \theta \cos^2 \phi) d\theta d\phi \\ &= a^4 \int_0^{2\pi} [\frac{1}{2} \sin^2 \theta + \frac{1}{4} \sin^4 \theta \cos^2 \phi]_0^{\pi/2} d\phi \\ &= a^4 \int_0^{2\pi} [\frac{1}{2} + \frac{1}{4} \cos^2 \phi] d\phi = \frac{5}{4} \pi a^4 \end{aligned}$$

thus confirming that

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \iiint_V \operatorname{div} \mathbf{F} dV$$

3.4.12 Stokes' theorem

Stokes' theorem is the generalization of Green's theorem, and relates line integrals in three dimensions with surface integrals. At the end of Section 3.3.3 we saw that the curl of the vector \mathbf{F} could be expressed in the form

$$\operatorname{curl} \mathbf{F} \cdot \hat{\mathbf{n}} = \lim_{\Delta S \rightarrow 0} \frac{\text{flow round } \Delta S}{\Delta S}$$

In terms of differentials, this becomes

$$\operatorname{curl} \mathbf{F} \cdot d\mathbf{S} = \text{flow round } dS$$

Consider the surface S shown in Figure 3.44, bounded by the curve C . Then the line integral $\oint_C \mathbf{F} \cdot d\mathbf{r}$ can be interpreted as the total flow of a fluid with velocity field \mathbf{F} around the curve C . Partitioning the surface S into elements ΔS_i ($i = 1, \dots, n$), we can write

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \lim_{n \rightarrow \infty} \sum_{i=1}^n (\text{flow round } \Delta S_i) = \lim_{n \rightarrow \infty} \sum_{i=1}^n (\operatorname{curl} \mathbf{F} \cdot \Delta S_i)$$

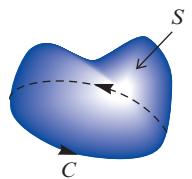


Figure 3.44 Surface S bounded by curve C .

so that

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_S (\operatorname{curl} \mathbf{F}) \cdot d\mathbf{S} \quad (3.37)$$

This result is known as **Stokes' theorem**. It provides a condition for a line integral to be independent of its path of integration. For, if the integral $\int_A^B \mathbf{F} \cdot d\mathbf{r}$ is independent of the path of integration then

$$\int_{C_1} \mathbf{F} \cdot d\mathbf{r} = \int_{C_2} \mathbf{F} \cdot d\mathbf{r}$$

where C_1 and C_2 are two different paths joining A and B as shown in Figure 3.45. Since

$$\int_{C_1} \mathbf{F} \cdot d\mathbf{r} = - \int_{-C_2} \mathbf{F} \cdot d\mathbf{r}$$

where $-C_2$ is the path C_2 traversed in the opposite direction, we have

$$\int_{C_1} \mathbf{F} \cdot d\mathbf{r} + \int_{-C_2} \mathbf{F} \cdot d\mathbf{r} = 0$$

That is,

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = 0$$

where C is the combined, closed curve formed from C_1 and $-C_2$. Stokes' theorem implies that if $\oint_C \mathbf{F} \cdot d\mathbf{r} = 0$ then

$$\iint_S (\operatorname{curl} \mathbf{F}) \cdot d\mathbf{S} = 0$$

for any surface S bounded by C . Since this is true for all surfaces bounded by C , we deduce that the integrand must be zero, that is $\operatorname{curl} \mathbf{F} = \mathbf{0}$. Writing $\mathbf{F} = (F_1, F_2, F_3)$, we then have that

$$\mathbf{F} \cdot d\mathbf{r} = F_1 dx + F_2 dy + F_3 dz$$

is an exact differential if $\operatorname{curl} \mathbf{F} = \mathbf{0}$; that is, if

$$\frac{\partial F_1}{\partial z} = \frac{\partial F_3}{\partial x}, \quad \frac{\partial F_1}{\partial y} = \frac{\partial F_2}{\partial x}, \quad \frac{\partial F_2}{\partial z} = \frac{\partial F_3}{\partial y}$$

Thus there is a function $f(x, y, z) = f(\mathbf{r})$ such that

$$F_1 = \frac{\partial f}{\partial x}, \quad F_2 = \frac{\partial f}{\partial y}, \quad F_3 = \frac{\partial f}{\partial z}$$

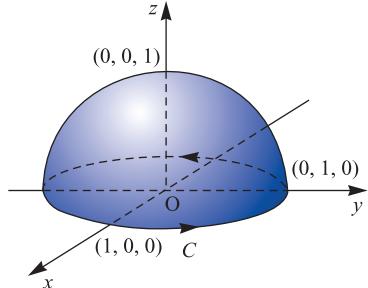
that is, such that $\mathbf{F}(\mathbf{r}) = \operatorname{grad} f$.

When $\mathbf{F}(\mathbf{r})$ represents a field of force, the field is said to be **conservative** (since it conserves rather than dissipates energy). When $\mathbf{F}(\mathbf{r})$ represents a velocity field for a fluid, the field is said to be curl-free or **irrotational**.

Example 3.32

Verify Stokes' theorem for $\mathbf{F} = (2x - y)\mathbf{i} - yz^2\mathbf{j} - y^2z\mathbf{k}$, where S is the upper half of the sphere $x^2 + y^2 + z^2 = 1$ and C is its boundary.

Figure 3.46
Hemispherical surface and boundary for Example 3.32.



Solution The surface and boundary involved are illustrated in Figure 3.46. We are required to show that

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_S \operatorname{curl} \mathbf{F} \cdot d\mathbf{S}$$

Since C is a circle of unit radius in the (x, y) plane, to evaluate $\oint_C \mathbf{F} \cdot d\mathbf{r}$, we take

$$x = \cos \phi, \quad y = \sin \phi$$

so that

$$\mathbf{r} = \cos \phi \mathbf{i} + \sin \phi \mathbf{j}$$

giving

$$d\mathbf{r} = -\sin \phi d\phi \mathbf{i} + \cos \phi d\phi \mathbf{j}$$

Also, on the boundary C , $z = 0$, so that

$$\mathbf{F} = (2x - y)\mathbf{i} = (2 \cos \phi - \sin \phi)\mathbf{i}$$

Thus

$$\begin{aligned} \oint_C \mathbf{F} \cdot d\mathbf{r} &= \int_0^{2\pi} (2 \cos \phi - \sin \phi)\mathbf{i} \cdot (-\sin \phi \mathbf{i} + \cos \phi \mathbf{j}) d\phi \\ &= \int_0^{2\pi} (-2 \sin \phi \cos \phi + \sin^2 \phi) d\phi = \int_0^{2\pi} [-\sin 2\phi + \frac{1}{2}(1 + \cos 2\phi)] d\phi \\ &= \pi \end{aligned}$$

$$\operatorname{curl} \mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ 2x - y & -yz^2 & -y^2z \end{vmatrix} = \mathbf{k}$$

The unit outward-drawn normal at a point (x, y, z) on the hemisphere is given by $(xi + yj + zk)$, since $x^2 + y^2 + z^2 = 1$. Thus

$$\begin{aligned}\iint_S \operatorname{curl} \mathbf{F} \cdot d\mathbf{S} &= \iint_S \mathbf{k} \cdot (xi + yj + zk) dS \\ &= \iint_S z dS \\ &= \int_0^{2\pi} \int_0^{\pi/2} \cos \theta \sin \theta d\theta d\phi \\ &= 2\pi [\frac{1}{2} \sin^2 \theta]_0^{\pi/2} = \pi\end{aligned}$$

Hence $\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_S (\operatorname{curl} \mathbf{F}) \cdot d\mathbf{S}$, and Stokes' theorem is verified.

3.4.13 Exercises

- 103 Evaluate $\iint_S \mathbf{F} \cdot d\mathbf{S}$, where $\mathbf{F} = (4xz, -y^2, yz)$ and S is the surface of the cube bounded by the planes $x = 0$, $x = 1$, $y = 0$, $y = 1$, $z = 0$ and $z = 1$.

- 104 Use the divergence theorem to evaluate the surface integral $\iint_S \mathbf{F} \cdot d\mathbf{S}$, where $\mathbf{F} = xzi + yzj + z^2k$ and S is the closed surface of the hemisphere $x^2 + y^2 + z^2 = 4$, $z > 0$. (Note that you are not required to verify the theorem.)

- 105 Verify the divergence theorem

$$\iint_S \mathbf{F} \cdot d\mathbf{S} = \iiint_V \operatorname{div} \mathbf{F} dV$$

for $\mathbf{F} = 4xi - 2y^2j + z^2k$ over the region bounded by $x^2 + y^2 = 4$, $z = 0$ and $z = 3$.

- 106 Prove that

$$\iiint_V (\operatorname{grad} \phi) \cdot (\operatorname{curl} \mathbf{F}) dV = \iint_S (\mathbf{F} \times \operatorname{grad} \phi) \cdot d\mathbf{S}$$

- 107 Verify the divergence theorem for $\mathbf{F} = (xy + y^2)i + x^2yj$ and the volume V in the first octant bounded by $x = 0$, $y = 0$, $z = 0$, $z = 1$ and $x^2 + y^2 = 4$.

- 108 Use Stokes' theorem to show that the value of the line integral $\int_A^B \mathbf{F} \cdot d\mathbf{r}$ for

$$\begin{aligned}\mathbf{F} = (36xz + 6y \cos x, 3 + 6 \sin x + z \sin y, \\ 18x^2 - \cos y)\end{aligned}$$

is independent of the path joining the points A and B.

- 109 Use Stokes' theorem to evaluate the line integral $\oint_C \mathbf{A} \cdot d\mathbf{r}$, where $\mathbf{A} = -yi + xj$ and C is the boundary of the ellipse $x^2/a^2 + y^2/b^2 = 1$, $z = 0$.

- 110 Verify Stokes' theorem by evaluating both sides of

$$\iint_S (\operatorname{curl} \mathbf{F}) \cdot d\mathbf{S} = \oint_C \mathbf{F} \cdot d\mathbf{r}$$

where $\mathbf{F} = (2x - y)i - yz^2j - y^2zk$ and S is the curved surface of the hemisphere $x^2 + y^2 + z^2 = 16$, $z \geq 0$.

- 111 By applying Stokes' theorem to the function $a\mathbf{f}(r)$, where a is a constant, deduce that

$$\iint_S (\mathbf{n} \times \operatorname{grad} f) \cdot d\mathbf{S} = \int_C f(\mathbf{r}) d\mathbf{r}$$

Verify this result for the function $f(\mathbf{r}) = 3xy^2$ and the rectangle in the plane $z = 0$ bounded by the lines $x = 0$, $x = 1$, $y = 0$ and $y = 2$.

- 112 Verify Stokes' theorem for $\mathbf{F} = (2y + z, x - z, y - x)$ for the part of $x^2 + y^2 + z^2 = 1$ lying in the positive octant.

3.5 Engineering application: streamlines in fluid dynamics

As we mentioned in Section 3.1.5, differentials often occur in mathematical modelling of practical problems. An example occurs in fluid dynamics. Consider the case of steady-state incompressible fluid flow in two dimensions. Using rectangular cartesian coordinates (x, y) to describe a point in the fluid, let u and v be the velocities of the fluid in the x and y directions respectively. Then by considering the flow in and flow out of a small rectangle, as shown in Figure 3.47, per unit time, we obtain a differential relationship between $u(x, y)$ and $v(x, y)$ that models the fact that no fluid is lost or gained in the rectangle; that is, the fluid is conserved.

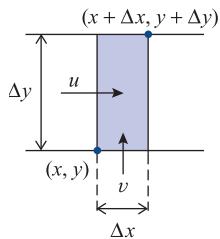


Figure 3.47
Fluid flow.

The velocity of the fluid \mathbf{q} is a vector point function. The values of its components u and v depend on the spatial coordinates x and y . The flow into the small rectangle in unit time is

$$u(x, \bar{y})\Delta y + v(\bar{x}, y)\Delta x$$

where \bar{x} lies between x and $x + \Delta x$, and \bar{y} lies between y and $y + \Delta y$. Similarly, the flow out of the rectangle is

$$u(x + \Delta x, \tilde{y})\Delta y + v(\tilde{x}, y + \Delta y)\Delta x$$

where \tilde{x} lies between x and $x + \Delta x$ and \tilde{y} lies between y and $y + \Delta y$. Because no fluid is created or destroyed within the rectangle, we may equate these two expressions, giving

$$u(x, \bar{y})\Delta y + v(\bar{x}, y)\Delta x = u(x + \Delta x, \tilde{y})\Delta y + v(\tilde{x}, y + \Delta y)\Delta x$$

Rearranging, we have

$$\frac{u(x + \Delta x, \tilde{y}) - u(x, \bar{y})}{\Delta x} + \frac{v(\tilde{x}, y + \Delta y) - v(\bar{x}, y)}{\Delta y} = 0$$

Letting $\Delta x \rightarrow 0$ and $\Delta y \rightarrow 0$ gives the **continuity equation**

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

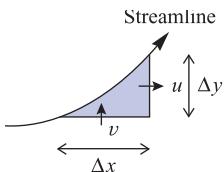


Figure 3.48
Streamline.

The fluid actually flows along paths called **streamlines** so that there is no flow across a streamline. Thus from Figure 3.48 we deduce that

$$v \Delta x = u \Delta y$$

and hence

$$v dx - u dy = 0$$

The condition for this expression to be an exact differential is

$$\frac{\partial}{\partial y}(v) = \frac{\partial}{\partial x}(-u)$$

or

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

This is satisfied for incompressible flow since it is just the continuity equation, so that we deduce that there is a function $\psi(x, y)$, called the **stream function**, such that

$$v = \frac{\partial \psi}{\partial x} \quad \text{and} \quad u = -\frac{\partial \psi}{\partial y}$$

It follows that if we are given u and v , as functions of x and y , that satisfy the continuity equation then we can find the equations of the streamlines given by $\psi(x, y) = \text{constant}$.

Example 3.33

Find the stream function $\psi(x, y)$ for the incompressible flow that is such that the velocity \mathbf{q} at the point (x, y) is

$$(-y/(x^2 + y^2), x/(x^2 + y^2))$$

Solution From the definition of the stream function, we have

$$u(x, y) = -\frac{\partial \psi}{\partial y} \quad \text{and} \quad v(x, y) = \frac{\partial \psi}{\partial x}$$

provided that

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

Here we have

$$u = \frac{-y}{x^2 + y^2} \quad \text{and} \quad v = \frac{x}{x^2 + y^2}$$

so that

$$\frac{\partial u}{\partial x} = \frac{2xy}{(x^2 + y^2)^2} \quad \text{and} \quad \frac{\partial v}{\partial y} = -\frac{2yx}{(x^2 + y^2)^2}$$

confirming that

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

Integrating

$$\frac{\partial \psi}{\partial y} = -u(x, y) = \frac{y}{x^2 + y^2}$$

with respect to y , keeping x constant, gives

$$\psi(x, y) = \frac{1}{2} \ln(x^2 + y^2) + g(x)$$

Differentiating partially with respect to x gives

$$\frac{\partial \psi}{\partial x} = \frac{x}{x^2 + y^2} + \frac{dg}{dx}$$

Since it is known that

$$\frac{\partial \psi}{\partial x} = v(x, y) = \frac{x}{x^2 + y^2}$$

we have

$$\frac{dg}{dx} = 0$$

which on integrating gives

$$g(x) = C$$

where C is a constant. Substituting back into the expression obtained for $\psi(x, y)$, we have

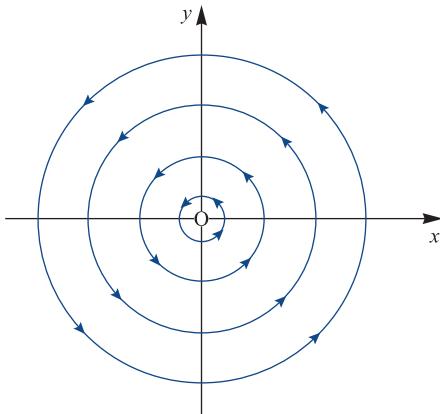
$$\psi(x, y) = \frac{1}{2} \ln(x^2 + y^2) + C$$

A streamline of the flow is given by the equation $\psi(x, y) = k$, where k is a constant. After a little manipulation this gives

$$x^2 + y^2 = a^2 \quad \text{and} \quad \ln a = k - C$$

and the corresponding streamlines are shown in Figure 3.49. This is an example of a **vortex**.

Figure 3.49
Streamline illustrating
a vortex.



3.6 Engineering application: heat transfer

In modelling heat transfer problems we make use of three experimental laws.

- (1) Heat flows from hot regions to cold regions of a body.
- (2) The rate at which heat flows through a plane section drawn in a body is proportional to its area and to the temperature gradient normal to the section.
- (3) The quantity of heat in a body is proportional to its mass and to its temperature.

In the simplest case we consider heat transfer in a medium for which the constants of proportionality in the above laws are independent of direction. Such a medium is called **thermally isentropic**. For any arbitrary region within such a medium we can obtain an equation that models such heat flows. The total amount $Q(t)$ of heat within the region V is

$$Q(t) = \iiint_V c \rho u(\mathbf{r}, t) dV$$

where c is the specific heat of the medium, ρ is the density and $u(\mathbf{r}, t)$ is the temperature at the point \mathbf{r} at time t . Heat flows out of the region through its bounding surface S . The experimental laws (1) and (2) above imply that the rate at which heat flows across an element ΔS of that surface is $-k\nabla u \cdot \Delta S$, where k is the thermal conductivity of the medium. (The minus sign indicates that heat flows from hot regions to cold.) Thus the rate at which heat flows across the whole surface of the region is given by

$$\iint_S (-k\nabla u) \cdot dS = -k \iint_S \nabla u \cdot dS$$

Using Gauss's theorem, we deduce that the rate at which heat flows out of the region is

$$-k \iiint_V \nabla^2 u \, dV$$

If there are no sources or sinks of heat within the region, this must equal the rate at which the region loses heat, $-dQ/dt$. Therefore

$$-\frac{d}{dt} \left[\iiint_V c\rho u(\mathbf{r}, t) \, dV \right] = -k \iiint_V \nabla^2 u \, dV$$

Since

$$\frac{d}{dt} \iiint_V u(\mathbf{r}, t) \, dV = \iiint_V \frac{\partial u}{\partial t} \, dV$$

this implies that

$$\iiint_V \left(k\nabla^2 u - c\rho \frac{\partial u}{\partial t} \right) \, dV = 0$$

This models the situation for any arbitrarily chosen region V . The arbitrariness in the choice of V implies that the value of the integral is independent of V and that the integrand is equal to zero. Thus

$$\nabla^2 u = \frac{c\rho}{k} \frac{\partial u}{\partial t}$$

The quantity $k/c\rho$ is termed the **thermal diffusivity** of the medium and is usually denoted by the Greek letter kappa, κ . The differential equation models heat flow within a medium. Its solution depends on the initial temperature distribution $u(\mathbf{r}, 0)$ and on the conditions pertaining at the boundary of the region. Methods for solving this equation are discussed in Chapter 9. This differential equation also occurs as a model for water percolation through a dam, for neutron transport in reactors and in charge transfer within charge-coupled devices. We shall now proceed to obtain its solution in a very special case.

Example 3.34

A large slab of material has an initial temperature distribution such that one half is at $-u_0$ and the other at $+u_0$. Obtain a mathematical model for this situation and solve it, stating explicitly the assumptions that are made.

Solution

When a problem is stated in such vague terms, it is difficult to know what approximations and simplifications may be reasonably made. Since we are dealing with heat transfer, we know that for an isentropic medium the temperature distribution satisfies the equation

$$\nabla^2 u = \frac{1}{\kappa} \frac{\partial u}{\partial t}$$

throughout the medium. We know that the region we are studying is divided so that at $t = 0$ the temperature in one part is $-u_0$ while that in the other is $+u_0$, as illustrated in Figure 3.50. We can deduce from this figure that the subsequent temperature at a point in the medium depends only on the perpendicular distance of the point from the dividing plane. We choose a coordinate system so that its origin lies on the dividing plane and the x axis is perpendicular to it, as shown in Figure 3.51. Then the differential equation simplifies, since $u(\mathbf{r}, t)$ is independent of y and z , and we have

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{\kappa} \frac{\partial u}{\partial t} \quad \text{with} \quad u(x, 0) = \begin{cases} -u_0 & (x < 0) \\ +u_0 & (x \geq 0) \end{cases}$$

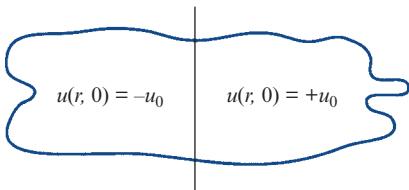


Figure 3.50 Region for Example 3.34.

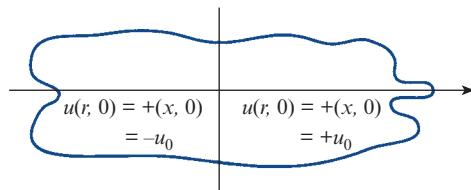


Figure 3.51 Coordinate system for Example 3.34.

Thinking about the physical problem also provides us with some further information. The heat flows from the hot region to the cold until (eventually) the temperature is uniform throughout the medium. In this case that terminal temperature is zero, since initially half the medium is at temperature $+u_0$ and the other half at $-u_0$. So we know that $u(x, t) \rightarrow 0$ as $t \rightarrow \infty$. We also deduce from the initial temperature distribution that $-u_0 \leq u(x, t) \leq u_0$ for all x and t , since there are no extra sources or sinks of heat in the medium. Summarizing, we have

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{\kappa} \frac{\partial u}{\partial t} \quad (-\infty < x < \infty, t \geq 0) \quad \text{with} \quad u(x, 0) = \begin{cases} -u_0 & (x < 0) \\ +u_0 & (x \geq 0) \end{cases}$$

$u(x, t) \quad \text{bounded for all } x$

$u(x, t) \rightarrow 0 \quad \text{as } t \rightarrow \infty$

There are many approaches to solving this problem (see Chapter 9). One is to investigate the effect of changing the scale of the independent variables x and t . Setting $x = \lambda X$ and $t = \mu T$, where λ and μ are positive constants, the problem becomes

$$\mu \frac{\partial^2 U}{\partial X^2} = \frac{\lambda^2}{\kappa} \frac{\partial U}{\partial T}$$

with $U(X, T) = u(x, t)$ and $U(X, 0) = u_0 \operatorname{sgn} X$. Choosing $\mu = \lambda^2$, we see that

$$\frac{\partial^2 U}{\partial X^2} = \frac{1}{\kappa} \frac{\partial U}{\partial T}, \quad \text{with } U(X, 0) = u_0 \operatorname{sgn} X$$

which implies that the solution $u(x, t)$ of the original equation is also a solution of the scaled equation. Thus

$$u(x, t) = u(\lambda x, \lambda^2 t)$$

which suggests that we should look for a solution expressed in terms of a new variable s that is proportional to the ratio of x to \sqrt{t} . Setting $s = ax/\sqrt{t}$, we seek a solution as a function of s :

$$u(x, t) = u_0 f(s)$$

This reduces the partial differential equation for u to an ordinary differential equation for f , since

$$\frac{\partial u}{\partial x} = \frac{au_0}{\sqrt{t}} \frac{df}{ds}, \quad \frac{\partial^2 u}{\partial x^2} = \frac{a^2 u_0}{t} \frac{d^2 f}{ds^2}, \quad \frac{\partial u}{\partial t} = -\frac{1}{2} \frac{axu_0}{t\sqrt{t}} \frac{df}{ds}$$

Thus the differential equation is transformed into

$$\frac{a^2}{t} \frac{d^2 f}{ds^2} = -\frac{ax}{2\kappa t\sqrt{t}} \frac{df}{ds}$$

giving

$$a^2 \frac{d^2 f}{ds^2} = -\frac{s}{2\kappa} \frac{df}{ds}$$

Choosing the constant a such that $a^2 = 1/(4\kappa)$ reduces this to the equation

$$\frac{d^2 f}{ds^2} = -2s \frac{df}{ds}$$

The initial condition is transformed into two conditions, since for $x < 0$, $s \rightarrow -\infty$ as $t \rightarrow 0$ and for $x > 0$, $s \rightarrow +\infty$ as $t \rightarrow 0$. So we have

$$f(s) \rightarrow 1 \quad \text{as } s \rightarrow \infty$$

$$f(s) \rightarrow -1 \quad \text{as } s \rightarrow -\infty$$

Integrating the differential equation once gives

$$\frac{df}{ds} = A e^{-s^2}, \quad \text{where } A \text{ is a constant}$$

and integrating a second time gives

$$f(s) = B + A \int e^{-s^2} ds$$

The integral occurring here is one that frequently arises in heat transfer problems, and is given a special name. We define the **error function**, $\text{erf}(x)$, by the integral

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-z^2} dz$$

Its name derives from the fact that it is associated with the normal distribution, which is a common model for the distribution of experimental errors (see Section 11.2.4). This is a well-tabulated function, and has the property that $\text{erf}(x) \rightarrow 1$ as $x \rightarrow \infty$.

Writing the solution obtained above in terms of the error function, we have

$$f(s) = A \text{erf}(s) + B$$

Letting $s \rightarrow \infty$ and $s \rightarrow -\infty$ gives two equations for A and B :

$$1 = A + B$$

$$-1 = -A + B$$

from which we deduce $A = 1$ and $B = 0$. Thus

$$f(s) = \text{erf}(s)$$

so that

$$u(x, t) = u_0 \text{erf}\left(\frac{x}{2\sqrt{t}}\right) = \frac{2u_0}{\sqrt{\pi}} \int_0^{x/2\sqrt{t}} e^{-z^2} dz$$

3.7 Review exercises (1–21)

- 1 Show that $u(x, y) = x^n f(t)$, $t = y/x$, satisfies the differential equations

$$(a) x \frac{\partial u}{\partial x} + y \frac{\partial u}{\partial y} = nu$$

$$(b) x^2 \frac{\partial^2 u}{\partial x^2} + 2xy \frac{\partial^2 u}{\partial x \partial y} + y^2 \frac{\partial^2 u}{\partial y^2} = n(n-1)u$$

Verify these results for the function

$$u(x, y) = x^4 + y^4 + 16x^2y^2.$$

- 2 Find the values of the numbers a and b such that the change of variables $u = x + ay$, $v = x + by$ transforms the differential equation

$$9 \frac{\partial^2 f}{\partial x^2} - 9 \frac{\partial^2 f}{\partial x \partial y} + 2 \frac{\partial^2 f}{\partial y^2} = 0$$

into

$$\frac{\partial^2 f}{\partial u \partial v} = 0$$

Hence deduce that the general solution of the equation is given by

$$u(x, y) = f(x + 3y) + g(x + \frac{3}{2}y)$$

where f and g are arbitrary functions.

Find the solution of the differential equation that satisfies the conditions

$$u(x, 0) = \sin x, \quad \frac{\partial u(x, 0)}{\partial y} = 3 \cos x$$

- 3 A differential $P(x, y, z) dx + Q(x, y, z) dy + R(x, y, z) dz$ is exact if there is a function $f(x, y, z)$ such that

$$\begin{aligned} P(x, y, z) dx + Q(x, y, z) dy + R(x, y, z) dz \\ = \nabla f \cdot (dx, dy, dz) \end{aligned}$$

Show that this implies $\nabla \times (P, Q, R) = 0$. Deduce that $\text{curl grad } f = 0$.

- 4** Find $\text{grad } f$, plot some level curves $f = \text{constant}$ and indicate $\text{grad } f$ by arrows at some points on the level curves for $f(\mathbf{r})$ given by

(a) xy (b) $x/(x^2 + y^2)$

- 5** Show that if ω is a constant vector then

(a) $\text{grad}(\omega \cdot \mathbf{r}) = \omega$
 (b) $\text{curl}(\omega \times \mathbf{r}) = 2\omega$

- 6** (a) Prove that if $f(\mathbf{r})$ is a scalar point function then

$$\text{curl grad } f = 0$$

- (b) Prove that if $\mathbf{v} = \text{grad}[zf(\mathbf{r})] + \alpha f(\mathbf{r})\mathbf{k}$ and $\nabla^2 f = 0$, where α is a constant and f is a scalar point function, then

$$\text{div } \mathbf{v} = (2 + \alpha)\frac{\partial f}{\partial z}, \quad \nabla^2 \mathbf{v} = \text{grad}\left(2\frac{\partial f}{\partial z}\right)$$

- 7** Show that if $\mathbf{F} = (x^2 - y^2 + x)\mathbf{i} - (2xy + y)\mathbf{j}$, then $\text{curl } \mathbf{F} = 0$, and find $f(\mathbf{r})$ such that $\mathbf{F} = \text{grad } f$.

Verify that

$$\int_{(1,2)}^{(2,1)} \mathbf{F} \cdot d\mathbf{r} = [f(\mathbf{r})]_{(1,2)}^{(2,1)}$$

- 8** A force \mathbf{F} acts on a particle that is moving in two dimensions along the semicircle $x = 1 - \cos \theta$, $y = \sin \theta$ ($0 \leq \theta \leq \pi$).

Find the work done when

(a) $\mathbf{F} = \sqrt{(x^2 + y^2)}\mathbf{i}$
 (b) $\mathbf{F} = \sqrt{(x^2 + y^2)}\hat{\mathbf{n}}$

$\hat{\mathbf{n}}$ being the unit vector *tangential* to the path.

- 9** A force $\mathbf{F} = (xy, -y, 1)$ acts on a particle as it moves along the straight line from $(0, 0, 0)$ to $(1, 1, 1)$. Calculate the work done.

- 10** The force \mathbf{F} per unit length of a conducting wire carrying a current I in a magnetic field \mathbf{B} is $\mathbf{F} = \mathbf{I} \times \mathbf{B}$. Find the force acting on a circuit whose shape is given by $x = \sin \theta$, $y = \cos \theta$, $z = \sin \frac{1}{2}\theta$, when current I flows in it and when it lies in a magnetic field $\mathbf{B} = x\mathbf{i} - y\mathbf{j} + \mathbf{k}$.

- 11** The velocity \mathbf{v} at the point (x, y) in a two-dimensional fluid flow is given by

$\mathbf{v} = (yi - xj)/(x^2 + y^2)$. Find the net circulation around the square $x = \pm 1$, $y = \pm 1$.

12

A metal plate has its boundary defined by $x = 0$, $y = x^2/c$ and $y = c$. The density at the point (x, y) is kxy (per unit area). Find the moment of inertia of the plate about an axis through $(0, 0)$ and perpendicular to the plate.

13

A right circular cone of height h and base radius a is cut into two pieces along a plane parallel to and distance c from the axis of the cone. Find the volume of the smaller piece.

14

The axes of two circular cylinders of radius a intersect at right angles. Show that the volume common to both cylinders may be expressed as the triple integral

$$8 \int_0^a dy \int_0^{\sqrt{(a^2-y^2)}} dx \int_0^{\sqrt{(a^2-y^2)}} dz$$

and hence evaluate it.

15

The elastic energy of a volume V of material is $q^2V/(2EI)$, where q is its stress and E and I are constants. Find the elastic energy of a cylindrical volume of radius r and length l in which the stress varies directly as the distance from its axis, being zero at the axis and q_0 at the outer surface.

16

The velocity of a fluid at the point (x, y, z) has components $(3x^2y, xy^2, 0)$. Find the flow rate out of the triangular prism bounded by $z = 0$, $z = 1$, $x = 0$, $y = 0$ and $x + y = 1$.

17

An electrostatic field has components $(2xy, -y^2, x + y)$ at the point (x, y, z) . Find the total flux out of the sphere $x^2 + y^2 + z^2 = a^2$.

18

Verify Stokes' theorem

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \iint_S (\text{curl } \mathbf{F}) \cdot d\mathbf{S}$$

where $\mathbf{F} = (x^2 + y - 4, 3xy, 2xz + z^2)$ and S is the surface of the hemisphere $x^2 + y^2 + z^2 = 16$ above the (x, y) plane.

19

Use the divergence theorem to evaluate the surface integral

$$\iint_S \mathbf{a} \cdot d\mathbf{S}$$

where $\mathbf{a} = xi + yj - 2zk$ and S is the surface of the sphere $x^2 + y^2 + z^2 = a^2$ above the (x, y) plane.

- 20 Evaluate the volume integral

$$\iiint_V xyz \, dV$$

where V denotes the wedge-shaped region bounded in the positive octant by the four planes $x = 0$, $y = 0$, $y = 1 - x$ and $z = 2 - x$.

21

Continuing the analysis of Section 3.5, show that the net circulation of fluid around the rectangular element shown in Figure 3.47 is given by

$$[u(x, y + \Delta y) - u(x, y)]\Delta x \\ - [v(x + \Delta x, y) - v(x, y)]\Delta y$$

Deduce that if the fluid motion is irrotational at (x, y) , then

$$\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x} = 0$$

Show that for irrotational incompressible flow, the stream function ψ satisfies Laplace equation

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = 0$$



4

Functions of a Complex Variable

Chapter 4 Contents

4.1	Introduction	250
4.2	Complex functions and mappings	251
4.3	Complex differentiation	274
4.4	Complex series	287
4.5	Singularities and zeros	300
4.6	Engineering application: analysing AC circuits	304
4.7	Engineering application: use of harmonic functions	305
4.8	Review exercises (1–19)	311

4.1

Introduction

In the theory of alternating currents, the application of quantities such as the complex impedance involves functions having complex numbers as independent variables. There are many other areas in engineering where this is the case; for example, the motion of fluids, the transfer of heat or the processing of signals. Some of these applications are discussed later in this book.

Traditionally, complex variable techniques have been important, and extensively used, in a wide variety of engineering situations. This has been especially the case in areas such as electromagnetic and electrostatic field theory, fluid dynamics, aerodynamics and elasticity. With the development of computer technology and the consequential use of sophisticated algorithms for analysis and design in engineering there has, over the last two decades or so, been less emphasis on the use of complex variable techniques and a shift towards numerical techniques applied directly to the underlying full partial differential equations model of the situation being investigated. However, even when this is the case there is still considerable merit in having an analytical solution, possibly for an idealized model, in order both to develop better understanding of the behaviour of the solution and to give confidence in the numerical estimates for the solution of enhanced models. Many sophisticated software packages now exist, many of which are available as freeware, downloadable from various internet sites. The older packages such as FLUENT and CFX are still available and still in use by engineering companies to solve problems such as fluid flow and heat transfer in real situations. The finite-element package TELEMAC is modular in style and is useful for larger-scale environmental problems; these types of software programs use a core plus optional add-ons tailored for specific applications. The best use of all such software still requires knowledge of mappings and use of complex variables. One should also mention the computer entertainment industry which makes use of such mathematics to enable accurate simulation of real life. The kind of mappings that used to be used extensively in aerodynamics are now used in the computer games industry. In particular the ability to analyse complicated flow patterns by mapping from a simple geometry to a complex one and back again remains very important. Examples at the end of the chapter illustrate the techniques that have been introduced. Many engineering mathematics texts have introduced programming segments that help the reader to use packages such as MATLAB or MAPLE to carry out the technicalities. This has not been done in this chapter since, in the latest version of MAPLE, the user simply opens the program and uses the menu to click on the application required (in this chapter a derivative or an integral), types in the problem and presses return to get the answer. Students are encouraged to use such software to solve any of the problems; the understanding of what the solutions mean is always more important than any tricks used to solve what are idealized problems.



Throughout engineering, transforms in one form or another play a major role in analysis and design. An area of continuing importance is the use of Laplace, z , Fourier and other transforms in areas such as control, communication and signal processing. Such transforms are considered later in the book where it will be seen that complex variables play a key role. This chapter is devoted to developing understanding of the standard techniques of complex variables so as to enable the reader to apply them with confidence in application areas.

4.2

Complex functions and mappings

The concept of a function involves two sets X and Y and a rule that assigns to each element x in the set X (written $x \in X$) precisely one element $y \in Y$. Whenever this situation arises, we say that there is a **function** f that **maps** the set X to the set Y , and represent this symbolically by

$$y = f(x) \quad (x \in X)$$

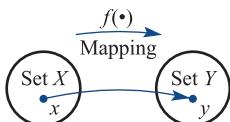


Figure 4.1 Real mapping $y = f(x)$.

Schematically we illustrate a function as in Figure 4.1. While x can take any value in the set X , the variable $y = f(x)$ depends on the particular element chosen for x . We therefore refer to x as the **independent** variable and y as the **dependent** variable. The set X is called the **domain** of the function, and the set of all images $y = f(x)$ ($x \in X$) is called the **image set** or **range** of f . Previously we were concerned with real functions, so that x and y were real numbers. If the independent variable is a complex variable $z = x + jy$, where x and y are real and $j = \sqrt{(-1)}$, then the function $f(z)$ of z will in general also be complex. For example, if $f(z) = z^2$ then, replacing z by $x + jy$ and expanding, we have

$$f(z) = (x + jy)^2 = (x^2 - y^2) + j2xy = u + jv \quad (\text{say})$$

where u and v are real. Such a function $f(z)$ is called a **complex function**, and we write

$$w = f(z)$$

where, in general, the dependent variable $w = u + jv$ is also complex.

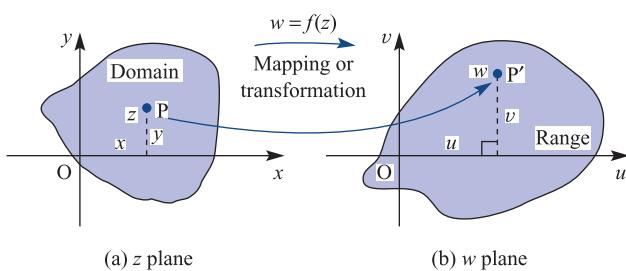
The reader will recall that a complex number $z = x + jy$ can be represented on a plane called the **Argand diagram**, as illustrated in Figure 4.2(a). However, we cannot plot the values of x , y and $f(z)$ on one set of axes, as we were able to do for real functions $y = f(x)$. We therefore represent the values of

$$w = f(z) = u + jv$$

on a second plane as illustrated in Figure 4.2(b). The plane containing the independent variable z is called the z plane and the plane containing the dependent variable w is called the w plane. Thus the complex function $w = f(z)$ may be regarded as a **mapping** or **transformation** of points P within a region in the z plane (called the **domain**) to corresponding image points P' within a region in the w plane (called the **range**).

It is this facility for mapping that gives the theory of complex functions much of its application in engineering. In most useful mappings the entire z plane is mapped onto the entire w plane, except perhaps for isolated points. Throughout this chapter the domain will be taken to be the entire z plane (that is, the set of all complex numbers, denoted by \mathbb{C}). This is analogous, for real functions, to the domain being the entire real

Figure 4.2 Complex mapping $w = f(z)$.



line (that is, the set of all real numbers \mathbb{R}). If this is not the case then the complex function is termed ‘not well defined’. In contrast, as for real functions, the range of the complex function may well be a proper subset of \mathbb{C} .

Example 4.1

Find the image in the w plane of the straight line $y = 2x + 4$ in the z plane, $z = x + jy$, under the mapping

$$w = 2z + 6$$

Solution Writing $w = u + jv$, where u and v are real, the mapping becomes

$$w = u + jv = 2(x + jy) + 6$$

or

$$u + jv = (2x + 6) + j2y$$

Equating real and imaginary parts then gives

$$u = 2x + 6, \quad v = 2y \quad (4.1)$$

which, on solving for x and y , leads to

$$x = \frac{1}{2}(u - 6), \quad y = \frac{1}{2}v$$

Thus the image of the straight line

$$y = 2x + 4$$

in the z plane is represented by

$$\frac{1}{2}v = 2 \times \frac{1}{2}(u - 6) + 4$$

or

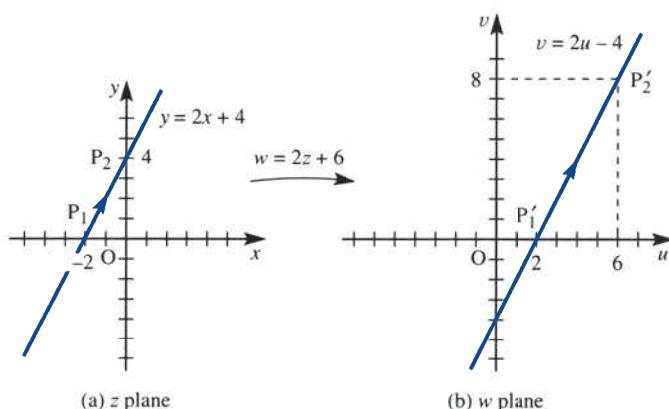
$$v = 2u - 4$$

which corresponds to a straight line in the w plane. The given line in the z plane and the mapped image line in the w plane are illustrated in Figures 4.3(a) and (b) respectively.

Note from (1.1) that, in particular, the point $P_1(-2 + j0)$ in the z plane is mapped to the point $P'_1(2 + j0)$ in the w plane, and that the point $P_2(0 + j4)$ in the z plane is mapped to the point $P'_2(6 + j8)$ in the w plane. Thus, as the point P moves from P_1 to P_2 along

Figure 4.3

The mapping of Example 4.1.



the line $y = 2x + 4$ in the z plane, the mapped point P' moves from P'_1 to P'_2 along the line $v = 2u - 4$ in the w plane. It is usual to indicate this with the arrowheads as illustrated in Figure 4.3.

4.2.1 Linear mappings

The mapping $w = 2z + 6$ in Example 4.1 is a particular example of a mapping corresponding to the general complex linear function

$$w = \alpha z + \beta \quad (4.2)$$

where w and z are complex-valued variables, and α and β are complex constants. In this section we shall investigate mappings of the z plane onto the w plane corresponding to (4.2) for different choices of the constants α and β . In so doing we shall also introduce some general properties of mappings.

Case (a) $\alpha = 0$

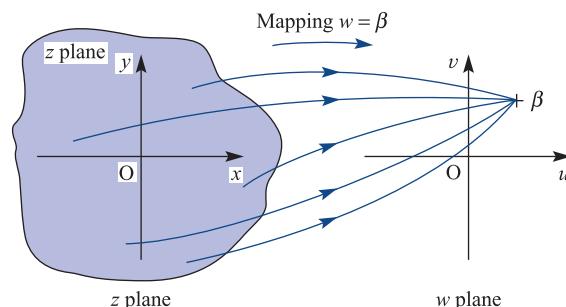
Letting $\alpha = 0$ (or $\alpha = 0 + j0$) in (4.2) gives

$$w = \beta$$

which implies that $w = \beta$, no matter what the value of z . This is quite obviously a degenerate mapping, with the entire z plane being mapped onto the one point $w = \beta$ in the w plane. If nothing else, this illustrates the point made earlier in this section, that the image set may only be part of the entire w plane. In this particular case the image set is a single point. Since the whole of the z plane maps onto $w = \beta$, it follows that, in particular, $z = \beta$ maps to $w = \beta$. The point β is thus a **fixed point** in this mapping, which is a useful concept in helping us to understand a particular mapping. A further question of interest when considering mappings is that of whether, given a point in the w plane, we can tell from which point in the z plane it came under the mapping. If it is possible to get back to a unique point in the z plane then it is said to have an **inverse mapping**. Clearly, for an inverse mapping $z = g(w)$ to exist, the point in the w plane has to be in the image set of the original mapping $w = f(z)$. Also, from the definition of a mapping, each point w in the w plane image set must lead to a single point z in the z plane under the inverse mapping $z = g(w)$. (Note the similarity to the requirements for the existence of an inverse function $f^{-1}(x)$ of a real function $f(x)$.) For the particular mapping $w = \beta$ considered here the image set is the single point $w = \beta$ in the w plane, and it is clear from Figure 4.4 that there is no way of getting back to just a single point in the z plane. Thus the mapping $w = \beta$ has no inverse.

Figure 4.4

The degenerate mapping $w = \beta$.



Case (b) $\beta = 0, \alpha \neq 0$

With such a choice for the constants α and β , the mapping corresponding to (4.2) becomes

$$w = \alpha z$$

Under this mapping, the origin is the only fixed point, there being no other fixed points that are finite. Also, in this case there exists an inverse mapping

$$z = \frac{1}{\alpha} w$$

that enables us to return from the w plane to the z plane to the very same point from which we started under $w = \alpha z$. To illustrate this mapping at work, let us choose $\alpha = 1 + j$, so that

$$w = (1 + j)z \quad (4.3)$$

and consider what happens to a general point z_0 in the z plane under this mapping. In general, there are two ways of doing this. We can proceed as in Example 4.1 and split both z and w into real and imaginary parts, equate real and imaginary parts and hence find the image curves in the w plane to specific curves (usually the lines $\operatorname{Re}(z) = \text{constant}$, $\operatorname{Im}(z) = \text{constant}$) in the z plane. Alternatively, we can rearrange the expression for w and deduce the properties of the mapping directly. The former course of action, as we shall see in this chapter, is the one most frequently used. Here, however, we shall take the latter approach and write $\alpha = 1 + j$ in polar form as

$$1 + j = \sqrt{2}e^{j\pi/4}$$

Then, if

$$z = r e^{j\theta}$$

in polar form it follows from (4.3) that

$$w = r\sqrt{2}e^{j(\theta + \pi/4)} \quad (4.4)$$

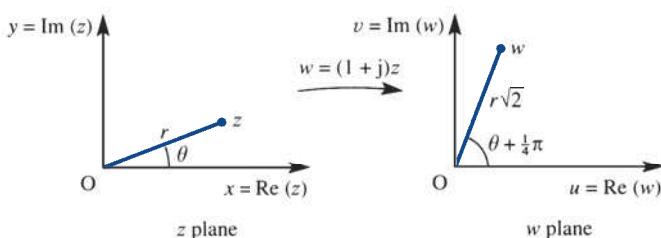
We can then readily deduce from (4.4) what the mapping does. The general point in the z plane with modulus r and argument θ is mapped onto an image point w , with modulus $r\sqrt{2}$ and argument $\theta + \frac{1}{4}\pi$ in the w plane as illustrated in Figure 4.5.

It follows that in general the mapping

$$w = \alpha z$$

maps the origin in the z plane to the origin in the w plane (fixed point), but effects an expansion by $|\alpha|$ and an anticlockwise rotation by $\arg \alpha$. Of course, $\arg \alpha$ need not be positive, and indeed it could even be zero (corresponding to α being real). The mapping can be loosely summed up in the phrase ‘magnification and rotation, but no translation’.

Figure 4.5
The mapping
 $w = (1 + j)z$.



Certain geometrical properties are also preserved, the most important being that straight lines in the z plane will be transformed to straight lines in the w plane. This is readily confirmed by noting that the equation of any straight line in the z plane can always be written in the form

$$|z - a| = |z - b|$$

where a and b are complex constants (this being the equation of the perpendicular bisector of the join of the two points representing a and b on the Argand diagram). Under the mapping $w = \alpha z$, the equation maps to

$$\left| \frac{w}{\alpha} - a \right| = \left| \frac{w}{\alpha} - b \right| \quad (\alpha \neq 0)$$

or

$$|w - a\alpha| = |w - b\alpha|$$

in the w plane, which is clearly another straight line.

We now return to the general linear mapping (4.2) and rewrite it in the form

$$w - \beta = \alpha z$$

This can be looked upon as two successive mappings: first,

$$\zeta = \alpha z$$

identical to $w = \alpha z$ considered earlier, but this time mapping points from the z plane to points in the ζ plane; secondly,

$$w = \zeta + \beta \tag{4.5}$$

mapping points in the ζ plane to points in the w plane. Elimination of ζ regains equation (4.2). The mapping (4.5) represents a translation in which the origin in the ζ plane is mapped to the point $w = \beta$ in the w plane, and the mapping of any other point in the ζ plane is obtained by adding β to the coordinates to obtain the equivalent point in the w plane. Geometrically, the mapping (4.5) is as if the ζ plane is picked up and, without rotation, the origin placed over the point β . The original axes then represent the w plane as illustrated in Figure 4.6. Obviously all curves, in particular straight lines, are preserved under this translation.

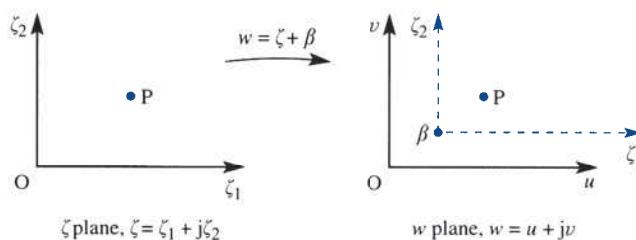
We are now in a position to interpret (4.2), the general linear mapping, geometrically as a combination of mappings that can be regarded as fundamental, namely

- translation
- rotation, and
- magnification

that is,

$$z \xrightarrow{\text{rotation}} e^{j\theta} z \xrightarrow{\text{magnification}} |\alpha| e^{j\theta} z \xrightarrow{\text{translation}} |\alpha| e^{j\theta} z + \beta = \alpha z + \beta = w$$

Figure 4.6
The mapping
 $w = \zeta + \beta$.



It clearly follows that a straight line in the z plane is mapped onto a corresponding straight line in the w plane under the linear mapping $w = \alpha z + \beta$. A second useful property of the linear mapping is that circles are mapped onto circles. To confirm this, consider the general circle

$$|z - z_0| = r$$

in the z plane, having the complex number z_0 as its centre and the real number r as its radius. Rearranging the mapping equation $w = \alpha z + \beta$ gives

$$z = \frac{w}{\alpha} - \frac{\beta}{\alpha} \quad (\alpha \neq 0)$$

so that

$$z - z_0 = \frac{w}{\alpha} - \frac{\beta}{\alpha} - z_0 = \frac{1}{\alpha}(w - w_0)$$

where $w_0 = \alpha z_0 + \beta$. Hence

$$|z - z_0| = r$$

implies

$$|w - w_0| = |\alpha|r$$

which is a circle, with centre w_0 given by the image of z_0 in the w plane and with radius $|\alpha|r$ given by the radius of the z plane circle magnified by $|\alpha|$.

We conclude this section by considering examples of linear mappings.

Example 4.2

Examine the mapping

$$w = (1 + j)z + 1 - j$$

as a succession of fundamental mappings: translation, rotation and magnification.

Solution The linear mapping can be regarded as the following sequence of simple mappings:

$$\begin{array}{ccccccc} z & \xrightarrow[\text{anticlockwise}]{\text{rotation}} & e^{j\pi/4}z & \xrightarrow[\text{by } \sqrt{2}]{\text{magnification}} & \sqrt{2}e^{j\pi/4}z & \xrightarrow[\text{translation}]{\begin{array}{l} 0 \rightarrow 1-j \text{ or} \\ (0, 0) \rightarrow (1, -1) \end{array}} & \sqrt{2}e^{j\pi/4}z + 1 - j = w \end{array}$$

Figure 4.7 illustrates this process diagrammatically. The shading in Figure 4.7 helps to identify how the z plane moves, turns and expands under this mapping. For example, the line joining the points $0 + j2$ and $1 + j0$ in the z plane has the cartesian equation

$$\frac{1}{2}y + x = 1$$

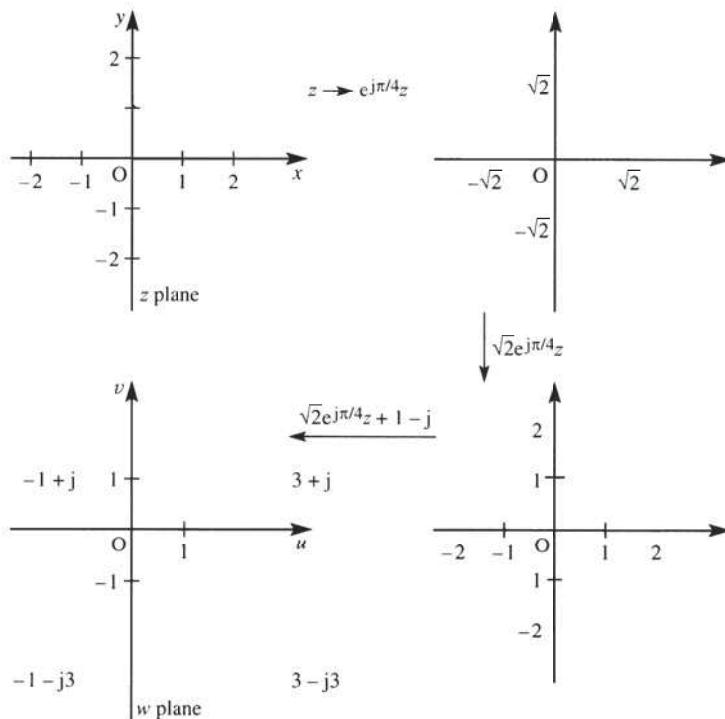
Taking $w = u + jv$ and $z = x + jy$, the mapping

$$w = (1 + j)z + 1 - j$$

becomes

$$u + jv = (1 + j)(x + jy) + 1 - j = (x - y + 1) + j(x + y - 1)$$

Figure 4.7
The mappings of Example 4.2.



Equating real and imaginary parts then gives

$$u = x - y + 1, \quad v = x + y - 1$$

which on solving for *x* and *y* gives

$$2x = u + v, \quad 2y = v - u + 2$$

Substituting for *x* and *y* into the equation $\frac{1}{2}y + x = 1$ then gives the image of this line in the *w* plane as the line

$$3v + u = 2$$

which crosses the real axis in the *w* plane at 2 and the imaginary axis at $\frac{2}{3}$. Both lines are shown dashed, in the *z* and *w* planes respectively, in Figure 4.7.

Example 4.3

The mapping $w = \alpha z + \beta$ (α, β constant complex numbers) maps the point $z = 1 + j$ to the point $w = j$, and the point $z = 1 - j$ to the point $w = -1$.

- Determine α and β .
- Find the region in the *w* plane corresponding to the right half-plane $\operatorname{Re}(z) \geq 0$ in the *z* plane.
- Find the region in the *w* plane corresponding to the interior of the unit circle $|z| < 1$ in the *z* plane.
- Find the fixed point(s) of the mapping.

In (b)–(d) use the values of α and β determined in (a).

Solution (a) The two values of z and w given as corresponding under the given linear mapping provide two equations for α and β as follows: $z = 1 + j$ mapping to $w = j$ implies

$$j = \alpha(1 + j) + \beta$$

while $z = 1 - j$ mapping to $w = -1$ implies

$$-1 = \alpha(1 - j) + \beta$$

Subtracting these two equations in α and β gives

$$j + 1 = \alpha(1 + j) - \alpha(1 - j)$$

so that

$$\alpha = \frac{1 + j}{j2} = \frac{1}{2}(1 - j)$$

Substituting back for β then gives

$$\beta = j - (1 + j)\alpha = j - \frac{1}{2}(1 - j^2) = j - 1$$

so that

$$w = \frac{1}{2}(1 - j)z + j - 1 = (1 - j)(\frac{1}{2}z - 1)$$

(b) The best way to find specific image curves in the w plane is first to express z ($= x + jy$) in terms of w ($= u + jv$) and then, by equating real and imaginary parts, to express x and y in terms of u and v . We have

$$w = (1 - j)(\frac{1}{2}z - 1)$$

which, on dividing by $1 - j$, gives

$$\frac{w}{1 - j} = \frac{1}{2}z - 1$$

Taking $w = u + jv$ and $z = x + jy$ and then rationalizing the left-hand side, we have

$$\frac{1}{2}(u + jv)(1 + j) = \frac{1}{2}(x + jy) - 1$$

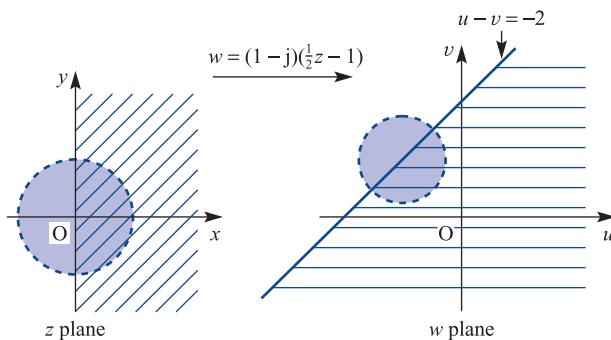
Equating real and imaginary parts then gives

$$u - v = x - 2, \quad u + v = y \tag{4.6}$$

The first of these can be used to find the image of $x \geq 0$. It is $u - v \geq -2$, which is also a region bordered by the straight line $u - v = -2$ and shown in Figure 4.8. Pick one point in the right half of the z plane, say $z = 2$, and the mapping gives $w = 0$ as the image of this point. This allays any doubts about which side of $u - v = -2$ corresponds to the right half of the z plane, $x \geq 0$. The two corresponding regions are shown ‘hatched’ in Figure 4.8.

Note that the following is always true, although we shall not prove it here. If a curve cuts the z plane in two then the corresponding curve in the w plane also cuts the w plane in two, and, further, points in one of the two distinct sets of the z plane partitioned by the curve correspond to points in just one of the similarly partitioned sets in the w plane.

Figure 4.8
The mappings of Example 4.3.



- (c) In cartesian form, with $z = x + jy$, the equation of the unit circle $|z| = 1$ is

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

Substituting for x and y from the mapping relationships (4.6) gives the image of this circle as

$$(u - v + 2)^2 + (u + v)^2 = 1$$

or

$$u^2 + v^2 + 2u - 2v + \frac{3}{2} = 0$$

which, on completing the squares, leads to

$$(u + 1)^2 + (v - 1)^2 = \frac{1}{2}$$

As expected, this is a circle, having in this particular case centre $(-1, 1)$ and radius $\sqrt{\frac{1}{2}}$. If $x^2 + y^2 < 1$ then $(u + 1)^2 + (v - 1)^2 < \frac{1}{2}$, so the region inside the circle $|z| = 1$ in the z plane corresponds to the region inside its image circle in the w plane. Corresponding regions are shown shaded in Figure 4.8.

- (d) The fixed point(s) of the mapping are obtained by putting $w = z$ in $w = \alpha z + \beta$, leading to

$$z = (\frac{1}{2}z - 1)(1 - j)$$

that is,

$$z = \frac{1}{2}z - \frac{1}{2}jz - 1 + j$$

so that

$$z = \frac{-1 + j}{\frac{1}{2} + \frac{1}{2}j} = j2$$

is the only fixed point.

One final point is in order before we leave this example. In Figure 4.8 the images of $x = 0$ and $x^2 + y^2 = 1$ can also be seen in the context of translation, rotation (the line in Figure 4.8 rotates about $z = 2j$) and magnification (in fact, shrinkage, as can be seen by the decrease in diameter of the circle compared with its image in the w plane).

4.2.2 Exercises

- 1 Find in the cartesian form $y = mx + c$ (m and c real constants) the equations of the following straight lines in the z plane, $z = x + jy$:
- $|z - 2 + j| = |z - j + 3|$
 - $z + z^* + 4j(z - z^*) = 6$
- where $*$ denotes the complex conjugate.
- 2 Find the point of intersection and the angle of intersection of the straight lines
- $$|z - 1 - j| = |z - 3 + j|$$
- $$|z - 1 + j| = |z - 3 - j|$$
- 3 The function $w = jz + 4 - 3j$ is a combination of translation and rotation. Show this diagrammatically, following the procedure used in Example 4.2. Find the image of the line $6x + y = 22$ ($z = x + jy$) in the w plane under this mapping.
- 4 Show that the mapping $w = (1 - j)z$, where $w = u + jv$ and $z = x + jy$, maps the region $y > 1$ in the z plane onto the region $u + v > 2$ in the w plane. Illustrate the regions in a diagram.
- 5 Under the mapping $w = jz + j$, where $w = u + jv$ and $z = x + jy$, show that the half-plane $x > 0$ in the z plane maps onto the half-plane $v > 1$ in the w plane.
- 6 For $z = x + jy$ find the image region in the w plane corresponding to the semi-infinite strip $x > 0$, $0 < y < 2$ in the z plane under the mapping $w = jz + 1$. Illustrate the regions in both planes.
- 7 Find the images of the following curves under the mapping
- $$w = (j + \sqrt{3})z + j\sqrt{3} - 1$$
- $y = 0$
 - $x = 0$
 - $x^2 + y^2 = 1$
 - $x^2 + y^2 + 2y = 1$
- where $z = x + jy$.
- 8 The mapping $w = \alpha z + \beta$ (α, β both constant complex numbers) maps the point $z = 1 + j$ to the point $w = j$ and the point $z = -1$ to the point $w = 1 + j$.
- Determine α and β .
 - Find the region in the w plane corresponding to the upper half-plane $\text{Im}(z) > 0$ and illustrate diagrammatically.
 - Find the region in the w plane corresponding to the disc $|z| < 2$ and illustrate diagrammatically.
 - Find the fixed point(s) of the mapping.
- In (b)–(d) use the values of α and β determined in (a).

4.2.3 Inversion

The inversion mapping is of the form

$$w = \frac{1}{z} \tag{4.7}$$

and in this subsection we shall consider the image of circles and straight lines in the z plane under such a mapping. Clearly, under this mapping the image in the w plane of the general circle

$$|z - z_0| = r$$

in the z plane, with centre at z_0 and radius r , is given by

$$\left| \frac{1}{w} - z_0 \right| = r \tag{4.8}$$

but it is not immediately obvious what shaped curve this represents in the w plane. To investigate, we take $w = u + jv$ and $z_0 = x_0 + jy_0$ in (4.8), giving

$$\left| \frac{u - jv}{u^2 + v^2} - x_0 - jy_0 \right| = r$$

Squaring we have

$$\left(\frac{u}{u^2 + v^2} - x_0 \right)^2 + \left(\frac{v}{u^2 + v^2} + y_0 \right)^2 = r^2$$

which on expanding leads to

$$\frac{u^2}{(u^2 + v^2)^2} - \frac{2ux_0}{u^2 + v^2} + x_0^2 + \frac{v^2}{(u^2 + v^2)^2} + \frac{2vy_0}{(u^2 + v^2)} + y_0^2 = r^2$$

or

$$\frac{u^2 + v^2}{(u^2 + v^2)^2} + \frac{2vy_0 - 2ux_0}{u^2 + v^2} = r^2 - x_0^2 - y_0^2$$

so that

$$(u^2 + v^2)(r^2 - x_0^2 - y_0^2) + 2ux_0 - 2vy_0 = 1 \quad (4.9)$$

The expression is a quadratic in u and v , with the coefficients of u^2 and v^2 equal and no term in uv . It therefore represents a circle, unless the coefficient of $u^2 + v^2$ is itself zero, which occurs when

$$x_0^2 + y_0^2 = r^2, \quad \text{or} \quad |z_0| = r$$

and we have

$$2ux_0 - 2vy_0 = 1$$

which represents a straight line in the w plane.

Summarizing, the inversion mapping $w = 1/z$ maps the circle $|z - z_0| = r$ in the z plane onto another circle in the w plane unless $|z_0| = r$, in which case the circle is mapped onto a straight line in the w plane that does not pass through the origin.

When $|z_0| \neq r$, we can divide the equation of the circle (4.9) in the w plane by the factor $r^2 - x_0^2 - y_0^2$ to give

$$\frac{u^2 + v^2}{r^2 - x_0^2 - y_0^2} - \frac{2y_0v}{r^2 - x_0^2 - y_0^2} = \frac{1}{r^2 - x_0^2 - y_0^2}$$

which can be written in the form

$$(u - u_0)^2 + (v - v_0)^2 = R^2$$

where (u_0, v_0) are the coordinates of the centre and R the radius of the w plane circle. It is left as an exercise for the reader to show that

$$(u_0, v_0) = \left(-\frac{x_0}{r^2 - |z_0|^2}, \frac{y_0}{r^2 - |z_0|^2} \right), \quad R = \frac{r}{r^2 - |z_0|^2}$$

Next we consider the general straight line

$$|z - a_1| = |z - a_2|$$

in the z plane, where a_1 and a_2 are constant complex numbers with $a_1 \neq a_2$. Under the mapping (4.7), this becomes the curve in the w plane represented by the equation

$$\left| \frac{1}{w} - a_1 \right| = \left| \frac{1}{w} - a_2 \right| \quad (4.10)$$

Again, it is not easy to identify this curve, so we proceed as before and take

$$w = u + jv, \quad a_1 = p + jq, \quad a_2 = r + js$$

where p, q, r and s are real constants. Substituting in (4.10) and squaring both sides, we have

$$\left(\frac{u}{u^2 + v^2} - p \right)^2 + \left(\frac{v}{u^2 + v^2} + q \right)^2 = \left(\frac{u}{u^2 + v^2} - r \right)^2 + \left(\frac{v}{u^2 + v^2} + s \right)^2$$

Expanding out each term, the squares of $u/(u^2 + v^2)$ and $v/(u^2 + v^2)$ cancel, giving

$$-\frac{2up}{u^2 + v^2} + p^2 + \frac{2vq}{u^2 + v^2} + q^2 = -\frac{2ur}{u^2 + v^2} + r^2 + \frac{2vs}{u^2 + v^2} + s^2$$

which on rearrangement becomes

$$(u^2 + v^2)(p^2 + q^2 - r^2 - s^2) + 2u(r - p) + 2v(q - s) = 0 \quad (4.11)$$

Again this represents a circle *through the origin* in the w plane, unless

$$p^2 + q^2 = r^2 + s^2$$

which implies $|a_1| = |a_2|$, when it represents a straight line, also through the origin, in the w plane. The algebraic form of the coordinates of the centre of the circle and its radius can be deduced from (4.11).

We can therefore make the important conclusion that the inversion mapping $w = 1/z$ takes circles or straight lines in the z plane onto circles or straight lines in the w plane. Further, since we have carried out the algebra, we can be more specific. If the circle in the z plane passes through the origin (that is, $|z_0| = r$ in (4.9)) then it is mapped onto a straight line that does *not* pass through the origin in the w plane. If the straight line in the z plane passes through the origin ($|a_1| = |a_2|$ in (4.11)) then it is mapped onto a straight line through the origin in the w plane. Figure 4.9 summarizes these conclusions.

To see why this is the case, we first note that the fixed points of the mapping, determined by putting $w = z$, are

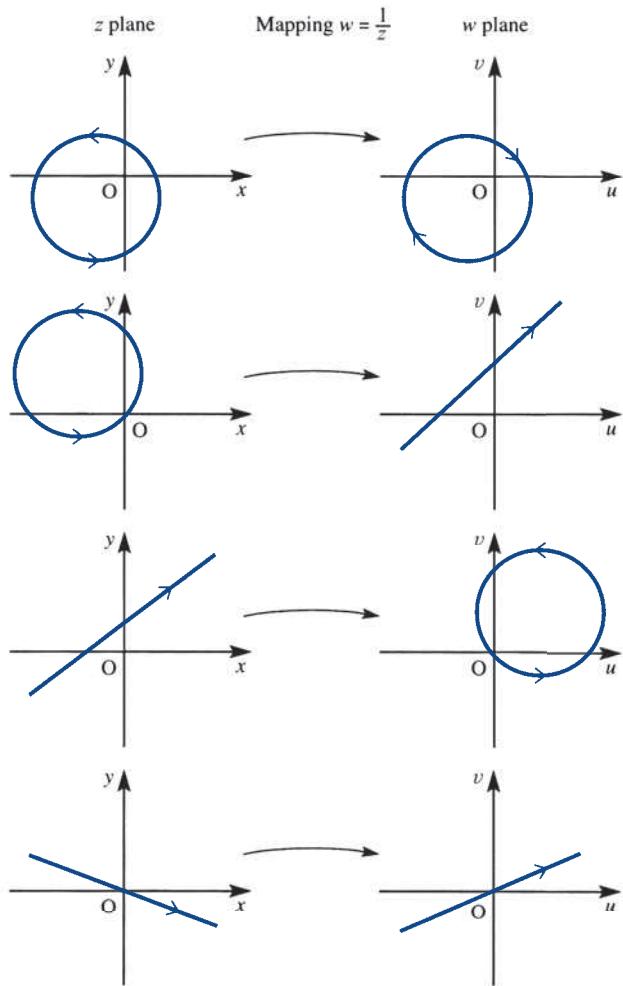
$$z = \frac{1}{z}, \quad \text{or} \quad z^2 = 1$$

so that $z = \pm 1$.

We also note that $z = 0$ is mapped to infinity in the w plane and $w = 0$ is mapped to infinity in the z plane and vice versa in both cases. Further, if we apply the mapping a second time, we get the identity mapping. That is, if

$$w = \frac{1}{z}, \quad \text{and} \quad \zeta = \frac{1}{w}$$

Figure 4.9
The inversion mapping $w = 1/z$.



then

$$\zeta = \frac{1}{1/z} = z$$

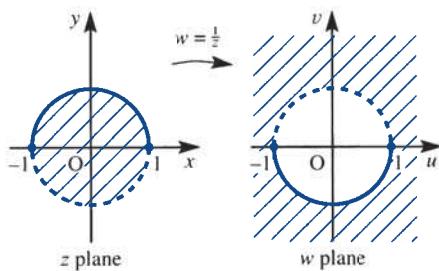
which is the identity mapping.

The inside of the unit circle in the z plane, $|z| < 1$, is mapped onto $|1/w| < 1$ or $|w| > 1$, the outside of the unit circle in the w plane. By the same token, therefore, the outside of the unit circle in the z plane $|z| > 1$ is mapped onto $|1/w| > 1$ or $|w| < 1$, the inside of the unit circle in the w plane. Points actually on $|z| = 1$ in the z plane are mapped to points on $|w| = 1$ in the w plane, with ± 1 staying fixed, as already shown. Figure 4.10 summarizes this property.

It is left as an exercise for the reader to show that the top half-boundary of $|z| = 1$ is mapped onto the bottom half-boundary of $|w| = 1$.

For any point z_0 in the z plane the point $1/z_0$ is called the **inverse of z_0 with respect to the circle $|z| = 1$** ; this is the reason for the name of the mapping. (Note the double meaning of inverse; here it means the reciprocal function and not the ‘reverse’)

Figure 4.10 Mapping of the unit circle under $w = 1/z$.



mapping.) The more general definition of inverse is that for any point z_0 in the z plane the point r^2/z_0 is the inverse of z_0 with respect to the circle $|z| = r$, where r is a real constant.

Example 4.4

Determine the image path in the w plane corresponding to the circle $|z - 3| = 2$ in the z plane under the mapping $w = 1/z$. Sketch the paths in both the z and w planes and shade the region in the w plane corresponding to the region inside the circle in the z plane.

Solution The image in the w plane of the circle $|z - 3| = 2$ in the z plane under the mapping $w = 1/z$ is given by

$$\left| \frac{1}{w} - 3 \right| = 2$$

which, on taking $w = u + jv$, gives

$$\left| \frac{u - jv}{u^2 + v^2} - 3 \right| = 2$$

Squaring both sides, we then have

$$\left(\frac{u}{u^2 + v^2} - 3 \right)^2 + \left(\frac{-v}{u^2 + v^2} \right)^2 = 4$$

or

$$\frac{u^2 + v^2}{(u^2 + v^2)^2} - \frac{6u}{u^2 + v^2} + 5 = 0$$

which reduces to

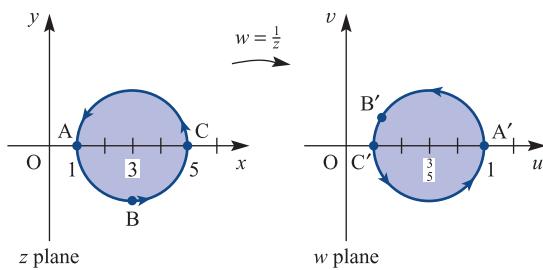
$$1 - 6u + 5(u^2 + v^2) = 0$$

or

$$(u - \frac{3}{5})^2 + v^2 = \frac{4}{25}$$

Thus the image in the w plane is a circle with centre $(\frac{3}{5}, 0)$ and radius $\frac{2}{5}$. The corresponding circles in the z and w planes are shown in Figure 4.11.

Figure 4.11
The mapping of Example 4.4.



Taking $z = x + jy$, the mapping $w = 1/z$ becomes

$$u + jv = \frac{1}{x + jy} = \frac{x - jy}{x^2 + y^2}$$

which, on equating real and imaginary parts, gives

$$u = \frac{x}{x^2 + y^2}, \quad v = \frac{-y}{x^2 + y^2}$$

We can now use these two relationships to determine the images of particular points under the mapping. In particular, the centre $(3, 0)$ of the circle in the z plane is mapped onto the point $u = \frac{1}{3}, v = 0$ in the w plane, which is inside the mapped circle. Thus, under the mapping, the region inside the circle in the z plane is mapped onto the region inside the circle in the w plane.

Further, considering three sample points $A(1 + j0)$, $B(3 - j2)$ and $C(5 + j0)$ on the circle in the z plane, we find that the corresponding image points on the circle in the w plane are $A'(1, 0)$, $B'(\frac{3}{13}, \frac{2}{13})$ and $C'(\frac{1}{5}, 0)$. Thus, as the point z traverses the circle in the z plane in an anticlockwise direction, the corresponding point w in the w plane will also traverse the mapped circle in an anticlockwise direction as indicated in Figure 4.11.

4.2.4 Bilinear mappings

A **bilinear mapping** is a mapping of the form

$$w = \frac{az + b}{cz + d} \quad (4.12)$$

where a, b, c and d are prescribed complex constants. It is called the bilinear mapping in z and w since it can be written in the form $Awz + Bw + Cz + D = 0$, which is linear in both z and w .

Clearly the bilinear mapping (4.12) is more complicated than the linear mapping given by (4.2). In fact, the general linear mapping is a special case of the bilinear mapping, since setting $c = 0$ and $d = 1$ in (4.12) gives (4.2). In order to investigate the bilinear mapping, we rewrite the right-hand side of (4.12) as follows:

$$w = \frac{az + b}{cz + d} = \frac{\frac{a}{c}(cz + d) - \frac{ad}{c} + b}{cz + d}$$

so that

$$w = \frac{a}{c} + \frac{bc - ad}{c(cz + d)} \quad (4.13)$$

This mapping clearly degenerates to $w = a/c$ unless we demand that $bc - ad \neq 0$. We therefore say that (4.12) represents a bilinear mapping provided the determinant

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

is non-zero. This is sometimes referred to as the **determinant of the mapping**. When the condition holds, the inverse mapping

$$z = \frac{-dw + b}{cw - a}$$

obtained by rearranging (4.12), is also bilinear, since

$$\begin{vmatrix} -d & b \\ c & -a \end{vmatrix} = da - cb \neq 0$$

Renaming the constants so that $\lambda = a/c$, $\mu = bc - ad$, $\alpha = c^2$ and $\beta = cd$, (4.13) becomes

$$w = \lambda + \frac{\mu}{\alpha z + \beta}$$

and we can break the mapping down into three steps as follows:

$$z_1 = \alpha z + \beta$$

$$z_2 = \frac{1}{z_1}$$

$$w = \lambda + \mu z_2$$

The first and third of these steps are linear mappings as considered in Section 4.2.1, while the second is the inversion mapping considered in Section 4.2.3. The bilinear mapping (4.12) can thus be generated from the following elementary mappings:

$$\begin{array}{ccccccc} z & \xrightarrow{\text{rotation}} & \alpha z & \xrightarrow{\text{translation}} & \alpha z + \beta & \xrightarrow{\text{inversion}} & \frac{1}{\alpha z + \beta} \\ & & \text{and} & & & & \\ & & \text{magnification} & & & & \\ \xrightarrow{\text{magnification}} & \frac{\mu}{\alpha z + \beta} & \xrightarrow{\text{translation}} & \lambda + \frac{\mu}{\alpha z + \beta} & = w & & \end{array}$$

We saw in Section 4.2.1 that the general linear transformation $w = \alpha z + \beta$ does not change the shape of the curve being mapped from the z plane onto the w plane. Also, in Section 4.2.3 we saw that the inversion mapping $w = 1/z$ maps circles or straight lines in the z plane onto circles or straight lines in the w plane. It follows that the bilinear mapping also exhibits this important property, in that it also will map circles or straight lines in the z plane onto circles or straight lines in the w plane.

Example 4.5

Investigate the mapping

$$w = \frac{z-1}{z+1}$$

by finding the images in the w plane of the lines $\operatorname{Re}(z) = \text{constant}$ and $\operatorname{Im}(z) = \text{constant}$. Find the fixed points of the mapping.

Solution

Since we are seeking specific image curves in the w plane, we first express z in terms of w and then express x and y in terms of u and v , where $z = x + jy$ and $w = u + jv$. Rearranging

$$w = \frac{z-1}{z+1}$$

gives

$$z = \frac{1+w}{1-w}$$

Taking $z = x + jy$ and $w = u + jv$, we have

$$\begin{aligned} x + jy &= \frac{1+u+jv}{1-u-jv} \\ &= \frac{1+u+jv}{1-u-jv} \frac{1-u+jv}{1-u+jv} \end{aligned}$$

which reduces to

$$x + jy = \frac{1 - u^2 - v^2}{(1-u)^2 + v^2} + j \frac{2v}{(1-u)^2 + v^2}$$

Equating real and imaginary parts then gives

$$x = \frac{1 - u^2 - v^2}{(1-u)^2 + v^2} \quad (4.14a)$$

$$y = \frac{2v}{(1-u)^2 + v^2} \quad (4.14b)$$

It follows from (4.14a) that the lines $\operatorname{Re}(z) = x = c_1$, which are parallel to the imaginary axis in the z plane, correspond to the curves

$$c_1 = \frac{1 - u^2 - v^2}{(1-u)^2 + v^2}$$

where c_1 is a constant, in the w plane. Rearranging this leads to

$$c_1(1 - 2u + u^2 + v^2) = 1 - u^2 - v^2$$

or, assuming that $1 + c_1 \neq 0$,

$$u^2 + v^2 - \frac{2c_1u}{1+c_1} + \frac{c_1-1}{c_1+1} = 0$$

which, on completing squares, gives

$$\left(u - \frac{c_1}{1 + c_1}\right)^2 + v^2 = \left(\frac{1}{1 + c_1}\right)^2$$

It is now clear that the corresponding curve in the w plane is a circle, centre $(u = c_1/(1 + c_1), v = 0)$ and radius $(1 + c_1)^{-1}$.

In the algebraic manipulation we assumed that $c_1 \neq -1$, in order to divide by $1 + c_1$. In the exceptional case $c_1 = -1$, we have $u = 1$, and the mapped curve is a straight line in the w plane parallel to the imaginary axis.

Similarly, it follows from (4.14b) that the lines $\text{Im}(z) = y = c_2$, which are parallel to the imaginary axis in the z plane, correspond to the curves

$$c_2 = \frac{2v}{(1-u)^2 + v^2}$$

where c_2 is a constant, in the w plane. Again, this usually represents a circle in the w plane, but exceptionally will represent a straight line. Rearranging the equation we have

$$(1-u)^2 + v^2 = \frac{2v}{c_2}$$

provided that $c_2 \neq 0$. Completing the square then leads to

$$(u-1)^2 + \left(v - \frac{1}{c_2}\right)^2 = \frac{1}{c_2^2}$$

which represents a circle in the w plane, centre $(u = 1, v = 1/c_2)$ and radius $1/c_2$.

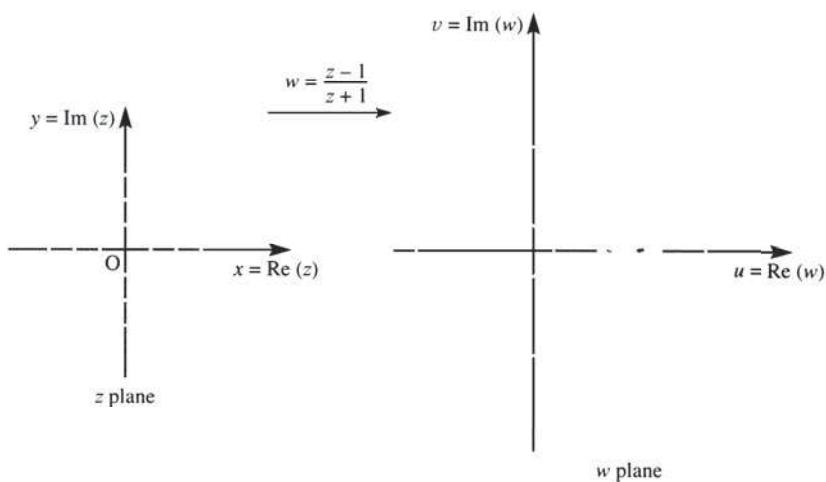
In the exceptional case $c_2 = 0, v = 0$ and we see that the real axis $y = 0$ in the z plane maps onto the real axis $v = 0$ in the w plane.

Putting a sequence of values to c_1 and then to c_2 , say -10 to $+10$ in steps of $+1$, enables us to sketch the mappings shown in Figure 4.12. The fixed points of the mapping are given by

$$z = \frac{z-1}{z+1}$$

Figure 4.12

The mapping of Example 4.5.



that is,

$$z^2 = -1, \quad \text{or} \quad z = \pm j$$

In general, all bilinear mappings will have two fixed points. However, although there are mathematically interesting properties associated with particular mappings having coincident fixed points, they do not impinge on engineering applications, so they only deserve passing reference here.

Example 4.6

Find the image in the w plane of the circle $|z| = 2$ in the z plane under the bilinear mapping

$$w = \frac{z - j}{z + j}$$

Sketch the curves in both the z and w planes and shade the region in the w plane corresponding to the region inside the circle in the z plane.

Solution Rearranging the transformation, we have

$$z = \frac{jw + j}{1 - w}$$

so that the image in the w plane of the circle $|z| = 2$ in the z plane is determined by

$$\left| \frac{jw + j}{1 - w} \right| = 2 \quad (4.15)$$

One possible way of proceeding now is to put $w = u + jv$ and proceed as in Example 4.4, but the algebra becomes a little messy. An alternative approach is to use the property of complex numbers that $|z_1/z_2| = |z_1|/|z_2|$, so that (4.15) becomes

$$|jw + j| = 2|1 - w|$$

Taking $w = u + jv$ then gives

$$|-v + j(u + 1)| = 2|(1 - u) - jv|$$

which on squaring both sides leads to

$$v^2 + (1 + u)^2 = 4[(1 - u)^2 + v^2]$$

or

$$u^2 + v^2 - \frac{10}{3}u + 1 = 0$$

Completing the square of the u term then gives

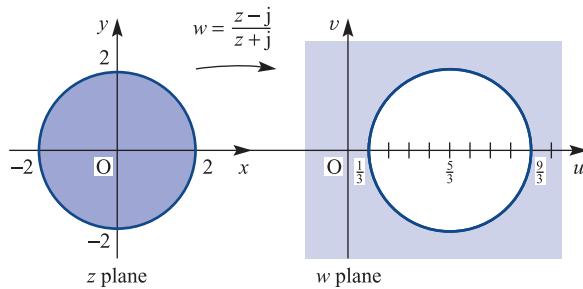
$$(u - \frac{5}{3})^2 + v^2 = \frac{16}{9}$$

indicating that the image curve in the w plane is a circle centre $(u = \frac{5}{3}, v = 0)$ and radius $\frac{4}{3}$. The corresponding circles in the z and w planes are illustrated in Figure 4.13. To identify corresponding regions, we consider the mapping of the point $z = 0 + j0$ inside the circle in the z plane. Under the given mapping, this maps to the point

$$w = \frac{0 - j}{0 + j} = -1 + j0$$

in the w plane. It then follows that the region inside the circle $|z| = 2$ in the z plane maps onto the region outside the mapped circle in the w plane.

Figure 4.13
The mapping of Example 4.6.



An interesting property of (4.12) is that there is just one bilinear transformation that maps three given distinct points z_1, z_2 and z_3 in the z plane onto three specified distinct points w_1, w_2 and w_3 respectively in the w plane. It is left as an exercise for the reader to show that the bilinear transformation is given by

$$\frac{(w - w_1)(w_2 - w_3)}{(w - w_3)(w_2 - w_1)} = \frac{(z - z_1)(z_2 - z_3)}{(z - z_3)(z_2 - z_1)} \quad (4.16)$$

The right-hand side of (4.16) is called the cross-ratio of z_1, z_2, z_3 and z . We shall illustrate with an example.

Example 4.7

Find the bilinear transformation that maps the three points $z = 0, -j$ and -1 onto the three points $w = j, 1, 0$ respectively in the w plane.

Solution Taking the transformation to be

$$w = \frac{az + b}{cz + d}$$

on using the given information on the three pairs of corresponding points we have

$$j = \frac{a(0) + b}{c(0) + d} = \frac{b}{d} \quad (4.17a)$$

$$1 = \frac{a(-j) + b}{c(-j) + d} \quad (4.17b)$$

$$0 = \frac{a(-1) + b}{c(-1) + d} \quad (4.17c)$$

From (4.17c) $a = b$; then from (4.17a)

$$d = \frac{b}{j} = -jb = -ja$$

and from (4.17b) $c = ja$. Thus

$$w = \frac{az + a}{jaz - ja} = \frac{1}{j} \frac{z + 1}{z - 1} = -j \frac{z + 1}{z - 1}$$

Alternatively, using (4.16) we can obtain

$$\frac{(w - j)(1 - 0)}{(w - 0)(1 - j)} = \frac{(z - 0)(-j + 1)}{(z + 1)(-j - 0)}$$

or

$$w = -j \frac{z + 1}{z - 1}$$

as before.

4.2.5 Exercises

- 9 Show that if $z = x + jy$, the image of the half-plane $y > c$ (c constant) under the mapping $w = 1/z$ is the interior of a circle, provided that $c > 0$. What is the image when $c = 0$ and when $c < 0$? Illustrate with sketches in the w plane.
- 10 Determine the image in the w plane of the circle $|z + \frac{3}{4} + j| = \frac{7}{4}$ under the inversion mapping $w = 1/z$.
- 11 Show that the mapping $w = 1/z$ maps the circle $|z - a| = a$, with a being a positive real constant, onto a straight line in the w plane. Sketch the corresponding curves in the z and w planes, indicating the region onto which the interior of the circle in the z plane is mapped.
- 12 Find a bilinear mapping that maps $z = 0$ to $w = j$, $z = -j$ to $w = 1$ and $z = -1$ to $w = 0$. Hence sketch the mapping by finding the images in the w plane of the lines $\operatorname{Re}(z) = \text{constant}$ and $\operatorname{Im}(z) = \text{constant}$ in the z plane. Verify that $z = \frac{1}{2}(j - 1)(-1 \pm \sqrt{3})$ are fixed points of the mapping.
- 13 The two complex variables w and z are related through the inverse mapping
- $$w = \frac{1 + j}{z} \quad 16$$
- (a) Find the images of the points $z = 1, 1 - j$ and 0 in the w plane.
(b) Find the region of the w plane corresponding to the interior of the unit circle $|z| < 1$ in the z plane.
- 14 Given the complex mapping
- $$w = \frac{z + 1}{z - 1}$$
- where $w = u + jv$ and $z = x + jy$, determine the image curve in the w plane corresponding to the semicircular arc $x^2 + y^2 = 1$ ($x \leq 0$) described from the point $(0, -1)$ to the point $(0, 1)$.
- 15 (a) Map the region in the z plane ($z = x + jy$) that lies between the lines $x = y$ and $y = 0$, with $x < 0$, onto the w plane under the bilinear mapping
- $$w = \frac{z + j}{z - 3}$$
- (Hint: Consider the point $w = \frac{2}{3}$ to help identify corresponding regions.)
- (b) Show that, under the same mapping as in (a), the straight line $3x + y = 4$ in the z plane corresponds to the unit circle $|w| = 1$ in the w plane and that the point $w = 1$ does not correspond to a finite value of z .
- If $w = (z - j)/(z + j)$, find and sketch the image in the w plane corresponding to the circle $|z| = 2$ in the z plane.
- 17 Show that the bilinear mapping
- $$w = e^{j\theta_0} \frac{z - z_0}{z - z_0^*}$$

where θ_0 is a real constant $0 \leq \theta_0 < 2\pi$, z_0 a fixed complex number and z_0^* its conjugate, maps the upper half of the z plane ($\operatorname{Im}(z) > 0$) onto the inside of the unit circle in the w plane ($|w| < 1$). Find the values of z_0 and θ_0 if $w = 0$ corresponds to $z = j$ and $w = -1$ corresponds to $z = \infty$.

- 18 Show that, under the mapping

$$w = \frac{2jz}{z + j}$$

circular arcs or the straight line through $z = 0$ and $z = j$ in the z plane are mapped onto circular arcs or the straight line through $w = 0$ and $w = j$ in the w plane. Find the images of the regions $|z - \frac{1}{2}| < \frac{1}{2}$ and $|z| < |z - j|$ in the w plane.

19

Find the most general bilinear mapping that maps the unit circle $|z| = 1$ in the z plane onto the unit circle $|w| = 1$ in the w plane and the point $z = z_0$ in the z plane to the origin $w = 0$ in the w plane.

4.2.6 The mapping $w = z^2$

There are a number of other mappings that are used by engineers. For example, in dealing with Laplace and z transforms, the subjects of Chapters 5 and 6 respectively, we are concerned with the polynomial mapping

$$w = a_0 + a_1 z + \cdots + a_n z^n$$

where a_0, a_1, \dots, a_n are complex constants, the rational function

$$w = \frac{P(z)}{Q(z)}$$

where P and Q are polynomials in z , and the exponential mapping

$$w = a e^{bz}$$

where $e = 2.71828\dots$, the base of natural logarithms. As is clear from the bilinear mapping in Section 4.2.4, even elementary mappings can be cumbersome to analyse. Fortunately, we have two factors on our side. First, very detailed tracing of specific curves and their images is not required, only images of points. Secondly, by using complex differentiation, the subject of Section 4.3, various facets of these more complicated mappings can be understood without lengthy algebra. As a prelude, in this subsection we analyse the mapping $w = z^2$, which is the simplest polynomial mapping.

Example 4.8

Investigate the mapping $w = z^2$ by plotting the images on the w plane of the lines $x = \text{constant}$ and $y = \text{constant}$ in the z plane.

Solution

There is some difficulty in inverting this mapping to get z as a function of w , since square roots lead to problems of uniqueness. However, there is no need to invert here, for taking $w = u + jv$ and $z = x + jy$, the mapping becomes

$$w = u + jv = (x + jy)^2 = (x^2 - y^2) + j2xy$$

which, on taking real and imaginary parts, gives

$$\begin{aligned} u &= x^2 - y^2 \\ v &= 2xy \end{aligned} \tag{4.18}$$

If $x = \alpha$, a real constant, then (4.18) becomes

$$u = \alpha^2 - y^2, \quad v = 2\alpha y$$

which, on eliminating y , gives

$$u = \alpha^2 - \frac{v^2}{4\alpha^2}$$

or

$$4\alpha^2 u = 4\alpha^4 - v^2$$

so that

$$v^2 = 4\alpha^4 - 4\alpha^2 u = 4\alpha^2(\alpha^2 - u)$$

This represents a parabola in the w plane, and, since the right-hand side must be positive, $\alpha^2 \geq u$ so the ‘nose’ of the parabola is at $u = \alpha^2$ on the positive real axis in the w plane.

If $y = \beta$, a real constant, then (4.18) becomes

$$u = x^2 - \beta^2, \quad v = 2x\beta$$

which, on eliminating x , gives

$$u = \frac{v^2}{4\beta^2} - \beta^2$$

or

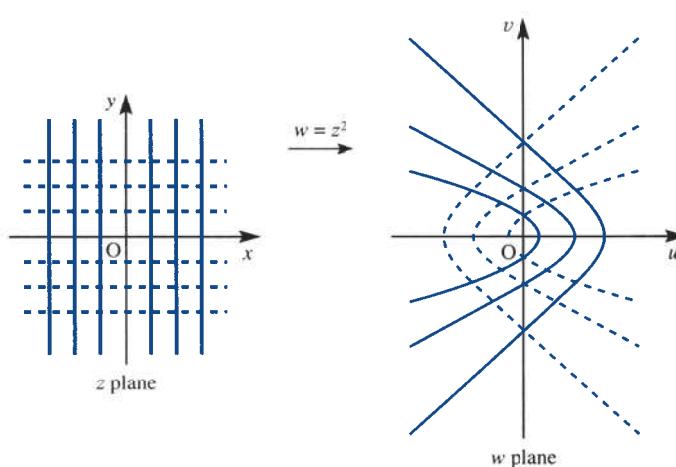
$$4\beta^2 = v^2 - 4\beta^4$$

so that

$$v^2 = 4\beta^2 u + 4\beta^4 = 4\beta^2(u + \beta^2)$$

This is also a parabola, but pointing in the opposite direction. The right-hand side, as before, must be positive, so that $u > -\beta^2$ and the ‘nose’ of the parabola is on the negative real axis. These curves are drawn in Figure 4.14.

Figure 4.14
The mapping of
Example 4.8.



We shall not dwell further on the finer points of the mapping $w = z^2$. Instead, we note that in general it is extremely difficult to plot images of curves in the z plane, even the straight lines parallel to the axes, under polynomial mappings. We also note that we do not often need to do so, and that we have done it only as an aid to understanding.

The exercises that follow should also help in understanding this topic. We shall then return to examine polynomial, rational and exponential mappings in Section 4.3.4, after introducing complex differentiation.

4.2.7 Exercises

- 20 Find the image region in the w plane corresponding to the region inside the triangle in the z plane having vertices at $0 + j0$, $2 + j0$ and $0 + j2$ under the mapping $w = z^2$. Illustrate with sketches.

- 21 Find the images of the lines $y = x$ and $y = -x$ under the mapping $w = z^2$. Also find the image of the general line through the origin $y = mx$. By putting $m = \tan \theta_0$, deduce that straight lines intersecting at the origin in the z plane map onto lines intersecting at the origin in the w plane, but that the angle between these image lines is double that between the original lines.

- 22 Consider the mapping $w = z^n$, where n is an integer (a generalization of the mapping $w = z^2$). Use the polar representation of complex numbers to show that

- (a) Circles centred at the origin in the z plane are mapped onto circles centred at the origin in the w plane.

- (b) Straight lines passing through the origin intersecting with angle θ_0 in the z plane are mapped onto straight lines passing through the origin in the w plane but intersecting at an angle $n\theta_0$.

- 23 If the complex function

$$w = \frac{1+z^2}{z}$$

is represented by a mapping from the z plane onto the w plane, find u in terms of x and y , and v in terms of x and y , where $z = x + jy$, $w = u + jv$. Find the image of the unit circle $|z| = 1$ in the w plane. Show that the circle centred at the origin, of radius r , in the z plane ($|z| = r$) is mapped onto the curve

$$\left(\frac{r^2 u}{r^2 + 1} \right)^2 + \left(\frac{r^2 v}{r^2 - 1} \right)^2 = r^2 \quad (r \neq 1)$$

in the w plane. What kind of curves are these? What happens for very large r ?

4.3 Complex differentiation

The derivative of a real function $f(x)$ of a single real variable x at $x = x_0$ is given by the limit

$$f'(x_0) = \lim_{x \rightarrow x_0} \left[\frac{f(x) - f(x_0)}{x - x_0} \right]$$

Here, of course, x_0 is a real number and so can be represented by a single point on the real line. The point representing x can then approach the fixed x_0 either from the left or from the right along this line. Let us now turn to complex variables and functions depending on them. We know that a plane is required to represent complex numbers, so z_0 is now a fixed point in the Argand diagram, somewhere in the plane. The definition of the derivative of the function $f(z)$ of the complex variable z at the point z_0 will thus be

$$f'(z_0) = \lim_{z \rightarrow z_0} \left[\frac{f(z) - f(z_0)}{z - z_0} \right]$$

It may appear that if we merely exchange z for x , the rest of this section will follow similar lines to the differentiation of functions of real variables. For real variables taking the limit could only be done from the left or from the right, and the existence of a unique limit was not difficult to establish. For complex variables, however, the point that represents the fixed complex number z_0 can be approached along an infinite number of curves in the z plane. The existence of a unique limit is thus a very stringent requirement. That most complex functions can be differentiated in the usual way is a remarkable property of the complex variable. Since $z = x + jy$, and x and y can vary independently, there are some connections with the calculus of functions of two real variables, but we shall not pursue this connection here.

Rather than use the word ‘differentiable’ to describe complex functions for which a derivative exists, if the function $f(z)$ has a derivative $f'(z)$ that exists at all points of a region R of the z plane then $f(z)$ is called **analytic** in R . Other terms such as **regular** or **holomorphic** are also used as alternatives to analytic. (Strictly, functions that have a power series expansion – see Section 4.4.1 – are called **analytic functions**. Since differentiable functions have a power series expansion they are referred to as analytic functions. However, there are examples of analytic functions that are not differentiable.)

4.3.1 Cauchy–Riemann equations

The following result is an important property of the analytic function.

If $z = x + jy$ and $f(z) = u(x, y) + jv(x, y)$, and $f(z)$ is analytic in some region R of the z plane, then the two equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \quad (4.19)$$

known as the **Cauchy–Riemann equations**, hold throughout R .

It is instructive to prove this result. Since $f'(z)$ exists at any point z_0 in R ,

$$f'(z_0) = \lim_{z \rightarrow z_0} \left[\frac{f(z) - f(z_0)}{z - z_0} \right]$$

where z can tend to z_0 along any path within R . Examination of (4.19) suggests that we might choose paths parallel to the x direction and parallel to the y direction, since these will lead to partial derivatives with respect to x and y . Thus, choosing $z - z_0 = \Delta x$, a real path, we see that

$$f'(z_0) = \lim_{\Delta x \rightarrow 0} \left[\frac{f(z_0 + \Delta x) - f(z_0)}{\Delta x} \right]$$

Since $f(z) = u + jv$, this means that

$$f'(z_0) = \lim_{\Delta x \rightarrow 0} \left[\frac{u(x_0 + \Delta x, y_0) + jv(x_0 + \Delta x, y_0) - u(x_0, y_0) - jv(x_0, y_0)}{\Delta x} \right]$$

or, on splitting into real and imaginary parts,

$$f'(z_0) = \lim_{\Delta x \rightarrow 0} \left[\frac{u(x_0 + \Delta x, y_0) - u(x_0, y_0)}{\Delta x} + j \frac{v(x_0 + \Delta x, y_0) - v(x_0, y_0)}{\Delta x} \right]$$

giving

$$f'(z_0) = \left[\frac{\partial u}{\partial x} + j \frac{\partial v}{\partial x} \right]_{x=x_0, y=y_0} \quad (4.20)$$

Starting again from the definition of $f'(z_0)$, but this time choosing $z - z_0 = j\Delta y$ for the path parallel to the y axis, we obtain

$$f'(z_0) = \lim_{j\Delta y \rightarrow 0} \left[\frac{f(z_0 + j\Delta y) - f(z_0)}{j\Delta y} \right]$$

Once again, using $f(z) = u + jv$ and splitting into real and imaginary parts, we see that

$$\begin{aligned} f'(z_0) &= \lim_{j\Delta y \rightarrow 0} \left[\frac{u(x_0, y_0 + \Delta y) + jv(x_0, y_0 + \Delta y) - u(x_0, y_0) - jv(x_0, y_0)}{j\Delta y} \right] \\ &= \lim_{\Delta y \rightarrow 0} \left[\frac{1}{j} \frac{u(x_0, y_0 + \Delta y) - u(x_0, y_0)}{\Delta y} + \frac{v(x_0, y_0 + \Delta y) - v(x_0, y_0)}{\Delta y} \right] \end{aligned}$$

giving

$$f'(z_0) = \left[\frac{1}{j} \frac{\partial u}{\partial y} + \frac{\partial v}{\partial y} \right]_{x=x_0, y=y_0} \quad (4.21)$$

Since $f'(z_0)$ must be the same no matter what path is followed, the two values obtained in (4.20) and (4.21) must be equal. Hence

$$\frac{\partial u}{\partial x} + j \frac{\partial v}{\partial x} = \frac{1}{j} \frac{\partial u}{\partial y} + \frac{\partial v}{\partial y} = -j \frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}$$

Equating real and imaginary parts then gives the required Cauchy–Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}$$

at the point $z = z_0$. However, z_0 is an arbitrarily chosen point in the region R ; hence the Cauchy–Riemann equations hold throughout R , and we have thus proved the required result.

It is tempting to think that should we choose more paths along which to let $z - z_0$ tend to zero, we could derive more relationships along the same lines as the Cauchy–Riemann equations. It turns out, however, that we merely reproduce them or expressions derivable from them, and it is possible to prove that satisfaction of the Cauchy–Riemann equations (4.19) is a necessary condition for a function $f(z) = u(x, y) + jv(x, y)$, $z = x + jy$, to be analytic in a specified region. At points where $f'(z)$ exists it may be obtained from either (4.20) or (4.21) as

$$f'(z) = \frac{\partial u}{\partial x} + j \frac{\partial v}{\partial x}$$

or

$$f'(z) = \frac{\partial v}{\partial y} - j \frac{\partial u}{\partial y}$$

If z is given in the polar form $z = r e^{j\theta}$ then

$$f(z) = u(r, \theta) + jv(r, \theta)$$

and the corresponding polar forms of the Cauchy–Riemann equations are

$$\frac{\partial u}{\partial r} = \frac{1}{r} \frac{\partial v}{\partial \theta}, \quad \frac{\partial v}{\partial r} = -\frac{1}{r} \frac{\partial u}{\partial \theta} \quad (4.22)$$

At points where $f'(z)$ exists it may be obtained from either of

$$f'(z) = e^{-j\theta} \left(\frac{\partial u}{\partial r} + j \frac{\partial v}{\partial r} \right) \quad (4.23a)$$

or

$$f'(z) = e^{-j\theta} \left(\frac{1}{r} \frac{\partial v}{\partial \theta} - j \frac{1}{r} \frac{\partial u}{\partial \theta} \right) \quad (4.23b)$$

Example 4.9

Verify that the function $f(z) = z^2$ satisfies the Cauchy–Riemann equations, and determine the derivative $f'(z)$.

Solution Since $z = x + jy$, we have

$$f(z) = z^2 = (x + jy)^2 = (x^2 - y^2) + j2xy$$

so if $f(z) = u(x, y) + jv(x, y)$ then

$$u = x^2 - y^2, \quad v = 2xy$$

giving the partial derivatives as

$$\frac{\partial u}{\partial x} = 2x, \quad \frac{\partial u}{\partial y} = -2y$$

$$\frac{\partial v}{\partial x} = 2y, \quad \frac{\partial v}{\partial y} = 2x$$

It is readily seen that the Cauchy–Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$$

are satisfied.

The derivative $f'(z)$ is then given by

$$f'(z) = \frac{\partial u}{\partial x} + j \frac{\partial v}{\partial x} = 2x + j2y = 2z$$

as expected.

Example 4.10

Verify that the exponential function $f(z) = e^{\alpha z}$, where α is a constant, satisfies the Cauchy–Riemann equations, and show that $f'(z) = \alpha e^{\alpha z}$.

Solution

$$f(z) = u + jv = e^{\alpha z} = e^{\alpha(x+jy)} = e^{\alpha x} e^{j\alpha y} = e^{\alpha x} (\cos \alpha y + j \sin \alpha y)$$

so, equating real and imaginary parts,

$$u = e^{\alpha x} \cos \alpha y, \quad v = e^{\alpha x} \sin \alpha y$$

The partial derivatives are

$$\frac{\partial u}{\partial x} = \alpha e^{\alpha x} \cos \alpha y, \quad \frac{\partial v}{\partial x} = \alpha e^{\alpha x} \sin \alpha y$$

$$\frac{\partial u}{\partial y} = -\alpha e^{\alpha x} \sin \alpha y, \quad \frac{\partial v}{\partial y} = \alpha e^{\alpha x} \cos \alpha y$$

confirming that the Cauchy–Riemann equations are satisfied. The derivative $f'(z)$ is then given by

$$f'(z) = \frac{\partial u}{\partial x} + j \frac{\partial v}{\partial x} = \alpha e^{\alpha x} (\cos \alpha y + j \sin \alpha y) = \alpha e^{\alpha z}$$

so that

$$\frac{d}{dz} e^{\alpha z} = \alpha e^{\alpha z} \quad (4.24)$$

As in the real variable case, we have (see Section 4.3.1)

$$e^{jz} = \cos z + j \sin z \quad (4.25)$$

so that $\cos z$ and $\sin z$ may be expressed as

$$\left. \begin{aligned} \cos z &= \frac{e^{jz} + e^{-jz}}{2} \\ \sin z &= \frac{e^{jz} - e^{-jz}}{2j} \end{aligned} \right\} \quad (4.26a)$$

Using result (4.24) from Example 4.10, it is then readily shown that

$$\frac{d}{dz} (\sin z) = \cos z$$

$$\frac{d}{dz} (\cos z) = -\sin z$$

Similarly, we define the hyperbolic functions $\sinh z$ and $\cosh z$ by

$$\left. \begin{aligned} \sinh z &= \frac{e^z - e^{-z}}{2} = -j \sin jz \\ \cosh z &= \frac{e^z + e^{-z}}{2} = \cos jz \end{aligned} \right\} \quad (4.26b)$$

from which, using (4.24), it is readily deduced that

$$\frac{d}{dz}(\sinh z) = \cosh z$$

$$\frac{d}{dz}(\cosh z) = \sinh z$$

We note from above that e^z has the following real and imaginary parts:

$$\operatorname{Re}(e^z) = e^x \cos y$$

$$\operatorname{Im}(e^z) = e^x \sin y$$

In real variables the exponential and circular functions are contrasted, one being monotonic, the other oscillatory. In complex variables, however, the real and imaginary parts of e^z are (two-variable) combinations of exponential and circular functions, which might seem surprising for an exponential function. Similarly, the circular functions of a complex variable have unfamiliar properties. For example, it is easy to see that $|\cos z|$ and $|\sin z|$ are unbounded for complex z by using the above relationships between circular and hyperbolic functions of complex variables. Contrast this with $|\cos x| \leq 1$ and $|\sin x| \leq 1$ for a real variable x .

In a similar way to the method adopted in Examples 4.9 and 4.10 it can be shown that the derivatives of the majority of functions $f(x)$ of a real variable x carry over to the complex variable case $f(z)$ at points where $f(z)$ is analytic. Thus, for example,

$$\frac{d}{dz} z^n = nz^{n-1}$$

for all z in the z plane, and

$$\frac{d}{dz} \ln z = \frac{1}{z}$$

for all z in the z plane except for points on the non-positive real axis, where $\ln z$ is non-analytic.

It can also be shown that the rules associated with derivatives of a function of a real variable, such as the sum, product, quotient and chain rules, carry over to the complex variable case. Thus,

$$\begin{aligned}\frac{d}{dz}[f(z) + g(z)] &= \frac{df(z)}{dz} + \frac{dg(z)}{dz} \\ \frac{d}{dz}[f(z)g(z)] &= f(z)\frac{dg(z)}{dz} + \frac{df(z)}{dz}g(z) \\ \frac{d}{dz}f(g(z)) &= \frac{df}{dg}\frac{dg}{dz} \\ \frac{d}{dz}\left[\frac{f(z)}{g(z)}\right] &= \frac{g(z)f'(z) - f(z)g'(z)}{[g(z)]^2}\end{aligned}$$

4.3.2 Conjugate and harmonic functions

A pair of functions $u(x, y)$ and $v(x, y)$ of the real variables x and y that satisfy the Cauchy–Riemann equations (4.19) are said to be **conjugate functions**. (Note here the different use of the word ‘conjugate’ to that used in complex number work, where $z^* = x - jy$ is the complex conjugate of $z = x + jy$.) Conjugate functions satisfy the orthogonality property in that the curves in the (x, y) plane defined by $u(x, y) = \text{constant}$ and $v(x, y) = \text{constant}$ are orthogonal curves. This follows since the gradient at any point on the curve $u(x, y) = \text{constant}$ is given by

$$\left[\frac{dy}{dx} \right]_u = -\frac{\partial u}{\partial y} / \frac{\partial u}{\partial x}$$

and the gradient at any point on the curve $v(x, y) = \text{constant}$ is given by

$$\left[\frac{dy}{dx} \right]_v = -\frac{\partial v}{\partial y} / \frac{\partial v}{\partial x}$$

It follows from the Cauchy–Riemann equations (4.19) that

$$\left[\frac{dy}{dx} \right]_u \left[\frac{dy}{dx} \right]_v = -1$$

so the curves are orthogonal.

A function that satisfies the Laplace equation in two dimensions is said to be **harmonic**; that is, $u(x, y)$ is a harmonic function if

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

It is readily shown (see Example 4.12) that if $f(z) = u(x, y) + jv(x, y)$ is analytic, so that the Cauchy–Riemann equations are satisfied, then both u and v are **harmonic** functions. Therefore u and v are **conjugate harmonic functions**. Harmonic functions have applications in such areas as stress analysis in plates, inviscid two-dimensional fluid flow and electrostatics.

Example 4.11

Given $u(x, y) = x^2 - y^2 + 2x$, find the conjugate function $v(x, y)$ such that $f(z) = u(x, y) + jv(x, y)$ is an analytic function of z throughout the z plane.

Solution

We are given $u(x, y) = x^2 - y^2 + 2x$, and, since $f(z) = u + jv$ is to be analytic, the Cauchy–Riemann equations must hold. Thus, from (4.19),

$$\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x} = 2x + 2$$

Integrating this with respect to y gives

$$v = 2xy + 2y + F(x)$$

where $F(x)$ is an arbitrary function of x , since the integration was performed holding x constant. Differentiating v partially with respect to x gives

$$\frac{\partial v}{\partial x} = 2y + \frac{dF}{dx}$$

but this equals $-\partial u/\partial y$ by the second of the Cauchy–Riemann equations (4.19). Hence

$$\frac{\partial u}{\partial y} = -2y - \frac{dF}{dx}$$

But since $u = x^2 - y^2 + 2x$, $\partial u/\partial y = -2y$, and comparison yields $F(x) = \text{constant}$. This constant is set equal to zero, since no conditions have been given by which it can be determined. Hence

$$u(x, y) + jv(x, y) = x^2 - y^2 + 2x + j(2xy + 2y)$$

To confirm that this is a function of z , note that $f(z)$ is $f(x + jy)$, and becomes just $f(x)$ if we set $y = 0$. Therefore we set $y = 0$ to obtain

$$f(x + j0) = f(x) = u(x, 0) + jv(x, 0) = x^2 + 2x$$

and it follows that

$$f(z) = z^2 + 2z$$

which can be easily checked by separation into real and imaginary parts.

Example 4.12

Show that the real and imaginary parts $u(x, y)$ and $v(x, y)$ of a complex analytic function $f(z)$ are harmonic.

Solution

Since

$$f(z) = u(x, y) + jv(x, y)$$

is analytic, the Cauchy–Riemann equations

$$\frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}, \quad \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}$$

are satisfied. Differentiating the first with respect to x gives

$$\frac{\partial^2 v}{\partial x^2} = -\frac{\partial^2 u}{\partial x \partial y} = -\frac{\partial^2 u}{\partial y \partial x} = -\frac{\partial}{\partial y} \left(\frac{\partial u}{\partial x} \right)$$

which is $-\partial^2 v/\partial y^2$, by the second Cauchy–Riemann equation. Hence

$$\frac{\partial^2 v}{\partial x^2} = -\frac{\partial^2 v}{\partial y^2}, \quad \text{or} \quad \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$$

and v is a harmonic function.

Similarly,

$$\frac{\partial^2 u}{\partial y^2} = -\frac{\partial^2 v}{\partial y \partial x} = -\frac{\partial}{\partial x} \left(\frac{\partial v}{\partial y} \right) = -\frac{\partial^2 u}{\partial x^2}$$

so that

$$\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} = 0$$

and u is also a harmonic function. We have assumed that both u and v have continuous second-order partial derivatives, so that

$$\frac{\partial^2 u}{\partial x \partial y} = \frac{\partial^2 u}{\partial y \partial x}, \quad \frac{\partial^2 v}{\partial x \partial y} = \frac{\partial^2 v}{\partial y \partial x}$$

4.3.3 Exercises

- 24 Determine whether the following functions are analytic, and find the derivative where appropriate:

- (a) $z e^z$ (b) $\sin 4z$
 (c) zz^* (d) $\cos 2z$

- 25 Determine the constants a and b in order that

$$w = x^2 + ay^2 - 2xy + j(bx^2 - y^2 + 2xy)$$

be analytic. For these values of a and b find the derivative of w , and express both w and dw/dz as functions of $z = x + jy$.

- 26 Find a function $v(x, y)$ such that, given $u = 2x(1 - y)$, $f(z) = u + jv$ is analytic in z .

- 27 Show that $\phi(x, y) = e^y(x \cos y - y \sin y)$ is a harmonic function, and find the conjugate harmonic function $\psi(x, y)$. Write $\phi(x, y) + j\psi(x, y)$ as a function of $z = x + jy$ only.

- 28 Show that $u(x, y) = \sin x \cosh y$ is harmonic. Find the harmonic conjugate $v(x, y)$ and express $w = u + jv$ as a function of $z = x + jy$.

- 29 Find the orthogonal trajectories of the following families of curves:

- (a) $x^3y - xy^3 = \alpha$ (constant α)
 (b) $e^{-x}\cos y + xy = \alpha$ (constant α)

- 30 Find the real and imaginary parts of the functions

- (a) $z^2 e^{2z}$
 (b) $\sin 2z$

Verify that they are analytic and find their derivatives.

- 31 Give a definition of the inverse sine function $\sin^{-1} z$ for complex z . Find the real and imaginary parts of $\sin^{-1} z$. (Hint: Put $z = \sin w$, split into real and imaginary parts, and with $w = u + jv$ and $z = x + jy$ solve for u and v in terms of x and y .) Is $\sin^{-1} z$ analytic? If so, what is its derivative?

- 32 Establish that if $z = x + jy$,
 $|\sinh y| \leq |\sin z| \leq \cosh y$.

4.3.4 Mappings revisited

In Section 4.2 we examined mappings from the z plane to the w plane, where in the main the relationship between w and z , $w = f(z)$ was linear or bilinear. There is an important property of mappings, hinted at in Example 4.8 when considering the mapping $w = z^2$. A mapping $w = f(z)$ that preserves angles is called **conformal**. Under such a mapping, the angle between two intersecting curves in the z plane is the same as the angle between the corresponding intersecting curves in the w plane. The sense of the angle is also preserved. That is, if θ is the angle between curves 1 and 2 taken in the anticlockwise sense in the z plane then θ is also the angle between the image of curve 1 and the image of curve 2 in the w plane, and it too is taken in the anticlockwise sense.

Figure 4.15
Conformal mappings.

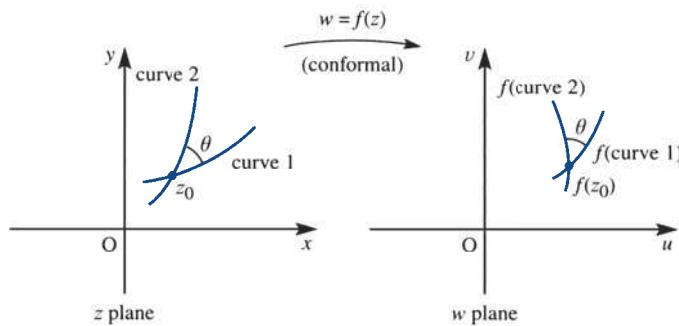


Figure 4.15 should make the idea of a conformal mapping clearer. If $f(z)$ is analytic then $w = f(z)$ defines a conformal mapping except at points where the derivative $f'(z)$ is zero.

Clearly the linear mappings

$$w = \alpha z + \beta \quad (\alpha \neq 0)$$

are conformal everywhere, since $dw/dz = \alpha$ and is not zero for any point in the z plane. Bilinear mappings given by (4.12) are not so straightforward to check. However, as we saw in Section 4.2.4, (4.12) can be rearranged as

$$w = \lambda + \frac{\mu}{\alpha z + \beta} \quad (\alpha, \mu \neq 0)$$

Thus

$$\frac{dw}{dz} = -\frac{\mu \alpha}{(\alpha z + \beta)^2}$$

which again is never zero for any point in the z plane. In fact, the only mapping we have considered so far that has a point at which it is not conformal everywhere is $w = z^2$ (cf. Example 4.8), which is not conformal at $z = 0$.

Example 4.13

Determine the points at which the mapping $w = z + 1/z$ is not conformal and demonstrate this by considering the image in the w plane of the real axis in the z plane.

Solution Taking $z = x + jy$ and $w = u + jv$, we have

$$w = u + jv = x + jy + \frac{x - jy}{x^2 + y^2}$$

which, on equating real and imaginary parts, gives

$$u = x + \frac{x}{x^2 + y^2}$$

$$v = y - \frac{y}{x^2 + y^2}$$

The real axis, $y = 0$, in the z plane corresponds to $v = 0$, the real axis in the w plane. Note, however, that the fixed point of the mapping is given by

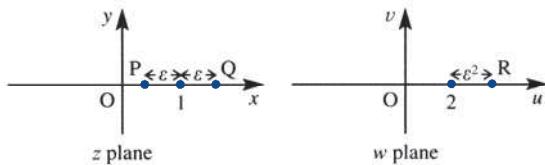
$$z = z + \frac{1}{z}$$

or $z = \infty$. From the Cauchy–Riemann equations it is readily shown that w is analytic everywhere except at $z = 0$. Also, $dw/dz = 0$ when

$$1 - \frac{1}{z^2} = 0, \quad \text{that is} \quad z = \pm 1$$

which are both on the real axis. Thus the mapping fails to be conformal at $z = 0$ and $z = \pm 1$. The image of $z = 1$ is $w = 2$, and the image of $z = -1$ is $w = -2$. Consideration of the image of the real axis is therefore perfectly adequate, since this is a curve passing through each point where $w = z + 1/z$ fails to be conformal. It would be satisfying if we could analyse this mapping in the same manner as we did with $w = z^2$ in Example 4.8. Unfortunately, we cannot do this, because the algebra gets unwieldy (and, indeed, our knowledge of algebraic curves is also too scanty). Instead, let us look at the image of the point $z = 1 + \varepsilon$, where ε is a small real number. $\varepsilon > 0$ corresponds to the point Q just to the right of $z = 1$ on the real axis in the z plane, and the point P just to the left of $z = 1$ corresponds to $\varepsilon < 0$ (Figure 4.16).

Figure 4.16 Image of $z = 1 + \varepsilon$ of Example 4.13.



If $z = 1 + \varepsilon$ then

$$\begin{aligned} w &= 1 + \varepsilon + \frac{1}{1 + \varepsilon} \\ &= 1 + \varepsilon + (1 + \varepsilon)^{-1} \\ &= 1 + \varepsilon + 1 - \varepsilon + \varepsilon^2 - \varepsilon^3 + \dots \\ &\approx 2 + \varepsilon^2 \end{aligned}$$

if $|\varepsilon|$ is much smaller than 1 (we shall discuss the validity of the power series expansion in Section 4.4). Whether ε is positive or negative, the point $w = 2 + \varepsilon^2$ is to the right of $w = 2$ in the w plane as indicated by the point R in Figure 4.16. Therefore, as $\varepsilon \rightarrow 0$, a curve (the real axis) that passes through $z = 1$ in the z plane making an angle $\theta = \pi$ corresponds to a curve (again the real axis) that approaches $w = 2$ in the w plane along the real axis from the right making an angle $\theta = 0$. Non-conformality has thus been confirmed. The treatment of $z = -1$ follows in an identical fashion, so the details are omitted. Note that when $y = 0$ ($v = 0$), $u = x + 1/x$ so, as the real axis in the z plane is traversed from $x = -\infty$ to $x = 0$, the real axis in the w plane is traversed from

$u = -\infty$ to -2 and back to $u = -\infty$ again (when $x = -1$, u reaches -2). As the real axis in the z plane is traversed from $x = 0$ through $x = 1$ to $x = +\infty$, so the real axis in the w plane is traversed from $u = +\infty$ to $u = +2$ ($x = 1$) back to $u = \infty$ again. Hence the points on the real axis in the w plane in the range $-2 < u < 2$ do not correspond to real values of z . Solving $u = x + 1/x$ for x gives

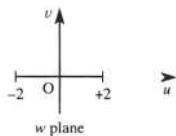


Figure 4.17 Image in w plane of the real axis in the z plane for Example 4.13.

$$x = \frac{1}{2} [u \pm \sqrt{(u^2 - 4)}]$$

which makes this point obvious. Figure 4.17 shows the image in the w plane of the real axis in the z plane. This mapping is very rich in interesting properties, but we shall not pursue it further here. Aeronautical engineers may well meet it again if they study the flow around an aerofoil in two dimensions, for this mapping takes circles centred at the origin in the z plane onto meniscus (lens-shaped) regions in the w plane, and only a slight alteration is required before these images become aerofoil-shaped.

Example 4.14

Examine the mapping

$$w = e^z$$

by (a) finding the images in the w plane of the lines $x = \text{constant}$ and $y = \text{constant}$ in the z plane, and (b) finding the image in the w plane of the left half-plane ($x < 0$) in the z plane.

Solution

Taking $z = x + jy$ and $w = u + jv$, for $w = e^z$ we have

$$u = e^x \cos y$$

$$v = e^x \sin y$$

Squaring and adding these two equations, we obtain

$$u^2 + v^2 = e^{2x}$$

On the other hand, dividing the two equations gives

$$\frac{v}{u} = \tan y$$

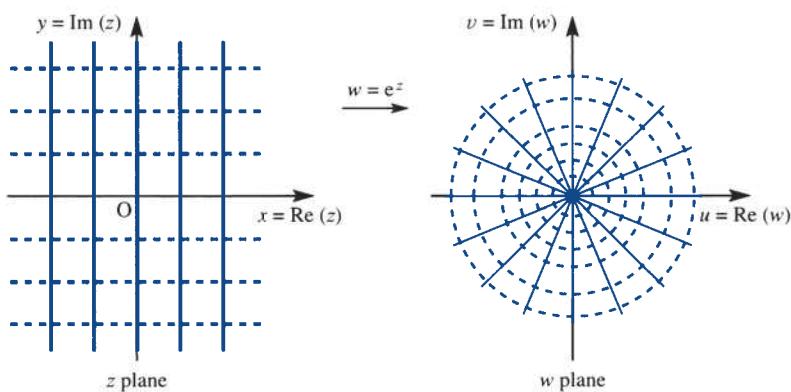
We can now tackle the questions.

- (a) Since $u^2 + v^2 = e^{2x}$, putting $x = \text{constant}$ shows that the lines parallel to the imaginary axis in the z plane correspond to circles centred at the origin in the w plane. The equation

$$\frac{v}{u} = \tan y$$

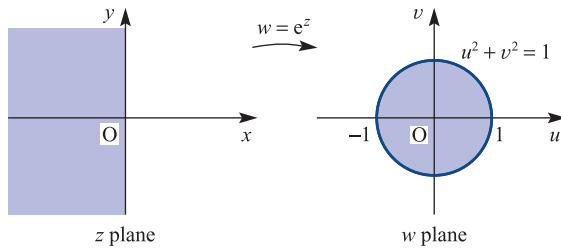
shows that the lines parallel to the real axis in the z plane correspond to straight lines through the origin in the w plane ($v = u \tan \alpha$ if $y = \alpha$, a constant). Figure 4.18 shows the general picture.

Figure 4.18 Mapping of lines under $w = e^z$ for Example 4.14.



- (b) Since $u^2 + v^2 = e^{2x}$, if $x = 0$ then $u^2 + v^2 = 1$, so the imaginary axis in the z plane corresponds to the unit circle in the w plane. If $x < 0$ then $e^{2x} < 1$, and as $x \rightarrow -\infty$, $e^{2x} \rightarrow 0$, so the left half of the z plane corresponds to the interior of the unit circle in the w plane, as illustrated in Figure 4.19.

Figure 4.19 Mapping of half-plane under $w = e^z$ for Example 4.14.



4.3.5 Exercises

33 Determine the points at which the following mappings are *not* conformal:

- (a) $w = z^2 - 1$ (b) $w = 2z^3 - 21z^2 + 72z + 6$
 (c) $w = 8z + \frac{1}{2z^2}$

34 Follow Example 4.13 for the mapping $w = z - 1/z$. Again determine the points at which the mapping is not conformal, but this time demonstrate this by looking at the image of the *imaginary* axis.

35 Find the region of the w plane corresponding to the following regions of the z plane under the exponential mapping $w = e^z$:

- (a) $0 \leq x < \infty$ (b) $0 \leq x \leq 1, 0 \leq y \leq 1$
 (c) $\frac{1}{2}\pi \leq y \leq \pi, 0 \leq x < \infty$

36 Consider the mapping $w = \sin z$. Determine the points at which the mapping is not conformal. By finding the images in the w plane of the lines $x = \text{constant}$ and $y = \text{constant}$ in the z plane ($z = x + jy$), draw the mapping along similar lines to Figures 4.14 and 4.18.

37 Show that the transformation

$$z = \zeta + \frac{a^2}{\zeta}$$

where $z = x + jy$ and $\zeta = R e^{j\theta}$ maps a circle, with centre at the origin and radius a , in the ζ plane, onto a straight-line segment in the z plane. What is the length of the line? What happens if the circle in the ζ plane is centred at the origin but is of radius b , where $b \neq a$?

4.4 Complex series

In *Modern Engineering Mathematics* (MEM) we saw that there were distinct advantages in being able to express a function $f(x)$, such as the exponential, trigonometric and logarithmic functions, of a real variable x in terms of its power series expansion

$$f(x) = \sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots + a_r x^r + \cdots \quad (4.27)$$

Power series are also very important in dealing with complex functions. In fact, any real function $f(x)$ which has a power series of the form in (4.27) has a corresponding complex function $f(z)$ having the same power series expansion, that is

$$f(z) = \sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + \cdots + a_r z^r + \cdots \quad (4.28)$$

This property enables us to extend real functions to the complex case, so that methods based on power series expansions have a key role to play in formulating the theory of complex functions. In this section we shall consider some of the properties of the power series expansion of a complex function by drawing, wherever possible, an analogy with the power series expansion of the corresponding real function.

4.4.1 Power series

A series having the form

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots + a_r(z - z_0)^r + \cdots \quad (4.29)$$

in which the coefficients a_n are real or complex and z_0 is a fixed point in the complex z plane is called a **power series** about z_0 or a power series centred on z_0 . Where $z_0 = 0$, the series (4.29) reduces to the series (4.28), which is a power series centred at the origin. In fact, on making the change of variable $z' = z - z_0$, (4.29) takes the form (4.28), so there is no loss of generality in considering the latter below.

Tests for the convergence or divergence of complex power series are similar to those used for power series of a real variable. However, in complex series it is essential that the modulus $|a_n|$ be used. For example, the geometric series

$$\sum_{n=0}^{\infty} z^n$$

has a sum to N terms

$$S_N = \sum_{n=0}^{N-1} z^n = \frac{1 - z^N}{1 - z}$$

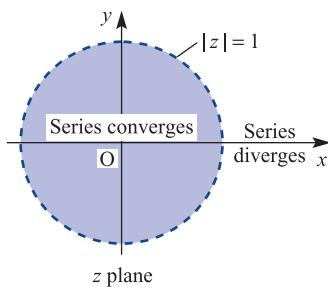
and converges, if $|z| < 1$, to the limit $1/(1-z)$ as $N \rightarrow \infty$. If $|z| \geq 1$, the series diverges. These results appear to be identical with the requirement that $|x| < 1$ to ensure convergence of the real power series

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$$

However, in the complex case the geometrical interpretation is different in that the condition $|z| < 1$ implies that z lies inside the circle centred at the origin and radius 1 in the z plane. Thus the series $\sum_{n=0}^{\infty} z^n$ converges if z lies inside this circle and diverges if z lies on or outside it. The situation is illustrated in Figure 4.20.

Figure 4.20

Region of convergence of $\sum_{n=0}^{\infty} z^n$.



The existence of such a circle leads to an important concept in that in general there exists a circle centred at the origin and of radius R such that the series

$$\sum_{n=0}^{\infty} a_n z^n \quad \begin{cases} \text{converges if } |z| < R \\ \text{diverges if } |z| > R \end{cases}$$

The radius R is called the **radius of convergence** of the power series; what happens when $|z| = R$ is normally investigated as a special case.

We have introduced the radius of convergence based on a circle centred at the origin, while the concept obviously does not depend on the location of the centre of the circle. If the series is centred on z_0 as in (4.29) then the convergence circle would be centred on z_0 . Indeed it could even be centred at infinity, when the power series becomes

$$\sum_{n=0}^{\infty} a_n z^{-n} = a_0 + \frac{a_1}{z} + \frac{a_2}{z^2} + \cdots + \frac{a_r}{z^r} + \cdots$$

which we shall consider further in Section 4.4.5.

In order to determine the radius of convergence R for a given series, various tests for convergence, such as those introduced in MEM for real series, may be applied. In particular, using d'Alembert's ratio test, it can be shown that the radius of convergence R of the complex series $\sum_{n=0}^{\infty} a_n z^n$ is given by

$$R = \lim_{n \rightarrow \infty} \left| \frac{a_n}{a_{n+1}} \right| \quad (4.30)$$

provided that the limit exists. Then the series is convergent within the disc $|z| < R$. In general, of course, the limit may not exist, and in such cases an alternative method must be used.

Example 4.15

Find the power series, in the form indicated, representing the function $1/(z - 3)$ in the following three regions:

$$(a) \quad |z| < 3; \quad \sum_{n=0}^{\infty} a_n z^n$$

$$(b) \quad |z - 2| < 1; \quad \sum_{n=0}^{\infty} a_n (z - 2)^n$$

$$(c) \quad |z| > 3; \quad \sum_{n=0}^{\infty} \frac{a_n}{z^n}$$

and sketch these regions on an Argand diagram.

Solution We know that the binomial series expansion

$$(1 + z)^n = 1 + nz + \frac{n(n-1)}{2!} z^2 + \dots + \frac{n(n-1)(n-2)\dots(n-r+1)}{r!} z^r + \dots$$

is valid for $|z| < 1$. To solve the problem, we exploit this result by expanding the function $1/(z - 3)$ in three different ways:

$$(a) \quad \frac{1}{z-3} = \frac{-\frac{1}{3}}{1-\frac{1}{3}z} = -\frac{1}{3}(1-\frac{1}{3}z)^{-1} = -\frac{1}{3}[1 + \frac{1}{3}z + (\frac{1}{3}z)^2 + \dots + (\frac{1}{3}z)^n + \dots]$$

for $|\frac{1}{3}z| < 1$, that is $|z| < 3$, giving the power series

$$\frac{1}{z-3} = -\frac{1}{3} - \frac{1}{9}z - \frac{1}{27}z^2 - \dots \quad (|z| < 3)$$

$$(b) \quad \frac{1}{z-3} = \frac{1}{(z-2)-1} = [(z-2)-1]^{-1} \\ = -[1 + (z-2) + (z-2)^2 + \dots] \quad (|z-2| < 1)$$

giving the power series

$$\frac{1}{z-3} = -1 - (z-2) - (z-2)^2 - \dots \quad (|z-2| < 1)$$

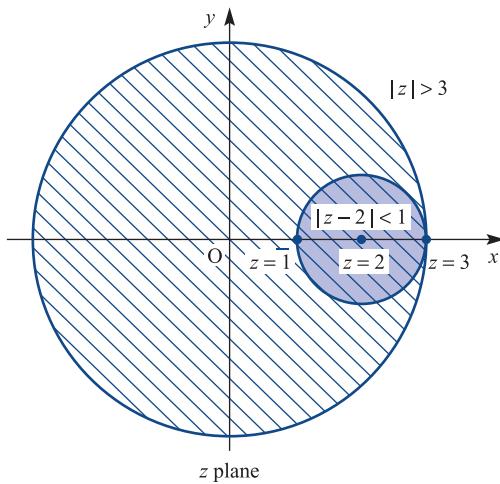
$$(c) \quad \frac{1}{z-3} = \frac{1/z}{1-3/z} = \frac{1}{z} \left[1 + \frac{3}{z} + \left(\frac{3}{z}\right)^2 + \dots \right]$$

giving the power series

$$\frac{1}{z-3} = \frac{1}{z} + \frac{3}{z^2} + \frac{9}{z^3} + \dots \quad (|z| > 3)$$

The three regions are sketched in Figure 4.21. Note that none of the regions includes the point $z = 3$, which is termed a **singularity** of the function, a concept we shall discuss in Section 4.5.1.

Figure 4.21 Regions of convergence for the series in Example 4.15.



In Example 4.15 the whole of the circle $|z| = 3$ was excluded from the three regions where the power series converge. In fact, it is possible to include any selected point in the z plane as a centre of the circle in which to define a power series that converges to $1/(z - 3)$ everywhere inside the circle, with the exception of the point $z = 3$. For example, the point $z = 4j$ would lead to the expansion of

$$\frac{1}{z - 4j + 4j - 3} = \frac{1}{4j - 3} \frac{1}{\frac{z - 4j}{4j - 3} + 1}$$

in a binomial series in powers of $(z - 4j)/(4j - 3)$, which converges to $1/(z - 3)$ inside the circle

$$|z - 4j| = |4j - 3| = \sqrt{(16 + 9)} = 5$$

We should not expect the point $z = 3$ to be included in any of the circles, since the function $1/(z - 3)$ is infinite there and hence not defined.

Example 4.16

Prove that both the power series $\sum_{n=0}^{\infty} a_n z^n$ and the corresponding series of derivatives $\sum_{n=0}^{\infty} n a_n z^{n-1}$ have the same radius of convergence.

Solution

Let R be the radius of convergence of the power series $\sum_{n=0}^{\infty} a_n z^n$. Since $\lim_{n \rightarrow \infty} (a_n z_0^n) = 0$ (otherwise the series has no chance of convergence), if $|z_0| < R$ for some complex number z_0 then it is always possible to choose

$$|a_n| < |z_0|^{-n}$$

for $n > N$, with N a fixed integer. We now use d'Alembert's ratio test, namely

$$\text{if } \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| < 1 \quad \text{then} \quad \sum_{n=0}^{\infty} a_n z^n \quad \text{converges}$$

$$\text{if } \lim_{n \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| > 1 \quad \text{then} \quad \sum_{n=0}^{\infty} a_n z^n \quad \text{diverges}$$

The differentiated series $\sum_{n=0}^{\infty} na_n z^{n-1}$ satisfies

$$\sum_{n=1}^{\infty} |na_n z^{n-1}| < \sum_{n=1}^{\infty} n|a_n| |z|^{n-1} < \sum_{n=1}^{\infty} n \frac{|z|^{n-1}}{|z_0|^n}$$

which, by the ratio test, converges if $0 < |z_0| < R$, since $|z| < |z_0|$ and $|z_0|$ can be as close to R as we choose. If, however, $|z| > R$ then $\lim_{n \rightarrow \infty} (a_n z^n) \neq 0$ and thus $\lim_{n \rightarrow \infty} (na_n z^{n-1}) \neq 0$ too. Hence R is also the radius of convergence of the differentiated series $\sum_{n=1}^{\infty} na_n z^{n-1}$.

The result obtained in Example 4.16 is important, since if the complex function

$$f(z) = \sum_{n=0}^{\infty} a_n z^n$$

converges in $|z| < R$ then the derivative

$$f'(z) = \sum_{n=1}^{\infty} na_n z^{n-1}$$

also converges in $|z| < R$. We can go on differentiating $f(z)$ through its power series and be sure that the differentiated function and the differentiated power series are equal inside the circle of convergence.

4.4.2 Exercises

- | | |
|---|--|
| <p>38 Find the power series representation for the function $1/(z - j)$ in the regions</p> <ul style="list-style-type: none"> (a) $z < 1$ (b) $z > 1$ (c) $z - 1 - j < \sqrt{2}$ | <p>39 Find the power series representation of the function $f(z) = \frac{1}{z^2 + 1}$ in the disc $z < 1$. Use Example 4.16 to deduce the power series for</p> |
|---|--|

Deduce that the radius of convergence of the power series representation of this function is $|z_0 - j|$, where $z = z_0$ is the centre of the circle of convergence ($z_0 \neq j$).

$$(a) \frac{1}{(z^2 + 1)^2} \quad (b) \frac{1}{(z^2 + 1)^3}$$

valid in this same disc.

4.4.3 Taylor series

In MEM we introduced the Taylor series expansion

$$f(x + a) = f(a) + \frac{x}{1!} f^{(1)}(a) + \frac{x^2}{2!} f^{(2)}(a) + \dots = \sum_{n=0}^{\infty} \frac{x^n}{n!} f^{(n)}(a) \quad (4.31)$$

of a function $f(x)$ of a real variable x about $x = a$ and valid within the interval of convergence of the power series. For the engineer the ability to express a function in such a power series expansion is seen to be particularly useful in the development of numerical methods and the assessment of errors. The ability to express a complex

function as a Taylor series is also important to engineers in many fields of applications, such as control and communications theory. The form of the Taylor series in the complex case is identical with that of (4.31).

If $f(z)$ is a complex function analytic inside and on a simple closed curve C (usually a circle) in the z plane then it follows from Example 4.16 that the higher derivatives of $f(z)$ also exist inside C . If z_0 and $z_0 + h$ are two fixed points inside C then

$$f(z_0 + h) = f(z_0) + hf^{(1)}(z_0) + \frac{h^2}{2!}f^{(2)}(z_0) + \cdots + \frac{h^n}{n!}f^{(n)}(z_0) + \cdots$$

where $f^{(k)}(z_0)$ is the k th derivative of $f(z)$ evaluated at $z = z_0$. Normally, $z = z_0 + h$ is introduced so that $h = z - z_0$, and the series expansion then becomes

$$\begin{aligned} f(z) &= f(z_0) + (z - z_0)f^{(1)}(z_0) + \frac{(z - z_0)^2}{2!}f^{(2)}(z_0) + \cdots \\ &\quad + \frac{(z - z_0)^n}{n!}f^{(n)}(z_0) + \cdots = \sum_{n=0}^{\infty} \frac{(z - z_0)^n}{n!}f^{(n)}(z_0) \end{aligned} \quad (4.32)$$

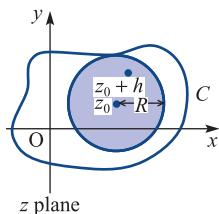


Figure 4.22 Region of convergence of the Taylor series.

The power series expansion (4.32) is called the **Taylor series expansion** of the complex function $f(z)$ about z_0 . The region of convergence of this series is $|z - z_0| < R$, a disc centred on $z = z_0$ and of radius R , the radius of convergence. Figure 4.22 illustrates the region of convergence. When $z_0 = 0$, as in real variables, the series expansion about the origin is often called a **Maclaurin series expansion**.

Since the proof of the Taylor series expansion does not add to our understanding of how to apply the result to the solution of engineering problems, we omit it at this stage.

Example 4.17

Determine the Taylor series expansion of the function

$$f(z) = \frac{1}{z(z - 2j)}$$

about the point $z = j$:

- (a) directly up to the term $(z - j)^4$,
- (b) using the binomial expansion.

Determine the radius of convergence.

Solution

- (a) The disadvantage with functions other than the most straightforward is that obtaining their derivatives is prohibitively complicated in terms of algebra. It is easier in this particular case to resolve the given function into partial fractions as

$$f(z) = \frac{1}{z(z - 2j)} = \left(\frac{1}{2j} \frac{1}{z - 2j} - \frac{1}{z} \right)$$

The right-hand side is now far easier to differentiate repeatedly. Proceeding to determine $f^{(k)}(j)$, we have

$$\begin{aligned} f(z) &= \left(\frac{1}{2j} \frac{1}{z-2j} - \frac{1}{z} \right), \quad \text{so that } f(j) = 1 \\ f^{(1)}(z) &= \frac{1}{2j} \left[-\frac{1}{(z-2j)^2} + \frac{1}{z^2} \right], \quad \text{so that } f^{(1)}(j) = 0 \\ f^{(2)}(z) &= \frac{1}{2j} \left[\frac{2}{(z-2j)^3} - \frac{2}{z^3} \right], \quad \text{so that } f^{(2)}(j) = -2 \\ f^{(3)}(z) &= \frac{1}{2j} \left[-\frac{6}{(z-2j)^4} + \frac{6}{z^4} \right], \quad \text{so that } f^{(3)}(j) = 0 \\ f^{(4)}(z) &= \frac{1}{2j} \left[\frac{24}{(z-2j)^5} - \frac{24}{z^5} \right], \quad \text{so that } f^{(4)}(j) = 24 \end{aligned}$$

leading from (4.32) to the Taylor series expansion

$$\begin{aligned} \frac{1}{z(z-2j)} &= 1 - \frac{2}{2!}(z-j)^2 + \frac{24}{4!}(z-j)^4 + \dots \\ &= 1 - (z-j)^2 + (z-j)^4 + \dots \end{aligned}$$

- (b) To use the binomial expansion, we first express $z(z-2j)$ as $(z-j+j)(z-j-j)$, which, being the difference of two squares $((z-j)^2 - j^2)$, leads to

$$f(z) = \frac{1}{z(z-2j)} = \frac{1}{(z-j)^2 + 1} = [1 + (z-j)^2]^{-1}$$

Use of the binomial expansion then gives

$$f(z) = 1 - (z-j)^2 + (z-j)^4 - (z-j)^6 + \dots$$

valid for $|z-j| < 1$, so the radius of convergence is 1.

The points where $f(z)$ is infinite (its singularities) are precisely at distance 1 away from $z=j$, so this value for the radius of convergence comes as no surprise.

Example 4.18

Suggest a function to represent the power series

$$1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots + \frac{z^n}{n!} + \dots$$

and determine its radius of convergence.

Solution Set

$$f(z) = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots = \sum_{n=0}^{\infty} \frac{z^n}{n!}$$

Assuming we can differentiate the series for $f(z)$ term by term, we obtain

$$f'(z) = \sum_{n=1}^{\infty} \frac{nz^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = f(z)$$

Hence $f(z)$ is its own derivative. Since e^x is its own derivative in real variables, and is the only such function, it seems sensible to propose that

$$f(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!} = e^z \quad (4.33)$$

the complex exponential function. Indeed the complex exponential e^z is defined by the power series (4.33). According to d'Alembert's ratio test the series $\sum_{n=0}^{\infty} a_n$ is convergent if $|a_{n+1}/a_n| \rightarrow L < 1$ as $n \rightarrow \infty$, where L is a real constant. If $a_n = z^n/n!$ then $|a_{n+1}/a_n| = |z|/(n+1)$ which is less than unity for sufficiently large n , no matter how big $|z|$ is. Hence $\sum_{n=0}^{\infty} z^n/n!$ is convergent for all z and so has an infinite radius of convergence. Note that this is confirmed from (4.30). Such functions are called **entire**.

In the same way as we define the exponential function e^z by the power series expansion (4.31), we can define the circular functions $\sin z$ and $\cos z$ by the power series expansions

$$\begin{aligned} \sin z &= z - \frac{z^3}{3!} + \frac{z^5}{5!} - \frac{z^7}{7!} + \cdots + (-1)^n \frac{z^{2n+1}}{(2n+1)!} + \cdots \\ \cos z &= 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - \frac{z^6}{6!} + \cdots + (-1)^n \frac{z^{2n}}{(2n)!} + \cdots \end{aligned}$$

both of which are valid for all z . Using these power series definitions, we can readily prove the result (4.25), namely

$$e^{iz} = \cos z + i \sin z$$

4.4.4 Exercises

- 40 Find the first four non-zero terms of the Taylor series expansions of the following functions about the points indicated, and determine the radius of convergence in each case:

(a) $\frac{1}{1+z}$ ($z = 1$) (b) $\frac{1}{z(z-4j)}$ ($z = 2j$)

(c) $\frac{1}{z^2}$ ($z = 1+j$)

- 41 Find the Maclaurin series expansion of the function

$$f(z) = \frac{1}{1+z+z^2}$$

up to and including the term in z^3 .

- 42 Without explicitly finding each Taylor series expansion, find the radius of convergence of the function

$$f(z) = \frac{1}{z^4 - 1}$$

about the three points $z = 0$, $z = 1+j$ and $z = 2+2j$. Why is there no Taylor series expansion of this function about $z = j$?

- 43 Determine a Maclaurin series expansion of $f(z) = \tan z$. What is its radius of convergence?

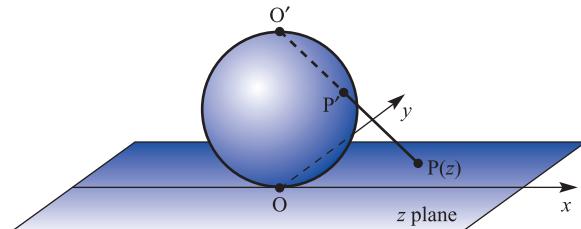
4.4.5 Laurent series

Let us now examine more closely the solution of Example 4.15(c), where the power series obtained was

$$\frac{1}{z-3} = \frac{1}{z} + \frac{3}{z^2} + \frac{9}{z^3} + \dots$$

valid for $|z| > 3$. In the context of the definition, this is a power series about ' $z = \infty$ ', the 'point at infinity'. Some readers, quite justifiably, may not be convinced that there is a single unique point at infinity. Figure 4.23 shows what is termed the **Riemann sphere**. A sphere lies on the complex z plane, with the contact point at the origin O . Let O' be the top of the sphere, at the diametrically opposite point to O . Now, for any arbitrarily chosen point P in the z plane, by joining O' and P we determine a unique point P' where the line $O'P$ intersects the sphere. There is thus exactly one point P' on the sphere corresponding to each P in the z plane. The point O' itself is the only point on the sphere that does not have a corresponding point on the (finite) z plane; we therefore say it corresponds to the **point at infinity** on the z plane.

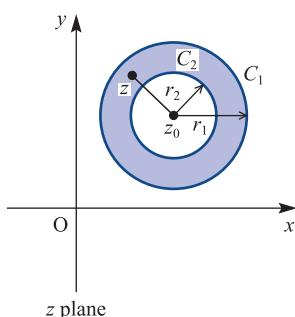
Figure 4.23
The Riemann sphere.



Returning to consider power series, we know that, inside the radius of convergence, a given function and its Taylor series expansion are identically equal. Points at which a function fails to be analytic are called **singularities**, which we shall discuss in Section 4.5.1. No Taylor series expansion is possible about a singularity. Indeed, a Taylor series expansion about a point z_0 at which a function is analytic is only valid within a circle, centre z_0 , up to the nearest singularity. Thus all singularities must be excluded in any Taylor series consideration. The Laurent series representation includes (or at least takes note of) the behaviour of the function in the vicinity of a singularity.

If $f(z)$ is a complex function analytic on concentric circles C_1 and C_2 of radii r_1 and r_2 (with $r_2 < r_1$), centred at z_0 , and also analytic throughout the region between the circles (that is, an annular region), then for each point z within the annulus (Figure 4.24) $f(z)$ may be represented by the **Laurent series**

Figure 4.24 Region of validity of the Laurent series.



$$\begin{aligned}
 f(z) &= \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n \\
 &= \cdots + \frac{c_{-r}}{(z - z_0)^r} + \frac{c_{-r+1}}{(z - z_0)^{r-1}} + \cdots + \frac{c_{-1}}{z - z_0} + c_0 \\
 &\quad + c_1(z - z_0) + \cdots + c_r(z - z_0)^r + \cdots
 \end{aligned} \tag{4.34}$$

where in general the coefficients c_r are complex. The annular shape of the region is necessary in order to exclude the point $z = z_0$, which may be a singularity of $f(z)$, from consideration. If $f(z)$ is analytic at $z = z_0$ then $c_n = 0$ for $n = -1, -2, \dots, -\infty$, and the Laurent series reduces to the Taylor series.

The Laurent series (4.34) for $f(z)$ may be written as

$$f(z) = \sum_{n=-\infty}^{-1} c_n (z - z_0)^n + \sum_{n=0}^{\infty} c_n (z - z_0)^n$$

and the first sum on the right-hand side, the ‘non-Taylor’ part, is called the **principal part** of the Laurent series.

Of course, we can seldom actually sum a series to infinity. There is therefore often more than theoretical interest in the so-called ‘remainder terms’, these being the difference between the first n terms of a power series and the exact value of the function. For both Taylor and Laurent series these remainder terms are expressed, as in the case of real variables, in terms of the $(n + 1)$ th derivative of the function itself.

Example 4.19

For $f(z) = 1/z^2(z + 1)$ find the Laurent series expansion about (a) $z = 0$ and (b) $z = -1$. Determine the region of validity in each case.

Solution

As with Example 4.15, problems such as this are tackled by making use of the binomial series expansion

$$(1 + z)^n = 1 + nz + \frac{n(n-1)}{2!}z^2 + \cdots + \frac{n(n-1)(n-2)\cdots(n-r+1)}{r!}z^r + \cdots$$

provided that $|z| < 1$.

- (a) In this case $z_0 = 0$, so we need a series in powers of z . Thus

$$\begin{aligned}
 \frac{1}{z^2(1+z)} &= \frac{1}{z^2}(1+z)^{-1} \\
 &= \frac{1}{z^2}(1 - z + z^2 - z^3 + z^4 - \cdots) \quad (0 < |z| < 1)
 \end{aligned}$$

Thus the required Laurent series expansion is

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

valid for $0 < |z| < 1$. The value $z = 0$ must be excluded because of the first two terms of the series. The region $0 < |z| < 1$ is an example of a **punctured disc**, a common occurrence in this branch of mathematics.

- (b) In this case $z_0 = -1$, so we need a series in powers of $(z + 1)$. Thus

$$\begin{aligned}\frac{1}{z^2(z+1)} &= \frac{1}{(z+1)}(z+1-1)^{-2} \\ &= \frac{1}{(z+1)}[1-(z+1)]^{-2} \\ &= \frac{1}{(z+1)}[1+2(z+1)+3(z+1)^2+\dots] \\ &= \frac{1}{z+1}+2+3(z+1)+4(z+1)^2+\dots\end{aligned}$$

valid for $0 < |z+1| < 1$. Note that in a meniscus-shaped region (that is, the region of overlap between the two circular regions $|z| < 1$ and $|z+1| < 1$) both Laurent series are simultaneously valid. This is quite typical, and not a cause for concern.

Example 4.20

Determine the Laurent series expansions of

$$f(z) = \frac{1}{(z+1)(z+3)}$$

valid for

- (a) $1 < |z| < 3$
- (b) $|z| > 3$
- (c) $0 < |z+1| < 2$
- (d) $|z| < 1$

Solution (a) Resolving into partial functions,

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

Since $|z| > 1$ and $|z| < 3$, we express this as

$$\begin{aligned}f(z) &= \frac{1}{2z}\left(\frac{1}{1+1/z}\right) - \frac{1}{6}\left(\frac{1}{1+\frac{1}{3}z}\right) \\ &= \frac{1}{2z}\left(1+\frac{1}{z}\right)^{-1} - \frac{1}{6}\left(1+\frac{1}{3}z\right)^{-1} \\ &= \frac{1}{2z}\left(1-\frac{1}{z}+\frac{1}{z^2}-\frac{1}{z^3}+\dots\right) - \frac{1}{6}\left(1-\frac{1}{3}z+\frac{1}{9}z^2-\frac{1}{27}z^3+\dots\right) \\ &= \dots - \frac{1}{2z^4} + \frac{1}{2z^3} - \frac{1}{2z^2} + \frac{1}{2z} - \frac{1}{6} + \frac{1}{18}z - \frac{1}{54}z^2 + \frac{1}{162}z^3 - \dots\end{aligned}$$

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

Since $|z| > 3$, we express this as

$$\begin{aligned} f(z) &= \frac{1}{2z} \left(\frac{1}{1+1/z} \right) - \frac{1}{2z} \left(\frac{1}{1+3/z} \right) \\ &= \frac{1}{2z} \left(1 + \frac{1}{z} \right)^{-1} - \frac{1}{2z} \left(1 + \frac{3}{z} \right)^{-1} \\ &= \frac{1}{2z} \left(1 - \frac{1}{z} + \frac{1}{z^2} - \frac{1}{z^3} + \dots \right) - \frac{1}{2z} \left(1 - \frac{3}{z} + \frac{9}{z^2} - \frac{27}{z^3} + \dots \right) \\ &= \frac{1}{z^2} - \frac{4}{z^3} + \frac{13}{z^4} - \frac{40}{z^5} + \dots \end{aligned}$$

- (c) We can proceed as in Example 4.18. Alternatively, we can take $z+1 = u$; then $0 < |u| < 2$ and

$$\begin{aligned} f(u) &= \frac{1}{u(u+2)} = \frac{1}{2u(1+\frac{1}{2}u)} \\ &= \frac{1}{2u} \left(1 - \frac{1}{2}u + \frac{1}{4}u^2 - \frac{1}{8}u^3 + \dots \right) \end{aligned}$$

giving

$$f(z) = \frac{1}{2(z+1)} - \frac{1}{4} + \frac{1}{8}(z+1) - \frac{1}{16}(z+1)^2 + \dots$$

(d) $f(z) = \frac{1}{2(z+1)} - \frac{1}{2(z+3)}$

Since $|z| < 1$, we express this as

$$\begin{aligned} f(z) &= \frac{1}{2(1+z)} - \frac{1}{6(1+\frac{1}{3}z)} \\ &= \frac{1}{2}(1+z)^{-1} - \frac{1}{6}(1+\frac{1}{3}z)^{-1} \\ &= \frac{1}{2}(1-z+z^2-z^3+\dots) - \frac{1}{6}(1-\frac{1}{3}z+\frac{1}{9}z^2-\frac{1}{27}z^3+\dots) \\ &= \frac{1}{3} - \frac{4}{9}z + \frac{13}{27}z^2 - \frac{40}{81}z^3 + \dots \end{aligned}$$

Example 4.21

Determine the Laurent series expansion of the function $f(z) = z^3 e^{1/z}$ about

- (a) $z = 0$
- (b) $z = a$, a finite, non-zero complex number
- (c) $z = \infty$

Solution (a) From (4.33),

$$e^z = 1 + z + \frac{z^2}{2!} + \dots \quad (0 \leq |z| < \infty)$$

Substituting $1/z$ for z , we obtain

$$e^{1/z} = 1 + \frac{1}{z} + \frac{1}{2!z^2} + \dots \quad (0 < |z| \leq \infty)$$

so that

$$z^3 e^{1/z} = z^3 + z^2 + \frac{z}{2!} + \frac{1}{3!} + \frac{1}{4!z} + \frac{1}{5!z^2} + \dots \quad (0 < |z| \leq \infty)$$

This series has infinitely many terms in its principal part, but stops at z^3 (it is written back to front). Series with never-ending principal parts are a problem, and fortunately are uncommon in engineering. Note also that the series is valid in an infinite punctured disc.

- (b) The value of $f(a)$ must be $a^3 e^{1/a}$, which is not infinite since $a \neq 0$. Therefore $f(z)$ has a Taylor series expansion

$$f(z) = f(a) + (z - a)f^{(1)}(a) + \frac{(z - a)^2}{2!}f^{(2)}(a) + \dots$$

about $z = a$. We have

$$f^{(1)}(z) = \frac{d}{dz}(z^3 e^{1/z}) = 3z^2 e^{1/z} - z e^{1/z}$$

$$f^{(2)}(z) = \frac{d}{dz}(3z^2 e^{1/z} - z e^{1/z}) = 6z e^{1/z} - 4 e^{1/z} + \frac{1}{z^2} e^{1/z}$$

giving the series as

$$\begin{aligned} z^3 e^{1/z} &= a^3 e^{1/a} + (z - a)(3a^2 e^{1/a} - a e^{1/a}) \\ &\quad + \frac{1}{2!}(z - a)^2 \left(6a e^{1/a} - 4 e^{1/a} + \frac{1}{a^2} e^{1/a} \right) + \dots \end{aligned}$$

which is valid in the region $|z - a| < R$, where R is the distance between the origin, where $f(z)$ is not defined, and the point a ; hence $R = |a|$. Thus the region of validity for this Taylor series is the disc $|z - a| < |a|$.

- (c) To expand about $z = \infty$, let $w = 1/z$, so that

$$f(z) = \frac{1}{w^3} e^w$$

Expanding about $w = 0$ then gives

$$\begin{aligned} f\left(\frac{1}{w}\right) &= \frac{1}{w^3} \left(1 + w + \frac{w^2}{2!} + \frac{w^3}{3!} + \dots \right) \\ &= \frac{1}{w^3} + \frac{1}{w^2} + \frac{1}{2!w} + \frac{1}{3!} + \frac{w}{4!} + \dots \quad (0 < |w| < \infty) \end{aligned}$$

Note that this time there are only three terms in the principal part of $f(z) (= f(1/w))$.

4.4.6 Exercises

44 Determine the Laurent series expansion of

$$f(z) = \frac{1}{z(z-1)^2}$$

about (a) $z=0$ and (b) $z=1$, and specify the region of validity for each.

45 Determine the Laurent series expansion of the function

$$f(z) = z^2 \sin \frac{1}{z}$$

about the points

- (a) $z=0$ (b) $z=\infty$

- (c) $z=a$, a finite non-zero complex number

(For (c), do *not* calculate the coefficients explicitly.)

46 Expand

$$f(z) = \frac{z}{(z-1)(2-z)}$$

in a Laurent series expansion valid for

- (a) $|z| < 1$ (b) $1 < |z| < 2$ (c) $|z| > 2$
 (d) $|z-1| > 1$ (e) $0 < |z-2| < 1$

4.5

Singularities and zeros

As indicated in Section 4.4.5 a **singularity** of a complex function $f(z)$ is a point of the z plane where $f(z)$ ceases to be analytic. Normally, this means $f(z)$ is infinite at such a point, but it can also mean that there is a choice of values, and it is not possible to pick a particular one. In this chapter we shall be mainly concerned with singularities at which $f(z)$ has an infinite value. A **zero** of $f(z)$ is a point in the z plane at which $f(z)=0$.

Singularities can be classified in terms of the Laurent series expansion of $f(z)$ about the point in question. If $f(z)$ has a Taylor series expansion, that is a Laurent series expansion with zero principal part, about the point $z=z_0$, then z_0 is a **regular point** of $f(z)$. If $f(z)$ has a Laurent series expansion with only a finite number of terms in its principal part, for example

$$f(z) = \frac{a_{-m}}{(z-z_0)^m} + \cdots + \frac{a_{-1}}{(z-z_0)} + a_0 + a_1(z-z_0) + \cdots + a_m(z-z_0)^m + \cdots$$

then $f(z)$ has a singularity at $z=z_0$ called a **pole**. If there are m terms in the principal part, as in this example, then the pole is said to be of **order m** . Another way of defining this is to say that z_0 is a pole of order m if

$$\lim_{z \rightarrow z_0} (z-z_0)^m f(z) = a_{-m} \quad (4.35)$$

where a_{-m} is finite and non-zero. If the principal part of the Laurent series for $f(z)$ at $z=z_0$ has infinitely many terms, which means that the above limit does not exist for any m , then $z=z_0$ is called an **essential singularity** of $f(z)$. (Note that in Example 4.20 the expansions given as representations of the function $f(z) = 1/[(z+1)(z+3)]$ in parts (a) and (b) are *not* valid at $z=0$. Hence, despite appearances, they do not represent a function which possesses an essential singularity at $z=0$. In this case $f(z)$ is **regular** at $z=0$ with a value $\frac{1}{3}$.)

If $f(z)$ appears to be singular at $z = z_0$, but it turns out to be possible to define a Taylor series expansion there, then $z = z_0$ is called a **removable singularity**. The following examples illustrate these cases.

- (a) $f(z) = z^{-1}$ has a pole of order one, called a **simple pole**, at $z = 0$.
- (b) $f(z) = (z - 1)^{-3}$ has a pole of order three at $z = 1$.
- (c) $f(z) = e^{1/(z-j)}$ has an essential singularity at $z = j$.
- (d) The function

$$f(z) = \frac{z-1}{(z+2)(z-3)^2}$$

has a zero at $z = 1$, a simple pole at $z = -2$ and a pole of order two at $z = 3$.

- (e) The function

$$f(z) = \frac{\sin z}{z}$$

is not defined at $z = 0$, and appears to be singular there. However, defining

$$\text{sinc } z = \begin{cases} (\sin z)/z & (z \neq 0) \\ 1 & (z = 0) \end{cases}$$

gives a function having a Taylor series expansion

$$\text{sinc } z = 1 - \frac{z^2}{3!} + \frac{z^4}{5!} - \dots$$

that is regular at $z = 0$. Therefore the (apparent) singularity at $z = 0$ has been removed, and thus $f(z) = (\sin z)/z$ has a removable singularity at $z = 0$.

Functions whose only singularities are poles are called **meromorphic** and, by and large, in engineering applications of complex variables most functions are meromorphic. To help familiarize the reader with these definitions, the following example should prove instructive.

Example 4.22

Find the singularities and zeros of the following complex functions:

- (a) $\frac{1}{z^4 - z^2(1+j)+j}$
- (b) $\frac{z-1}{z^4 - z^2(1+j)+j}$
- (c) $\frac{\sin(z-1)}{z^4 - z^2(1+j)+j}$
- (d) $\frac{1}{[z^4 - z^2(1+j)+j]^3}$

Solution (a) For

$$f(z) = \frac{1}{z^4 - z^2(1+j)+j}$$

the numerator is never zero, and the denominator is only infinite when z is infinite. Thus $f(z)$ has no zeros in the finite z plane. The denominator is zero when

$$z^4 - z^2(1 + j) + j = 0$$

which factorizes to give

$$(z^2 - 1)(z^2 - j) = 0$$

leading to

$$z^2 = 1 \text{ or } j$$

so that the singularities are at

$$z = +1, -1, (1 + j)/\sqrt{2}, (-1 - j)/\sqrt{2} \quad (4.36)$$

all of which are simple poles since none of the roots are repeated.

(b) The function

$$f(z) = \frac{z - 1}{z^4 - z^2(1 + j) + j}$$

is similar to $f(z)$ in (a), except that it has the additional term $z - 1$ in the numerator. Therefore, at first glance, it seems that the singularities are as in (4.36). However, a closer look indicates that $f(z)$ can be rewritten as

$$f(z) = \frac{z - 1}{(z - 1)(z + 1)[z + \sqrt{\frac{1}{2}}(1 + j)][z - \sqrt{\frac{1}{2}}(1 + j)]}$$

and the factor $z - 1$ cancels, rendering $z = 1$ a removable singularity, and reducing $f(z)$ to

$$f(z) = \frac{1}{(z + 1)[z + \sqrt{\frac{1}{2}}(1 + j)][z - \sqrt{\frac{1}{2}}(1 + j)]}$$

which has no (finite) zeros and $z = -1$, $\sqrt{\frac{1}{2}}(1 + j)$ and $\sqrt{\frac{1}{2}}(-1 - j)$ as simple poles.

(c) In the case of

$$f(z) = \frac{\sin(z - 1)}{z^4 - z^2(1 + j) + j}$$

the function may be rewritten as

$$f(z) = \frac{\sin(z - 1)}{z - 1} \frac{1}{(z + 1)[z + \sqrt{\frac{1}{2}}(1 + j)][z - \sqrt{\frac{1}{2}}(1 + j)]}$$

Now

$$\frac{\sin(z - 1)}{z - 1} \rightarrow 1 \quad \text{as } z \rightarrow 1$$

so once again $z = 1$ is a removable singularity. Also, as in (b), $z = -1$, $\sqrt{\frac{1}{2}}(1 + j)$ and $\sqrt{\frac{1}{2}}(-1 - j)$ are simple poles and the only singularities. However,

$$\sin(z - 1) = 0$$

has the general solution $z = 1 + N\pi$ ($N = 0, \pm 1, \pm 2, \dots$). Thus, apart from $N = 0$, all of these are zeros of $f(z)$.

(d) For

$$f(z) = \frac{1}{[z^4 - z^2(1+j) + j]^3}$$

factorizing as in (b), we have

$$f(z) = \frac{1}{(z-1)^3(z+1)^3[z + \sqrt{\frac{1}{2}}(1+j)]^3[z - \sqrt{\frac{1}{2}}(1+j)]^3}$$

so $-1, +1, \sqrt{\frac{1}{2}}(1+j)$ and $\sqrt{\frac{1}{2}}(-1-j)$ are still singularities, but this time they are triply repeated. Hence they are all poles of order three. There are no zeros.

4.5.1 Exercises

- 47 Determine the location of, and classify, the singularities and zeros of the following functions. Specify also any zeros that may exist.

$$\begin{array}{lll} (a) \frac{\cos z}{z^2} & (b) \frac{1}{(z+j)^2(z-j)} & (c) \frac{z}{z^4-1} \\ (d) \coth z & (e) \frac{\sin z}{z^2+\pi^2} & (f) e^{z/(1-z)} \\ (g) \frac{z-1}{z^2+1} & (h) \frac{z+j}{(z+2)^3(z-3)} & \\ (i) \frac{1}{z^2(z^2-4z+5)} & & \end{array}$$

- 48 Expand each of the following functions in a Laurent series about $z = 0$, and give the type of singularity (if any) in each case:

$$\begin{array}{ll} (a) \frac{1-\cos z}{z} & \\ (b) \frac{e^{z^2}}{z^3} & \\ (c) z^{-1} \cosh z^{-1} & \\ (d) \tan^{-1}(z^2+2z+2) & \end{array}$$

- 49 Show that if $f(z)$ is the ratio of two polynomials then it cannot have an essential singularity.

4.6 Engineering application: analysing AC circuits

In the circuit shown in Figure 4.25 we wish to find the variation in impedance Z and admittance Y as the capacitance C of the capacitor varies from 0 to ∞ . Here



$$\frac{1}{Z} = \frac{1}{R} + j\omega C, \quad Y = \frac{1}{Z}$$

Writing

Figure 4.25
AC circuit of
Section 4.6.

$$\frac{1}{Z} = \frac{1 + j\omega CR}{R}$$

we clearly have

$$Z = \frac{R}{1 + j\omega CR} \quad (4.37)$$

Equation (4.37) can be interpreted as a bilinear mapping with Z and C as the two variables. We examine what happens to the real axis in the C plane (C varies from 0 to ∞ and, of course, is real) under the inverse of the mapping given by (4.37). Rearranging (4.37), we have

$$C = \frac{R - Z}{j\omega R Z} \quad (4.38)$$

Taking $Z = x + jy$

$$C = \frac{R - x - jy}{j\omega R(x + jy)} = \frac{x + jy - R}{\omega R(y - jx)} = \frac{(x + jy - R)(y + jx)}{\omega R(x^2 + y^2)} \quad (4.39)$$

Equating imaginary parts, and remembering that C is real, gives

$$0 = x^2 + y^2 - Rx \quad (4.40)$$

which represents a circle, with centre at $\left(\frac{1}{2}R, 0\right)$ and of radius $\frac{1}{2}R$. Thus the real axis in the C plane is mapped onto the circle given by (4.40) in the Z plane. Of course, C is positive. If $C = 0$, (4.40) indicates that $Z = R$. The circuit of Figure 4.25 confirms that the impedance is R in this case. If $C \rightarrow \infty$ then $Z \rightarrow 0$, so the positive real axis in the plane is mapped onto either the upper or lower half of the circle. Equating real parts in (4.39) gives

$$C = \frac{-y}{\omega(x^2 + y^2)}$$

so $C > 0$ gives $y < 0$, implying that the lower half of the circle is the image in the Z plane of the positive real axis in the C plane, as indicated in Figure 4.26. A diagram

Figure 4.26 Mapping for the impedance Z .

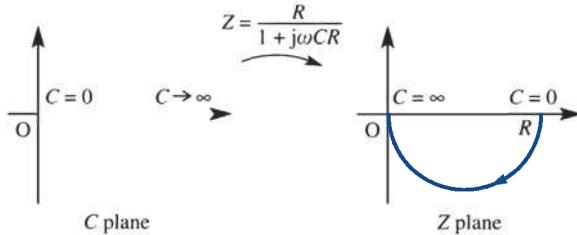
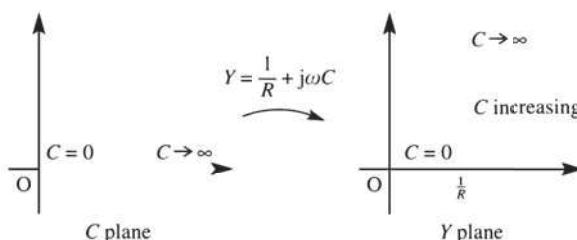


Figure 4.27 Mapping for the admittance Y .



such as Figure 4.26 gives an immediate visual impression of how the impedance Z varies as C varies.

The admittance $Y = 1/Z$ is given by

$$Y = \frac{1}{R} + j\omega C$$

which represents a linear mapping as shown in Figure 4.27.

4.7 Engineering application: use of harmonic functions

In this section we discuss two engineering applications where use is made of the properties of harmonic functions.

4.7.1 A heat transfer problem

We saw in Section 4.3.2 that every analytic function generates a pair of harmonic functions. The problem of finding a function that is harmonic in a specified region and satisfies prescribed boundary conditions is one of the oldest and most important problems in science-based engineering. Sometimes the solution can be found by means of a conformal mapping defined by an analytic function. This, essentially, is a consequence of the ‘function of a function’ rule of calculus, which implies that every harmonic function of x and y transforms into a harmonic function of u and v under the mapping

$$w = u + jv = f(x + jy) = f(z)$$

where $f(z)$ is analytic. Furthermore, the level curves of the harmonic function in the z plane are mapped onto corresponding level curves in the w plane, so that a harmonic

function that has a constant value along part of the boundary of a region or has a zero normal derivative along part of the boundary is mapped onto a harmonic function with the same property in the w plane.

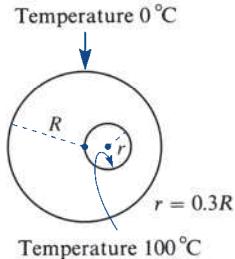


Figure 4.28
Schematic diagram of heat transfer problem.

For heat transfer problems the level curves of the harmonic function correspond to isotherms, and a zero normal derivative corresponds to thermal insulation. To illustrate these ideas, consider the simple steady-state heat transfer problem shown schematically in Figure 4.28. There is a cylindrical pipe with an offset cylindrical cavity through which steam passes at 100°C . The outer temperature of the pipe is 0°C . The radius of the inner circle is $\frac{3}{10}$ of that of the outer circle, so by choosing the outer radius as the unit of length the problem can be stated as that of finding a harmonic function $T(x, y)$ such that

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$$

in the region between the circles $|z| = 1$ and $|z - 0.3| = 0.3$, and $T = 0$ on $|z| = 1$ and $T = 100$ on $|z - 0.3| = 0.3$.

The mapping

$$w = \frac{z - 3}{3z - 1}$$

transforms the circle $|z| = 1$ onto the circle $|w| = 1$ and the circle $|z - 0.3| = 0.3$ onto the circle $|w| = 3$ as shown in Figure 4.29. Thus the problem is transformed into the axially symmetric problem in the w plane of finding a harmonic function $T(u, v)$ such that $T(u, v) = 100$ on $|w| = 1$ and $T(u, v) = 0$ on $|w| = 3$. Harmonic functions with such axial symmetry have the general form

$$T(u, v) = A \ln(u^2 + v^2) + B$$

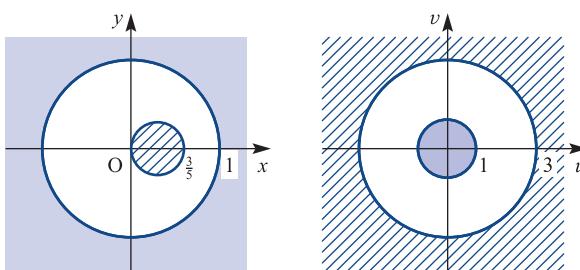
where A and B are constants.

Here we require, in addition to the axial symmetry, that $T(u, v) = 100$ on $u^2 + v^2 = 1$ and $T(u, v) = 0$ on $u^2 + v^2 = 9$. Thus $B = 100$ and $A = -100 \ln 9$, and the solution on the w plane is

$$T(u, v) = \frac{100[1 - \ln(u^2 + v^2)]}{\ln 9}$$

We need the solution on the z plane, which means in general we have to obtain u and v in terms of x and y . Here, however, it is a little easier, since $u^2 + v^2 = |w|^2$ and

Figure 4.29
The mapping
 $w = (z - 3)/(3z - 1)$.



$$|w|^2 = \left| \frac{z-3}{3z-1} \right|^2 = \frac{|z-3|^2}{|3z-1|^2} = \frac{(x-3)^2 + y^2}{(3x-1)^2 + 9y^2}$$

Thus

$$T(x, y) = \frac{100}{\ln 9} \{ 1 - \ln [(x-3)^2 + y^2] - \ln [(3x-1)^2 + 9y^2] \}$$

4.7.2 Current in a field-effect transistor

The fields (E_x, E_y) in an insulated-gate field-effect transistor are harmonic conjugates that satisfy a nonlinear boundary condition. For the transistor shown schematically in Figure 4.30 we have

$$\frac{\partial E_y}{\partial x} = \frac{\partial E_x}{\partial y}, \quad \frac{\partial E_y}{\partial y} = -\frac{\partial E_x}{\partial x}$$

with conditions

$$E_x = 0 \quad \text{on the electrodes}$$

$$E_x \left(E_y + \frac{V_0}{h} \right) = -\frac{I}{2\mu\epsilon_0\epsilon_r} \quad \text{on the channel}$$

$$E_y \rightarrow -\frac{V_g}{h} \quad \text{as } x \rightarrow -\infty \quad (0 < y < h)$$

$$E_y \rightarrow \frac{V_d - V_g}{h} \quad \text{as } x \rightarrow \infty \quad (0 < y < h)$$

where V_0 is a constant with dimensions of potential, h is the insulator thickness, I is the current in the channel, which is to be found, μ, ϵ_0 and ϵ_r have their usual meanings, and the gate potential V_g and the drain potential V_d are taken with respect to the source potential.

The key to the solution of this problem is the observation that the nonlinear boundary condition

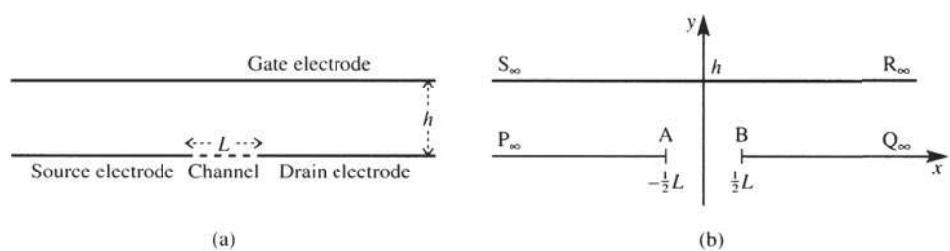
$$2E_x \left(E_y + \frac{V_0}{h} \right) = -\frac{I}{\mu\epsilon_0\epsilon_r}$$

contains the harmonic function (now of E_x and E_y)

$$H(E_x, E_y) = 2E_x \left(E_y + \frac{V_0}{h} \right)$$

Figure 4.30

- (a) Schematic diagram for an insulated-gate field-effect transistor;
- (b) an appropriate coordinate system for the application.



A harmonic conjugate of H is the function

$$G(E_x, E_y) = \left(E_y + \frac{V_0}{h} \right)^2 - E_x^2$$

Since E_x and E_y are harmonic conjugates with respect to x and y , so are G and H . Thus the problem may be restated as that of finding harmonic conjugates G and H such that

$$H = 0 \quad \text{on the electrodes}$$

$$H = -\frac{I}{\mu\epsilon_0\epsilon_r} \quad \text{on the channel}$$

$$G \rightarrow \left(\frac{V_0 - V_g}{h} \right)^2 \quad \text{as } x \rightarrow \infty \quad (0 < y < h)$$

$$G \rightarrow \left(\frac{V_0 + V_d - V_g}{h} \right)^2 \quad \text{as } x \rightarrow -\infty \quad (0 < y < h)$$

Using the sequence of mappings shown in Figure 4.31, which may be composed into the single formula

$$w = \frac{a e^{bz} - a^2}{a e^{bz} - 1}$$

where $a = e^{bL/2}$ and $b = \pi/h$, the problem is transformed into finding harmonic-conjugate functions G and H (on the w plane) such that

$$H = 0 \quad \text{on } v = 0 \quad (u > 0) \tag{4.41}$$

$$H = -\frac{I}{\mu\epsilon_0\epsilon_r} \quad \text{on } v = 0 \quad (u < 0) \tag{4.42}$$

$$G = \left(\frac{V_0 - V_g}{h} \right)^2 \quad \text{at } w = e^{bL} \tag{4.43}$$

$$G = \left(\frac{V_0 + V_d - V_g}{h} \right)^2 \quad \text{at } w = 1 \tag{4.44}$$

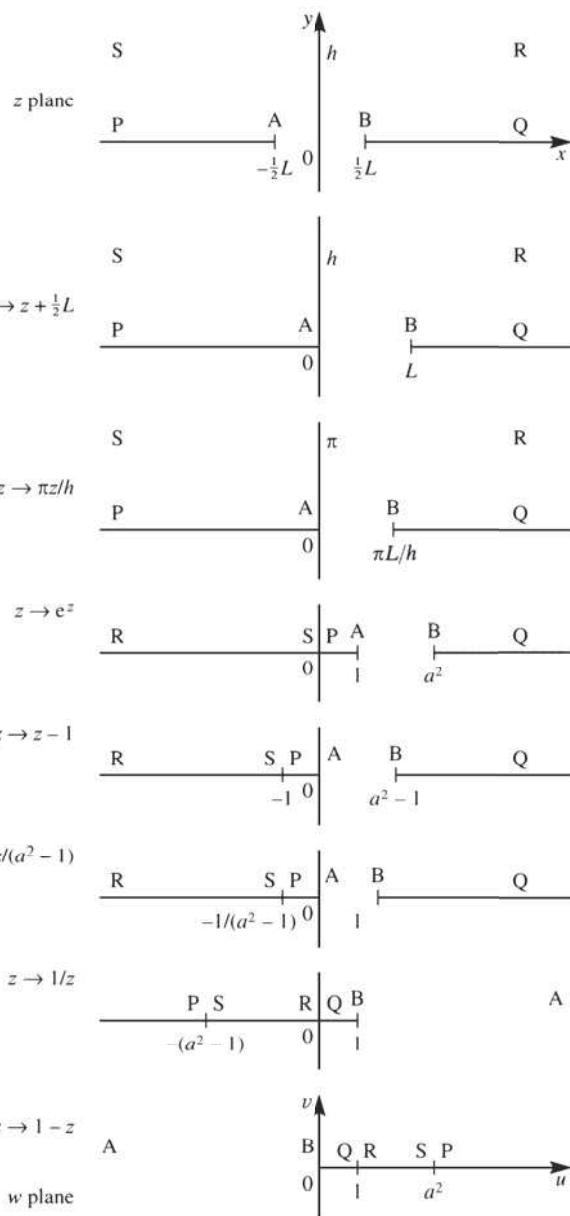
The conditions (4.41), (4.42) and (4.44) are sufficient to determine H and G completely

$$H = -\frac{I \arg(w)}{\pi \mu \epsilon_0 \epsilon_r}$$

$$G = \frac{I \ln|w|}{\pi \mu \epsilon_0 \epsilon_r} + \left(\frac{V_0 + V_d - V_g}{h} \right)^2$$

Figure 4.31

Sequence of mappings to simplify the problem.



while the condition (4.43) determines the values of I

$$I = \frac{\mu \epsilon_0 \epsilon_r}{Lh} (2V_0 - 2V_g + V_d)V_d$$

This example shows the power of complex variable methods for solving difficult problems arising in engineering mathematics. The following exercises give some simpler examples for the reader to investigate.

4.7.3 Exercises

- 50 Show that the transformation $w = 1/z$, $w = u + jv$, $z = x + jy$, transforms the circle $x^2 + y^2 = 2ax$ in the z plane into the straight line $u = 1/2a$ in the w plane. Two long conducting wires of radius a are placed adjacent and parallel to each other, so that their cross-section appears as in Figure 4.32. The wires are separated at O by an insulating gap of negligible dimensions, and carry potentials $\pm V_0$ as indicated. Find an expression for the potential at a general point (x, y) in the plane of the cross-section and sketch the equipotentials.

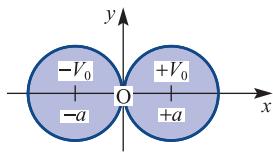


Figure 4.32 Conducting wires of Exercise 50.

- 51 Find the images under the mapping

$$w = \frac{z+1}{1-z}$$

$z = x + jy$, of

- (a) the points A(-1, 0), B(0, 1), C($\frac{24}{25}, \frac{7}{25}$) and D($\frac{3}{4}, 0$) in the z plane,
- (b) the straight line $y = 0$,
- (c) the circle $x^2 + y^2 = 1$.

Illustrate your answer with a diagram showing the z and w planes and shade on the w plane the region corresponding to $x^2 + y^2 < 1$.

A semicircular disc of unit radius, $[(x, y): x^2 + y^2 \leq 1, y \geq 0]$, has its straight boundary at temperature 0°C and its curved boundary at 100°C . Prove that the temperature at the point (x, y) is

$$T = \frac{200}{\pi} \tan^{-1} \left(\frac{2y}{1-x^2-y^2} \right)$$

- 52 (a) Show that the function

$$G(x, y) = 2x(1-y)$$

satisfies the Laplace equation and construct its harmonic conjugate $H(x, y)$ that satisfies $H(0, 0) = 0$. Hence obtain, in terms of z , where $z = x + jy$, the function F such that $W = F(z)$ where $W = G + jH$.

- (b) Show that under the mapping $w = \ln z$, the harmonic function $G(x, y)$ defined in (a) is mapped into the function

$$G(u, v) = 2e^u \cos v - e^{2u} \sin 2v$$

Verify that $G(u, v)$ is harmonic.

- (c) Generalize the result (b) to prove that under the mapping $w = f(z)$, where $f'(z)$ exists, a harmonic function of (x, y) is transformed into a harmonic function of (u, v) .

53

- Show that if $w = (z+3)/(z-3)$, $w = u + jv$, $z = x + jy$, the circle $u^2 + v^2 = k^2$ in the w plane is the image of the circle

$$x^2 + y^2 + 6 \frac{1+k^2}{1-k^2} x + 9 = 0 \quad (k^2 \neq 1)$$

in the z plane.

Two long cylindrical wires, each of radius 4 mm, are placed parallel to each other with their axes 10 mm apart, so that their cross-section appears as in Figure 4.33. The wires carry potentials $\pm V_0$ as shown. Show that the potential $V(x, y)$ at the point (x, y) is given by

$$V = \frac{V_0}{\ln 4} \{ \ln [(x+3)^2 + y^2] - \ln [(x-3)^2 + y^2] \}$$

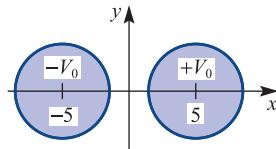


Figure 4.33 Cylindrical wires of Exercise 53.

54

- Find the image under the mapping

$$w = \frac{j(1-z)}{1+z}$$

$z = x + jy$, $w = u + jv$, of

- (a) the points A(1, 0), B(0, 1), C(0, -1) in the z plane,
- (b) the straight line $y = 0$,
- (c) the circle $x^2 + y^2 = 1$.

A circular plate of unit radius, $[(x, y): x^2 + y^2 \leq 1]$, has one half (with $y > 0$) of its rim, $x^2 + y^2 = 1$, at temperature 0°C and the other half (with $y < 0$) at temperature 100°C . Using the above mapping, prove that the steady-state temperature at the point (x, y) is

$$T = \frac{100}{\pi} \tan^{-1} \left(\frac{1-x^2-y^2}{2y} \right)$$

- 55** The problem shown schematically in Figure 4.34 arose during a steady-state heat transfer investigation. T is the temperature. By applying the successive mappings

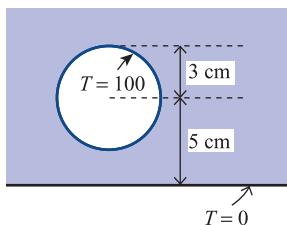


Figure 4.34 Schematic representation of Exercise 55.

$$z_1 = \frac{z + j4}{z - j4}, \quad w = \ln z_1$$

show that the temperature at the point (x, y) in the shaded region in the figure is given by

$$T(x, y) = \frac{50}{\ln 3} \ln \left[\frac{x^2 + (4+y)^2}{x^2 + (4-y)^2} \right]$$

- 56** The functions

$$w = z + \frac{1}{z}, \quad w = \frac{z+1}{z-1}$$

perform the mappings shown in Figure 4.35. A long bar of semicircular cross-section has the temperature of the part of its curved surface corresponding to the arc PQ in Figure 4.36 kept at 100°C while the rest of the surface is kept at 0°C . Show that the temperature T at the point (x, y) is given by

$$T = \frac{100}{\pi} [\arg(z^2 + z + 1) - \arg(z^2 - z + 1)]$$

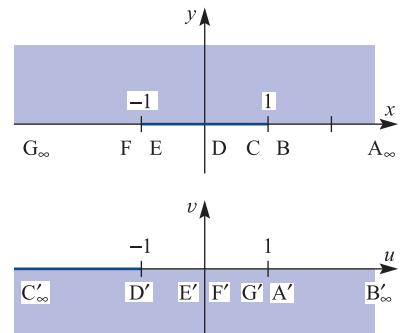
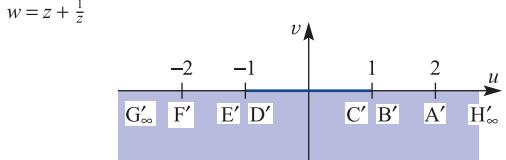
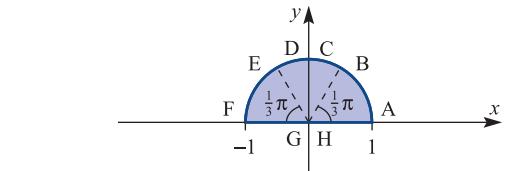


Figure 4.35 Mappings of Exercise 56.

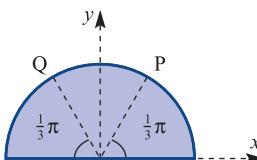


Figure 4.36 Cross-section of bar of Exercise 56.

4.8 Review exercises (1–19)

- 1** Find the images of the following points under the mappings given:
- $z = 1 + j$ under $w = (1 + j)z + j$
 - $z = 1 - j2$ under $w = j3z + j + 1$
 - $z = 1$ under $w = \frac{1}{2}(1 - j)z + \frac{1}{2}(1 + j)$
 - $z = j2$ under $w = \frac{1}{2}(1 - j)z + \frac{1}{2}(1 + j)$
- 2** Under each of the mappings given in Review exercise 1, find the images in the w plane of the two straight lines
- $y = 2x$
 - $x + y = 1$
- in the z plane, $z = x + jy$.

- 3** The linear mapping $w = \alpha z + \beta$, where α and β are complex constants, maps the point $z = 2 - j$ in the z plane to the point $w = 1$ in the w plane, and the point $z = 0$ to the point $w = 3 + j$.
- Determine α and β
 - Find the region in the w plane corresponding to the left half-plane $\operatorname{Re}(z) \leq 0$ in the z plane.
 - Find the region in the w plane corresponding to the circular region $5|z| \leq 1$ in the z plane.
 - Find the fixed point of the mapping.
- 4** Map the following straight lines from the z plane, $z = x + jy$, to the w plane under the inverse mapping $w = j/z$:
- $x = y + 1$
 - $y = 3x$
 - the line joining $A(1 + j)$ to $B(2 + j3)$ in the z plane
 - $y = 4$
- In each case sketch the image curve.
- 5** Two complex variables w and z are related by the mapping
- $$w = \frac{z+1}{z-1}$$
- Sketch this mapping by finding the images in the w plane of the lines $\operatorname{Re}(z) = \text{constant}$ and $\operatorname{Im}(z) = \text{constant}$. Find the fixed points of the mapping.
- 6** The mapping
- $$w = \frac{1-z^2}{z}$$
- takes points from the z plane to the w plane. Find the fixed points of the mapping, and show that the circle of radius r with centre at the origin in the z plane is transformed to the ellipse
- $$\left(\frac{ur^2}{r^2-1}\right)^2 + \left(\frac{vr^2}{r^2+1}\right)^2 = r^2$$
- in the w plane, where $w = u + jv$. Investigate what happens when $r = 1$.
- 7** Find the real and imaginary parts of the complex function $w = z^3$, and verify the Cauchy–Riemann equations.
- 8** Find a function $v(x, y)$ such that, given
- $$u(x, y) = x \sin x \cosh y - y \cos x \sinh y$$
- $$f(z) = u + jv$$
- is an analytic function of
- z
- ,
- $f(0) = 0$
- .
- 9** Find the bilinear transformation that maps the three points $z = 0, j$ and $\frac{1}{2}(1+j)$ in the z plane to the three points $w = \infty, -j$ and $1-j$ respectively in the w plane. Check that the transformation will map
- the lower half of the z plane onto the upper half of the w plane
 - the interior of the circle with centre $z = j\frac{1}{2}$ and radius $\frac{1}{2}$ in the z plane onto the half-plane $\operatorname{Im}(w) < -1$ in the w plane.
- 10** Show that the mapping
- $$z = \zeta + \frac{a^2}{4\zeta}$$
- where $z = x + jy$ and $\zeta = R e^{j\theta}$ maps the circle $R = \text{constant}$ in the ζ plane onto an ellipse in the z plane. Suggest a possible use for this mapping.
- 11** Find the power series representation of the function
- $$\frac{1}{1+z^3}$$
- in the disc $|z| < 1$. Deduce the power series for
- $$\frac{1}{(1+z^3)^2}$$
- valid in the same disc.
- 12** Find the first four non-zero terms of the Taylor series expansion of the following functions about the point indicated, and determine the radius of convergence of each:
- $\frac{1-z}{1+z}$ ($z = 0$)
 - $\frac{1}{z^2+1}$ ($z = 1$)
 - $\frac{z}{z+1}$ ($z = j$)
- 13** Find the radius of convergence of each Taylor series expansion of the following function about the points indicated, *without* finding the series itself:
- $$f(z) = \frac{1}{z(z^2+1)}$$
- at the points $z = 1, -1, 1+j, 1+j\frac{1}{2}$ and $2+j3$.

- 14** Determine the Laurent series expansion of the function

$$f(z) = \frac{1}{(z^2 + 1)z}$$

about the points (a) $z = 0$ and (b) $z = 1$, and determine the region of validity of each.

- 15** Find the Laurent series expansion of the function

$$f(z) = e^z \sin\left(\frac{1}{1-z}\right)$$

about (a) $z = 0$, (b) $z = 1$ and (c) $z = \infty$, indicating the range of validity in each case. (Do *not* find terms explicitly; indicate only the form of the principal part.)

- 16** Find the real and imaginary parts of the functions

(a) $e^z \sinh z$ (b) $\cos 2z$
 (c) $\frac{\sin z}{z}$ (d) $\tan z$

- 17** Determine whether the following mappings are conformal, and, if not, find the non-conformal points:

(a) $w = \frac{1}{z^2}$

(b) $w = 2z^3 + 3z^2 + 6(1 - j)z + 1$

(c) $w = 64z + \frac{1}{z^3}$

- 18**

Consider the mapping $w = \cos z$. Determine the points where the mapping is not conformal. By finding the images in the w plane of the lines $x = \text{constant}$ and $y = \text{constant}$ in the z plane ($z = x + jy$), draw the mapping similarly to Figures 4.14 and 4.18.

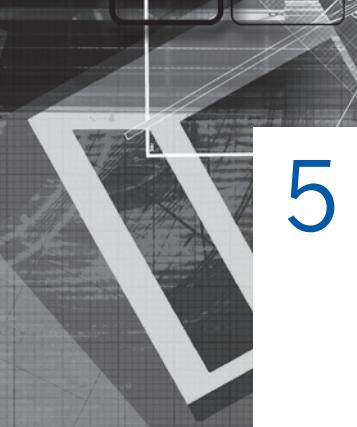
- 19**

Determine the location of and classify the singularities of the following functions:

(a) $\frac{\sin z}{z^2}$ (b) $\frac{1}{(z^3 - 8)^2}$

(c) $\frac{z+1}{z^4 - 1}$ (d) $\operatorname{sech} z$

(e) $\sinh z$ (f) $\sin\left(\frac{1}{z}\right)$ (g) z^z



5

Laplace Transforms

Chapter 5 Contents

5.1	Introduction	316
5.2	Step and impulse functions	320
5.3	Transfer functions	356
5.4	Solution of state-space equations	378
5.5	Engineering application: frequency response	390
5.6	Engineering application: pole placement	398
5.7	Review exercises (1–18)	401

5.1 Introduction

The Laplace transform was introduced in Chapter 11 of *Modern Engineering Mathematics* (MEM), and a brief summary of that material serves to introduce further useful properties and their applications to engineering. In most engineering applications it is useful at the outset to think of the Laplace transform in terms of an input to a system and the subsequent response, see Figure 5.1.

Figure 5.1 Schematic representation of a system.



5.1.1 Definition and notation

Definition 5.1:

The Laplace transform of a function $f(t)$ can be defined as

$$\mathcal{L}\{f(t)\} = \int_0^{\infty} e^{-st} f(t) dt \quad (5.1)$$

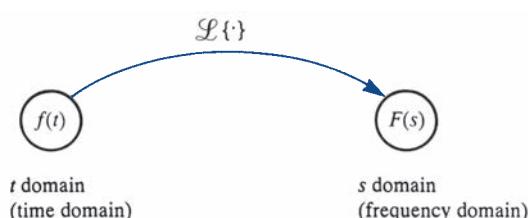
where the symbol \mathcal{L} is the Laplace transform operator.

It is an operator performed upon the function $f(t)$ and the output is a function $F(s)$ where

$$F(s) = \int_0^{\infty} e^{-st} f(t) dt \quad (5.2)$$

This relationship is depicted graphically in Figure 5.2.

Figure 5.2
The Laplace transform operator.

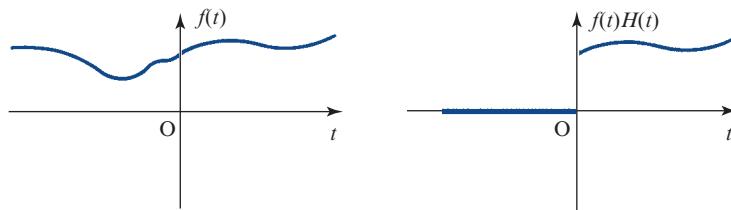


The Heaviside step function is in fact where this chapter actually starts in earnest a bit later with new applications; however, it is defined as

$$H(t) = \begin{cases} 0 & (t < 0) \\ 1 & (t > 0) \end{cases} \quad (5.3)$$

and it is useful in representing any function $f(t)$ as $H(t)f(t)$ where $f(t)$ is defined for all t but, for practical applications, focus is on values $t > 0$, see Figure 5.3.

Figure 5.3
Graph of $f(t)$ and its causal equivalent function.



In MEM the two-sided Laplace transform is defined for the cases where the behaviour of $f(t)$ for $t < 0$ is of interest. Behaviour at $t = 0$ is addressed later when impulses are introduced. In MEM various elementary properties of the Laplace transform are derived. A short list of the Laplace transforms of some elementary function is given in the table below.

$F(s)$	$f(t)$
$\frac{1}{s}$	1
$\frac{1}{s^n}, n = 1, 2, \dots$	$\frac{t^{n-1}}{(n-1)!}$
$\frac{1}{s-k}$	c^{kt}
$\frac{s}{s^2 + a^2}$	$\cos(at)$
$\frac{a}{s^2 + a^2}$	$\sin(at)$
$\frac{s}{(s+k)^2 + a^2}$	$c^{-kt} \cos(at)$
$\frac{a}{(s+k)^2 + a^2}$	$c^{-kt} \sin(at)$

A more extensive one is available in specialist books (see, for example, Phil Dyke, *An Introduction to Laplace Transforms and Fourier Series*, second edition, London, Springer, 2014, or online).

Of course it is also important to know if the Laplace transform of a given function actually exists, and this is assured as long as the following inequality is true: that there exists a real number σ and positive constants M and T such that

$$|f(t)| < M e^{\sigma t}$$

for all $t > T$. This is called $f(t)$ being of *exponential order*. This means almost all reasonable functions are included, but functions that grow faster than exponential such as e^{t^2} are excluded. Functions that have jumps are not excluded, see later, but of course there cannot be gaps, that is values of t where $f(t)$ is not defined.

The Laplace transform is linear, in other words \mathcal{L} obeys the rule

$$\mathcal{L}\{\alpha f(t) + \beta g(t)\} = \alpha \mathcal{L}\{f(t)\} + \beta \mathcal{L}\{g(t)\}$$

This enables results to be combined, finding the Laplace transform of $2 \sin 3t + 3 \cos 2t$ for example. Another useful property is the first shift theorem (MEM Theorem 11.2).

Theorem 5.1

First shift theorem

If $f(t)$ is a function having Laplace transform $F(s)$, with $\text{Re}(s) > \sigma_c$ then the function $e^{at} f(t)$ has a Laplace transform given by

$$\mathcal{L}\{e^{at} f(t)\} = F(s - a) \quad \text{Re}(s) > \sigma_c + \text{Re}(a)$$

[end of theorem](#)

This is a useful result as shown through the examples in MEM.

5.1.2 Other results from MEM

Treating the Laplace transform as a mapping from the t domain to the s domain, one needs to consider the reverse mapping from the s domain to the t domain. The formal process involves complex variable theory and is too technical (but again see specialist books such as that by Phil Dyke referred to above), so in these two texts the process is to use tables either in the old fashioned way, or to use MATLAB or MAPLE to help with the details. This is covered in MEM.

One of the useful applications of Laplace transforms is to the solution of differential equations, so taking the Laplace transform of a derivative is required. Here is a useful result:

$$\mathcal{L}\left\{\frac{df}{dt}\right\} = sF(s) - f(0)$$

Straight away it is apparent that the process of using the Laplace transform operator eliminates the derivative. In MEM this result is proved using integration by parts to integrate out the derivative. The Laplace transform of the second derivative uses integration by parts twice and the result is

$$\mathcal{L}\left\{\frac{d^2f}{dt^2}\right\} = s^2F(s) - sf(0) - f'(0)$$

where the dash denotes differentiation with respect to t . These results can be applied to solving both single-variable differential equations and simultaneous differential equations. The application to the solution of partial differential equations is covered in Section 9.3.3.

Another result, perhaps not quite so useful, is the Laplace transform of an integral:

$$\mathcal{L} \left\{ \int_0^t f(\tau) d\tau \right\} = \frac{1}{s} F(s)$$

One of the most important applications to the solution of either single or simultaneous differential equations is the detection of resonance and other vibrations in either electrical or mechanical systems, and there is a lot of space devoted to this in MEM. This brings us up to speed and we are now ready to embark on new results and applications of the very powerful Laplace transform. We start with more about the step function.

(Section 5.2 follows on the next page)

5.2 Step and impulse functions

5.2.1 The Heaviside step function

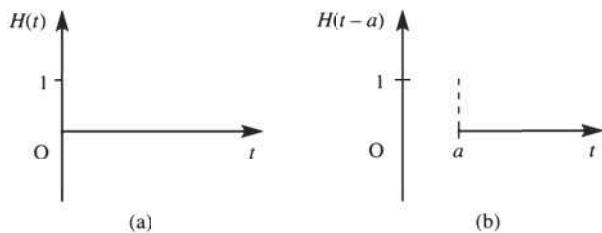
In Sections 11.3.3 and 11.3.5 of MEM we considered linear differential equations in which the forcing functions were continuous. In many engineering applications the forcing function may frequently be discontinuous, for example a square wave resulting from an on/off switch. In order to accommodate such discontinuous functions, we use the Heaviside unit step function $H(t)$, which, as we saw in Section 5.2.1, or in section 11.2.1 of MEM, is defined by

$$H(t) = \begin{cases} 0 & (t < 0) \\ 1 & (t \geq 0) \end{cases}$$

and is illustrated graphically in Figure 5.4(a). The Heaviside function is also frequently referred to simply as the **unit step function**. A function representing a unit step at $t = a$ may be obtained by a horizontal translation of duration a . This is depicted graphically in Figure 5.4(b), and defined by

$$H(t-a) = \begin{cases} 0 & (t < a) \\ 1 & (t \geq a) \end{cases}$$

Figure 5.4
Heaviside unit
step function.

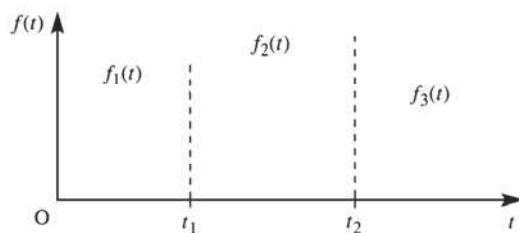


The product function $f(t)H(t-a)$ takes values

$$f(t)H(t-a) = \begin{cases} 0 & (t < a) \\ f(t) & (t \geq a) \end{cases}$$

so the function $H(t-a)$ may be interpreted as a device for ‘switching on’ the function $f(t)$ at $t = a$. In this way the unit step function may be used to write a concise formulation of piecewise-continuous functions. To illustrate this, consider the piecewise-continuous function $f(t)$ illustrated in Figure 5.5 and defined by

Figure 5.5
Piecewise-continuous
function.



$$f(t) = \begin{cases} f_1(t) & (0 \leq t < t_1) \\ f_2(t) & (t_1 \leq t < t_2) \\ f_3(t) & (t \geq t_2) \end{cases}$$

To construct this function $f(t)$, we could use the following ‘switching’ operations:

- (a) switch on the function $f_1(t)$ at $t = 0$;
- (b) switch on the function $f_2(t)$ at $t = t_1$ and at the same time switch off the function $f_1(t)$;
- (c) switch on the function $f_3(t)$ at $t = t_2$ and at the same time switch off the function $f_2(t)$.

In terms of the unit step function, the function $f(t)$ may thus be expressed as

$$f(t) = f_1(t)H(t) + [f_2(t) - f_1(t)]H(t - t_1) + [f_3(t) - f_2(t)]H(t - t_2)$$

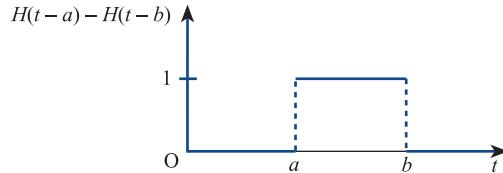
Alternatively, $f(t)$ may be constructed using the **top hat function** $H(t - a) - H(t - b)$. Clearly,

$$H(t - a) - H(t - b) = \begin{cases} 1 & (a \leq t < b) \\ 0 & \text{otherwise} \end{cases} \quad (5.4)$$

which, as illustrated in Figure 5.6, gives

$$f(t)[H(t - a) - H(t - b)] = \begin{cases} f(t) & (a \leq t < b) \\ 0 & \text{otherwise} \end{cases}$$

Figure 5.6
Top hat function.



Using this approach, the function $f(t)$ of Figure 5.5 may be expressed as

$$f(t) = f_1(t)[H(t) - H(t - t_1)] + f_2(t)[H(t - t_1) - H(t - t_2)] + f_3(t)H(t - t_2)$$

giving, as before,

$$f(t) = f_1(t)H(t) + [f_2(t) - f_1(t)]H(t - t_1) + [f_3(t) - f_2(t)]H(t - t_2)$$

It is easily checked that this corresponds to the given formulation, since for $0 \leq t < t_1$

$$H(t) = 1, \quad H(t - t_1) = H(t - t_2) = 0$$

giving

$$f(t) = f_1(t) \quad (0 \leq t < t_1)$$

while for $t_1 \leq t < t_2$

$$H(t) = H(t - t_1) = 1, \quad H(t - t_2) = 0$$

giving

$$f(t) = f_1(t) + [f_2(t) - f_1(t)] = f_2(t) \quad (t_1 \leq t \leq t_2)$$

and finally for $t \geq t_2$

$$H(t) = H(t - t_1) = H(t - t_2) = 1$$

giving

$$f(t) = f_1(t) + [f_2(t) - f_1(t)] + [f_3(t) - f_2(t)] = f_3(t) \quad (t \geq t_2)$$

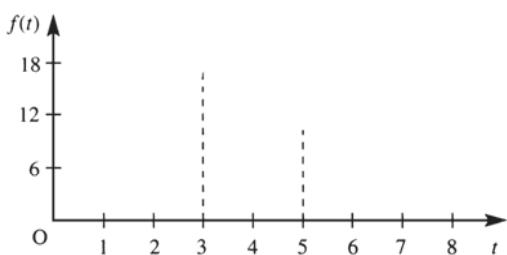
Example 5.1

Express in terms of unit step functions the piecewise-continuous causal function

$$f(t) = \begin{cases} 2t^2 & (0 \leq t < 3) \\ t+4 & (3 \leq t < 5) \\ 9 & (t \geq 5) \end{cases}$$

Figure 5.7

Piecewise-continuous function of Example 5.1.



Solution $f(t)$ is depicted graphically in Figure 5.7, and in terms of unit step functions it may be expressed as

$$f(t) = 2t^2H(t) + (t+4 - 2t^2)H(t-3) + (9 - t - 4)H(t-5)$$

That is,

$$f(t) = 2t^2H(t) + (4 + t - 2t^2)H(t-3) + (5 - t)H(t-5)$$

Example 5.2

Express in terms of unit step functions the piecewise-continuous causal function

$$f(t) = \begin{cases} 0 & (t < 1) \\ 1 & (1 \leq t < 3) \\ 3 & (3 \leq t < 5) \\ 2 & (5 \leq t < 6) \\ 0 & (t \geq 6) \end{cases}$$

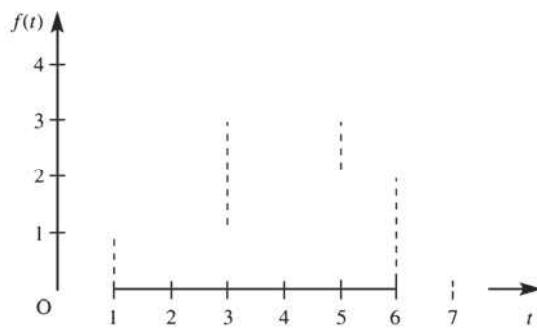
Solution $f(t)$ is depicted graphically in Figure 5.8, and in terms of unit step functions it may be expressed as

$$f(t) = 1H(t-1) + (3-1)H(t-3) + (2-3)H(t-5) + (0-2)H(t-6)$$

That is,

$$f(t) = 1H(t-1) + 2H(t-3) - 1H(t-5) - 2H(t-6)$$

Figure 5.8
Piecewise-continuous function of Example 5.2.



5.2.2 Laplace transform of unit step function

By definition of the Laplace transform, the transform of $H(t - a)$, $a \geq 0$, is given by

$$\begin{aligned}\mathcal{L}\{H(t - a)\} &= \int_0^{\infty} H(t - a) e^{-st} dt = \int_0^a 0 e^{-st} dt + \int_a^{\infty} 1 e^{-st} dt \\ &= \left[\frac{e^{-st}}{-s} \right]_a^{\infty} = \frac{e^{-as}}{s}\end{aligned}$$

That is,

$$\mathcal{L}\{H(t - a)\} = \frac{e^{-as}}{s} \quad (a \geq 0) \quad (5.5)$$

and in the particular case of $a = 0$

$$\mathcal{L}\{H(t)\} = \frac{1}{s} \quad (5.6)$$



This may be implemented in MATLAB using the commands

```
syms s t
H=sym('Heaviside(t)')
laplace(H)
```

which return

```
ans=1/s
```

It may also be obtained directly using the command

```
laplace(sym('Heaviside(t)'))
```

Likewise to obtain the Laplace transform of $H(t - 2)$ we use the commands

```
H2=sym('Heaviside(t-2)')
laplace(H2)
```

which return

```
ans=exp(-2*s)/s
```

In MAPLE the results are obtained using the commands:

```
with(inttrans):
laplace(Heaviside(t), t, s);
laplace(Heaviside(t-2), t, s);
```

Example 5.3

Determine the Laplace transform of the rectangular pulse

$$f(t) = \begin{cases} 0 & (t < a) \\ K & (a \leq t < b) \\ 0 & (t \geq b) \end{cases} \quad K \text{ constant}, \quad b > a > 0$$

Solution

The pulse is depicted graphically in Figure 5.9. In terms of unit step functions, it may be expressed, using the top hat function, as

$$f(t) = K[H(t-a) - H(t-b)]$$

Then, taking Laplace transforms,

$$\mathcal{L}\{f(t)\} = K\mathcal{L}\{H(t-a)\} - K\mathcal{L}\{H(t-b)\}$$

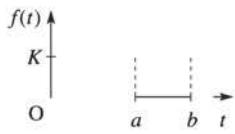


Figure 5.9

Rectangular pulse.

which, on using the elementary properties of the Laplace transform, specifically (11.23) in MEM, gives

$$\mathcal{L}\{f(t)\} = K \frac{e^{-as}}{s} - K \frac{e^{-bs}}{s}$$

That is,

$$\mathcal{L}\{f(t)\} = \frac{K}{s} (e^{-as} - e^{-bs})$$

Example 5.4

Determine the Laplace transform of the piecewise-constant function $f(t)$ shown in Figure 5.8.

Solution

From Example 5.2 $f(t)$ may be expressed as

$$f(t) = 1H(t-1) + 2H(t-3) - 1H(t-5) - 2H(t-6)$$

Taking Laplace transforms,

$$\mathcal{L}\{f(t)\} = 1\mathcal{L}\{H(t-1)\} + 2\mathcal{L}\{H(t-3)\} - 1\mathcal{L}\{H(t-5)\} - 2\mathcal{L}\{H(t-6)\}$$

which, on using the result (5.5), gives

$$\mathcal{L}\{f(t)\} = \frac{e^{-s}}{s} + 2 \frac{e^{-3s}}{s} - \frac{e^{-5s}}{s} - 2 \frac{e^{-6s}}{s}$$

That is,

$$\mathcal{L}\{f(t)\} = \frac{1}{s} (e^{-s} + 2e^{-3s} - e^{-5s} - 2e^{-6s})$$



Check that the same answer is obtained using the MATLAB sequence of commands

```
syms s t
H1=sym('Heaviside(t-1)');
H3=sym('Heaviside(t-3)');
H5=sym('Heaviside(t-5)');
H6=sym('Heaviside(t-6)');
laplace(H1-2*H3-H5-2*H6)
```

In MAPLE the commands

```
with(inttrans):
laplace(Heaviside(t-1)+Heaviside(t-3)*2 - Heaviside(t-5)
       - Heaviside(t-6)*2,t,s);
```

return the answer

$$\frac{e^{(-s)} + 2e^{(-3s)} - e^{(-5s)} - 2e^{(-6s)}}{s}$$

5.2.3 The second shift theorem

This theorem is dual to the first shift theorem given as Theorem 5.1, and is sometimes referred to as the **Heaviside** or **delay theorem**.

Theorem 5.2

If $\mathcal{L}\{f(t)\} = F(s)$ then for a positive constant a

$$\mathcal{L}\{f(t-a)H(t-a)\} = e^{-as}F(s)$$

Proof By definition,

$$\begin{aligned}\mathcal{L}\{f(t-a)H(t-a)\} &= \int_0^\infty f(t-a)H(t-a) e^{-st} dt \\ &= \int_a^\infty f(t-a) e^{-st} dt\end{aligned}$$

Making the substitution $T = t - a$,

$$\begin{aligned}\mathcal{L}\{f(t-a)H(t-a)\} &= \int_0^\infty f(T) e^{-s(T+a)} dT \\ &= e^{-sa} \int_0^\infty f(T) e^{-sT} dT\end{aligned}$$

Since $F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty f(T) e^{-sT}$, it follows that

$$\mathcal{L}\{f(t-a)H(t-a)\} = e^{-as}F(s)$$

[end of theorem](#)

It is important to distinguish between the two functions $f(t)H(t-a)$ and $f(t-a)H(t-a)$. As we saw earlier, $f(t)H(t-a)$ simply indicates that the function $f(t)$ is ‘switched on’ at time $t = a$, so that

$$f(t)H(t-a) = \begin{cases} 0 & (t < a) \\ f(t) & (t \geq a) \end{cases}$$

On the other hand, $f(t-a)H(t-a)$ represents a translation of the function $f(t)$ by a units to the right (to the right, since $a > 0$), so that

$$f(t-a)H(t-a) = \begin{cases} 0 & (t < a) \\ f(t-a) & (t \geq a) \end{cases}$$

The difference between the two is illustrated graphically in Figure 5.10. $f(t-a)H(t-a)$ may be interpreted as representing the function $f(t)$ delayed in time by a units. Thus, when considering its Laplace transform $e^{-as}F(s)$, where $F(s)$ denotes the Laplace transform of $f(t)$, the component e^{-as} may be interpreted as a delay operator on the transform $F(s)$, indicating that the response of the system characterized by $F(s)$ will be delayed in time by a units. Since many practically important systems have some form of delay inherent in their behaviour, it is clear that the result of this theorem is very useful.

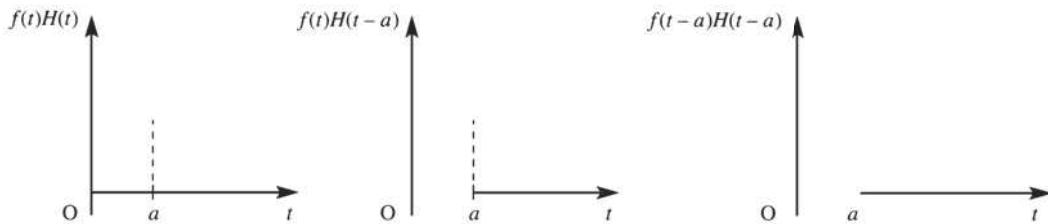


Figure 5.10 Illustration of $f(t-a)H(t-a)$.

Example 5.5

Determine the Laplace transform of the causal function $f(t)$ defined by

$$f(t) = \begin{cases} t & (0 \leq t < b) \\ 0 & (t \geq b) \end{cases}$$

Solution

$f(t)$ is illustrated graphically in Figure 5.11, and is seen to characterize a sawtooth pulse of duration b . In terms of unit step functions,

$$f(t) = tH(t) - tH(t-b)$$

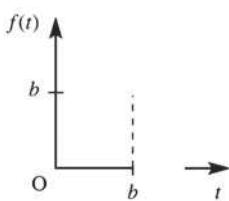
In order to apply the second shift theorem, each term must be rearranged to be of the form $f(t-a)H(t-a)$; that is, the time argument $t-a$ of the function must be the same as that of the associated step function. In this particular example this gives

$$f(t) = tH(t) - (t-b)H(t-b) - bH(t-b)$$

Figure 5.11
Sawtooth pulse.

Taking Laplace transforms,

$$\mathcal{L}\{f(t)\} = \mathcal{L}\{tH(t)\} - \mathcal{L}\{(t-b)H(t-b)\} - b\mathcal{L}\{H(t-b)\}$$



which, on using the second shift theorem, Theorem 5.2, leads to

$$\mathcal{L}\{f(t)\} = \frac{1}{s^2} - e^{-bs} \mathcal{L}(t) - b \frac{e^{-bs}}{s} = \frac{1}{s^2} - \frac{e^{-bs}}{s^2} - b \frac{e^{-bs}}{s}$$

giving

$$\mathcal{L}\{f(t)\} = \frac{1}{s^2}(1 - e^{-bs}) - \frac{b}{s} e^{-bs}$$

It should be noted that this result could have been obtained without the use of the second shift theorem, since, directly from the definition of the Laplace transform,

$$\begin{aligned}\mathcal{L}\{f(t)\} &= \int_0^\infty f(t) e^{-st} dt = \int_0^b t e^{-st} dt + \int_b^\infty 0 e^{-st} dt \\ &= \left[-\frac{t e^{-st}}{s} \right]_0^b + \int_0^b \frac{e^{-st}}{s} dt = \left[-\frac{t e^{-st}}{s} - \frac{e^{-st}}{s^2} \right]_0^b \\ &= \left(-\frac{b e^{-sb}}{s} - \frac{e^{-sb}}{s^2} \right) - \left(-\frac{1}{s^2} \right) = \frac{1}{s^2}(1 - e^{-bs}) - \frac{b}{s} e^{-bs}\end{aligned}$$

as before.

Example 5.6

Obtain the Laplace transform of the piecewise-continuous causal function

$$f(t) = \begin{cases} 2t^2 & (0 \leq t < 3) \\ t+4 & (3 \leq t < 5) \\ 9 & (t \geq 5) \end{cases}$$

considered in Example 5.1.

Solution In Example 5.1 we saw that $f(t)$ may be expressed in terms of unit step functions as

$$f(t) = 2t^2 H(t) - (2t^2 - t - 4)H(t - 3) - (t - 5)H(t - 5)$$

Before we can find $\mathcal{L}\{f(t)\}$, the function $2t^2 - t - 4$ must be expressed as a function of $t - 3$. This may be readily achieved as follows. Let $z = t - 3$. Then

$$\begin{aligned}2t^2 - t - 4 &= 2(z+3)^2 - (z+3) - 4 \\ &= 2z^2 + 11z + 11 \\ &= 2(t-3)^2 + 11(t-3) + 11\end{aligned}$$

Hence

$$f(t) = 2t^2 H(t) - [2(t-3)^2 + 11(t-3) + 11]H(t-3) - (t-5)H(t-5)$$

Taking Laplace transforms,

$$\begin{aligned}\mathcal{L}\{f(t)\} &= 2\mathcal{L}\{t^2 H(t)\} - \mathcal{L}\{[2(t-3)^2 + 11(t-3) + 11]H(t-3)\} \\ &\quad - \mathcal{L}\{(t-5)H(t-5)\}\end{aligned}$$

which, on using Theorem 5.2, leads to

$$\begin{aligned}\mathcal{L}\{f(t)\} &= 2 \frac{2}{s^3} - e^{-3s} \mathcal{L}\{2t^2 + 11t + 11\} - e^{-5s} \mathcal{L}\{t\} \\ &= \frac{4}{s^3} - e^{-3s} \left(\frac{4}{s^3} + \frac{11}{s^2} + \frac{11}{s} \right) - \frac{e^{-5s}}{s^2}\end{aligned}$$

Again this result could have been obtained directly from the definition of the Laplace transform, but in this case the required integration by parts is a little more tedious.



Having set up s and t as symbolic variables and specified H , $H1$ and $H5$ then the MATLAB commands

```
laplace(2*t^2*H - (2*t^2-t-4)*H3 - (t-5)*H5);
pretty(ans)
```

generate

```
ans = 4/s^3 - 11*exp(-3s)/s - 11*exp(-3s)/s^2 - 4*exp(-3s)/s^3 - exp(-5s)/s^2
```

In MAPLE the commands

```
with(inttrans):
laplace(Heaviside(t)*2*t^2 - Heaviside(t-3)*(2*t^2-t-4)
       - Heaviside(t-5)*(t-5), t, s);
```

return the answer

$$-\frac{e^{(-5s)}}{s^2} + \frac{4 - e^{(-3s)}(11s^2 + 11s + 4)}{s^3}$$

5.2.4 Inversion using the second shift theorem

We have seen in Examples 5.3 and 5.4 that, to obtain the Laplace transforms of piecewise-continuous functions, use of the second shift theorem could be avoided, since it is possible to obtain such transforms directly from the definition of the Laplace transform.

In practice, the importance of the theorem lies in determining *inverse* transforms, since, as indicated earlier, delays are inherent in most practical systems and engineers are interested in knowing how these influence the system response. Consequently, by far the most useful form of the second shift theorem is

$$\mathcal{L}^{-1}\{e^{-as}F(s)\} = f(t-a)H(t-a) \quad (5.7)$$

Comparing (5.7) with the result (11.11) (see p. 918 of MEM), namely

$$\mathcal{L}^{-1}\{F(s)\} = f(t)H(t)$$

we see that

$$\mathcal{L}^{-1}\{e^{-as}F(s)\} = [f(t-a)H(t-a)] \quad \text{with } t-a \text{ instead of } t$$

indicating that the response $f(t)$ has been delayed in time by a units. This is why the theorem is sometimes called the delay theorem.



This is readily implemented in MATLAB using the command `ilaplace`.

Example 5.7

Determine $\mathcal{L}^{-1}\left\{\frac{4e^{-4s}}{s(s+2)}\right\}$.

Solution

This may be written as $\mathcal{L}^{-1}\{e^{-4s}F(s)\}$, where

$$F(s) = \frac{4}{s(s+2)}$$

First we obtain the inverse transform $f(t)$ of $F(s)$. Resolving into partial fractions,

$$F(s) = \frac{2}{s} - \frac{2}{s+2}$$

which, on inversion, gives

$$f(t) = 2 - 2e^{-2t}$$

a graph of which is shown in Figure 5.12(a). Then, using (5.7), we have

$$\begin{aligned}\mathcal{L}^{-1}\left\{e^{-4s}\frac{4}{s(s+2)}\right\} &= \mathcal{L}^{-1}\{e^{-4s}F(s)\} = f(t-4)H(t-4) \\ &= (2 - 2e^{-2(t-4)})H(t-4)\end{aligned}$$

giving

$$\mathcal{L}^{-1}\left\{\frac{4e^{-4s}}{s(s+2)}\right\} = \begin{cases} 0 & (t < 4) \\ 2(1 - e^{-2(t-4)}) & (t \geq 4) \end{cases}$$

which is plotted in Figure 5.12(b).



Using MATLAB confirm that the commands

```
ilaplace(4*exp(-4*s)/(s*(s+2)));
pretty(ans)
```

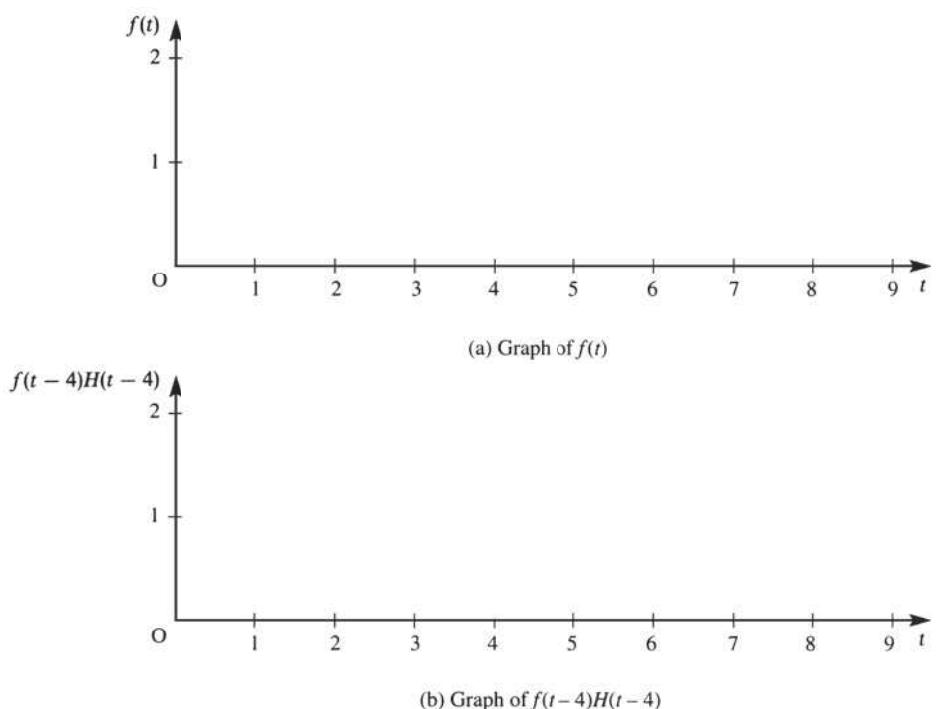
generate the answer

```
2H(t-4)(1-exp(-2t+8))
```

The same answer is obtained in MAPLE using the commands

```
with(inttrans):
invlaplace(4*exp(-4*s)/(s*(s+2)), s, t);
```

Figure 5.12 Inverse transforms of Example 5.7.



Example 5.8

Determine $\mathcal{L}^{-1}\left\{\frac{e^{-s\pi}(s+3)}{s(s^2+1)}\right\}$.

Solution This may be written as $\mathcal{L}^{-1}\{e^{-s\pi}F(s)\}$, where

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

Resolving into partial fractions,

$$F(s) = \frac{3}{s} - \frac{3s}{s^2+1} + \frac{1}{s^2+1}$$

which, on inversion, gives

$$f(t) = 3 - 3 \cos t + \sin t$$

a graph of which is shown in Figure 5.13(a). Then, using (5.7), we have

$$\begin{aligned} \mathcal{L}^{-1}\left\{\frac{e^{-s\pi}(s+3)}{s(s^2+1)}\right\} &= \mathcal{L}^{-1}\{e^{-s\pi}F(s)\} = f(t-\pi)H(t-\pi) \\ &= [3 - 3 \cos(t-\pi) + \sin(t-\pi)]H(t-\pi) \\ &= (3 + 3 \cos t - \sin t)H(t-\pi) \end{aligned}$$

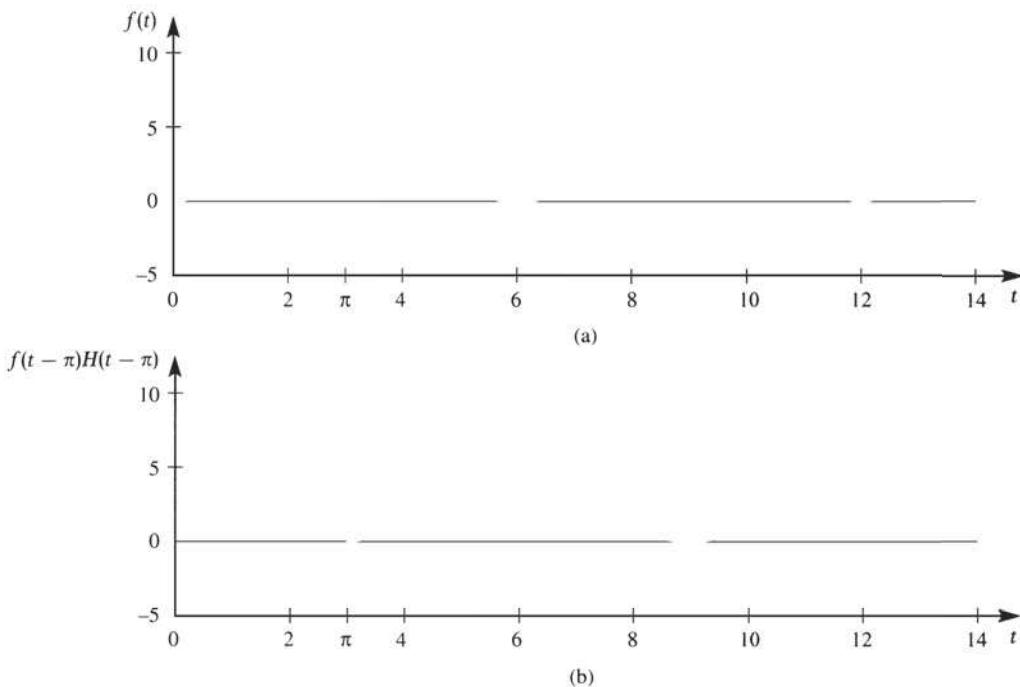


Figure 5.13 Inverse transforms of Example 5.8.

giving

$$\mathcal{L}^{-1}\left\{\frac{e^{-s\pi}(s+3)}{s(s^2+1)}\right\} = \begin{cases} 0 & (t < \pi) \\ 3 + 3 \cos t - \sin t & (t \geq \pi) \end{cases}$$

which is plotted in Figure 5.13(b).

5.2.5 Differential equations

We now return to the solution of linear differential equations for which the forcing function $f(t)$ is piecewise-continuous, like that illustrated in Figure 5.5. One approach to solving a differential equation having such a forcing function is to solve it separately for each of the continuous components $f_1(t)$, $f_2(t)$, and so on, comprising $f(t)$, using the fact that in this equation all the derivatives, except the highest, must remain continuous so that values at the point of discontinuity provide the initial conditions for the next section. This approach is obviously rather tedious, and a much more direct one is to make use of Heaviside step functions to specify $f(t)$. Then the method of solution follows that used in Section 11.3 of MEM and we shall simply illustrate it by examples.

Example 5.9

Obtain the solution $x(t)$, $t \geq 0$, of the differential equation

$$\frac{d^2x}{dt^2} + 5 \frac{dx}{dt} + 6x = f(t) \quad (5.8)$$

where $f(t)$ is the pulse function

$$f(t) = \begin{cases} 3 & (0 \leq t < 6) \\ 0 & (t \geq 6) \end{cases}$$

and subject to the initial conditions $x(0) = 0$ and $\dot{x}(0) = 2$.

Solution To illustrate the advantage of using a step function formulation of the forcing function $f(t)$, we shall first solve separately for each of the time ranges.

Method 1 For $0 \leq t < 6$, (5.8) becomes

$$\frac{d^2x}{dt^2} + 5 \frac{dx}{dt} + 6x = 3$$

with $x(0) = 0$ and $\dot{x}(0) = 2$.

Taking Laplace transforms gives

$$(s^2 + 5s + 6)X(s) = sx(0) + \dot{x}(0) + 5x(0) + \frac{3}{s} = 2 + \frac{3}{s}$$

That is,

$$X(s) = \frac{2s+3}{s(s+2)(s+3)} = \frac{\frac{1}{2}}{s} + \frac{\frac{1}{2}}{s+2} - \frac{1}{s+3}$$

which, on inversion, gives

$$x(t) = \frac{1}{2} + \frac{1}{2} e^{-2t} - e^{-3t} \quad (0 \leq t < 6)$$

We now determine the values of $x(6)$ and $\dot{x}(6)$ in order to provide the initial conditions for the next stage:

$$x(6) = \frac{1}{2} + \frac{1}{2} e^{-12} - e^{-18} = \alpha, \quad \dot{x}(6) = -e^{-12} + 3e^{-18} = \beta$$

For $t \geq 6$ we make the change of independent variable $T = t - 6$, whence (5.8) becomes

$$\frac{d^2x}{dT^2} + 5 \frac{dx}{dT} + 6x = 0$$

subject to $x(T=0) = \alpha$ and $\dot{x}(T=0) = \beta$.

Taking Laplace transforms gives

$$(s^2 + 5s + 6)X(s) = sx(T=0) + \dot{x}(T=0) + 5x(T=0) = \alpha s + 5\alpha + \beta$$

That is,

$$X(s) = \frac{\alpha s + 5\alpha + \beta}{(s+2)(s+3)} = \frac{\beta + 3\alpha}{s+2} - \frac{\beta + 2\alpha}{s+3}$$

which, on taking inverse transforms, gives

$$x(T) = (\beta + 3\alpha)e^{-2T} - (\beta + 2\alpha)e^{-3T}$$

Substituting the values of α and β and reverting to the independent variable t gives

$$x(t) = \left(\frac{3}{2} + \frac{1}{2}e^{-12}\right)e^{-2(t-6)} - (1 + e^{-18})e^{-3(t-6)} \quad (t \geq 6)$$

That is,

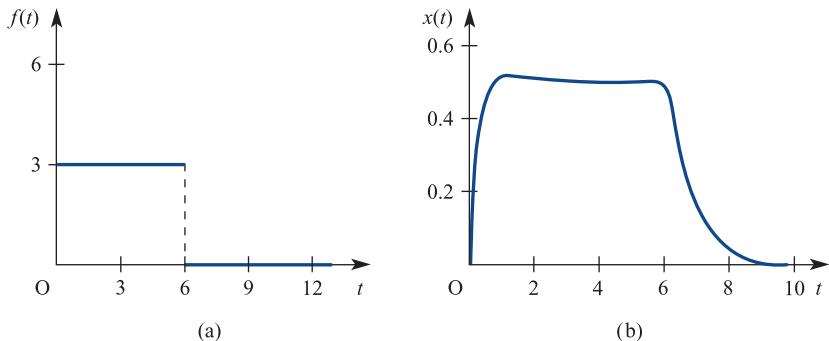
$$x(t) = \left(\frac{1}{2}e^{-2t} - e^{-3t}\right) + \left(\frac{3}{2}e^{-2(t-6)} - e^{-3(t-6)}\right) \quad (t \geq 6)$$

Thus the solution of the differential equation is

$$x(t) = \begin{cases} \frac{1}{2} + \frac{1}{2}e^{-2t} - e^{-3t} & (0 \leq t < 6) \\ \left(\frac{1}{2}e^{-2t} - e^{-3t}\right) + \left(\frac{3}{2}e^{-2(t-6)} - e^{-3(t-6)}\right) & (t \geq 6) \end{cases}$$

The forcing function $f(t)$ and response $x(t)$ are shown in Figures 5.14(a) and (b) respectively.

Figure 5.14
Forcing function
and response of
Example 5.9.



Method 2 In terms of Heaviside step functions,

$$f(t) = 3H(t) - 3H(t - 6)$$

so that, using (5.5),

$$\mathcal{L}\{f(t)\} = \frac{3}{s} - \frac{3}{s}e^{-6s}$$

Taking Laplace transforms in (5.8) then gives

$$(s^2 + 5s + 6)X(s) = sx(0) + \dot{x}(0) + 5x(0) + \mathcal{L}\{f(t)\} = 2 + \frac{3}{s} - \frac{3}{s}e^{-6s}$$

That is,

$$X(s) = \frac{2s+3}{s(s+2)(s+3)} - e^{-6s} \frac{3}{s(s+2)(s+3)}$$

$$= \left(\frac{\frac{1}{2}}{s} + \frac{\frac{1}{2}}{s+2} - \frac{1}{s+3}\right) - e^{-6s} \left(\frac{\frac{1}{2}}{s} - \frac{\frac{3}{2}}{s+2} + \frac{1}{s+3}\right)$$

Taking inverse Laplace transforms and using the result (5.7) gives

$$x(t) = \left(\frac{1}{2} + \frac{1}{2}e^{-2t} - e^{-3t}\right) - \left(\frac{1}{2} + \frac{3}{2}e^{-2(t-6)} + e^{-3(t-6)}\right)H(t - 6)$$

which is the required solution. This corresponds to that obtained in Method 1, since, using the definition of $H(t - 6)$, it may be written as

$$x(t) = \begin{cases} \frac{1}{2} + \frac{1}{2} e^{-2t} - e^{-3t} & (0 \leq t < 6) \\ \left(\frac{1}{2} e^{-2t} - e^{-3t}\right) + \left(\frac{3}{2} e^{-2(t-6)} - e^{-3(t-6)}\right) & (t \geq 6) \end{cases}$$

This approach is clearly less tedious, since the initial conditions at the discontinuities are automatically taken account of in the solution.



It seems that the standard `dsolve` command is unable to deal with differential equations having such Heaviside functions as their forcing function. To resolve this problem use can be made of the `maple` command in MATLAB, which lets us access MAPLE commands directly. Confirm that the following commands produce the correct solution:

```
maple('de:=diff(x(t),t$2)+5*diff(x(t),t)+6*x(t)
      =3*Heaviside-3*Heaviside(t-6);')
ans=
de := diff(x(t), '$'(t, 2)) + 5*diff(x(t), t) + 6*x(t)
= 3*Heaviside-3*Heaviside(t-6)
maple('dsolve({de,x(0)=0,D(x)(0)=2},x(t)),method=laplace;')
```

In MAPLE the answer may be obtained directly using the commands

```
with(inttrans):
de:=diff(x(t),t$2)+5*diff(x(t),t)+6*x(t)
-3*Heaviside-3*Heaviside(t-6);
dsolve({de,x(0)=0,D(x)(0)=2},x(t)),method=laplace;
```

Example 5.10

Determine the solution $x(t)$, $t \geq 0$, of the differential equation

$$\frac{d^2x}{dt^2} + 2 \frac{dx}{dt} + 5x = f(t) \quad (5.9)$$

where

$$f(t) = \begin{cases} t & (0 \leq t \leq \pi) \\ 0 & (t \geq \pi) \end{cases}$$

and subject to the initial conditions $x(0) = 0$ and $\dot{x}(0) = 3$.

Solution Following the procedures of Example 5.5, we have

$$\begin{aligned} f(t) &= tH(t) - tH(t - \pi) \\ &= tH(t) - (t - \pi)H(t - \pi) - \pi H(t - \pi) \end{aligned}$$

so that, using Theorem 5.2,

$$\mathcal{L}\{f(t)\} = \frac{1}{s^2} - \frac{e^{-\pi s}}{s^2} - \frac{\pi e^{-\pi s}}{s} = \frac{1}{s^2} - e^{-\pi s} \left(\frac{1}{s^2} + \frac{\pi}{s} \right)$$

Taking Laplace transforms in (5.9) then gives

$$\begin{aligned}(s^2 + 2s + 5)X(s) &= sx(0) + \dot{x}(0) + 2x(0) + \mathcal{L}\{f(t)\} \\ &= 3 + \frac{1}{s^2} - e^{-\pi s} \left(\frac{1}{s^2} + \frac{\pi}{s} \right)\end{aligned}$$

using the given initial conditions.

Thus

$$X(s) = \frac{3s^2 + 1}{s^2(s^2 + 2s + 5)} - e^{-\pi s} \frac{1 + s\pi}{s^2(s^2 + 2s + 5)}$$

which, on resolving into partial fractions, leads to

$$\begin{aligned}X(s) &= \frac{1}{25} \left[-\frac{2}{s} + \frac{5}{s^2} + \frac{2s + 74}{(s+1)^2 + 4} \right] - \frac{e^{-\pi s}}{25} \left[\frac{5\pi - 2}{s} + \frac{5}{s^2} - \frac{(5\pi - 2)s + (10\pi + 1)}{(s+1)^2 + 4} \right] \\ &= \frac{1}{25} \left[-\frac{2}{s} + \frac{5}{s^2} + \frac{2(s+1) + 72}{(s+1)^2 + 4} \right] \\ &\quad - \frac{e^{-\pi s}}{25} \left[\frac{5\pi - 2}{s} + \frac{5}{s^2} - \frac{(5\pi - 2)(s+1) + (5\pi + 3)}{(s+1)^2 + 4} \right]\end{aligned}$$

Taking inverse Laplace transforms and using (5.7) gives the desired solution:

$$\begin{aligned}x(t) &= \frac{1}{25} (-2 + 5t + 2e^{-t} \cos 2t + 36e^{-t} \sin 2t) \\ &\quad - \frac{1}{25} [(5\pi - 2) + 5(t - \pi) - (5\pi - 2)e^{-(t-\pi)} \cos 2(t - \pi) \\ &\quad - \frac{1}{2}(5\pi + 3)e^{-(t-\pi)} \sin 2(t - \pi)] H(t - \pi)\end{aligned}$$

That is,

$$\begin{aligned}x(t) &= \frac{1}{25} [5t - 2 + 2e^{-t}(\cos 2t + 18 \sin 2t)] \\ &\quad - \frac{1}{25} \{5t - 2 - e^\pi e^{-t} [(5\pi - 2) \cos 2t + \frac{1}{2}(5\pi + 3) \sin 2t]\} H(t - \pi)\end{aligned}$$

or, in alternative form,

$$x(t) = \begin{cases} \frac{1}{25}[5t - 2 + 2e^{-t}(\cos 2t + 18 \sin 2t)] & (0 \leq t < \pi) \\ \frac{1}{25}e^{-t}\{(2 + (5\pi - 2)e^\pi) \cos 2t + [36 + \frac{1}{2}(5\pi + 3)e^\pi] \sin 2t\} & (t \geq \pi) \end{cases}$$

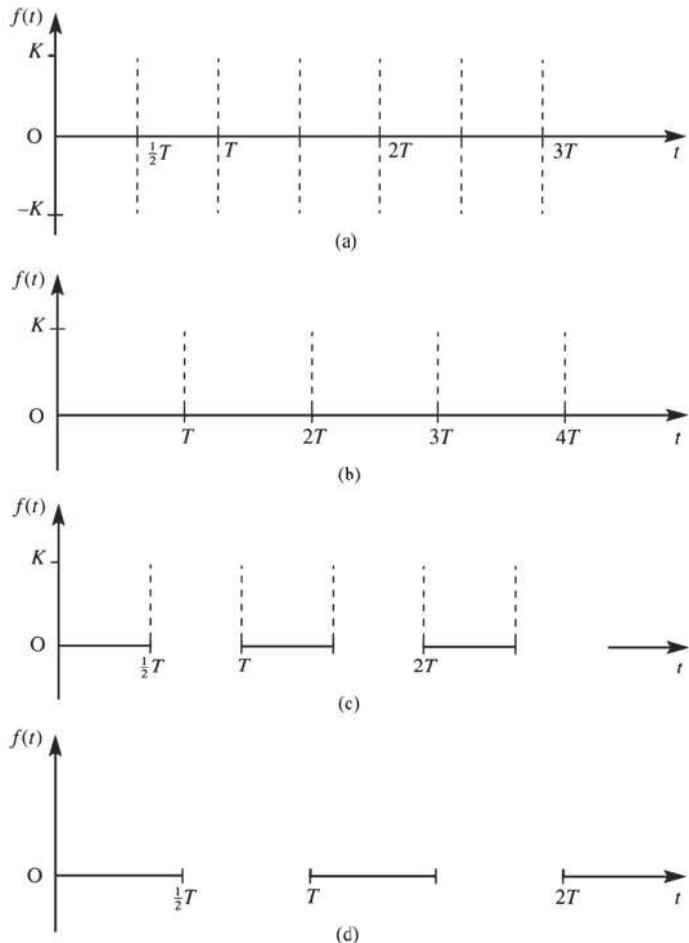
5.2.6 Periodic functions

We have already determined the Laplace transforms of periodic functions, such as $\sin \omega t$ and $\cos \omega t$, which are smooth (differentiable) continuous functions. In many engineering applications, however, one frequently encounters periodic functions that exhibit discontinuous behaviour. Examples of typical periodic functions of practical importance are shown in Figure 5.15.

Such periodic functions may be represented as infinite series of terms involving step functions; once expressed in such a form, the result (5.5) may then be used to obtain their Laplace transforms.

Figure 5.15

Typical practically important periodic functions: (a) square wave; (b) sawtooth wave; (c) repeated pulse wave; (d) half-wave rectifier.

**Example 5.11**

Obtain the Laplace transform of the square wave illustrated in Figure 5.15(a).

Solution

In terms of step functions, the square wave may be expressed in the form

$$\begin{aligned} f(t) &= KH(t) - 2KH\left(t - \frac{1}{2}T\right) + 2KH(t - T) - 2KH\left(t - \frac{3}{2}T\right) + 2KH(t - 2T) + \dots \\ &= K[H(t) - 2H\left(t - \frac{1}{2}T\right) + 2H(t - T) - 2H\left(t - \frac{3}{2}T\right) + 2H(t - 2T) + \dots] \end{aligned}$$

Taking Laplace transforms and using the result (5.5) gives

$$\begin{aligned} \mathcal{L}\{f(t)\} &= F(s) = K\left(\frac{1}{s} - \frac{2}{s}e^{-sT/2} + \frac{2}{s}e^{-sT} - \frac{2}{s}e^{-3sT/2} + \frac{2}{s}e^{-2sT} + \dots\right) \\ &= \frac{2K}{s} [1 - e^{-sT/2} + (e^{-sT/2})^2 - (e^{-sT/2})^3 + (e^{-sT/2})^4 - \dots] - \frac{K}{s} \end{aligned}$$

The series inside the square brackets is an infinite geometric progression with first term 1 and common ratio $-e^{-sT/2}$, and therefore has sum $(1 + e^{-sT/2})^{-1}$. Thus,

$$F(s) = \frac{2K}{s} \frac{1}{1 + e^{-sT/2}} - \frac{K}{s} = \frac{K}{s} \frac{1 - e^{-sT/2}}{1 + e^{-sT/2}}$$

That is,

$$\mathcal{L}\{f(t)\} = F(s) = \frac{K}{s} \tanh \frac{1}{4}sT$$

The approach used in Example 5.11 may be used to prove the following theorem, which provides an explicit expression for the Laplace transform of a periodic function.

Theorem 5.3

If $f(t)$, defined for all positive t , is a periodic function with period T , that is $f(t + nT) = f(t)$ for all integers n , then

$$\mathcal{L}\{f(t)\} = \frac{1}{1 - e^{-sT}} \int_0^T e^{-st} f(t) dt$$

Proof If, as illustrated in Figure 5.16, the periodic function $f(t)$ is piecewise-continuous over an interval of length T , then its Laplace transform exists and can be expressed as a series of integrals over successive periods; that is,

$$\begin{aligned} \mathcal{L}\{f(t)\} &= \int_0^\infty f(t) e^{-st} dt \\ &= \int_0^T f(t) e^{-st} dt + \int_T^{2T} f(t) e^{-st} dt + \int_{2T}^{3T} f(t) e^{-st} dt + \dots \\ &\quad + \int_{(n-1)T}^{nT} f(t) e^{-st} dt + \dots \end{aligned}$$

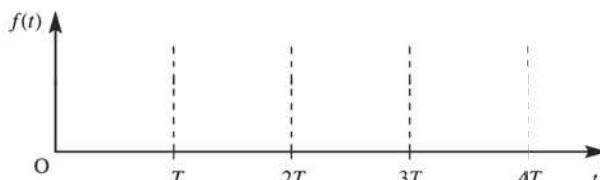
If in successive integrals we make the substitutions

$$t = \tau + nT \quad (n = 0, 1, 2, 3, \dots)$$

then

$$\mathcal{L}\{f(t)\} = \sum_{n=0}^{\infty} \int_0^T f(\tau + nT) e^{-s(\tau+nT)} d\tau$$

Figure 5.16
Periodic function
having period T .



Since $f(t)$ is periodic with period T ,

$$f(\tau + nT) = f(t) \quad (n = 0, 1, 2, 3, \dots)$$

so that

$$\mathcal{L}\{f(t)\} = \sum_{n=0}^{\infty} \int_0^T f(\tau) e^{-s\tau} e^{-snT} d\tau = \left(\sum_{n=0}^{\infty} e^{-snT} \right) \int_0^T f(\tau) e^{-s\tau} d\tau$$

The series $\sum_{n=0}^{\infty} e^{-snT} = 1 + e^{-sT} + e^{-2sT} + e^{-3sT} + \dots$ is an infinite geometric progression with first term 1 and common ratio e^{-sT} . Its sum is given by $(1 - e^{-sT})^{-1}$, so that

$$\mathcal{L}\{f(t)\} = \frac{1}{1 - e^{-sT}} \int_0^T f(\tau) e^{-s\tau} d\tau$$

Since, within the integral, τ is a ‘dummy’ variable, it may be replaced by t to give the desired result.

[end of theorem](#)

We note that, in terms of the Heaviside step function, Theorem 5.3 may be stated as follows:

If $f(t)$, defined for all positive t , is a periodic function with period T and

$$f_1(t) = f(t)(H(t) - H(t - T))$$

then

$$\mathcal{L}\{f(t)\} = (1 - e^{-sT})^{-1} \mathcal{L}\{f_1(t)\}$$

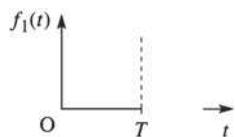


Figure 5.17
Plot of periodic function within one period.

This formulation follows since $f(t)$ is periodic and $f_1(t) = 0$ for $t > T$. For the periodic function $f(t)$ shown in Figure 5.16 the corresponding function $f_1(t)$ is shown in Figure 5.17. We shall see from the following examples that this formulation simplifies the process of obtaining Laplace transforms of periodic functions.

Example 5.12

Confirm the result obtained in Example 5.11 using Theorem 5.3.

Solution

For the square wave $f(t)$ illustrated in Figure 5.15(a), $f(t)$ is defined over the period $0 < t < T$ by

$$f(t) = \begin{cases} K & (0 < t < \frac{1}{2}T) \\ -K & (\frac{1}{2}T < t < T) \end{cases}$$

Hence we can write $f_1(t) = K[H(t) - 2H(t - \frac{1}{2}T) + H(t - T)]$, and thus

$$\mathcal{L}\{f_1(t)\} = K \left(\frac{1}{s} - \frac{2}{s} e^{-sT/2} + \frac{1}{s} e^{-sT} \right) = \frac{K}{s} (1 - e^{-sT/2})^2$$

Using the result of Theorem 5.3,

$$\begin{aligned}\mathcal{L}\{f(t)\} &= \frac{K(1 - e^{-sT/2})^2}{s(1 - e^{-sT})} = \frac{K(1 - e^{-sT/2})^2}{s(1 - e^{-sT/2})(1 + e^{-sT/2})} \\ &= \frac{K}{s} \frac{1 - e^{-sT/2}}{1 + e^{-sT/2}} = \frac{K}{s} \tanh \frac{1}{4}sT\end{aligned}$$

confirming the result obtained in Example 5.11.

Example 5.13

Determine the Laplace transform of the rectified half-wave defined by

$$f(t) = \begin{cases} \sin \omega t & (0 < t < \pi/\omega) \\ 0 & (\pi/\omega < t < 2\pi/\omega) \end{cases}$$

$$f(t + 2n\pi/\omega) = f(t) \quad \text{for all integers } n$$

Solution $f(t)$ is illustrated in Figure 5.15(d), with $T = 2\pi/\omega$. We can express $f_1(t)$ as

$$\begin{aligned}f_1(t) &= \sin \omega t [H(t) - H(t - \pi/\omega)] \\ &= \sin \omega t H(t) + \sin \omega(t - \pi/\omega) H(t - \pi/\omega)\end{aligned}$$

So

$$\mathcal{L}\{f_1(t)\} = \frac{\omega}{s^2 + \omega^2} + e^{-s\pi/\omega} \frac{\omega}{s^2 + \omega^2} = \frac{\omega}{s^2 + \omega^2}(1 + e^{-s\pi/\omega})$$

Then, by the result of Theorem 5.3,

$$\mathcal{L}\{f(t)\} = \frac{\omega}{s^2 + \omega^2} \frac{1 + e^{-s\pi/\omega}}{1 - e^{-2s\pi/\omega}} = \frac{\omega}{(s^2 + \omega^2)(1 - e^{-s\pi/\omega})}$$

5.2.7 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 1 A function $f(t)$ is defined by

$$f(t) = \begin{cases} t & (0 \leq t \leq 1) \\ 0 & (t > 1) \end{cases}$$

Express $f(t)$ in terms of Heaviside unit step functions and show that

$$\mathcal{L}\{f(t)\} = \frac{1}{s^2} (1 - e^{-s}) - \frac{1}{s} e^{-s}$$

- 2 Express in terms of Heaviside unit step functions the following piecewise-continuous causal functions. In each case obtain the Laplace transform of the function.

$$(a) f(t) = \begin{cases} 3t^2 & (0 < t \leq 4) \\ 2t - 3 & (4 < t < 6) \\ 5 & (t > 6) \end{cases}$$

$$(b) g(t) = \begin{cases} t & (0 \leq t < 1) \\ 2 - t & (1 < t < 2) \\ 0 & (t > 2) \end{cases}$$

- 3 Obtain the inverse Laplace transforms of the following:

$$(a) \frac{e^{-5s}}{(s-2)^4} \quad (b) \frac{3e^{-2s}}{(s+3)(s+1)}$$

$$\begin{array}{ll} \text{(c)} \frac{s+1}{s^2(s^2+1)} e^{-s} & \text{(d)} \frac{s+1}{s^2+s+1} e^{-\pi s} \\ \text{(e)} \frac{s}{s^2+25} e^{-4\pi s/5} & \text{(f)} \frac{e^{-s}(1-e^{-s})}{s^2(s^2+1)} \end{array}$$

- 4** Given that $x = 0$ when $t = 0$, obtain the solution of the differential equation

$$\frac{dx}{dt} + x = f(t) \quad (t \geq 0)$$

where $f(t)$ is the function defined in Exercise 1. Sketch a graph of the solution.

- 5** Given that $x = 1$ and $dx/dt = 0$, obtain the solution of the differential equation

$$\frac{d^2x}{dt^2} + \frac{dx}{dt} + x = g(t) \quad (t \geq 0)$$

where $g(t)$ is the piecewise-continuous function defined in Exercise 2(b).

- 6** Show that the function

$$f(t) = \begin{cases} 0 & (0 \leq t < \frac{1}{2}\pi) \\ \sin t & (t \geq \frac{1}{2}\pi) \end{cases}$$

may be expressed in the form $f(t) = \cos(t - \frac{1}{2}\pi)H(t - \frac{1}{2}\pi)$, where $H(t)$ is the Heaviside unit step function. Hence solve the differential equation

$$\frac{d^2x}{dt^2} + 3 \frac{dx}{dt} + 2x = f(t)$$

where $f(t)$ is given above, and $x = 1$ and $dx/dt = -1$ when $t = 0$.

- 7** Express the function

$$f(t) = \begin{cases} 3 & (0 \leq t < 4) \\ 2t - 5 & (t \geq 4) \end{cases}$$

in terms of Heaviside unit step functions and obtain its Laplace transform. Obtain the response of the harmonic oscillator

$$\ddot{x} + x = f(t)$$

to such a forcing function, given that $x = 1$ and $dx/dt = 0$ when $t = 0$.

- 8** The response $\theta_o(t)$ of a system to a forcing function $\theta_i(t)$ is determined by the second-order differential equation

$$\ddot{\theta}_o + 6\dot{\theta}_o + 10\theta_o = \theta_i \quad (t \geq 0)$$

Suppose that $\theta_i(t)$ is a constant stimulus applied for a limited period and characterized by

$$\theta_i(t) = \begin{cases} 3 & (0 \leq t < a) \\ 0 & (t \geq a) \end{cases}$$

Determine the response of the system at time t given that the system was initially in a quiescent state. Show that the transient response at time $T (> a)$ is

$$-\frac{3}{10}e^{-3T}\{\cos T + 3 \sin T - e^{3a}[\cos(T-a) + 3 \sin(T-a)]\}$$

9

The input $\theta_i(t)$ and output $\theta_o(t)$ of a servomechanism are related by the differential equation

$$\ddot{\theta}_o + 8\dot{\theta}_o + 16\theta_o = \theta_i \quad (t \geq 0)$$

and initially $\theta_o(0) = \dot{\theta}_o(0) = 0$. For $\theta_i = f(t)$, where

$$f(t) = \begin{cases} 1-t & (0 < t < 1) \\ 0 & (t \geq 1) \end{cases}$$

Show that

$$\mathcal{L}\{\theta_i(t)\} = \frac{s-1}{s^2} + \frac{1}{s^2}e^{-s}$$

and hence obtain an expression for the response of the system at time t .

10

During the time interval t_1 to t_2 , a constant electromotive force e_0 acts on the series RC circuit shown in Figure 5.18. Assuming that the circuit is initially in a quiescent state, show that the current in the circuit at time t is

$$i(t) = \frac{e_0}{R} [e^{-(t-t_1)/RC} H(t-t_1) - e^{-(t-t_2)/RC} H(t-t_2)]$$

Sketch this as a function of time.

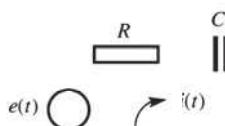


Figure 5.18 Circuit of Exercise 10.

11

A periodic function $f(t)$, with period 4 units, is defined within the interval $0 \leq t < 4$ by

$$f(t) = \begin{cases} 3t & (0 \leq t < 2) \\ 6 & (2 \leq t < 4) \end{cases}$$

Sketch a graph of the function for $0 \leq t < 12$ and obtain its Laplace transform.

12

Obtain the Laplace transform of the periodic sawtooth wave with period T , illustrated in Figure 5.15(b).

5.2.8 The impulse function

Suppose a hammer is used to strike a nail then the hammer will be in contact with the nail for a very short period of time, indeed almost instantaneously. A similar situation arises when a golfer strikes a golf ball. In both cases the force applied, during this short period of time, builds up rapidly to a large value and then rapidly decreases to zero. Such short sharp forces are known as **impulsive forces** and are of interest in many engineering applications. In practice it is not the duration of contact that is important but the momentum transmitted, this being proportional to the time integral of the force applied. Mathematically such forcing functions are represented by the **impulse function**. To develop a mathematical formulation of the impulse function and obtain some insight into its physical interpretation, consider the pulse function $\phi(t)$ defined by

$$\phi(t) = \begin{cases} 0 & (0 < t < a - \frac{1}{2}T) \\ A/T & (a - \frac{1}{2}T \leq t < a + \frac{1}{2}T) \\ 0 & (t \geq a + \frac{1}{2}T) \end{cases}$$

and illustrated in Figure 5.19(a). Since the height of the pulse is A/T and its duration (or width) is T , the area under the pulse is A ; that is,

$$\int_{-\infty}^{\infty} \phi(t) dt = \int_{a-T/2}^{a+T/2} \frac{A}{T} dt = A$$

If we now consider the limiting process in which the duration of the pulse approaches zero, in such a way that the area under the pulse remains A , then we obtain a formulation of the impulse function of magnitude A occurring at time $t = a$. It is important to appreciate that the magnitude of the impulse function is measured by its area.

The impulse function whose magnitude is unity is called the **unit impulse function** or **Dirac delta function** (or simply **delta function**). The unit impulse occurring at $t = a$ is the limiting case of the pulse $\phi(t)$ of Figure 5.19(a) with A having the value unity. It is denoted by $\delta(t - a)$ and has the properties

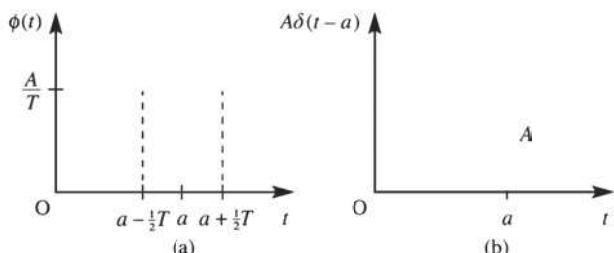
$$\delta(t - a) = 0 \quad (t \neq a)$$

$$\int_{-\infty}^{\infty} \delta(t - a) dt = 1$$

Likewise, an impulse function of magnitude A occurring at $t = a$ is denoted by $A\delta(t - a)$ and may be represented diagrammatically as in Figure 5.19(b).

An impulse function is not a function in the usual sense, but is an example of a class of what are called **generalized functions**, which may be analysed using the theory of

Figure 5.19
Impulse function.



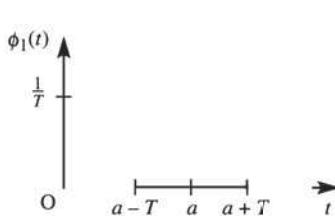


Figure 5.20 Approximation to a unit pulse.

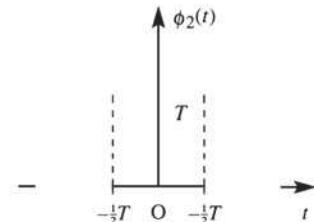


Figure 5.21 Pulse at the origin.

generalized calculus. (It may also be regarded mathematically as a **distribution** and investigated using the **theory of distributions**.) However, its properties are such that, used with care, it can lead to results that have physical or practical significance and which in many cases cannot be obtained by any other method. In this context it provides engineers with an important mathematical tool. Although, clearly, an impulse function is not physically realizable, it follows from the above formulation that physical signals can be produced that closely approximate it.

We noted that the magnitude of the impulse function is determined by the area under the limiting pulse. The actual shape of the limiting pulse is not really important, provided that the area contained within it remains constant as its duration approaches zero. Physically, therefore, the unit impulse function at $t = a$ may equally well be regarded as the pulse $\phi_1(t)$ of Figure 5.20 in the limiting case as T approaches zero.

In some applications we need to consider a unit impulse function at time $t = 0$. This is denoted by $\delta(t)$ and is defined as the limiting case of the pulse $\phi_2(t)$ illustrated in Figure 5.21 as T approaches zero. It has the properties

$$\delta(t) = 0 \quad (t \neq 0)$$

$$\int_{-\infty}^{\infty} \delta(t) dt = 1$$

5.2.9 The sifting property

An important property of the unit impulse function that is of practical significance is the so-called **sifting property**, which states that if $f(t)$ is continuous at $t = a$ then

$$\int_{-\infty}^{\infty} f(t) \delta(t-a) dt = f(a) \tag{5.10}$$

This is referred to as the sifting property because it provides a method of isolating, or sifting out, the value of a function at any particular point.

For theoretical reasons it is convenient to use infinite limits in (5.10), while in reality finite limits can be substituted. This follows since for $\alpha < a < \beta$, where α and β are constants,

$$\int_{\alpha}^{\beta} f(t) \delta(t-a) dt = f(a) \tag{5.11}$$

For example,

$$\int_0^{2\pi} \cos t \delta(t - \frac{1}{3}\pi) dt = \cos \frac{1}{3}\pi = \frac{1}{2}$$

5.2.10 Laplace transforms of impulse functions

By the definition of the Laplace transform, we have for any $a > 0$

$$\mathcal{L}\{\delta(t-a)\} = \int_0^\infty \delta(t-a) e^{-st} dt$$

which, using the sifting property, gives the important result

$$\mathcal{L}\{\delta(t-a)\} = e^{-as} \quad (5.12)$$

or, in terms of the inverse transform,

$$\mathcal{L}^{-1}\{e^{-as}\} = \delta(t-a) \quad (5.13)$$

As mentioned earlier, in many applications we may have an impulse function $\delta(t)$ at $t = 0$, and it is in order to handle such a function that we must carefully specify whether the lower limit in the Laplace integral defined in Section 5.1.1 is 0^- or 0^+ . Adopting the notation

$$\mathcal{L}_+\{f(t)\} = \int_{0^+}^\infty f(t) e^{-st} dt$$

$$\mathcal{L}_-\{f(t)\} = \int_{0^-}^\infty f(t) e^{-st} dt$$

we have

$$\mathcal{L}_-\{f(t)\} = \int_{0^-}^{0^+} f(t) e^{-st} dt + \int_{0^+}^\infty f(t) e^{-st} dt$$

If $f(t)$ does not involve an impulse function at $t = 0$ then clearly $\mathcal{L}_+\{f(t)\} = \mathcal{L}_-\{f(t)\}$. However, if $f(t)$ does involve an impulse function at $t = 0$ then

$$\int_{0^-}^{0^+} f(t) dt \neq 0$$

and it follows that

$$\mathcal{L}_+\{f(t)\} \neq \mathcal{L}_-\{f(t)\}$$

We adopt the definition (see Section 11.2 of MEM)

$$\mathcal{L}\{f(t)\} = \mathcal{L}_-\{f(t)\}$$

so that (5.12) and (5.13) hold for $a = 0$, giving

$$\mathcal{L}\{\delta(t)\} = \int_{0^-}^\infty \delta(t) e^{-st} dt = e^{-s0} = 1$$

so that

$$\mathcal{L}\{\delta(t)\} = 1 \quad (5.14)$$

or, in inverse form,

$$\mathcal{L}^{-1}\{1\} = \delta(t) \quad (5.15)$$



This transform can be implemented in MATLAB using the sequence of commands

```
syms s t
del=sym('Dirac(t)');
laplace(del)
```

Likewise for (5.12); for example, if $a = 2$ then the Laplace transform of $\delta(t - 2)$ is generated by the commands

```
del2=sym('Dirac(t-2)');
laplace(del2)
```

or directly using the command

```
laplace(sym('Dirac(t-2)'))
```

giving the answer $\exp(-2*s)$ in each case.

In MAPLE the commands

```
with(inttrans):
laplace(Dirac(t-2), t, s);
return the answer e^{(-2s)}.
```

Example 5.14

Determine $\mathcal{L}^{-1}\left\{\frac{s^2}{s^2 + 4}\right\}$.

Solution Since

$$\frac{s^2}{s^2 + 4} = \frac{s^2 + 4 - 4}{s^2 + 4} = 1 - \frac{4}{s^2 + 4}$$

we have

$$\mathcal{L}^{-1}\left\{\frac{s^2}{s^2 + 4}\right\} = \mathcal{L}^{-1}\{1\} - \mathcal{L}^{-1}\left\{\frac{4}{s^2 + 4}\right\}$$

giving

$$\mathcal{L}^{-1}\left\{\frac{s^2}{s^2 + 4}\right\} = \delta(t) - 2 \sin 2t$$



In MATLAB this is obtained directly, with the commands

```
ilaplace(s^2/(s^2+4));
pretty(ans)
```

generating the answer

```
Dirac(t)-2sin2t
```

The answers may also be obtained in MAPLE using the commands

```
with(inttrans):
invlaplace(s^2/(s^2+4), s, t);
```

Example 5.15

Determine the solution of the differential equation

$$\frac{d^2x}{dt^2} + 3 \frac{dx}{dt} + 2x = 1 + \delta(t - 4) \quad (5.16)$$

subject to the initial conditions $x(0) = \dot{x}(0) = 0$.

Solution

Taking Laplace transforms in (5.16) gives

$$[s^2X(s) - sx(0) - \dot{x}(0)] + 3[sX(s) - x(0)] + 2X(s) = \mathcal{L}\{1\} + \mathcal{L}\{\delta(t - 4)\}$$

which, on incorporating the given initial conditions and using (5.12), leads to

$$(s^2 + 3s + 2)X(s) = \frac{1}{s} + e^{-4s}$$

giving

$$X(s) = \frac{1}{s(s+2)(s+1)} + e^{-4s} \frac{1}{(s+2)(s+1)}$$

Resolving into partial fractions, we have

$$X(s) = \frac{1}{2} \left(\frac{1}{s} + \frac{1}{s+2} - \frac{2}{s+1} \right) + e^{-4s} \left(\frac{1}{s+1} - \frac{1}{s+2} \right)$$

which, on taking inverse transforms and using the result (5.7), gives the required response:

$$x(t) = \frac{1}{2} (1 + e^{-2t} - 2e^{-t}) + (e^{-(t-4)} - e^{-2(t-4)})H(t - 4)$$

or, in an alternative form,

$$x(t) = \begin{cases} \frac{1}{2}(1 + e^{-2t} - 2e^{-t}) & (0 \leq t < 4) \\ \frac{1}{2} + (e^4 - 1)e^{-t} - (e^8 - \frac{1}{2})e^{-2t} & (t \geq 4) \end{cases}$$

We note that, although the response $x(t)$ is continuous at $t = 4$, the consequence of the impulsive input at $t = 4$ is a step change in the derivative $\dot{x}(t)$.



As was the case in Example 5.9, when considering Heaviside functions as forcing terms, it seems that the `dsolve` command in MATLAB cannot be used directly in this case. Using the `maple` command the following commands:

```
maple('de:=diff(x(t),t$2)+3*diff(x(t),t)+2*x(t)
      = 1+Dirac(t-4);')
ans=
de := diff(x(t), '$'(t, 2)) + 3*diff(x(t), t) + 2*x(t)
= 1+Dirac(t-4)
maple('dsolve({de,x(0)=0,D(x)(0)=0},x(t)),
method=laplace;')

output the required answer:
x(t)=1/2-exp(-t)+1/2*exp(-2*t)-Heaviside(t-4)*
exp(-2*t+8)+Heaviside(t-4)*exp(-t+4)
```

5.2.11 Relationship between Heaviside step and impulse functions

From the definitions of $H(t)$ and $\delta(t)$, it can be argued that

$$H(t) = \int_{-\infty}^t \delta(\tau) d\tau \quad (5.17)$$

since the interval of integration contains zero if $t > 0$ but not if $t < 0$. Conversely, (5.17) may be written as

$$\delta(t) = \frac{d}{dt} H(t) = H'(t) \quad (5.18)$$

which expresses the fact that $H'(t)$ is zero everywhere except at $t = 0$, when the jump in $H(t)$ occurs.

While this argument may suffice in practice, since we are dealing with generalized functions a more formal proof requires the development of some properties of generalized functions. In particular, we need to define what is meant by saying that two generalized functions are equivalent.

One method of approach is to use the concept of a **test function** $\theta(t)$, which is a continuous function that has continuous derivatives of all orders and that is zero outside a finite interval. One class of testing function, adopted by R. R. Gabel and R. A. Roberts (*Signals and Linear Systems*, New York, Wiley, 1973), is

$$\theta(t) = \begin{cases} e^{-d^2/(d^2-t^2)} & (|t| < d), \text{ where } d = \text{constant} \\ 0 & \text{otherwise} \end{cases}$$

For a generalized function $g(t)$ the integral

$$G(\theta) = \int_{-\infty}^{\infty} \theta(t)g(t) dt$$

is evaluated. This integral assigns the number $G(\theta)$ to each function $\theta(t)$, so that $G(\theta)$ is a generalization of the concept of a function: it is a **linear functional** on the space of test functions $\theta(t)$. For example, if $g(t) = \delta(t)$ then

$$G(\theta) = \int_{-\infty}^{\infty} \theta(t) \delta(t) dt = \theta(0)$$

so that in this particular case, for each weighting function $\theta(t)$, the value $\theta(0)$ is assigned to $G(\theta)$.

We can now use the concept of a test function to define what is meant by saying that two generalized functions are equivalent or ‘equal’.

Definition 5.2: The equivalence property

If $g_1(t)$ and $g_2(t)$ are two generalized functions then $g_1(t) = g_2(t)$ if and only if

$$\int_{-\infty}^{\infty} \theta(t)g_1(t) dt = \int_{-\infty}^{\infty} \theta(t)g_2(t) dt$$

for all test functions $\theta(t)$ for which the integrals exist.

The test function may be regarded as a ‘device’ for examining the generalized function. Gabel and Roberts draw a rough parallel with the role of using the output of a measuring instrument to deduce properties about what is being measured. In such an analogy $g_1(t) = g_2(t)$ if the measuring instrument can detect no differences between them.

Using the concept of a test function $\theta(t)$, the Dirac delta function $\delta(t)$ may be defined in the generalized form

$$\int_{-\infty}^{\infty} \theta(t) \delta(t) dt = \theta(0)$$

Interpreted as an ordinary integral, this has no meaning. The integral and the function $\delta(t)$ are merely defined by the number $\theta(0)$. In this sense we can handle $\delta(t)$ as if it were an ordinary function, except that we never talk about the value of $\delta(t)$; rather we talk about the value of integrals involving $\delta(t)$.

Using the equivalence property, we can now confirm the result (5.18), namely that

$$\delta(t) = \frac{d}{dt} H(t) = H'(t)$$

To prove this, we must show that

$$\int_{-\infty}^{\infty} \theta(t) \delta(t) dt = \int_{-\infty}^{\infty} \theta(t) H'(t) dt \quad (5.19)$$

Integrating the right-hand side of (5.19) by parts, we have

$$\begin{aligned}
 \int_{-\infty}^{\infty} \theta(t) H'(t) dt &= [H(t) \theta(t)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} H(t) \theta'(t) dt \\
 &= 0 - \int_{-\infty}^{\infty} \theta'(t) dt \quad (\text{by the definitions of } \theta(t) \text{ and } H(t)) \\
 &= -[\theta(t)]_{-\infty}^{\infty} = \theta(0)
 \end{aligned}$$

Since the left-hand side of (5.19) is also $\theta(0)$, the equivalence of $\delta(t)$ and $H'(t)$ is proved. Likewise, it can be shown that

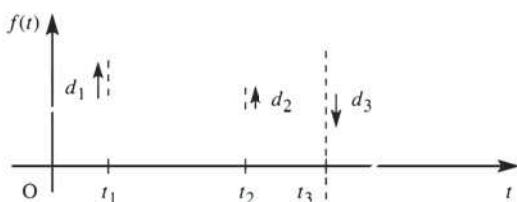
$$\delta(t-a) = \frac{d}{dt} H(t-a) = H'(t-a) \quad (5.20)$$

The results (5.18) and (5.20) may be used to obtain the **generalized derivatives** of piecewise-continuous functions having jump discontinuities d_1, d_2, \dots, d_n at times t_1, t_2, \dots, t_n respectively, as illustrated in Figure 5.22. On expressing $f(t)$ in terms of Heaviside step functions as in Section 5.2.1, and differentiating using the product rule, use of (5.18) and (5.20) leads to the result

$$f'(t) = g'(t) + \sum_{i=1}^n d_i \delta(t-t_i) \quad (5.21)$$

where $g'(t)$ denotes the ordinary derivative of $f(t)$ where it exists. The result (5.21) tells us that the derivative of a piecewise-continuous function with jump discontinuities is the ordinary derivative where it exists plus the sum of delta functions at the discontinuities multiplied by the magnitudes of the respective jumps.

Figure 5.22
Piecewise-continuous
function with jump
discontinuities.



By the magnitude d_i of a jump in a function $f(t)$ at a point t_i , we mean the difference between the right-hand and left-hand limits of $f(t)$ at t_i ; that is,

$$d_i = f(t_i+0) - f(t_i-0)$$

It follows that an upward jump, such as d_1 and d_2 in Figure 5.22, is positive, while a downward jump, such as d_3 in Figure 5.22, is negative.

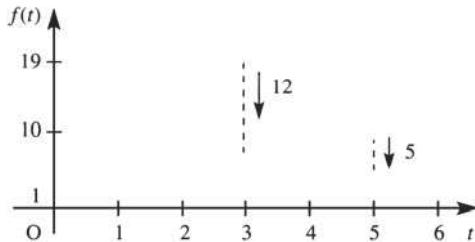
The result (5.21) gives an indication as to why the use of differentiators in practical systems is not encouraged, since the introduction of impulses means that derivatives increase noise levels in signal reception. In contrast, integrators have a smoothing effect on signals, and are widely used.

Example 5.16

Obtain the generalized derivative of the piecewise-continuous function

$$f(t) = \begin{cases} 2t^2 + 1 & (0 \leq t < 3) \\ t + 4 & (3 \leq t < 5) \\ 4 & (t \geq 5) \end{cases}$$

Figure 5.23 Piecewise-continuous function of Example 5.16.



Solution $f(t)$ is depicted graphically in Figure 5.23, and it has jump discontinuities of magnitudes 1, -12 and -5 at times $t = 0, 3$ and 5 respectively. Using (5.21), the generalized derivative is

$$f'(t) = g'(t) + 1\delta(t) - 12\delta(t-3) - 5\delta(t-5)$$

where

$$g'(t) = \begin{cases} 4t & (0 \leq t < 3) \\ 1 & (3 \leq t < 5) \\ 0 & (t \geq 5) \end{cases}$$

Example 5.17

A system is characterized by the differential equation model

$$\frac{d^2x}{dt^2} + 5\frac{dx}{dt} + 6x = u + 3\frac{du}{dt} \quad (5.22)$$

Determine the response of the system to a forcing function $u(t) = e^{-t}$ applied at time $t = 0$, given that it was initially in a quiescent state.

Solution Since the system is initially in a quiescent state, the transformed equation corresponding to (5.22) is

$$(s^2 + 5s + 6)X(s) = (3s + 1)U(s)$$

giving

$$X(s) = \frac{3s+1}{s^2 + 5s + 6} U(s)$$

In the particular case when $u(t) = e^{-t}$, $U(s) = 1/(s+1)$, so that

$$X(s) = \frac{(3s+1)}{(s+1)(s+2)(s+3)} = \frac{-1}{s+1} + \frac{5}{s+2} - \frac{4}{s+3}$$

which, on taking inverse transforms, gives the desired response as

$$x(t) = -e^{-t} + 5e^{-2t} - 4e^{-3t} \quad (t \geq 0)$$

One might have been tempted to adopt a different approach and substitute for $u(t)$ directly in (5.22) before taking Laplace transforms. This leads to

$$\frac{d^2x}{dt^2} + 5\frac{dx}{dt} + 6x = e^{-t} - 3e^{-t} = -2e^{-t}$$

which, on taking Laplace transforms, leads to

$$(s^2 + 5s + 6)X(s) = -\frac{2}{s+1}$$

giving

$$X(s) = \frac{-2}{(s+1)(s+2)(s+3)} = \frac{-1}{s+1} + \frac{2}{s+2} - \frac{1}{s+3}$$

which, on inversion, gives

$$x(t) = -e^{-t} + 2e^{-2t} - e^{-3t} \quad (t \geq 0)$$

Clearly this approach results in a different solution, and therefore appears to lead to a paradox. However, this apparent paradox can be resolved by noting that the second approach is erroneous in that it ignores the important fact that we are dealing with causal functions. Strictly speaking,

$$u(t) = e^{-t}H(t)$$

and, when determining du/dt , the product rule of differential calculus should be employed, giving

$$\begin{aligned}\frac{du}{dt} &= -e^{-t}H(t) + e^{-t}\frac{d}{dt}H(t) \\ &= -e^{-t}H(t) + e^{-t}\delta(t)\end{aligned}$$

Substituting this into (5.22) and taking Laplace transforms gives

$$(s^2 + 5s + 6)X(s) = \frac{1}{s+1} + 3\left(-\frac{1}{s+1} + 1\right) = \frac{3s+1}{s+1}$$

That is,

$$X(s) = \frac{3s+1}{(s+1)(s^2 + 5s + 6)}$$

leading to the same response

$$x(t) = -e^{-t} + 5e^{-2t} - 4e^{-3t} \quad (t \geq 0)$$

as in the first approach above.

The differential equation used in Example 5.17 is of a form that occurs frequently in practice, so it is important that the causal nature of the forcing term be recognized.

The derivative $\delta'(t)$ of the impulse function is also a generalized function, and, using the equivalence property, it is readily shown that

$$\int_{-\infty}^{\infty} f(t) \delta'(t) dt = -f'(0)$$

or, more generally,

$$\int_{-\infty}^{\infty} f(t) \delta'(t-a) dt = -f'(a)$$

provided that $f'(t)$ is continuous at $t = a$.

Likewise, the n th derivative satisfies

$$\int_{-\infty}^{\infty} f(t) \delta^n(t-a) dt = (-1)^n f^{(n)}(a)$$

provided that $f^{(n)}(t)$ is continuous at $t = a$.

Using the definition of the Laplace transform, it follows that

$$\mathcal{L}\{\delta^{(n)}(t-a)\} = s^n e^{-as}$$

and, in particular,

$$\mathcal{L}\{\delta^{(n)}(t)\} = s^n \quad (5.23)$$

5.2.12 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 13 Obtain the inverse Laplace transforms of the following:

(a) $\frac{2s^2+1}{(s+2)(s+3)}$ (b) $\frac{s^2-1}{s^2+4}$ (c) $\frac{s^2+2}{s^2+2s+5}$

(c) $\frac{d^2x}{dt^2} + 7 \frac{dx}{dt} + 12x = \delta(t-3)$

subject to $x = 1$ and $\frac{dx}{dt} = 1$ at $t = 0$

- 14 Solve for $t \geq 0$ the following differential equations, subject to the specified initial conditions:

(a) $\frac{d^2x}{dt^2} + 7 \frac{dx}{dt} + 12x = 2 + \delta(t-2)$

subject to $x = 0$ and $\frac{dx}{dt} = 0$ at $t = 0$

(b) $\frac{d^2x}{dt^2} + 6 \frac{dx}{dt} + 13x = \delta(t-2\pi)$

subject to $x = 0$ and $\frac{dx}{dt} = 0$ at $t = 0$

15

Obtain the generalized derivatives of the following piecewise-continuous functions:

(a) $f(t) = \begin{cases} 3t^2 & (0 \leq t < 4) \\ 2t-3 & (4 \leq t < 6) \\ 5 & (t \geq 6) \end{cases}$

(b) $g(t) = \begin{cases} t & (0 \leq t < 1) \\ 2-t & (1 \leq t < 2) \\ 0 & (t \geq 2) \end{cases}$

$$(c) f(t) = \begin{cases} 2t+5 & (0 \leq t < 2) \\ 9-3t & (2 \leq t < 4) \\ t^2-t & (t \geq 4) \end{cases}$$

$$\frac{d^2x}{dt^2} + \omega^2 x = f(t) \quad (t \geq 0)$$

Show that

- 16 Solve for $t \geq 0$ the differential equation

$$\frac{d^2x}{dt^2} + 7 \frac{dx}{dt} + 10x = 2u + 3 \frac{du}{dt}$$

subject to $x = 0$ and $dx/dt = 2$ at $t = 0$ and where $u(t) = e^{-2t}H(t)$.

- 17 A periodic function $f(t)$ is an infinite train of unit impulses at $t = 0$ and repeated at intervals of $t = T$. Show that

$$\mathcal{L}\{f(t)\} = \frac{1}{1 - e^{-sT}}$$

The response of a harmonic oscillator to such a periodic stimulus is determined by the differential equation

$$x(t) = \frac{1}{\omega} \sum_{n=0}^{\infty} H(t-nT) \sin \omega(t-nT) \quad (t \geq 0)$$

and sketch the responses from $t = 0$ to $t = 6\pi/\omega$ for the two cases (a) $T = \pi/\omega$ and (b) $T = 2\pi/\omega$

- 18 An impulse voltage $E\delta(t)$ is applied at time $t = 0$ to a circuit consisting of a resistor R , a capacitor C and an inductor L connected in series. Prior to application of this voltage, both the charge on the capacitor and the resulting current in the circuit are zero. Determine the charge $q(t)$ on the capacitor and the resulting current $i(t)$ in the circuit at time t .

5.2.13 Bending of beams

So far, we have considered examples in which Laplace transform methods have been used to solve initial-value-type problems. These methods may also be used to solve boundary-value problems, and, to illustrate, we consider in this section the application of Laplace transform methods to determine the transverse deflection of a uniform thin beam due to loading.

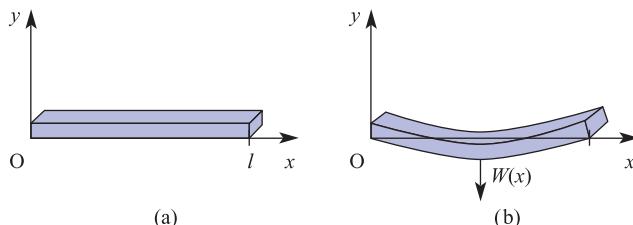
Consider a thin uniform beam of length l and let $y(x)$ be its transverse displacement, at distance x measured from one end, from the original position due to loading. The situation is illustrated in Figure 5.24, with the displacement measured upwards. Then, from the elementary theory of beams, we have

$$EI \frac{d^4y}{dx^4} = -W(x) \quad (5.24)$$

where $W(x)$ is the transverse force per unit length, with a downwards force taken to be positive, and EI is the flexural rigidity of the beam (E is Young's modulus of elasticity and I is the moment of inertia of the beam about its central axis). It is assumed that the beam has uniform elastic properties and a uniform cross-section over its length, so that both E and I are taken to be constants.

Figure 5.24

Transverse deflection of a beam: (a) initial position; (b) displaced position.



Equation (5.24) is sometimes written as

$$EI \frac{d^4y}{dx^4} = W(x)$$

where $y(x)$ is the transverse displacement measured downwards and not upwards as in (5.24).

In cases when the loading is uniform along the full length of the beam, that is $W(x) = \text{constant}$, (5.24) may be readily solved by the normal techniques of integral calculus. However, when the loading is non-uniform, the use of Laplace transform methods has a distinct advantage, since by making use of Heaviside unit functions and impulse functions, the problem of solving (5.24) independently for various sections of the beam may be avoided.

Taking Laplace transforms throughout in (5.24) gives

$$EI[s^4Y(s) - s^3y(0) - s^2y_1(0) - sy_2(0) - y_3(0)] = -W(s) \quad (5.25)$$

where

$$y_1(0) = \left(\frac{dy}{dx} \right)_{x=0}, \quad y_2(0) = \left(\frac{d^2y}{dx^2} \right)_{x=0}, \quad y_3(0) = \left(\frac{d^3y}{dx^3} \right)_{x=0}$$

and may be interpreted physically as follows:

- $EIy_3(0)$ is the shear at $x = 0$
- $EIy_2(0)$ is the bending moment at $x = 0$
- $y_1(0)$ is the slope at $x = 0$
- $y(0)$ is the deflection at $x = 0$

Solving (5.25) for $y(s)$ leads to

$$Y(s) = -\frac{W(s)}{EIs^4} + \frac{y(0)}{s} + \frac{y_1(0)}{s^2} + \frac{y_2(0)}{s^3} + \frac{y_3(0)}{s^4} \quad (5.26)$$

Thus four boundary conditions need to be found, and ideally they should be the shear, bending moment, slope and deflection at $x = 0$. However, in practice these boundary conditions are not often available. While some of them are known, other boundary conditions are specified at points along the beam other than at $x = 0$, for example conditions at the far end, $x = l$, or conditions at possible points of support along the beam. That is, we are faced with a boundary-value problem rather than an initial-value problem.

To proceed, known conditions at $x = 0$ are inserted, while the other conditions among $y(0)$, $y_1(0)$, $y_2(0)$ and $y_3(0)$ that are not specified are carried forward as undetermined constants. Inverse transforms are taken throughout in (5.7) to obtain the deflection $y(x)$, and the outstanding undetermined constants are obtained using the boundary conditions specified at points along the beam other than at $x = 0$.

The boundary conditions are usually embodied in physical conditions such as the following:

- (a) The beam is freely, or simply, supported at both ends, indicating that both the bending moments and deflection are zero at both ends, so that $y = d^2y/dx^2 = 0$ at both $x = 0$ and $x = l$.
- (b) At both ends the beam is clamped, or built into a wall. Thus the beam is horizontal at both ends, so that $y = dy/dx = 0$ at both $x = 0$ and $x = l$.

- (c) The beam is a cantilever with one end free (that is, fixed horizontally at one end, with the other end free). At the fixed end (say $x = 0$)

$$y = \frac{dy}{dx} = 0 \quad \text{at } x = 0$$

and at the free end ($x = l$), since both the shearing force and bending moment are zero,

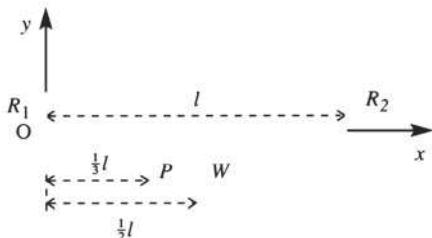
$$\frac{d^2y}{dx^2} = \frac{d^3y}{dx^3} = 0 \quad \text{at } x = l$$

If the load is not uniform along the full length of the beam, use is made of Heaviside step functions and impulse functions in specifying $W(x)$ in (5.24). For example, a uniform load w per unit length over the portion of the beam $x = x_1$ to $x = x_2$ is specified as $wH(x - x_1) - wH(x - x_2)$, and a point load w at $x = x_1$ is specified as $w\delta(x - x_1)$.

Example 5.18

Figure 5.25 illustrates a uniform beam of length l , freely supported at both ends, bending under uniformly distributed self-weight W and a concentrated point load P at $x = \frac{1}{3}l$. Determine the transverse deflection $y(x)$ of the beam.

Figure 5.25
Loaded beam of Example 5.18.



Solution

As in Figure 5.24, the origin is taken at the left-hand end of the beam, and the deflection $y(x)$ measured upwards from the horizontal at the level of the supports. The deflection $y(x)$ is then given by (5.24), with the force function $W(x)$ having contributions from the weight W , the concentrated load P and the support reactions R_1 and R_2 . However, since we are interested in solving (5.24) for $0 \leq x \leq l$, point loads or reactions at the end $x = l$ may be omitted from the force function.

As a preliminary, we need to determine R_1 . This is done by taking static moments about the end $x = l$, assuming the weight W to be concentrated at the centroid $x = \frac{1}{2}l$, giving

$$R_1 l = \frac{1}{2} W l + P \frac{2}{3} l$$

or

$$R_1 = \frac{1}{2} W + \frac{2}{3} P$$

The force function $W(x)$ may then be expressed as

$$W(x) = \frac{W}{l} H(x) + P \delta(x - \frac{1}{3}l) - (\frac{1}{2} W + \frac{2}{3} P) \delta(x)$$

with a Laplace transform

$$W(s) = \frac{W}{ls} + P e^{-ls/3} - (\frac{1}{2} W + \frac{2}{3} P)$$

Since the beam is freely supported at both ends, the deflection and bending moments are zero at both ends, so we take the boundary conditions as

$$y = 0 \quad \text{at } x = 0 \text{ and } x = l$$

$$\frac{d^2y}{dx^2} = 0 \quad \text{at } x = 0 \text{ and } x = l$$

The transformed equation (5.26) becomes

$$Y(s) = -\frac{1}{EI} \left[\frac{W}{ls^5} + \frac{P}{s^4} e^{-ls/3} - \left(\frac{1}{2}W + \frac{2}{3}P \right) \frac{1}{s^4} \right] + \frac{y_1(0)}{s^2} + \frac{y_3(0)}{s^4}$$

Taking inverse transforms, making use of the second shift theorem (Theorem 5.2), gives the deflection $y(x)$ as

$$y(x) = -\frac{1}{EI} \left[\frac{1}{24} \frac{W}{l} x^4 + \frac{1}{6} P \left(x - \frac{1}{3}l \right)^3 H(x - \frac{1}{3}l) - \frac{1}{6} \left(\frac{1}{2}W + \frac{2}{3}P \right) x^3 \right] \\ + y_1(0)x + \frac{1}{6} y_3(0)x^3$$

To obtain the value of the undetermined constants $y_1(0)$ and $y_3(0)$, we employ the unused boundary conditions at $x = l$, namely $y(l) = 0$ and $y_2(l) = 0$. For $x > \frac{1}{3}l$

$$y(x) = -\frac{1}{EI} \left[\frac{1}{24} \frac{W}{l} x^4 + \frac{1}{6} P \left(x - \frac{1}{3}l \right)^3 - \frac{1}{6} \left(\frac{1}{2}W + \frac{2}{3}P \right) x^3 \right] + y_1(0)x + \frac{1}{6} y_3(0)x^3$$

$$\frac{d^2y}{dx^2} = y_2(x) = -\frac{1}{EI} \left[\frac{Wx^2}{2l} + P \left(x - \frac{1}{3}l \right) - \left(\frac{1}{3}W + \frac{2P}{3} \right) x \right] + y_3(0).$$

Thus taking $y_2(l) = 0$ gives $y_3(0) = 0$, and taking $y(l) = 0$ gives

$$-\frac{1}{EI} \left(\frac{1}{24} Wl^3 + \frac{4}{81} Pl^3 - \frac{1}{12} Wl^3 - \frac{1}{9} Pl^3 \right) + y_1(0)l = 0$$

so that

$$y_1(0) = -\frac{l^2}{El} \left(\frac{1}{24} W + \frac{5}{81} P \right)$$

Substituting back, we find that the deflection $y(x)$ is given by

$$y(x) = -\frac{W}{EI} \left(\frac{x^4}{24l} - \frac{1}{12} x^3 + \frac{1}{24} l^2 x \right) - \frac{P}{EI} \left(\frac{5}{81} l^2 x - \frac{1}{9} x^3 \right) - \frac{P}{6EI} \left(x - \frac{1}{3}l \right)^3 H \left(x - \frac{1}{3}l \right)$$

or, for the two sections of the beam,

$$y(x) = \begin{cases} -\frac{W}{EI} \left(\frac{x}{24l} - \frac{1}{12} x^3 + \frac{1}{24} l^2 x \right) - \frac{P}{EI} \left(\frac{5}{81} l^2 x - \frac{1}{9} x^3 \right) & \left(0 < x < \frac{1}{3}l \right) \\ -\frac{W}{EI} \left(\frac{x^4}{24l} - \frac{1}{12} x^3 + \frac{1}{24} l^2 x \right) - \frac{P}{EI} \left(\frac{19}{162} l^2 x + \frac{1}{18} x^3 - \frac{1}{6} x^2 l - \frac{1}{162} l^3 \right) & \left(\frac{1}{3}l < x < l \right) \end{cases}$$

5.2.14 Exercises

- 19 Find the deflection of a beam simply supported at its ends $x = 0$ and $x = l$, bending under a uniformly distributed self-weight M and a concentrated load W at $x = \frac{1}{2}l$.
- 20 A cantilever beam of negligible weight and of length l is clamped at the end $x = 0$. Determine the deflection of the beam when it is subjected to a load

per unit length, w , over the section $x = x_1$ to $x = x_2$. What is the maximum deflection if $x_1 = 0$ and $x_2 = l$?

- 21 A uniform cantilever beam of length l is subjected to a concentrated load W at a point distance b from the fixed end. Determine the deflection of the beam, distinguishing between the sections $0 < x \leq b$ and $b < x \leq l$.

5.3 Transfer functions

5.3.1 Definitions

The **transfer function** of a linear time-invariant system is defined to be the ratio of the Laplace transform of the system output (or response function) to the Laplace transform of the system input (or forcing function), *under the assumption that all the initial conditions are zero* (that is, the system is initially in a **quiescent state**).

Transfer functions are frequently used in engineering to characterize the input-output relationships of linear time-invariant systems, and play an important role in the analysis and design of such systems.

Consider a linear time-invariant system characterized by the differential equation

$$a_n \frac{d^n x}{dt^n} + a_{n-1} \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_0 x = b_m \frac{d^m u}{dt^m} + \cdots + b_0 u \quad (5.27)$$

where $n \geq m$, the as and bs are constant coefficients, and $x(t)$ is the system response or output to the input or forcing term $u(t)$ applied at time $t = 0$. Taking Laplace transforms throughout in (5.27) will lead to the transformed equation. Since all the initial conditions are assumed to be zero, using the relationship between the Laplace transform and derivatives as shown in MEM (11.14) we see that, in order to obtain the transformed equation, we simply replace d/dt by s , giving

$$(a_n s^n + a_{n-1} s^{n-1} + \cdots + a_0) X(s) = (b_m s^m + \cdots + b_0) U(s)$$

where $X(s)$ and $U(s)$ denote the Laplace transforms of $x(t)$ and $u(t)$ respectively.

The system transfer function $G(s)$ is then defined to be

$$G(s) = \frac{X(s)}{U(s)} = \frac{b_m s^m + \cdots + b_0}{a_n s^n + \cdots + a_0} \quad (5.28)$$

with (5.28) being referred to as the **transfer function** model of the system characterized by the differential equation model (5.27). Diagrammatically this may be represented by the **input-output block diagram** of Figure 5.26.

Writing

$$P(s) = b_m s^m + \cdots + b_0$$

$$Q(s) = a_n s^n + \cdots + a_0$$

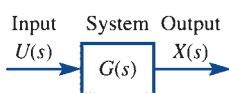


Figure 5.26
Transfer function block diagram.

the transfer function may be expressed as

$$G(s) = \frac{P(s)}{Q(s)}$$

where, in order to make the system physically realizable, the degrees m and n of the polynomials $P(s)$ and $Q(s)$ must be such that $n \geq m$. This is because it follows from (5.23) that if $m > n$ then the system response $x(t)$ to a realistic input $u(t)$ will involve impulses.

The equation $Q(s) = 0$ is called the **characteristic equation** of the system; its order determines the **order of the system**, and its roots are referred to as the **poles** of the transfer function. Likewise, the roots of $P(s) = 0$ are referred to as the **zeros** of the transfer function.

It is important to realize that, in general, a transfer function is only used to characterize a linear time-invariant system. It is a property of the system itself, and is independent of both system input and output.

Although the transfer function characterizes the dynamics of the system, it provides no information concerning the actual physical structure of the system, and in fact systems that are physically different may have identical transfer functions; for example, the mass–spring–damper system of Figure 11.12 in MEM and the *LCR* circuit of Figure 11.8 in MEM both have the transfer function

$$G(s) = \frac{X(s)}{U(s)} = \frac{1}{\alpha s^2 + \beta s + \gamma}$$

In the mass–spring–damper system $X(s)$ determines the displacement $x(t)$ of the mass and $U(s)$ represents the applied force $F(t)$, while α denotes the mass, β the damping coefficient and γ the spring constant. On the other hand, in the *LCR* circuit $X(s)$ determines the charge $q(t)$ on the condenser and $U(s)$ represents the applied emf $e(t)$, while α denotes the inductance, β the resistance and γ the reciprocal of the capacitance.

In practice, an overall system may be made up of a number of components each characterized by its own transfer function and related operation box. The overall system input–output transfer function is then obtained by the rules of **block diagram algebra**.

Since $G(s)$ may be written as

$$G(s) = \frac{b_m}{a_m} \frac{(s - z_1)(s - z_2) \cdots (s - z_m)}{(s - p_1)(s - p_2) \cdots (s - p_n)}$$

where the z_i s and p_i s are the transfer function zeros and poles respectively, we observe that $G(s)$ is known, apart from a constant factor, if the positions of all the poles and zeros are known. Consequently, a plot of the poles and zeros of $G(s)$ is often used as an aid in the graphical analysis of the transfer function (a common convention is to mark the position of a zero by a circle \circ and that of a pole by a cross \times). Since the coefficients of the polynomials $P(s)$ and $Q(s)$ are real, all complex roots always occur in complex conjugate pairs, so that the **pole–zero plot** is symmetrical about the real axis.

Example 5.19

The response $x(t)$ of a system to a forcing function $u(t)$ is determined by the differential equation

$$9\frac{d^2x}{dt^2} + 12\frac{dx}{dt} + 13x = 2\frac{du}{dt} + 3u$$

- (a) Determine the transfer function characterizing the system.
- (b) Write down the characteristic equation of the system. What is the order of the system?
- (c) Determine the transfer function poles and zeros, and illustrate them diagrammatically in the s plane.

Solution (a) Assuming all the initial conditions to be zero, taking Laplace transforms throughout in the differential equation

$$9\frac{d^2x}{dt^2} + 12\frac{dx}{dt} + 13x = 2\frac{du}{dt} + 3u$$

leads to

$$(9s^2 + 12s + 13)X(s) = (2s + 3)U(s)$$

so that the system transfer function is given by

$$G(s) = \frac{X(s)}{U(s)} = \frac{2s + 3}{9s^2 + 12s + 13}$$

- (b) The characteristic equation of the system is

$$9s^2 + 12s + 13 = 0$$

and the system is of order 2.

- (c) The transfer function poles are the roots of the characteristic equation

$$9s^2 + 12s + 13 = 0$$

which are

$$s = \frac{-12 \pm \sqrt{(144 - 468)}}{18} = \frac{-2 \pm j3}{3}$$

That is, the transfer function has simple poles at

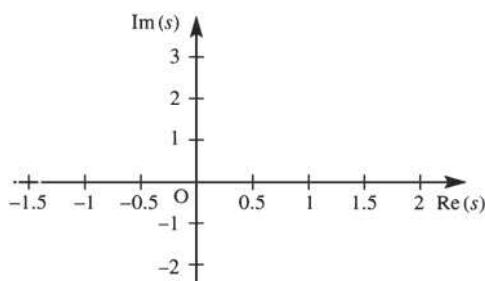
$$s = -\frac{2}{3} + j \quad \text{and} \quad s = -\frac{2}{3} - j$$

The transfer function zeros are determined by equating the numerator polynomial $2s + 3$ to zero, giving a single zero at

$$s = -\frac{3}{2}$$

The corresponding pole-zero plot in the s plane is shown in Figure 5.27.

Figure 5.27
Pole (x)-zero (o) plot
for Example 5.19.





A transfer function (`tf`) is implemented within MATLAB using the commands

```
s = tf('s')
G = G(s)
```

Thus, entering `G=(2*s+3) / (9*s^2+12*s+13)` generates

$$\text{transfer function} = \frac{2s + 3}{9s^2 + 12s + 13}$$

The command `poly(G)` generates the characteristic polynomial, whilst the commands `pole(G)` and `zero(G)` generate the poles and zeros respectively. The command `pzmap(G)` draws the pole-zero map.

5.3.2 Stability

The stability of a system is a property of vital importance to engineers. Intuitively, we may regard a stable system as one that will remain at rest unless it is excited by an external source, and will return to rest if all such external influences are removed. Thus a stable system is one whose response, in the absence of an input, will approach zero as time approaches infinity. This then ensures that any bounded input produces a bounded output; this property is frequently taken to be the definition of a **stable linear system**.

Clearly, stability is a property of the system itself, and does not depend on the system input or forcing function. Since a system may be characterized in the s domain by its transfer function $G(s)$, it should be possible to use the transfer function to specify conditions for the system to be stable.

In considering the time response of

$$X(s) = G(s)U(s), \quad G(s) = \frac{P(s)}{Q(s)}$$

to any given input $u(t)$, it is necessary to factorize the denominator polynomial

$$Q(s) = a_n s^n + a_{n-1} s^{n-1} + \dots + a_0$$

and various forms of factors can be involved.

Simple factor of the form $s + \alpha$, with α real

This corresponds to a simple pole at $s = -\alpha$, and will in the partial-fractions expansion of $G(s)$ lead to a term of the form $c/(s + \alpha)$ having corresponding time response $c e^{-\alpha t} H(t)$, using the strict form of the inverse given in (11.11) of MEM. If $\alpha > 0$, so that the pole is in the left half of the s plane, the time response will tend to zero as $t \rightarrow \infty$. If $\alpha < 0$, so that the pole is in the right half of the s plane, the time response will increase without bound as $t \rightarrow \infty$. It follows that a stable system must have real-valued simple poles of $G(s)$ in the left half of the s plane.

$\alpha = 0$ corresponds to a simple pole at the origin, having a corresponding time response that is a step $cH(t)$. A system having such a pole is said to be **marginally**

stable; this does not ensure that a bounded input will lead to a bounded output, since, for example, if such a system has an input that is a step d applied at time $t = 0$ then the response will be a ramp $cdtH(t)$, which is unbounded as $t \rightarrow \infty$.

Repeated simple factors of the form $(s + \alpha)^n$, with α real

This corresponds to a multiple pole at $s = -\alpha$, and will lead in the partial-fractions expansion of $G(s)$ to a term of the form $c/(s + \alpha)^n$ having corresponding time response $[c/(n - 1)!]t^{n-1} e^{-\alpha t} H(t)$. Again the response will decay to zero as $t \rightarrow \infty$ only if $\alpha > 0$, indicating that a stable system must have all real-valued repeated poles of $G(s)$ in the left half of the s plane.

Quadratic factors of the form $(s + \alpha)^2 + \beta^2$, with α and β real

This corresponds to a pair of complex conjugate poles at $s = -\alpha + j\beta$, $s = -\alpha - j\beta$, and will lead in the partial-fractions expansion of $G(s)$ to a term of the form

$$\frac{c(s + \alpha) + d\beta}{(s + \alpha)^2 + \beta^2}$$

having corresponding time response

$$e^{-\alpha t}(c \cos \beta t + d \sin \beta t) \equiv A e^{-\alpha t} \sin(\beta t + \gamma)$$

where $A = \sqrt{(c^2 + d^2)}$ and $\gamma = \tan^{-1}(c/d)$.

Again we see that poles in the left half of the s plane (corresponding to $\alpha > 0$) have corresponding time responses that die away, in the form of an exponentially damped sinusoid, as $t \rightarrow \infty$. A stable system must therefore have complex conjugate poles located in the left half of the s plane; that is, all complex poles must have a negative real part.

If $\alpha = 0$, the corresponding time response will be a periodic sinusoid, which will not die away as $t \rightarrow \infty$. Again this corresponds to a marginally stable system, and will, for example, give rise to a response that increases without bound as $t \rightarrow \infty$ when the input is a sinusoid at the same frequency β .

A summary of the responses corresponding to the various types of poles is given in Figure 5.28.

The concept of stability may be expressed in the form of Definition 5.3.

Definition 5.3

A physically realizable causal time-invariant linear system with transfer function $G(s)$ is stable provided that all the poles of $G(s)$ are in the left half of the s plane.

The requirement in the definition that the system be physically realizable, that is $n \geq m$ in the transfer function $G(s)$ of (5.28), avoids terms of the form s^{m-n} in the partial-fractions expansion of $G(s)$. Such a term would correspond to differentiation of degree $m - n$, and were an input such as $\sin \omega t$ used to excite the system then the response would include a term such as $\omega^{m-n} \sin \omega t$ or $\omega^{m-n} \cos \omega t$, which could be made as large as desired by increasing the input frequency ω .

Figure 5.28
Relationship between transfer function poles and time response.

Poles of $G(s)$ in form $\sigma + j\omega$	Poles in complex s plane	Corresponding time response	Nature of response
$\sigma = \omega = 0$			Constant
$\sigma = \omega = 0$ (multiplicity 2)			Ramp
$\sigma < 0, \omega = 0$			Exponential decay
$\sigma > 0, \omega = 0$			Exponential growth
$\sigma = 0, \omega > 0$			Sinusoidal
$\sigma = 0, \omega > 0$ (multiplicity 2)			Linearly growing sinusoidal
$\sigma < 0, \omega > 0$			Exponentially decaying sinusoidal
$\sigma > 0, \omega > 0$			Exponentially growing sinusoidal

In terms of the poles of the transfer function $G(s)$, its abscissa of convergence σ_c corresponds to the real part of the pole located furthest to the right in the s plane. For example, if

$$G(s) = \frac{s+1}{(s+3)(s+2)}$$

then the abscissa of convergence $\sigma_c = -2$.

It follows from Definition 5.3 that the transfer function $G(s)$ of a stable system has an abscissa of convergence $\sigma_c = -\alpha$, with $\alpha > 0$. Thus its region of convergence includes the imaginary axis, so that $G(s)$ exists when $s = j\omega$. We shall return to this result when considering the relationship between Laplace and Fourier transforms in Section 8.4.1.

According to Definition 5.3, in order to prove stability, we need to show that all the roots of the characteristic equation

$$Q(s) = a_n s^n + a_{n-1} s^{n-1} + \cdots + a_1 s + a_0 = 0 \quad (5.29)$$

have negative real parts (that is, they lie in the left half of the s plane). Various criteria exist to show that all the roots satisfy this requirement, and it is not necessary to solve the equation to prove stability. One widely used criterion is the **Routh–Hurwitz criterion**, which can be stated as follows:

A necessary and sufficient condition for all the roots of equation (5.29) to have negative real parts is that the determinants $\Delta_1, \Delta_2, \dots, \Delta_n$ are all positive, where

$$\Delta_r = \begin{vmatrix} a_{n-1} & a_n & 0 & 0 & \dots & 0 \\ a_{n-3} & a_{n-2} & a_{n-1} & a_n & \dots & 0 \\ a_{n-5} & a_{n-4} & a_{n-3} & a_{n-2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{n-(2r-1)} & a_{n-2r} & a_{n-2r-1} & a_{n-2r-2} & \dots & a_{n-r} \end{vmatrix} \quad (5.30)$$

it being understood that in each determinant all the a s with subscripts that are either negative or greater than n are to be replaced by zero.

Example 5.20

Show that the roots of the characteristic equation

$$s^4 + 9s^3 + 33s^2 + 51s + 26 = 0$$

all have negative real parts.

Solution

In this case $n = 4$, $a_0 = 26$, $a_1 = 51$, $a_2 = 33$, $a_3 = 9$, $a_4 = 1$ and $a_r = 0$ ($r > 4$). The determinants of the Routh–Hurwitz criterion are

$$\Delta_1 = |a_{n-1}| = |a_3| = |9| = 9 > 0$$

$$\Delta_2 = \begin{vmatrix} a_{n-1} & a_n \\ a_{n-3} & a_{n-2} \end{vmatrix} = \begin{vmatrix} a_3 & a_4 \\ a_1 & a_2 \end{vmatrix}$$

$$= \begin{vmatrix} 9 & 1 \\ 51 & 33 \end{vmatrix} = 246 > 0$$

$$\Delta_3 = \begin{vmatrix} a_{n-1} & a_n & 0 \\ a_{n-3} & a_{n-2} & a_{n-1} \\ a_{n-5} & a_{n-4} & a_{n-3} \end{vmatrix} = \begin{vmatrix} a_3 & a_4 & 0 \\ a_1 & a_2 & a_3 \\ a_{-1} & a_0 & a_1 \end{vmatrix}$$

$$\begin{aligned}\Delta_4 &= \begin{vmatrix} a_{n-1} & a_n & 0 & 0 \\ a_{n-3} & a_{n-2} & a_{n-1} & a_n \\ a_{n-5} & a_{n-4} & a_{n-3} & a_{n-2} \\ a_{n-7} & a_{n-6} & a_{n-5} & a_{n-4} \end{vmatrix} = \begin{vmatrix} a_3 & a_4 & 0 & 0 \\ a_1 & a_2 & a_3 & a_4 \\ a_{-1} & a_0 & a_1 & a_2 \\ a_{-3} & a_{-2} & a_{-1} & a_0 \end{vmatrix} \\ &= \begin{vmatrix} 9 & 1 & 0 & 0 \\ 51 & 33 & 9 & 1 \\ 0 & 26 & 51 & 37 \\ 0 & 0 & 0 & 26 \end{vmatrix} = 26\Delta_3 > 0\end{aligned}$$

Thus $\Delta_1 > 0$, $\Delta_2 > 0$, $\Delta_3 > 0$ and $\Delta_4 > 0$, so that all the roots of the given characteristic equation have negative real parts. This is readily checked, since the roots are -2 , -1 , $-3 + j2$ and $-3 - j2$.

Example 5.21

The steady motion of a steam-engine governor is modelled by the differential equations

$$m\ddot{\eta} + b\dot{\eta} + d\eta - e\omega = 0 \quad (5.31)$$

$$I_0\dot{\omega} = -f\eta \quad (5.32)$$

where η is a small fluctuation in the angle of inclination, ω a small fluctuation in the angular velocity of rotation, and m , b , d , e , f and I_0 are all positive constants. Show that the motion of the governor is stable provided that

$$\frac{bd}{m} > \frac{ef}{I_0}$$

Solution Differentiating (5.31) gives

$$m\ddot{\eta} + b\ddot{\eta} + d\dot{\eta} - e\dot{\omega} = 0$$

which, on using (5.32), leads to

$$m\ddot{\eta} + b\ddot{\eta} + d\dot{\eta} + \frac{ef}{I_0}\eta = 0$$

for which the corresponding characteristic equation is

$$ms^3 + bs^2 + ds + \frac{ef}{I_0} = 0$$

This is a cubic polynomial, so the parameters of (5.29) are

$$n = 3, \quad a_0 = \frac{ef}{I_0}, \quad a_1 = d, \quad a_2 = b, \quad a_3 = m \quad (a_r = 0, r > 3)$$

The determinants (5.30) of the Routh–Hurwitz criterion are

$$\Delta_1 = |a_2| = b > 0$$

$$\Delta_2 = \begin{vmatrix} a_2 & a_3 \\ a_0 & a_1 \end{vmatrix} = \begin{vmatrix} b & m \\ ef/I_0 & d \end{vmatrix} = bd - \frac{mef}{I_0}$$

(and so $\Delta_2 > 0$ provided that $bd - mef/I_0 > 0$ or $bd/m > ef/I_0$), and

$$\Delta_3 = \begin{vmatrix} a_2 & a_3 & 0 \\ a_0 & a_1 & a_2 \\ 0 & 0 & a_0 \end{vmatrix} = a_0\Delta_2 > 0 \quad \text{if } \Delta_2 > 0$$

Thus the action of the governor is stable provided that $\Delta_2 > 0$; that is,

$$\frac{bd}{m} > \frac{ef}{I_0}$$

5.3.3 Impulse response

From (5.28), we find that for a system having transfer function $G(s)$ the response $x(t)$ of the system, initially in a quiescent state, to an input $u(t)$ is determined by the transformed relationship

$$X(s) = G(s)U(s)$$

If the input $u(t)$ is taken to be the unit impulse function $\delta(t)$ then the system response will be determined by

$$X(s) = G(s)\mathcal{L}\{\delta(t)\} = G(s)$$

Taking inverse Laplace transforms leads to the corresponding time response $h(t)$, which is called the **impulse response** of the system (it is also sometimes referred to as the **weighting function** of the system); that is, the impulse response is given by

$$h(t) = \mathcal{L}^{-1}\{X(s)\} = \mathcal{L}^{-1}\{G(s)\} \tag{5.33}$$

We therefore have the following definition.

Definition 5.4: Impulse response

The impulse response $h(t)$ of a linear time-invariant system is the response of the system to a unit impulse applied at time $t = 0$ when all the initial conditions are zero. It is such that $\mathcal{L}\{h(t)\} = G(s)$, where $G(s)$ is the system transfer function.

Since the impulse response is the inverse Laplace transform of the transfer function, it follows that both the impulse response and the transfer function carry the same information about the dynamics of a linear time-invariant system. Theoretically, therefore, it is possible to determine the complete information about the system by exciting it with an impulse and measuring the response. For this reason, it is common practice in engineering to regard the transfer function as being the Laplace transform of the impulse response, since this places greater emphasis on the parameters of the system when considering system design.

We saw in Section 5.3.2 that, since the transfer function $G(s)$ completely characterizes a linear time-invariant system, it can be used to specify conditions for system stability, which are that all the poles of $G(s)$ lie in the left half of the s plane. Alternatively, characterizing the system by its impulse response, we can say that the system is stable provided that its impulse response decays to zero as $t \rightarrow \infty$.

Example 5.22

Determine the impulse response of the linear system whose response $x(t)$ to an input $u(t)$ is determined by the differential equation

$$\frac{d^2x}{dt^2} + 5\frac{dx}{dt} + 6x = 5u(t) \quad (5.34)$$

Solution

The impulse response $h(t)$ is the system response to $u(t) = \delta(t)$ when all the initial conditions are zero. It is therefore determined as the solution of the differential equation

$$\frac{d^2h}{dt^2} + 5\frac{dh}{dt} + 6h = 5\delta(t) \quad (5.35)$$

subject to the initial conditions $h(0) = \dot{h}(0) = 0$. Taking Laplace transforms in (5.35) gives

$$(s^2 + 5s + 6)H(s) = 5\mathcal{L}\{\delta(t)\} = 5$$

so that

$$H(s) = \frac{5}{(s+3)(s+2)} = \frac{5}{s+2} - \frac{5}{s+3}$$

which, on inversion, gives the desired impulse response

$$h(t) = 5(e^{-2t} - e^{-3t})$$

Alternatively, the transfer function $G(s)$ of the system determined by (5.34) is

$$G(s) = \frac{5}{s^2 + 5s + 6}$$

so that $h(t) = \mathcal{L}^{-1}\{G(s)\} = 5(e^{-2t} - e^{-3t})$ as before.

Note: This example serves to illustrate the necessity for incorporating 0^- as the lower limit in the Laplace transform integral, in order to accommodate for an impulse applied at $t = 0$. The effect of the impulse is to cause a step change in $\dot{x}(t)$ at $t = 0$, with the initial condition accounting for what happens up to 0^- .



In MATLAB a plot of the impulse response is obtained using the commands

```
s=tf('s')
G=G(s)
impulse(G)
```

5.3.4 Initial- and final-value theorems

The initial- and final-value theorems are two useful theorems that enable us to predict system behaviour as $t \rightarrow 0$ and $t \rightarrow \infty$ without actually inverting Laplace transforms.

Theorem 5.4

The initial-value theorem

If $f(t)$ and $f'(t)$ are both Laplace-transformable and if $\lim_{s \rightarrow \infty} sF(s)$ exists then

$$\lim_{t \rightarrow 0^+} f(t) = f(0^+) = \lim_{s \rightarrow \infty} sF(s)$$

Proof From MEM (11.12) or simply by direct integration,

$$\mathcal{L}\{f'(t)\} = \int_{0^-}^{\infty} f'(t) e^{-st} dt = sF(s) - f(0^-)$$

where we have highlighted the fact that the lower limit is 0^- . Hence

$$\begin{aligned} \lim_{s \rightarrow \infty} [sF(s) - f(0^-)] &= \lim_{s \rightarrow \infty} \int_{0^-}^{\infty} f'(t) e^{-st} dt \\ &= \lim_{s \rightarrow \infty} \int_{0^-}^{0^+} f'(t) e^{-st} dt + \lim_{s \rightarrow \infty} \int_{0^+}^{\infty} f'(t) e^{-st} dt \end{aligned} \quad (5.36)$$

If $f(t)$ is discontinuous at the origin, so that $f(0^+) \neq f(0^-)$, then, from (5.21), $f'(t)$ contains an impulse term $[f(0^+) - f(0^-)]\delta(t)$, so that

$$\lim_{s \rightarrow \infty} \int_{0^-}^{0^+} f'(t) e^{-st} dt = f(0^+) - f(0^-)$$

Also, since the Laplace transform of $f'(t)$ exists, it is of exponential order and we have

$$\lim_{s \rightarrow \infty} \int_{0^+}^{\infty} f'(t) e^{-st} dt = 0$$

so that (5.36) becomes

$$\lim_{s \rightarrow \infty} sF(s) - f(0^-) = f(0^+) - f(0^-)$$

giving the required result:

$$\lim_{s \rightarrow \infty} sF(s) = f(0^+)$$

If $f(t)$ is continuous at the origin then $f'(t)$ does not contain an impulse term, and the right-hand side of (5.36) is zero, giving

$$\lim_{s \rightarrow \infty} sF(s) = f(0^-) = f(0^+)$$

[end of theorem](#)

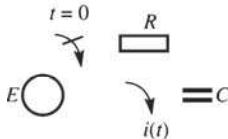
It is important to recognize that the initial-value theorem does not give the initial value $f(0^-)$ used when determining the Laplace transform, but rather gives the value of $f(t)$ as $t \rightarrow 0^+$. This distinction is highlighted in the following example.

Example 5.23

The circuit of Figure 5.29 consists of a resistance R and a capacitance C connected in series together with constant voltage source E . Prior to closing the switch at time $t = 0$, both the charge on the capacitor and the resulting current in the circuit are zero. Determine the current $i(t)$ in the circuit at time t after the switch is closed, and investigate the use of the initial-value theorem.

Solution Applying Kirchhoff's law to the circuit of Figure 5.29, we have

$$Ri + \frac{1}{C} \int i dt = E_0$$



which, on taking Laplace transforms, gives the transformed equation

$$RI(s) + \frac{1}{C} \frac{I(s)}{s} = \frac{E_0}{s}$$

Therefore

Figure 5.29
RC circuit of
Example 5.23.

$$I(s) = \frac{E_0/R}{s + 1/RC}$$

Taking inverse transforms gives the current $i(t)$ at $t \geq 0$ as

$$i(t) = \frac{E_0}{R} e^{-t/RC} \quad (5.37)$$

Applying the initial-value theorem,

$$\lim_{t \rightarrow 0^+} i(t) = \lim_{s \rightarrow \infty} sI(s) = \lim_{s \rightarrow \infty} \frac{sE_0/R}{s + 1/RC} = \lim_{s \rightarrow \infty} \frac{E_0/R}{1 + 1/RCs} = \frac{E_0}{R}$$

That is,

$$i(0^+) = \frac{E_0}{R}$$

a result that is readily confirmed by allowing $t \rightarrow 0^+$ in (5.37). We note that this is not the same as the initial state $i(0) = 0$ owing to the fact that there is a step change in $i(t)$ at $t = 0$.

Theorem 5.5

The final-value theorem

If $f(t)$ and $f'(t)$ are both Laplace-transformable and $\lim_{t \rightarrow \infty} f(t)$ exists then

$$\lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0} sF(s)$$

Proof From (11.12) of MEM, the Laplace transform of a derivative,

$$\mathcal{L}\{f'(t)\} = \int_{0^-}^{\infty^+} f'(t) e^{-st} dt = sF(s) - f(0^-)$$

Taking limits, we have

$$\begin{aligned} \lim_{s \rightarrow 0} [sF(s) - f(0^-)] &= \lim_{s \rightarrow 0} \int_{0^-}^{\infty} f'(t) e^{-st} dt = \int_{0^-}^{\infty} f'(t) dt = [f(t)]_{0^-}^{\infty} \\ &= \lim_{t \rightarrow \infty} f(t) - f(0^-) \end{aligned}$$

giving the required result:

$$\lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0} sF(s)$$

[end of theorem](#)

The restriction that $\lim_{t \rightarrow \infty} f(t)$ must exist means that the theorem does not hold for functions such as e^t , which tends to infinity as $t \rightarrow \infty$, or $\sin \omega t$, whose limit is undefined. Since in practice the final-value theorem is used to obtain the behaviour of $f(t)$ as $t \rightarrow \infty$ from knowledge of the transform $F(s)$, it is more common to express the restriction in terms of restrictions on $F(s)$, which are that $sF(s)$ must have all its poles in the left half of the s plane; that is, $sF(s)$ must represent a stable transfer function. It is important that the theorem be used with caution and that this restriction be fully recognized, since the existence of $\lim_{t \rightarrow \infty} sF(s)$ does *not* imply that $f(t)$ has a limiting value as $t \rightarrow \infty$.

Example 5.24

Investigate the application of the final-value theorem to the transfer function

$$F(s) = \frac{1}{(s+2)(s-3)} \quad (5.38)$$

Solution

$$\lim_{s \rightarrow 0} sF(s) = \lim_{s \rightarrow 0} \frac{s}{(s+2)(s-3)} = 0$$

so the use of the final-value theorem implies that for the time function $f(t)$ corresponding to $F(s)$ we have

$$\lim_{t \rightarrow \infty} f(t) = 0$$

However, taking inverse transforms in (5.38) gives

$$f(t) = \frac{1}{s}(e^{3t} - e^{-2t})$$

implying that $f(t)$ tends to infinity as $t \rightarrow \infty$. This implied contradiction arises since the theorem is not valid in this case. Although $\lim_{s \rightarrow 0} sF(s)$ exists, $sF(s)$ has a pole at $s = 3$, which is not in the left half of the s plane.

The final-value theorem provides a useful vehicle for determining a system's **steady-state gain (SSG)** and the **steady-state errors, or offsets**, in feedback control systems, both of which are important features in control system design.

The SSG of a stable system is the system's steady-state response, that is the response as $t \rightarrow \infty$, to a unit step input. For a system with transfer function $G(s)$ we have, from (5.28), that its response $x(t)$ is related to the input $u(t)$ by the transformed equation

$$X(s) = G(s)U(s)$$

For a unit step input

$$u(t) = 1H(t) \quad \text{giving} \quad U(s) = \frac{1}{s}$$

so that

$$X(s) = \frac{G(s)}{s}$$

From the final-value theorem, the steady-state gain is

$$\text{SSG} = \lim_{t \rightarrow \infty} x(t) = \lim_{s \rightarrow 0} sX(s) = \lim_{s \rightarrow 0} G(s)$$

Example 5.25

Determine the steady-state gain of a system having transfer function

$$G(s) = \frac{20(1 + 3s)}{s^2 + 7s + 10}$$

Solution The response $x(t)$ to a unit step input $u(t) = 1H(t)$ is given by the transformed equation

$$X(s) = G(s)U(s) = \frac{20(1 + 3s)}{s^2 + 7s + 10} \frac{1}{s}$$

Then, by the final-value theorem, the steady-state gain is given by

$$\text{SSG} = \lim_{t \rightarrow \infty} x(t) = \lim_{s \rightarrow 0} sX(s) = \lim_{s \rightarrow 0} \frac{20(1 + 3s)}{s^2 + 7s + 10} = 2$$

Note that for a step input of magnitude K , that is $u(t) = KH(t)$, the steady-state response will be $\lim_{s \rightarrow 0} kG(s) = 2K$; that is,

$$\text{steady-state response to step input} = \text{SSG} \times \text{magnitude of step input}$$

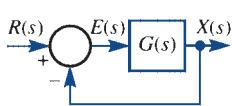


Figure 5.30 Unity feedback control system.

A unity feedback control system having forward-path transfer function $G(s)$, reference input or desired output $r(t)$ and actual output $x(t)$ is illustrated by the block diagram of Figure 5.30. Defining the error to be $e(t) = r(t) - x(t)$, it follows that

$$G(s)E(s) = X(s) = R(s) - E(s)$$

giving

$$E(s) = \frac{R(s)}{1 + G(s)}$$

Thus, from the final-value theorem, the steady-state error (SSE) is

$$\text{SSE} = \lim_{t \rightarrow \infty} e(t) = \lim_{s \rightarrow 0} sE(s) = \lim_{s \rightarrow 0} \frac{sR(s)}{1 + G(s)} \quad (5.39)$$

Example 5.26

Determine the SSE for the system of Figure 5.30 when $G(s)$ is the same as in Example 5.19 and $r(t)$ is a step of magnitude K .

Solution

Since $r(t) = KH(t)$, we have $R(s) = K/s$, so, using (5.39),

$$\text{SSE} = \lim_{s \rightarrow 0} \frac{sK/s}{1 + G(s)} = \frac{K}{1 + \text{SSG}}$$

where SSG = 2 as determined in Example 5.25. Thus

$$\text{SSE} = \frac{1}{3}K$$

It is clear from Example 5.26 that if we are to reduce the SSE, which is clearly desirable in practice, then the SSG needs to be increased. However, such an increase could lead to an undesirable transient response, and in system design a balance must be achieved. Detailed design techniques for alleviating such problems are not considered here; for such a discussion the reader is referred to specialist texts (see, for example J. Schwarzenbach and K. F. Gill, *System Modelling and Control*, third edition, Oxford, Butterworth-Heinemann, 1992).

5.3.5 Exercises

- 22 The response $x(t)$ of a system to a forcing function $u(t)$ is determined by the differential equation model

$$\frac{d^2x}{dt^2} + 2\frac{dx}{dt} + 5x = 3\frac{du}{dt} + 2u$$

- (a) Determine the transfer function characterizing the system.
- (b) Write down the characteristic equation of the system. What is the order of the system?
- (c) Determine the transfer function poles and zeros, and illustrate them diagrammatically in the s plane.

- 23 Repeat Exercise 22 for a system whose response $x(t)$ to an input $u(t)$ is determined by the differential equation

$$\frac{d^3x}{dt^3} + 5\frac{d^2x}{dt^2} + 17\frac{dx}{dt} + 13x = \frac{d^2u}{dt^2} + 5\frac{du}{dt} + 6$$

- 24 Which of the following transfer functions represent stable systems and which represent unstable systems?

- (a) $\frac{s - 1}{(s + 2)(s^2 + 4)}$
- (b) $\frac{(s + 2)(s - 2)}{(s + 1)(s - 1)(s + 4)}$
- (c) $\frac{s - 1}{(s + 2)(s + 4)}$
- (d) $\frac{6}{(s^2 + s + 1)(s + 1)^2}$
- (e) $\frac{5(s + 10)}{(s + 5)(s^2 - s + 10)}$

- 25 Which of the following characteristic equations are representative of stable systems?

- (a) $s^2 - 4s + 13 = 0$
- (b) $5s^3 + 13s^2 + 31s + 15 = 0$

(c) $s^3 + s^2 + s + 1 = 0$

(d) $24s^4 + 11s^3 + 26s^2 + 45s + 36 = 0$

(e) $s^3 + 2s^2 + 2s + 1 = 0$

- 26 The differential equation governing the motion of a mass–spring–damper system with controller is

$$m\frac{d^3x}{dt^3} + c\frac{d^2x}{dt^2} + K\frac{dx}{dt} + Kr x = 0$$

where m, c, K and r are positive constants. Show that the motion of the system is stable provided that $r < c/m$.

- 27 The behaviour of a system having a gain controller is characterized by the characteristic equation

$$s^4 + 2s^3 + (K + 2)s^2 + 7s + K = 0$$

where K is the controller gain. Show that the system is stable provided that $K > 2.1$.

- 28 A feedback control system has characteristic equation

$$s^3 + 15Ks^2 + (2K - 1)s + 5K = 0$$

where K is a constant gain factor. Determine the range of positive values of K for which the system will be stable.

- 29 Determine the impulse responses of the linear systems whose response $x(t)$ to an input $u(t)$ is determined by the following differential equations:

(a) $\frac{d^2x}{dt^2} + 15\frac{dx}{dt} + 56x = 3u(t)$

(b) $\frac{d^2x}{dt^2} + 8\frac{dx}{dt} + 25x = u(t)$

(c) $\frac{d^2x}{dt^2} - 2\frac{dx}{dt} - 8x = 4u(t)$

(d) $\frac{d^2x}{dt^2} - 4\frac{dx}{dt} + 13x = u(t)$

What can be said about the stability of each of the systems?

- 30 The response of a given system to a unit step $u(t) = 1H(t)$ is given by

$$x(t) = 1 - \frac{7}{3}e^{-t} + \frac{3}{2}e^{-2t} - \frac{1}{6}e^{-4t}$$

What is the transfer function of the system?

- 31 Verify the initial-value theorem for the functions
 (a) $2 - 3 \cos t$ (b) $(3t - 1)^2$ (c) $t + 3 \sin 2t$

- 32 Verify the final-value theorem for the functions
 (a) $1 + 3e^{-t} \sin 2t$ (b) $t^2 e^{-2t}$
 (c) $3 - 2e^{-3t} + e^{-t} \cos 2t$

- 33 Use the initial- and final-value theorems to find the jump at $t = 0$ and the limiting value as $t \rightarrow \infty$ for the solution of the initial-value problem

$$7\frac{dy}{dt} + 5y = 4 + e^{-3t} + 2\delta(t)$$

with $y(0^-) = -1$.

5.3.6 Convolution

Convolution is a useful concept that has many applications in various fields of engineering. In Section 5.3.7 we shall use it to obtain the response of a linear system to any input in terms of the impulse response.

Definition 5.5: Convolution

Given two piecewise-continuous functions $f(t)$ and $g(t)$, the **convolution** of $f(t)$ and $g(t)$, denoted by $f * g(t)$, is defined as

$$f * g(t) = \int_{-\infty}^{\infty} f(\tau)g(t - \tau) d\tau$$

In the particular case when $f(t)$ and $g(t)$ are causal functions

$$f(\tau) = g(\tau) = 0 \quad (\tau < 0), \quad g(t - \tau) = 0 \quad (\tau > t)$$

and we have

$$f * g(t) = \int_0^t f(\tau)g(t - \tau) d\tau \tag{5.40}$$

The notation $f * g(t)$ indicates that the convolution $f * g$ is a function of t ; that is, it could also be written as $(f * g)(t)$. The integral $\int_{-\infty}^{\infty} f(\tau)g(t - \tau) d\tau$ is called the **convolution integral**. Alternative names are the **superposition integral**, **Duhamel integral**, **folding integral** and **faltung integral**.

Convolution can be considered as a generalized function, and as such it has many of the properties of multiplication. In particular, the commutative law is satisfied, so that

$$f * g(t) = g * f(t)$$

or, for causal functions,

$$\int_0^t f(\tau)g(t - \tau) d\tau = \int_0^t f(t - \tau)g(\tau) d\tau \quad (5.41)$$

This means that the convolution can be evaluated by time-shifting either of the two functions. The result (5.41) is readily proved, since by making the substitution $\tau_1 = t - \tau$ in (5.40) we obtain

$$f * g(t) = \int_t^0 f(t - \tau_1)g(\tau_1)(-d\tau_1) = \int_0^t f(t - \tau_1)g(\tau_1) d\tau_1 = g * f(t)$$

Example 5.27

For the two causal functions

$$f(t) = tH(t), \quad g(t) = \sin 2t H(t)$$

show that $f * g(t) = g * f(t)$.

Solution

$$f * g(t) = \int_0^t f(\tau)g(t - \tau) d\tau = \int_0^t \tau \sin 2(t - \tau) d\tau$$

Integrating by parts gives

$$\begin{aligned} f * g(t) &= [\frac{1}{2}\tau \cos 2(t - \tau) + \frac{1}{4} \sin 2(t - \tau)]_0^t = \frac{1}{2}t - \frac{1}{4} \sin 2t \\ g * f(t) &= \int_0^t f(t - \tau)g(\tau) d\tau = \int_0^t (t - \tau) \sin 2\tau d\tau \\ &= [-\frac{1}{2}(t - \tau) \cos 2\tau - \frac{1}{4} \sin 2\tau]_0^t = \frac{1}{2}t - \frac{1}{4} \sin 2t \end{aligned}$$

so that $f * g(t) = g * f(t)$.

The importance of convolution in Laplace transform work is that it enables us to obtain the inverse transform of the product of two transforms. The necessary result for doing this is contained in the following theorem.

Theorem 5.6

Convolution theorem for Laplace transforms

If $f(t)$ and $g(t)$ are of exponential order σ , piecewise-continuous on $t \geq 0$ and have Laplace transforms $F(s)$ and $G(s)$ respectively, then, for $s > \sigma$,

$$\mathcal{L}\left\{\int_0^t f(t)g(t - \tau) dt\right\} = \mathcal{L}\{f * g(t)\} = F(s)G(s)$$

or, in the more useful inverse form,

$$\mathcal{L}^{-1}\{F(s)G(s)\} = f * g(t) \quad (5.42)$$

Proof By definition,

$$F(s)G(s) = \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\} = \left[\int_0^\infty e^{-sx} f(x) dx \right] \left[\int_0^\infty e^{-sy} g(y) dy \right]$$

where we have used the ‘dummy’ variables x and y , rather than t , in the integrals to avoid confusion. This may now be expressed in the form of the double integral

$$F(s)G(s) = \int_0^\infty \int_0^\infty e^{-s(x+y)} f(x)g(y) dx dy = \iint_R e^{-s(x+y)} f(x)g(y) dx dy$$

where R is the first quadrant in the (x, y) plane, as shown in Figure 5.31(a). On making the substitution

$$x + y = t, \quad y = \tau$$

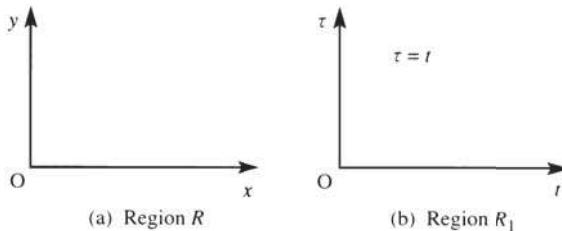
the double integral is transformed into

$$F(s)G(s) = \iint_{R_1} e^{-st} f(t - \tau) g(\tau) dt d\tau$$

where R_1 is the semi-infinite region in the (τ, t) plane bounded by the lines $\tau = 0$ and $\tau = t$, as shown in Figure 5.31(b). This may be written as

$$F(s)G(s) = \int_0^\infty e^{-st} \left(\int_0^t f(t - \tau) g(\tau) d\tau \right) dt = \int_0^\infty e^{-st} [g * f(t)] dt = \mathcal{L}\{g * f(t)\}$$

Figure 5.31
Regions of integration.



and, since convolution is commutative, we may write this as

$$F(s)G(s) = \mathcal{L}\{f * g(t)\}$$

which concludes the proof.

end of theorem

Example 5.28

Using the convolution theorem, determine $\mathcal{L}^{-1}\left\{\frac{1}{s^2(s+2)^2}\right\}$.

Solution We express $1/s^2(s+2)^2$ as $(1/s^2)[1/(s+2)^2]$; then, since

$$\mathcal{L}\{t\} = \frac{1}{s^2}, \quad \mathcal{L}\{te^{-2t}\} = \frac{1}{(s+2)^2}$$

taking $f(t) = t$ and $g(t) = t e^{-2t}$ in the convolution theorem gives

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2} \frac{1}{(s+2)^2}\right\} = \int_0^t f(t-\tau)g(\tau) d\tau = \int_0^t (t-\tau)\tau e^{-2\tau} d\tau$$

which on integration by parts gives

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2} \frac{1}{(s+2)^2}\right\} = \left[-\frac{1}{2}e^{-2\tau}[(t-\tau)\tau + \frac{1}{2}(t-2\tau) - \frac{1}{2}]\right]_0^t = \frac{1}{4}[t-1+(t+1)e^{-2t}]$$

We can check this result by first expressing the given transform in partial-fractions form and then inverting to give

$$\frac{1}{s^2(s+2)^2} = \frac{-\frac{1}{4}}{s} + \frac{\frac{1}{4}}{s^2} + \frac{\frac{1}{4}}{s+2} + \frac{\frac{1}{4}}{(s+2)^2}$$

so that

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2(s+2)^2}\right\} = -\frac{1}{4} + \frac{1}{4}t + \frac{1}{4}e^{-2t} + \frac{1}{4}te^{-2t} = \frac{1}{4}[t-1+(t+1)e^{-2t}]$$

as before.

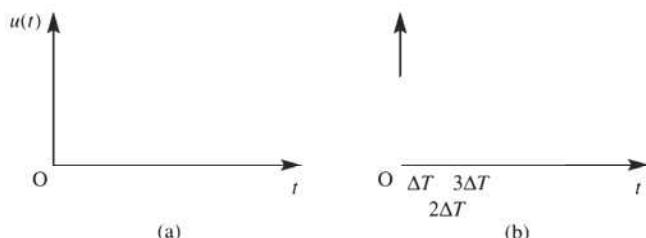
5.3.7 System response to an arbitrary input

The impulse response of a linear time-invariant system is particularly useful in practice in that it enables us to obtain the response of the system to an arbitrary input using the convolution integral. This provides engineers with a powerful approach to the analysis of dynamical systems.

Let us consider a linear system characterized by its impulse response $h(t)$. Then we wish to determine the response $x(t)$ of the system to an arbitrary input $u(t)$ such as that illustrated in Figure 5.32(a). We first approximate the continuous function $u(t)$ by an infinite sequence of impulses of magnitude $u(n\Delta T)$, $n = 0, 1, 2, \dots$, as shown in Figure 5.32(b). This approximation for $u(t)$ may be written as

$$u(t) \approx \sum_{n=0}^{\infty} u(n\Delta T) \delta(t - n\Delta T) \Delta T \quad (5.43)$$

Figure 5.32
Approximation to a continuous input.



Since the system is linear, the **principle of superposition** holds, so that the response of the system to the sum of the impulses is equal to the sum of the responses of the system to each of the impulses acting separately. Depicting the impulse response $h(t)$ of the linear system by Figure 5.33, the responses due to the individual impulses forming the sum in (5.43) are illustrated in the sequence of plots in Figure 5.34.

Figure 5.33
Impulse response
of a linear system.

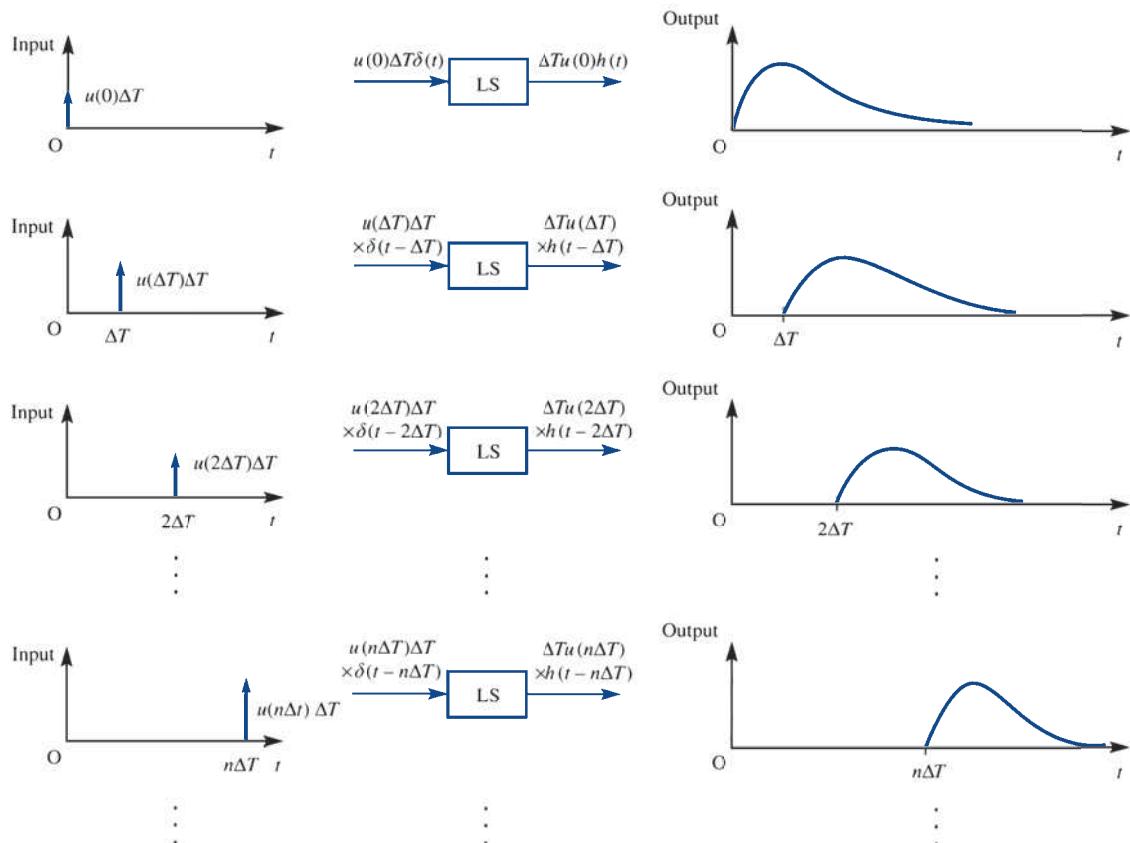
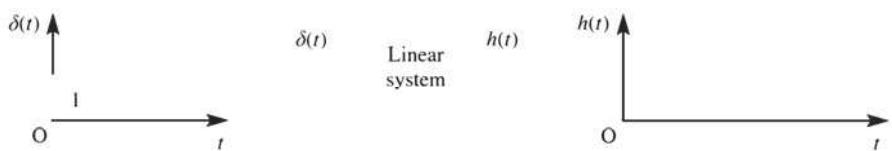


Figure 5.34 Responses due to individual impulses.

Summing the individual responses, we find that the response due to the sum of the impulses is

$$\sum_{n=0}^{\infty} u(n\Delta T)h(t - n\Delta T) \Delta T \quad (5.44)$$

Allowing $\Delta T \rightarrow 0$, so that $n\Delta T$ approaches a continuous variable τ , the above sum will approach an integral that will be representative of the system response $x(t)$ to the continuous input $u(t)$. Thus

$$x(t) = \int_0^{\infty} u(\tau)h(t - \tau) d\tau = \int_0^t u(\tau)h(t - \tau) d\tau \quad (\text{since } h(t) \text{ is a causal function})$$

That is,

$$x(t) = u * h(t)$$

Since convolution is commutative, we may also write

$$x(t) = h * u(t) = \int_0^t h(\tau)u(t - \tau) d\tau$$

In summary, we have the result that if the impulse response of a linear time-invariant system is $h(t)$ then its response to an arbitrary input $u(t)$ is

$$x(t) = \int_0^t u(\tau)h(t - \tau) d\tau = \int_0^t h(\tau)u(t - \tau) d\tau \quad (5.45)$$

It is important to realize that this is the response of the system to the input $u(t)$ assuming it to be initially in a quiescent state.

Example 5.29

The response $\theta_o(t)$ of a system to a driving force $\theta_i(t)$ is given by the linear differential equation

$$\frac{d^2\theta_o}{dt^2} + \frac{2d\theta_o}{dt} + 5\theta_o = \theta_i$$

Determine the impulse response of the system. Hence, using the convolution integral, determine the response of the system to a unit step input at time $t = 0$, assuming that it is initially in a quiescent state. Confirm this latter result by direct calculation.

Solution

The impulse response $h(t)$ is the solution of

$$\frac{d^2h}{dt^2} + 2\frac{dh}{dt} + 5h = \delta(t)$$

subject to the initial conditions $h(0) = \dot{h}(0) = 0$. Taking Laplace transforms gives

$$(s^2 + 2s + 5)H(s) = \mathcal{L}\{\delta(t)\} = 1$$

so that

$$H(s) = \frac{1}{s^2 + 2s + 5} = \frac{1}{2} \frac{2}{(s + 1)^2 + 2^2}$$

which, on inversion, gives the impulse response as

$$h(t) = \frac{1}{2} e^{-t} \sin 2t$$

Using the convolution integral

$$\theta_o(t) = \int_0^t h(\tau)\theta_i(t - \tau) d\tau$$

with $\theta_i(t) = 1H(t)$ gives the response to the unit step as

$$\theta_o(t) = \frac{1}{2} \int_0^t e^{-\tau} \sin 2\tau d\tau$$

Integrating by parts twice gives

$$\begin{aligned}\theta_o(t) &= -\frac{1}{2} e^{-t} \sin 2t - e^{-t} \cos 2t + 1 - 2 \int_0^t e^{-\tau} \sin 2\tau d\tau \\ &= -\frac{1}{2} e^{-t} \sin 2t - e^{-t} \cos 2t + 1 - 4\theta_o(t)\end{aligned}$$

Hence

$$\theta_o(t) = \frac{1}{5} (1 - e^{-t} \cos 2t - \frac{1}{2} e^{-t} \sin 2t)$$

(Note that in this case, because of the simple form of $\theta_i(t)$, the convolution integral $\int_0^t h(\tau) \theta_i(t-\tau) d\tau$ is taken in preference to $\int_0^t \theta_i(\tau) h(t-\tau) d\tau$.)

To obtain the step response directly, we need to solve for $t \geq 0$ the differential equation

$$\frac{d^2\theta_o}{dt^2} + 2 \frac{d\theta_o}{dt} + 5\theta_o = 1$$

subject to the initial conditions $\theta_o(0) = \dot{\theta}_o(0) = 0$. Taking Laplace transforms gives

$$(s^2 + 2s + 5)\Theta(s) = \frac{1}{s}$$

so that

$$\Theta = \frac{1}{s(s^2 + 2s + 5)} = \frac{\frac{1}{5}}{s} - \frac{\frac{1}{5}}{s} \frac{s+2}{(s+1)^2 + 4}$$

which, on inversion, gives

$$\theta_o(t) = \frac{1}{5} - \frac{1}{5} e^{-t} (\cos 2t + \frac{1}{2} \sin 2t) = \frac{1}{5} (1 - e^{-t} \cos 2t - \frac{1}{2} e^{-t} \sin 2t)$$

confirming the previous result.

We therefore see that a linear time-invariant system may be characterized in the frequency domain (or s domain) by its transfer function $G(s)$ or in the time domain by its impulse response $h(t)$, as depicted in Figures 5.35(a) and (b) respectively. The response in the frequency domain is obtained by algebraic multiplication, while the time-domain response involves a convolution. This equivalence of the operation of convolution in the time domain with algebraic multiplication in the frequency domain is clearly a powerful argument for the use of frequency-domain techniques in engineering design.

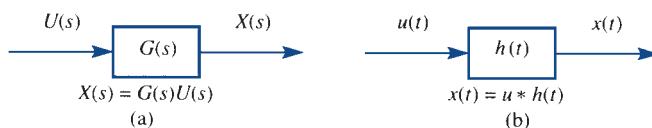


Figure 5.35 (a) Frequency-domain and (b) time-domain representations of a linear time-invariant system.

5.3.8 Exercises

- 34 For the following pairs of causal functions $f(t)$ and $g(t)$ show that $f * g(t) = g * f(t)$:

$$\begin{array}{ll} (a) \quad f(t) = t, & g(t) = \cos 3t \\ (b) \quad f(t) = t + 1, & g(t) = e^{-2t} \\ (c) \quad f(t) = t^2, & g(t) = \sin 2t \\ (d) \quad f(t) = e^{-t}, & g(t) = \sin t \end{array}$$

is

$$y(t) = (t - 1) + e^{-t}.$$

- 35 Using the convolution theorem, determine the following inverse Laplace transforms. Check your results by first expressing the given transform in partial-fractions form and then inverting using the standard results:

$$\begin{array}{l} (a) \quad \mathcal{L}^{-1}\left\{\frac{1}{s(s+3)^3}\right\} \\ (b) \quad \mathcal{L}^{-1}\left\{\frac{1}{(s-2)^2(s+3)^2}\right\} \\ (c) \quad \mathcal{L}^{-1}\left\{\frac{1}{s^2(s+4)}\right\} \end{array}$$

- 36 Taking $f(\lambda) = \lambda$ and $g(\lambda) = e^{-\lambda}$, use the inverse form (5.42) of the convolution theorem to show that the solution of the integral equation

$$y(t) = \int_0^t \lambda e^{-(t-\lambda)} d\lambda$$

- 37 Find the impulse response of the system characterized by the differential equation

$$\frac{d^2x}{dt^2} + 7\frac{dx}{dt} + 12x = u(t)$$

and hence find the response of the system to the pulse input $u(t) = A[H(t) - H(t - T)]$, assuming that it is initially in a quiescent state.

- 38 The response $\theta_o(t)$ of a servomechanism to a driving force $\theta_i(t)$ is given by the second-order differential equation

$$\frac{d^2\theta_o}{dt^2} + 4\frac{d\theta_o}{dt} + 5\theta_o = \theta_i \quad (t \geq 0)$$

Determine the impulse response of the system, and hence, using the convolution integral, obtain the response of the servomechanism to a unit step driving force, applied at time $t = 0$, given that the system is initially in a quiescent state.

Check your answer by directly solving the differential equation

$$\frac{d^2\theta_o}{dt^2} + 4\frac{d\theta_o}{dt} + 5\theta_o = 1$$

subject to the initial conditions $\theta_o = \dot{\theta}_o = 0$ when $t = 0$.

5.4 Solution of state-space equations

In this section we return to consider further the state-space model of dynamical systems introduced in Section 1.9. In particular we consider how Laplace transform methods may be used to solve the state-space equations.

5.4.1 SISO systems

In Section 1.9.1 we saw that the single-input-single-output system characterized by the differential equation (1.63) may be expressed in the state-space form

$$\dot{x} = Ax + bu \tag{5.46a}$$

$$y = c^T x \tag{5.46b}$$

where $\mathbf{x} = \mathbf{x}(t) = [x_1 \ x_2 \ \dots \ x_n]^T$ is the state vector and y the scalar output, the corresponding input-output transfer function model being

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_m s^m + \dots + b_0}{s^n + a_{n-1} s^{n-1} + \dots + a_0} \quad \begin{matrix} \leftarrow \mathbf{c} \\ \leftarrow \mathbf{A} \end{matrix} \quad (5.47)$$

where $Y(s)$ and $U(s)$ are the Laplace transforms of $y(t)$ and $u(t)$ respectively. Defining \mathbf{A} and \mathbf{b} as in (1.63), that is, we take \mathbf{A} to be the companion matrix of the left-hand side of (1.63) and take $\mathbf{b} = [0 \ 0 \ \dots \ 0 \ 1]^T$. In order to achieve the desired response, the vector \mathbf{c} is chosen to be

$$\mathbf{c} = [b_0 \ b_1 \ \dots \ b_m \ 0 \ \dots \ 0]^T \quad (5.48)$$

a structure we can confirm to be appropriate using Laplace transform notation. Defining $X_i(s) = \mathcal{L}\{x_i(t)\}$ and taking

$$X_1(s) = \frac{1}{s^n + a_{n-1} s^{n-1} + \dots + a_0} U(s)$$

we have

$$X_2(s) = sX_1(s), X_3(s) = sX_2(s) = s^2X_1(s), \dots, X_n(s) = sX_{n-1}(s) = s^{n-1}X_1(s)$$

so that

$$\begin{aligned} Y(s) &= b_0 X_1(s) + b_1 X_2(s) + \dots + b_m X_{m+1}(s) \\ &= \frac{b_0 + b_1 s + b_2 s^2 + \dots + b_m s^m}{s^n + a_{n-1} s^{n-1} + \dots + a_0} U(s) \end{aligned}$$

which confirms (5.48).

Note that adopting this structure for the state-space representation the last row in \mathbf{A} and the vector \mathbf{c} may be obtained directly from the transfer function (5.47) by reading the coefficients of the denominator and numerator backwards as indicated by the arrows, and negating those in the denominator.

Example 5.30

For the system characterized by the differential equation model

$$\frac{d^3y}{dt^3} + 6\frac{d^2y}{dx^2} + 11\frac{dy}{dt} + 3y = 5\frac{d^2u}{dt^2} + \frac{du}{dt} + u \quad (5.49)$$

considered in Example 1.40, obtain

- (a) a transfer function model;
- (b) a state-space model

Solution (a) Assuming all initial conditions to be zero, taking Laplace transforms throughout in (5.49) leads to

$$(s^3 + 6s^2 + 11s + 3)Y(s) = (5s^2 + s + 1)U(s)$$

so that the transfer-function model is given by

$$G(s) = \frac{Y(s)}{U(s)} = \frac{5s^2 + s + 1}{s^3 + 6s^2 + 11s + 3} \quad \begin{matrix} \leftarrow \mathbf{c} \\ \leftarrow \mathbf{A} \end{matrix}$$

(b) Taking \mathbf{A} to be the companion matrix $\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -3 & -11 & -6 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ then

$\mathbf{c} = [1 \ 1 \ 5]^T$ and the corresponding state-space model is given by (5.46).

Note: The eigenvalues of the state matrix \mathbf{A} are given by the roots of the characteristic equation $|\lambda\mathbf{I} - \mathbf{A}| = \lambda^3 + 6\lambda^2 + 11\lambda + 3 = 0$, which are the same as the poles of the transfer function $G(s)$.

Defining

$$\mathcal{L}\{\mathbf{x}(t)\} = \begin{bmatrix} \mathcal{L}\{x_1(t)\} \\ \mathcal{L}\{x_2(t)\} \\ \vdots \\ \mathcal{L}\{x_n(t)\} \end{bmatrix} = \begin{bmatrix} X_1(s) \\ X_2(s) \\ \vdots \\ X_n(s) \end{bmatrix} = \mathbf{X}(s)$$

and then taking the Laplace transform throughout in the state equation (5.46a) gives

$$s\mathbf{X}(s) - \mathbf{x}(0) = \mathbf{AX}(s) + \mathbf{b}U(s)$$

which on rearranging gives

$$(s\mathbf{I} - \mathbf{A})\mathbf{X}(s) = \mathbf{x}(0) + \mathbf{b}U(s)$$

where \mathbf{I} is the identity matrix. Premultiplying throughout by $(s\mathbf{I} - \mathbf{A})^{-1}$ gives

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}(0) + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}U(s) \quad (5.50)$$

which on taking inverse Laplace transforms gives the response as

$$\mathbf{x}(t) = \mathcal{L}^{-1}\{(s\mathbf{I} - \mathbf{A})^{-1}\}\mathbf{x}(0) + \mathcal{L}^{-1}\{(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}U(s)\} \quad (5.51)$$

Having obtained an expression for the system state $\mathbf{x}(t)$ its output, or response, $y(t)$ may be obtained from the linear output equation (5.46b).

Taking the Laplace transform throughout in (5.46b) gives

$$Y(s) = \mathbf{c}^T\mathbf{X}(s) \quad (5.52)$$

Assuming zero initial conditions in (5.50) we have

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}U(s)$$

which, on substitution in (5.52), gives the input-output relationship

$$Y(s) = \mathbf{c}^T(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}U(s) \quad (5.53)$$

From (5.53) it follows that the system transfer function $G(s)$ may be expressed in the form

$$G(s) = \mathbf{c}^T(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} = \frac{\mathbf{c}^T \text{adj}(s\mathbf{I} - \mathbf{A})\mathbf{b}}{\det(s\mathbf{I} - \mathbf{A})}$$

which indicates that the eigenvalues of \mathbf{A} are the same as the poles of $G(s)$, as noted at the end of Example 5.30. It follows, from Definition 5.2, that the system is stable provided all the eigenvalues of the state matrix \mathbf{A} have negative real parts.

On comparing the solution (5.51) with that given in (1.78), we find that the transition matrix $\Phi(t) = e^{\mathbf{A}t}$ may also be written in the form

$$\Phi(t) = \mathcal{L}^{-1}\{(sI - \mathbf{A})^{-1}\}$$

As mentioned in Section 1.10.3, having obtained $\Phi(t)$,

$$\Phi(t, t_0) = e^{\mathbf{A}(t-t_0)}$$

may be obtained by simply replacing t by $t - t_0$.

Example 5.31

Using the Laplace transform approach, obtain an expression for the state $\mathbf{x}(t)$ of the system characterized by the state equation

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & -3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t)$$

when the input $u(t)$ is the unit step function

$$u(t) = H(t) = \begin{cases} 0 & (t < 0) \\ 1 & (t \geq 0) \end{cases}$$

and subject to the initial condition $\mathbf{x}(0) = [1 \quad 1]^T$.

Solution In this case

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 1 & -3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad u(t) = H(t), \quad \mathbf{x}_0 = [1 \quad 1]^T$$

Thus

$$sI - \mathbf{A} = \begin{bmatrix} s+1 & 0 \\ -1 & s+3 \end{bmatrix}, \quad \det(sI - \mathbf{A}) = (s+1)(s+3)$$

giving

$$(sI - \mathbf{A})^{-1} = \frac{1}{(s+1)(s+3)} \begin{bmatrix} s+3 & 0 \\ 1 & s+1 \end{bmatrix} = \begin{bmatrix} \frac{1}{s+1} & 0 \\ \frac{1}{2(s+1)} - \frac{1}{2(s+3)} & \frac{1}{s+3} \end{bmatrix}$$

which, on taking inverse transforms, gives the transition matrix as

$$e^{\mathbf{A}t} = \mathcal{L}^{-1}\{(sI - \mathbf{A})^{-1}\} = \begin{bmatrix} e^{-t} & 0 \\ \frac{1}{2}e^{-t} - \frac{1}{2}e^{-3t} & e^{-3t} \end{bmatrix}$$

so that the first term in the solution (5.51) becomes

$$\mathcal{L}^{-1}\{(sI - \mathbf{A})^{-1}\}\mathbf{x}_0 = \begin{bmatrix} e^{-t} & 0 \\ \frac{1}{2}e^{-t} - \frac{1}{2}e^{-3t} & e^{-3t} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} e^{-t} \\ \frac{1}{2}e^{-t} + \frac{1}{2}e^{-3t} \end{bmatrix} \quad (5.54)$$

Since $U(s) = \mathcal{L}\{H(t)\} = 1/s$,

$$\begin{aligned}(sI - A)^{-1}bU(s) &= \frac{1}{(s+1)(s+3)} \begin{bmatrix} s+3 & 0 \\ 1 & s+1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \frac{1}{s} \\ &= \frac{1}{s(s+1)(s+3)} \begin{bmatrix} s+3 \\ s+2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{s} - \frac{1}{s+1} \\ \frac{2}{3s} - \frac{1}{2(s+1)} - \frac{1}{6(s+3)} \end{bmatrix}\end{aligned}$$

so that the second term in (5.51) becomes

$$\mathcal{L}^{-1}\{(sI - A)^{-1}bU(s)\} = \begin{bmatrix} 1 - e^{-t} \\ \frac{2}{3} - \frac{1}{2}e^{-t} - \frac{1}{6}e^{-3t} \end{bmatrix} \quad (5.55)$$

Combining (5.54) and (5.55), the response $x(t)$ is given by

$$x(t) = \begin{bmatrix} e^{-t} \\ \frac{1}{2}e^{-t} + \frac{1}{2}e^{-3t} \end{bmatrix} + \begin{bmatrix} 1 - e^{-t} \\ \frac{2}{3} - \frac{1}{2}e^{-t} - \frac{1}{6}e^{-3t} \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{2}{3} + \frac{1}{3}e^{-3t} \end{bmatrix}$$

5.4.2 Exercises

- 39 A system is modelled by the following differential equations

$$\begin{aligned}\dot{x}_1 + 5x_1 + x_2 &= 2u \\ \dot{x}_2 - 3x_1 + x_2 &= 5u\end{aligned}$$

coupled with the output equation

$$y = x_1 + 2x_2$$

Express the model in state-space form and obtain the transfer function of the system.

- 40 Find the state-space representation of the second order system modelled by the transfer function

$$G(s) = \frac{Y(s)}{U(s)} = \frac{s+1}{s^2 + 7s + 6}$$

- 41 Obtain the dynamic equations in state-space form for the systems having transfer-function models

$$(a) \frac{s^2 + 3s + 5}{s^3 + 6s^2 + 5s + 7} \quad (b) \frac{s^2 + 3s + 2}{s^3 + 4s^2 + 3s}$$

using the companion form of the system matrix in each case.

- 42 In formulating the state-space model (5.46) it is sometimes desirable to specify the output y to be the state variable x_1 ; that is, we take $c^T = [1 \ 0 \dots 0]^T$. If A is again taken to be the companion matrix of the denominator then it can be shown that the coefficients b_1, b_2, \dots, b_n of the vector b are determined as the first n coefficients in the series in s^{-1} obtained by dividing the denominator of the transfer function (5.47) into the numerator. Illustrate this approach for the transfer-function model of Figure 5.36.

$$\frac{5s^2 + s + 1}{s^3 + 6s^2 + 11s + 6}$$

Figure 5.36 Transfer-function model of Exercise 44.

- 43 A system is governed by the vector-matrix differential equation

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} u(t) \quad (t \geq 0)$$

where $\mathbf{x}(t)$ and $u(t)$ are respectively the state and input vectors of the system. Use Laplace transforms to obtain the state vector $\mathbf{x}(t)$ for the input $u(t) = [4 \ 3]^T$ and subject to the initial condition $\mathbf{x}(0) = [1 \ 2]^T$.

- 44 Given that the differential equations modelling a certain control system are

$$\dot{x}_1 = x_1 - 3x_2 + u$$

$$\dot{x}_2 = 2x_1 - 4x_2 + u$$

use (5.51) to determine the state vector $\mathbf{x} = [x_1 \ x_2]^T$ for the control input $u = e^{-3t}$, applied at time $t = 0$, given that $x_1 = x_2 = 1$ at time $t = 0$.

- 45 Using the Laplace transform approach, obtain an expression for the state $\mathbf{x}(t)$ of the system characterized by the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ 0 \end{bmatrix} u \quad (t \geq 0)$$

where the input is

$$u(t) = \begin{cases} 0 & (t < 0) \\ e^{-t} & (t \geq 0) \end{cases}$$

and subject to the initial condition $\mathbf{x}(0) = [1 \ 0]^T$.

- 46 A third-order single-input-single-output system is characterized by the transfer-function model

$$\frac{Y(s)}{U(s)} = \frac{3s^2 + 2s + 1}{s^3 + 6s^2 + 11s + 6}$$

Express the system model in the state-space form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + bu \quad (5.56a)$$

$$y = \mathbf{c}^T \mathbf{x} \quad (5.56b)$$

where \mathbf{A} is in the companion form. By making a suitable transformation $\mathbf{x} = Mz$, reduce the state-space model to its canonical form, and comment on the stability, controllability and observability of the system.

Given that

- (i) a necessary and sufficient condition for the system (5.56) to be controllable is that the rank of the **Kalman matrix** $[\mathbf{b} \ \mathbf{Ab} \ \mathbf{A}^2\mathbf{b} \dots \mathbf{A}^{n-1}\mathbf{b}]$ be the same as the order of \mathbf{A} , and
- (ii) a necessary and sufficient condition for it to be observable is that the rank of the Kalman matrix $[\mathbf{c} \ \mathbf{A}^T\mathbf{c} \ (\mathbf{A}^T)^2\mathbf{c} \dots (\mathbf{A}^T)^{n-1}\mathbf{c}]$ be the same as the order of \mathbf{A} ,

evaluate the ranks of the relevant Kalman matrices to confirm your earlier conclusions on the controllability and observability of the given system.

Repeat Exercise 46 for the system characterized by the transfer-function model

$$\frac{s^2 + 3s + 5}{s^3 + 6s^2 + 5s}$$

5.4.3 MIMO systems

As indicated in (1.66) the general form of the state-space model representation of an n th-order **multi-input–multi-output** system subject to r inputs and l outputs is

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu} \quad (5.57a)$$

$$\mathbf{y} = \mathbf{Cx} + \mathbf{Du} \quad (5.57b)$$

where \mathbf{x} is the n -state vector, \mathbf{u} is the r -input vector, \mathbf{y} is the l -output vector, \mathbf{A} is the $n \times n$ system matrix, \mathbf{B} is the $n \times r$ control (or input) matrix and \mathbf{C} and \mathbf{D} are respectively $l \times n$ and $l \times r$ output matrices, with the matrix \mathbf{D} relating to the part of the input that is applied directly into the output.

Defining

$$\mathcal{L}\{\mathbf{y}(t)\} = \begin{bmatrix} \mathcal{L}\{y_1(t)\} \\ \mathcal{L}\{y_2(t)\} \\ \vdots \\ \mathcal{L}\{y_l(t)\} \end{bmatrix} = \begin{bmatrix} Y_1(s) \\ Y_2(s) \\ \vdots \\ Y_l(s) \end{bmatrix} = \mathbf{Y}(s)$$

$$\mathcal{L}\{\mathbf{u}(t)\} = \begin{bmatrix} \mathcal{L}\{u_1(t)\} \\ \mathcal{L}\{u_2(t)\} \\ \vdots \\ \mathcal{L}\{u_r(t)\} \end{bmatrix} = \begin{bmatrix} U_1(s) \\ U_2(s) \\ \vdots \\ U_r(s) \end{bmatrix} = \mathbf{U}(s)$$

and taking Laplace transforms throughout in the state equation (5.57a), following the same procedure as for the SISO case, gives

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}(0) + (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s) \quad (5.58)$$

Taking inverse Laplace transforms in (5.58) gives

$$\mathbf{x}(t) = \mathcal{L}^{-1}\{(s\mathbf{I} - \mathbf{A})^{-1}\}\mathbf{x}(0) + \mathcal{L}^{-1}\{(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s)\} \quad (5.59)$$

The output, or response, vector $\mathbf{y}(t)$ may then be obtained directly from (5.57b).

We can also use the Laplace transform formulation to obtain the **transfer matrix** $\mathbf{G}(s)$, between the input and output vectors, for a multivariable system. Taking Laplace transforms throughout in the output equation (5.57b) gives

$$\mathbf{Y}(s) = \mathbf{C}\mathbf{X}(s) + \mathbf{D}\mathbf{U}(s) \quad (5.60)$$

Assuming zero initial conditions in (5.58) we have

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s)$$

Substituting in (5.60), gives the system input–output relationship

$$\mathbf{Y}(s) = [\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}]\mathbf{U}(s)$$

Thus the transfer matrix $\mathbf{G}(s)$ model of a state-space model defined by the quadruple $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ is

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \quad (5.61)$$

The reverse problem of obtaining a state-space model from a given transfer matrix is not uniquely solvable. For example, in Section 1.10.6 we showed that a state-space model can be reduced to canonical form and indicated that this was without affecting the input–output behaviour. In Section 1.10.6 it was shown that under the transformation $\mathbf{x} = \mathbf{Tz}$, where \mathbf{T} is a non-singular matrix, (5.57) may be reduced to the form

$$\begin{aligned} \dot{\mathbf{z}} &= \tilde{\mathbf{A}}\mathbf{z} + \tilde{\mathbf{B}}\mathbf{u} \\ \mathbf{y} &= \tilde{\mathbf{C}}\mathbf{z} + \tilde{\mathbf{D}}\mathbf{u} \end{aligned} \quad (5.62)$$

where \mathbf{z} is now a state vector and

$$\tilde{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \tilde{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B}, \tilde{\mathbf{C}} = \mathbf{C}\mathbf{T}, \tilde{\mathbf{D}} = \mathbf{D}$$

From (5.61), the input–output transfer matrix corresponding to (5.62) is

$$\begin{aligned}
 \mathbf{G}_1(s) &= \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \\
 &= \mathbf{C}\mathbf{T}(s\mathbf{I} - \mathbf{T}^{-1}\mathbf{AT})^{-1}\mathbf{T}^{-1}\mathbf{B} + \mathbf{D} \\
 &= \mathbf{C}\mathbf{T}(s\mathbf{T}^{-1}\mathbf{IT} - \mathbf{T}^{-1}\mathbf{AT})^{-1}\mathbf{T}^{-1}\mathbf{B} + \mathbf{D} \\
 &= \mathbf{C}\mathbf{T}[\mathbf{T}^{-1}(s\mathbf{I} - \mathbf{A})\mathbf{T}]^{-1}\mathbf{T}^{-1}\mathbf{B} + \mathbf{D} \\
 &= \mathbf{C}\mathbf{T}[\mathbf{T}^{-1}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{T}]\mathbf{T}^{-1}\mathbf{B} + \mathbf{D} \quad (\text{using the commutative property}) \\
 &= \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \\
 &= \mathbf{G}(s)
 \end{aligned}$$

where $\mathbf{G}(s)$ is the transfer matrix corresponding to (5.57), confirming that the input–output behaviour of the state-space model defined by the quadruple $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ is the same as that defined by the quadruple $\{\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}, \tilde{\mathbf{D}}\}$. The problem of finding state-space models that have a specified transfer-function matrix is known as the **realization problem**.

It follows from (5.61) that

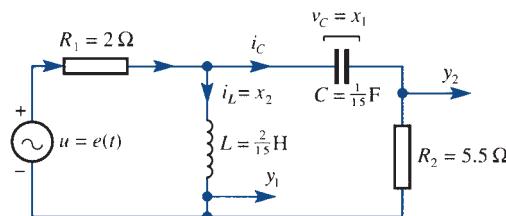
$$\mathbf{G}(s) = \frac{\mathbf{C}\text{adj}(s\mathbf{I} - \mathbf{A})\mathbf{B} + \mathbf{D}}{\det(s\mathbf{I} - \mathbf{A})}$$

Clearly, if $s = p$ is a pole of $\mathbf{G}(s)$ then it must necessarily be an eigenvalue of the state matrix \mathbf{A} , but the converse is not necessarily true. It can be shown that the poles of $\mathbf{G}(s)$ are identical to the eigenvalues of \mathbf{A} when it is impossible to find a state-space model with a smaller state dimension than n having the same transfer-function matrix. In such cases the state-space model is said to be in **minimal form**.

Example 5.32

- Obtain the state-space model characterizing the network of Figure 5.37. Take the inductor current and the voltage drop across the capacitor as the state variables, take the input variable to be the output of the voltage source, and take the output variables to be the currents through L and R_2 respectively.
- Find the transfer-function matrix relating the output variables y_1 and y_2 to the input variable u . Thus find the system response to the unit step $u(t) = H(t)$, assuming that the circuit is initially in a quiescent state.

Figure 5.37 Network of Example 5.32.



Solution

- The current i_C in the capacitor is given by

$$i_C = C\dot{v}_C = C\dot{x}_1$$

Applying Kirchhoff's second law to the outer loop gives

$$e = R_1(i_L + i_C) + v_C + R_2i_C = R_1(x_2 + C\dot{x}_1) + x_1 + R_2C\dot{x}_1$$

leading to

$$\dot{x}_1 = -\frac{1}{C(R_1 + R_2)}x_1 - \frac{R_1}{C(R_1 + R_2)}x_2 + \frac{e}{C(R_1 + R_2)}$$

Applying Kirchhoff's second law to the left-hand loop gives

$$e = R_1(i_L + i_C) + L\dot{i}_L = R_1(x_2 + C\dot{x}_1) + L\dot{x}_2$$

leading to

$$\dot{x}_2 = \frac{R_1}{L(R_1 + R_2)}x_1 - \frac{R_1 R_2}{L(R_1 + R_2)}x_2 + \frac{e}{L} \frac{R_2}{R_1 + R_2}$$

Also,

$$y_1 = x_2$$

$$y_2 = C\dot{x}_1 = -\frac{1}{R_1 + R_2}x_1 - \frac{R_1}{R_1 + R_2}x_2 + \frac{e}{R_1 + R_2}$$

Substituting the given parameter values leads to the state-space representation

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -2 & -4 \\ 2 & -11 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ \frac{11}{2} \end{bmatrix} u$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{2}{15} & -\frac{4}{15} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2}{15} \end{bmatrix} u$$

which is of the standard form

$$\dot{x} = Ax + bu$$

$$y = Cx + du$$

- (b) From (5.61), the transfer-function matrix $\mathbf{G}(s)$ relating the output variables y_1 and y_2 to the input u is

$$\mathbf{G}(s) = \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{b} + \mathbf{d}$$

Now

$$sI - \mathbf{A} = \begin{bmatrix} s+2 & 4 \\ -2 & s+11 \end{bmatrix}$$

giving

$$(sI - \mathbf{A})^{-1} = \frac{1}{(s+3)(s+10)} \begin{bmatrix} s+11 & -4 \\ 2 & s+2 \end{bmatrix}$$

$$\begin{aligned} \mathbf{C}(sI - \mathbf{A})^{-1}\mathbf{b} &= \frac{1}{(s+3)(s+10)} \begin{bmatrix} 0 & 1 \\ -\frac{2}{15} & -\frac{4}{15} \end{bmatrix} \begin{bmatrix} s+11 & -4 \\ 2 & s+2 \end{bmatrix} \begin{bmatrix} 2 \\ \frac{11}{2} \end{bmatrix} \\ &= \frac{1}{(s+3)(s+10)} \begin{bmatrix} \frac{11}{2}s + 15 \\ -\frac{26}{15}s - 4 \end{bmatrix} \end{aligned}$$

so that

$$\mathbf{G}(s) = \frac{1}{(s+3)(s+10)} \begin{bmatrix} \frac{11}{2}s + 15 \\ -\frac{26}{15}s - 4 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{2}{15} \end{bmatrix} = \begin{bmatrix} \frac{\frac{11}{2}s + 15}{(s+3)(s+10)} \\ \frac{-\frac{26}{15}s - 4}{(s+3)(s+10)} + \frac{2}{15} \end{bmatrix}$$

The output variables y_1 and y_2 are then given by the inverse Laplace transform of

$$\mathbf{Y}(s) = \mathbf{G}(s)U(s)$$

where $U(s) = \mathcal{L}[u(t)] = \mathcal{L}[H(t)] = 1/s$; that is,

$$\begin{aligned} \mathbf{Y}(s) &= \begin{bmatrix} \frac{\frac{11}{2}s + 15}{s(s+3)(s+10)} \\ \frac{-\frac{26}{15}s - 4}{s(s+3)(s+10)} + \frac{2}{15s} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\frac{1}{2}}{s} + \frac{\frac{1}{14}}{s+3} - \frac{\frac{4}{7}}{s+10} \\ \frac{-\frac{2}{15}}{s} - \frac{\frac{2}{35}}{s+3} + \frac{\frac{4}{21}}{s+10} + \frac{\frac{2}{15}}{s} \end{bmatrix} \end{aligned}$$

which on taking inverse Laplace transforms gives the output variables as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} + \frac{1}{14}e^{-3t} - \frac{4}{7}e^{-10t} \\ -\frac{2}{35}e^{-3t} + \frac{4}{21}e^{-10t} \end{bmatrix} \quad (t \geq 0)$$



In MATLAB the function `tf2ss` can be used to convert a transfer function **to** state-space form for SISO systems. At present there appears to be no equivalent function for MIMO systems. Thus the command

$$[A, B, C, D] = \text{tf2ss}(b, a)$$

returns the A, B, C, D matrices of the state-form representation of the transfer function

$$G(s) = C(sI - A)^{-1}B + D = \frac{b_1 s^{m-1} + \dots + b_{m-1} s + b_m}{a_1 s^{n-1} + \dots + a_{n-1} s + a_n}$$

where the input vector a contains the denominator coefficients and the vector b contains the numerator coefficients, both in ascending powers of s .

(Note: The function `tfss` can also be used in the case of single-input–multi-output systems. In such cases the matrix numerator must contain the numerator coefficients with as many rows as there are outputs.)

To illustrate consider the system of Example 5.30, for which

$$G(s) = \frac{5s^2 + s + 1}{s^3 + 6s^2 + 11s + 3}$$

In this case the commands

```
b = [5 1 1];
a = [1 6 11 3];
[A,B,C,D] = tf2ss(b,a)

return

A = -6 -11 -3
    1     0     0
    0     1     0
B = 1
    0
    0
C = 5 1 1
D = 0
```

giving the state-space model

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -6 & -11 & -3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u; y = [5 \ 1 \ 1] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

(Note: This state-space model differs from the one given in the answer to Example 5.30. Both forms are equivalent to the given transfer function model, with an alternative companion form taken as indicated in Section 1.9.1.)

Likewise, in MATLAB the function `ss2tf` converts the state-space representation to the equivalent transfer function/matrix representation (this being applicable to both SISO and MIMO systems). The command

```
[b,a] = ss2tf(A,B,C,D,iu)
```

returning the transfer function/matrix

$$\mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

from the `iu`-th input. Again the vector `a` contains the coefficients of the denominator in ascending powers of s and the numerator coefficients are returned in array `b` with as many rows as there are outputs.

(Note: The entry `iu` in the command can be omitted when considering SISO systems so, for example, the commands

```
A = [-6 -11 -3; 1 0 0; 0 1 0];
B=[1;0;0];
C=[5 1 1];
D=0;
[b,a]=ss2tf(A,B,C,D)

return

b = 0 5.0000 1.0000 1.0000
a = 1.0000 6.0000 11.0000 3.0000
```

giving the transfer function representation

$$\mathbf{G}(s) = \frac{5s^2 + s + 1}{s^3 + 6s^2 + 11s + 3}$$

which confirms the answer to the above example. As an exercise confirm that the state-space model obtained in the answer to Example 5.30 is also equivalent to this transfer function representation.)

To illustrate a MIMO system consider the system in Exercise 49, in which the state-space model is

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \end{aligned}$$

and we wish to determine the equivalent transfer matrix. The commands

```
A = [0 1 0 0;-1 -1 0 1;0 0 0 1;0 1 -1 -1];
B=[0 0;1 0;0 0;0 1];
C=[1 0 0 0 ; 0 0 1 0];
D=[0 0 ;0 0];
[b1,a] = ss2tf(A,B,C,D,1)
```

return the response to u_1

```
b1 = 0 0 1.0000 1.0000 1.0000
      0 0.0000 0.0000 1.0000 0.0000
a = 1.0000 2.0000 2.0000 2.0000 1.0000
```

and the additional command

```
[b2,a] = ss2tf(A,B,C,D,2)
```

returns response to u_2

```
b2 = 0 0.0000 0.0000 1.0000 0.0000
      0 0.0000 1.0000 1.0000 1.0000
A = 1.0000 2.0000 2.0000 2.0000 1.0000
```

leading to the transfer matrix model

$$\begin{aligned} \mathbf{G}(s) &= \frac{1}{s^4 + 2s^3 + 2s^2 + 2s + 1} \begin{bmatrix} s^2 + s + 1 & s \\ s & s^2 + s + 1 \end{bmatrix} \\ &= \frac{1}{(s+1)^2(s^2+1)} \begin{bmatrix} s^2 + s + 1 & s \\ s & s^2 + s + 1 \end{bmatrix} \end{aligned}$$

5.4.4 Exercises

- 48 Determine the response $y = x_1$ of the system governed by the differential equations

$$\begin{cases} \dot{x}_1 = -2x_2 + u_1 - u_2 \\ \dot{x}_2 = x_1 - 3x_2 + u_1 + u_2 \end{cases} \quad (t \geq 0)$$

to an input $\mathbf{u} = [u_1 \ u_2]^T = [1 \ t]^T$ and subject to the initial conditions $\mathbf{x}_1(0) = 0, \mathbf{x}_2(0) = 1$.

- 49 Consider the 2-input–2-output system modelled by the pair of simultaneous differential equations

$$\begin{aligned} \ddot{y}_1 + \dot{y}_1 - \dot{y}_2 + y_1 &= u_1 \\ \ddot{y}_2 + \dot{y}_2 - \dot{y}_1 + y_2 &= u_2 \end{aligned}$$

Taking the state vector to be $\mathbf{x} = [y_1 \ \dot{y}_1 \ y_2 \ \dot{y}_2]^T$ express the model as a state-space model of the form

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$

$$\mathbf{y} = \mathbf{Cx}$$

Determine the transfer matrix and verify that its poles are identical to the eigenvalues of the state matrix \mathbf{A} .

- 50 Considering the network of Figure 5.38

- (a) Determine the state-space model in the form

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$

$$\mathbf{y} = \mathbf{Cx}$$

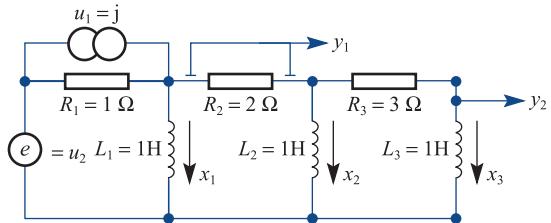


Figure 5.38 Network of Exercise 50.

Take the inductor currents in L_1, L_2 and L_3 as the state variables x_1, x_2, x_3 respectively; take the input variables u_1 and u_2 to be the outputs of the current and voltage sources respectively; and take the output variables y_1 and y_2 to be the voltage across R_2 and the current through L_3 respectively.

- (b) Determine the transfer matrix $\mathbf{G}(s)$ relating the output vector to the input vector.
(c) Assuming that the circuit is initially in a quiescent state, determine the response $\mathbf{y}(t)$ to the input pair

$$u_1(t) = H(t)$$

$$u_2(t) = tH(t)$$

where $H(t)$ denotes the Heaviside function.

5.5 Engineering application: frequency response

Frequency-response methods provide a graphical approach for the analysis and design of systems. Traditionally these methods have evolved from practical considerations, and as such are still widely used by engineers, providing tremendous insight into overall system behaviour. In this section we shall illustrate how the frequency response can be readily obtained from the system transfer function $G(s)$ by simply replacing s by $j\omega$. Methods of representing it graphically will also be considered.

Consider the system depicted in Figure 5.26, with transfer function

$$G(s) = \frac{K(s - z_1)(s - z_2) \dots (s - z_m)}{(s - p_1)(s - p_2) \dots (s - p_n)} \quad (m \leq n) \quad (5.63)$$

When the input is the sinusoidally varying signal

$$u(t) = A \sin \omega t$$

applied at time $t = 0$, the system response $x(t)$ for $t \geq 0$ is determined by

$$X(s) = G(s)\mathcal{L}\{A \sin \omega t\}$$

That is,

$$\begin{aligned} X(s) &= G(s) \frac{A\omega}{s^2 + \omega^2} \\ &= \frac{KA\omega(s - z_1)(s - z_2) \cdots (s - z_m)}{(s - p_1)(s - p_2) \cdots (s - p_n)(s - j\omega)(s + j\omega)} \end{aligned}$$

which, on expanding in partial fractions, gives

$$X(s) = \frac{\alpha_1}{s - j\omega} + \frac{\alpha_2}{s + j\omega} + \sum_{i=1}^n \frac{\beta_i}{s - p_i}$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2, \dots, \beta_n$ are constants. Here the first two terms in the summation are generated by the input and determine the steady-state response, while the remaining terms are generated by the transfer function and determine the system transient response.

Taking inverse Laplace transforms, the system response $x(t)$, $t \geq 0$, is given by

$$x(t) = \alpha_1 e^{j\omega t} + \alpha_2 e^{-j\omega t} + \sum_{i=1}^n \beta_i e^{p_i t} \quad (t \geq 0)$$

In practice we are generally concerned with systems that are stable, for which the poles p_i , $i = 1, 2, \dots, n$, of the transfer function $G(s)$ lie in the left half of the s plane. Consequently, for practical systems the time-domain terms $\beta_i e^{p_i t}$, $i = 1, 2, \dots, n$, decay to zero as t increases, and will not contribute to the steady-state response $x_{ss}(t)$ of the system. Thus for stable linear systems the latter is determined by the first two terms as

$$x_{ss}(t) = \alpha_1 e^{j\omega t} + \alpha_2 e^{-j\omega t}$$

Using the ‘cover-up’ rule for determining the coefficients α_1 and α_2 in the partial-fraction expansions gives

$$\begin{aligned} \alpha_1 &= \left[\frac{(s - j\omega)G(s)A\omega}{(s - j\omega)(s + j\omega)} \right]_{s=j\omega} = \frac{A}{2j}G(j\omega) \\ \alpha_2 &= \left[\frac{(s + j\omega)G(s)A\omega}{(s - j\omega)(s + j\omega)} \right]_{s=-j\omega} = -\frac{A}{2j}G(-j\omega) \end{aligned}$$

so that the steady-state response becomes

$$x_{ss}(t) = \frac{A}{2j}G(j\omega)e^{j\omega t} - \frac{A}{2j}G(-j\omega)e^{-j\omega t} \quad (5.64)$$

$G(j\omega)$ can be expressed in the polar form

$$G(j\omega) = |G(j\omega)| e^{j\arg G(j\omega)}$$

where $|G(j\omega)|$ denotes the magnitude (or modulus) of $G(j\omega)$. (Note that both the magnitude and argument vary with frequency ω) Then, assuming that the system has real parameters,

$$G(-j\omega) = |G(j\omega)| e^{-j\arg G(j\omega)}$$

and the steady-state response (5.64) becomes

$$\begin{aligned}x_{ss}(t) &= \frac{A}{2j} [|G(j\omega)| e^{j\arg G(j\omega)}] e^{j\omega t} - \frac{A}{2j} [|G(j\omega)| e^{-j\arg G(j\omega)}] e^{-j\omega t} \\&= \frac{A}{2j} |G(j\omega)| [e^{j[\omega t + \arg G(j\omega)]} - e^{-j[\omega t + \arg G(j\omega)]}]\end{aligned}$$

That is,

$$x_{ss}(t) = A |G(j\omega)| \sin [\omega t + \arg G(j\omega)] \quad (5.65)$$

This indicates that if a stable linear system with transfer function $G(s)$ is subjected to a sinusoidal input then

- (a) the steady-state system response is also a sinusoid having the same frequency ω as the input;
- (b) the amplitude of this response is $|G(j\omega)|$ times the amplitude A of the input sinusoid; the input is said to be **amplified** if $|G(j\omega)| > 1$ and **attenuated** if $|G(j\omega)| < 1$;
- (c) the phase shift between input and output is $\arg G(j\omega)$. The system is said to **lead** if $\arg G(j\omega) > 0$ and **lag** if $\arg G(j\omega) < 0$.

The variations in both the magnitude $|G(j\omega)|$ and argument $\arg G(j\omega)$ as the frequency ω of the input sinusoid is varied constitute the **frequency response of the system**, the magnitude $|G(j\omega)|$ representing the **amplitude gain** or **amplitude ratio** of the system for sinusoidal input with frequency ω , and the argument $\arg G(j\omega)$ representing the **phase shift**.

The result (5.65) implies that the function $G(j\omega)$ may be found experimentally by subjecting a system to sinusoidal excitations and measuring the amplitude gain and phase shift between output and input as the input frequency is varied over the range $0 < \omega < \infty$. In principle, therefore, frequency-response measurements may be used to determine the system transfer function $G(s)$.

In Chapters 7 and 8, dealing with Fourier series and Fourier transforms, we shall see that most functions can be written as sums of sinusoids, and consequently the response of a linear system to almost any input can be deduced in the form of the corresponding sinusoidal responses. It is important, however, to appreciate that the term ‘response’ in the expression ‘frequency response’ only relates to the steady-state response behaviour of the system.

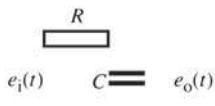
The information contained in the system frequency response may be conveniently displayed in graphical form. In practice it is usual to represent it by two graphs: one showing how the amplitude $|G(j\omega)|$ varies with frequency and one showing how the phase shift $\arg G(j\omega)$ varies with frequency.

Example 5.33

Determine the frequency response of the RC filter shown in Figure 5.39. Sketch the amplitude and phase-shift plots.

Solution The input–output relationship is given by

$$E_o(s) = \frac{1}{RCs + 1} E_i(s)$$



so that the filter is characterized by the transfer function

$$e_i(t) \quad C = e_o(t) \quad G(s) = \frac{1}{RCs + 1}$$

Figure 5.39 RC filter. Therefore

$$\begin{aligned} G(j\omega) &= \frac{1}{RCj\omega + 1} = \frac{1 - jRC\omega}{1 + R^2C^2\omega^2} \\ &= \frac{1}{1 + R^2C^2\omega^2} - j \frac{RC\omega}{1 + R^2C^2\omega^2} \end{aligned}$$

giving the frequency-response characteristics

$$\begin{aligned} \text{amplitude ratio} &= |G(j\omega)| \\ &= \sqrt{\left[\frac{1}{(1 + R^2C^2\omega^2)^2} + \frac{R^2C^2\omega^2}{(1 + R^2C^2\omega^2)^2} \right]} \\ &= \frac{1}{\sqrt{1 + R^2C^2\omega^2}} \end{aligned}$$

$$\text{phase shift} = \arg G(j\omega) = -\tan^{-1}(RC\omega)$$

Note that for $\omega = 0$

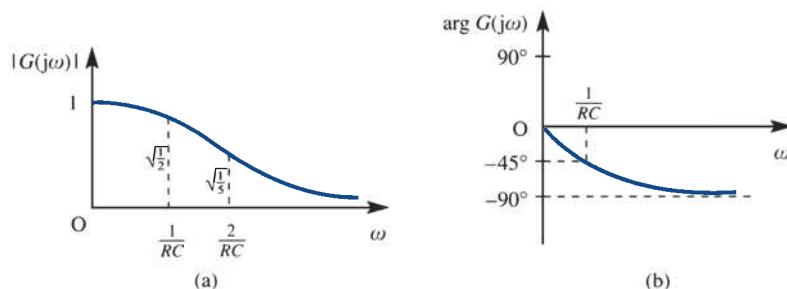
$$|G(j\omega)| = 1, \quad \arg G(j\omega) = 0$$

and as $\omega \rightarrow \infty$

$$|G(j\omega)| \rightarrow 0, \quad \arg G(j\omega) \rightarrow -\frac{1}{2}\pi$$

Plots of the amplitude and phase-shift curves are shown in Figures 5.40(a) and (b) respectively.

Figure 5.40
Frequency-response plots for Example 5.33:
(a) amplitude plot;
(b) phase-shift plot.



For the simple transfer function of Example 5.33, plotting the amplitude and phase-shift characteristics was relatively easy. For higher-order transfer functions it can be a rather tedious task, and it would be far more efficient to use a suitable computer

package. However, to facilitate the use of frequency-response techniques in system design, engineers adopt a different approach, making use of **Bode plots** to display the relevant information. This approach is named after H. W. Bode, who developed the techniques at the Bell Laboratories in the late 1930s. Again it involves drawing separate plots of amplitude and phase shift, but in this case on semi-logarithmic graph paper, with frequency plotted on the horizontal logarithmic axis and amplitude, or phase, on the vertical linear axis. It is also normal to express the amplitude gain in decibels (dB); that is,

$$\text{amplitude gain in dB} = 20 \log |G(j\omega)|$$

and the phase shift $\arg G(j\omega)$ in degrees. Thus the Bode plots consist of

- (a) a plot of amplitude in decibels versus $\log \omega$, and
- (b) a plot of phase shift in degrees versus $\log \omega$.

Note that with the amplitude gain measured in decibels, the input signal will be amplified if the gain is greater than zero and attenuated if it is less than zero.

The advantage of using Bode plots is that the amplitude and phase information can be obtained from the constituent parts of the transfer function by graphical addition. It is also possible to make simplifying approximations in which curves can be replaced by straight-line asymptotes. These can be drawn relatively quickly, and provide sufficient information to give an engineer a ‘feel’ for the system behaviour. Desirable system characteristics are frequently specified in terms of frequency-response behaviour, and since the approximate Bode plots permit quick determination of the effect of changes, they provide a good test for the system designer.

Example 5.34

Draw the approximate Bode plots corresponding to the transfer function

$$G(s) = \frac{4 \times 10^3 (5 + s)}{s(100 + s)(20 + s)} \quad (5.66)$$

Solution First we express the transfer function in what is known as the **standard form**, namely

$$G(s) = \frac{10(1 + 0.2s)}{s(1 + 0.01s)(1 + 0.05s)}$$

giving

$$G(j\omega) = \frac{10(1 + j0.2\omega)}{j\omega(1 + j0.01\omega)(1 + j0.05\omega)}$$

Taking logarithms to base 10,

$$\begin{aligned} 20 \log |G(j\omega)| &= 20 \log 10 + 20 \log |1 + j0.2\omega| - 20 \log |j\omega| \\ &\quad - 20 \log |1 + j0.01\omega| - 20 \log |1 + j0.05\omega| \\ \arg G(j\omega) &= \arg 10 + \arg(1 + j0.2\omega) - \arg j\omega - \arg(1 + j0.01\omega) \\ &\quad - \arg(1 + j0.05\omega) \end{aligned} \quad (5.67)$$

The transfer function involves constituents that are again a simple zero and simple poles (including one at the origin). We shall now illustrate how the Bode plots can be built up from those of the constituent parts.

Consider first the amplitude gain plot, which is a plot of $20 \log |G(j\omega)|$ versus $\log \omega$:

- for a simple gain k a plot of $20 \log k$ is a horizontal straight line, being above the 0 dB axis if $k > 1$ and below it if $k < 1$;
- for a simple pole at the origin a plot of $-20 \log \omega$ is a straight line with slope -20 dB/decade and intersecting the 0 dB axis at $\omega = 1$;
- for a simple zero or pole not at the origin we see that

$$20 \log |1 + j\tau\omega| \rightarrow \begin{cases} 0 & \text{as } \omega \rightarrow 0 \\ 20 \log \tau\omega = 20 \log \omega - 20 \log(1/\tau) & \text{as } \omega \rightarrow \infty \end{cases}$$

Note that the graph of $20 \log \tau\omega$ is a straight line with slope 20 dB/decade and intersecting the 0 dB axis at $\omega = 1/\tau$. Thus the plot of $20 \log |1 + j\tau\omega|$ may be approximated by two straight lines: one for $\omega < 1/\tau$ and one for $\omega > 1/\tau$. The frequency at intersection $\omega = 1/\tau$ is called the **breakpoint** or **corner frequency**; here $|1 + j\tau\omega| = \sqrt{2}$, enabling the true curve to be indicated at this frequency. Using this approach, straight-line approximations to the amplitude plots of a simple zero and a simple pole, neither at zero, are shown in Figures 5.41(a) and (b) respectively (actual plots are also shown).

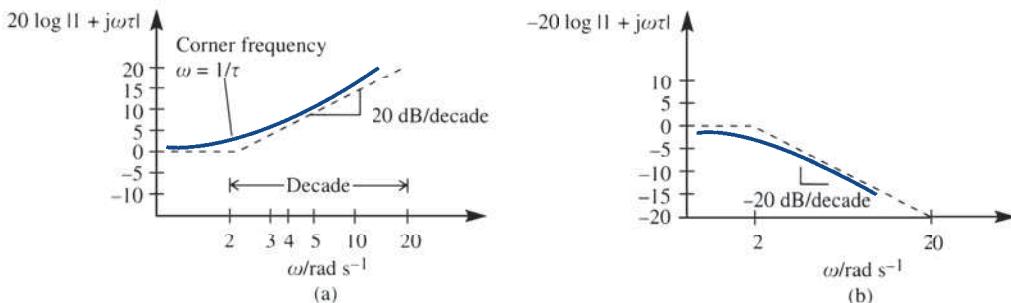


Figure 5.41 Straight-line approximations to Bode amplitude plots: (a) simple zero; (b) simple pole.

Using the approximation plots for the constituent parts as indicated in (a)–(c) earlier, we can build up the approximate amplitude gain plot corresponding to (5.66) by graphical addition as indicated in Figure 5.42. The actual amplitude gain plot, produced using a software package, is also shown.

The idea of using asymptotes can also be used to draw the phase-shift Bode plots, again taking account of the accumulated effects of the individual components making up the transfer function, namely that

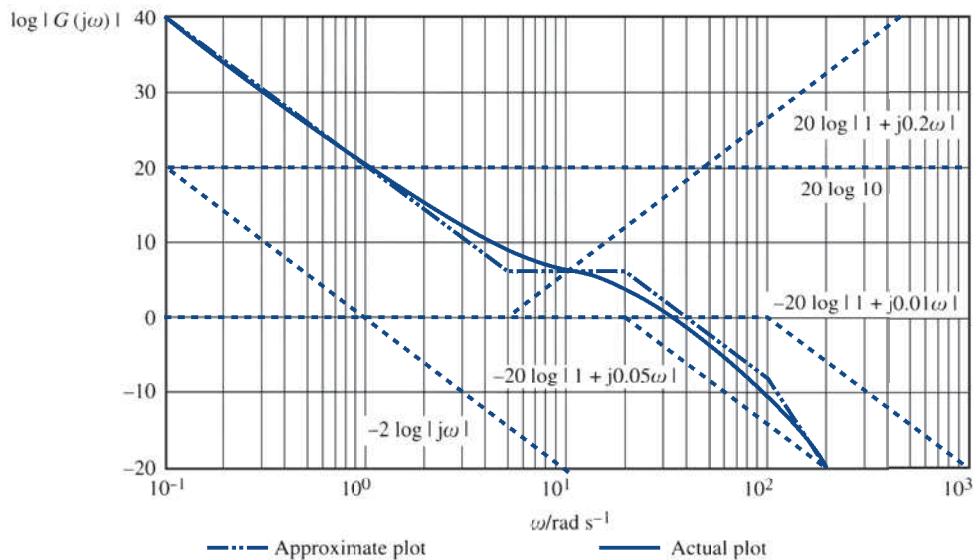
- the phase shift associated with a constant gain k is zero;
- the phase shift associated with a simple pole or zero at the origin is $+90^\circ$ or -90° respectively;
- for a simple zero or pole not at the origin

$$\tan^{-1}(\omega\tau) \rightarrow \begin{cases} 0 & \text{as } \omega \rightarrow 0 \\ 90^\circ & \text{as } \omega \rightarrow \infty \end{cases}$$

$$\tan^{-1}(\omega\tau) = 45^\circ \quad \text{when } \omega\tau = 1$$

Figure 5.42

Amplitude Bode plots for the $G(s)$ of Example 5.34.



With these observations in mind, the following approximations are made. For frequencies ω less than one-tenth of the corner frequency $\omega = 1/\tau$ (that is, for $\omega < 1/10\tau$) the phase shift is assumed to be 0° , and for frequencies greater than ten times the corner frequency (that is, for $\omega > 10/\tau$) the phase shift is assumed to be $\pm 90^\circ$. For frequencies between these limits (that is, $1/10\tau < \omega < 10/\tau$) the phase-shift plot is taken to be a straight line that passes through 0° at $\omega = 1/10\tau$, $\pm 45^\circ$ at $\omega = 1/\tau$, and $\pm 90^\circ$ at $\omega = 10/\tau$. In each case the plus sign is associated with a zero and the minus sign with a pole. With these assumptions, straight-line approximations to the phase-shift plots for a simple zero and pole, neither located at the origin, are shown in Figures 5.43(a) and (b) respectively (the actual plots are represented by the broken curves).

Using these approximations, a straight-line approximate phase-gain plot corresponding to (5.67) is shown in Figure 5.44. Again, the actual phase-gain plot, produced using a software package, is shown.

Figure 5.43

Approximate Bode phase-shift plots:
(a) simple zero;
(b) simple pole.

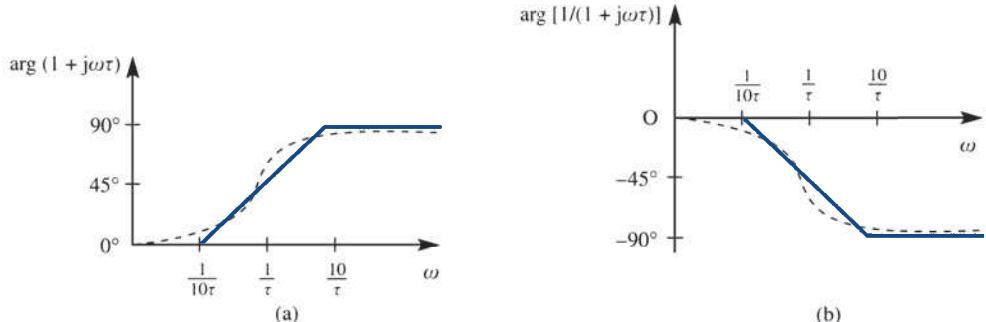
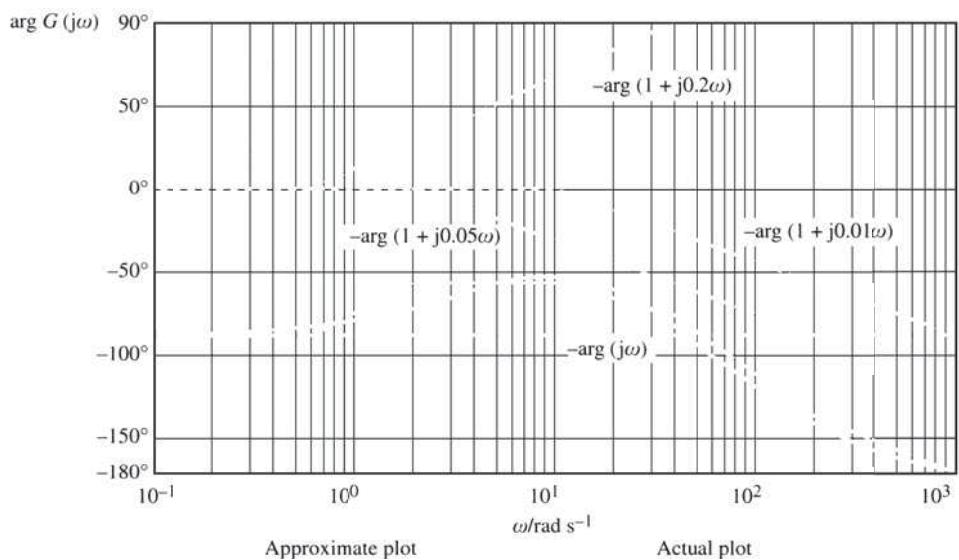


Figure 5.44
Phase-shift Bode plot for the $G(s)$ of Example 5.34.



In MATLAB the amplitude and phase-gain plots are generated using the commands

```
s=tf('s')
G=4*10^3*(s+5)/(s*(100+s)*(20+s));
bode(G)
```

In the graphical approach adopted in this section, separate plots of amplitude gain and phase shift versus frequency have been drawn. It is also possible to represent the frequency response graphically using only one plot. When this is done using the pair of polar coordinates ($|G(j\omega)|$, $\arg G(j\omega)$) and allowing the frequency ω to vary, the resulting Argand diagram is referred to as the **polar plot** or **frequency-response plot**. Such a graphical representation of the transfer function forms the basis of the **Nyquist approach** to the analysis of feedback systems. In fact, the main use of frequency-response methods in practice is in the analysis and design of closed-loop control systems. For the unity feedback system of Figure 5.30 the frequency-response plot of the forward-path transfer function $G(s)$ is used to infer overall closed-loop system behaviour. The Bode plots are perhaps the quickest plots to construct, especially when straight-line approximations are made, and are useful when attempting to estimate a transfer function from a set of physical frequency-response measurements. Other plots used in practice are the **Nichols diagram** and the **inverse Nyquist** (or **polar**) **plot**, the first of these being useful for designing feedforward compensators and the second for designing feedback compensators. Although there is no simple mathematical relationship, it is also worth noting that transient behaviour may also be inferred from the various frequency-response plots. For example, the reciprocal of the inverse M circle centred on the -1 point in the inverse Nyquist plot gives an indication of the peak overshoot in the transient behaviour (see, for example, G. Franklin, D. Powell and A. Naeini-Emami, *Feedback Control of Dynamic Systems*, seventh edn, Boston, MA, Pearson, 2015).



Investigation of such design tools may be carried out in MATLAB, incorporating Control Toolbox, using the command `rltool(G)`.

5.6 Engineering application: pole placement

In Chapter 1 we examined the behaviour of linear continuous-time systems modelled in the form of vector-matrix (or state-space) differential equations. In this chapter we have extended this, concentrating on the transform domain representation using the Laplace transform. In Chapter 6 we shall extend the approach to discrete-time systems using the z transform. So far we have concentrated on system *analysis*; that is, the question ‘Given the system, how does it behave?’ In this section we turn our attention briefly to consider the design or synthesis problem, and while it is not possible to produce an exhaustive treatment, it is intended to give the reader an appreciation of the role of mathematics in this task.

5.6.1 Poles and eigenvalues

By now the reader should be convinced that there is an association between system poles as deduced from the system transfer function and the eigenvalues of the system matrix in state-space form. Thus, for example, the system modelled by the second-order differential equation

$$\frac{d^2y}{dt^2} + \frac{1}{2}\frac{dy}{dt} - \frac{1}{2}y = u$$

has transfer function

$$G(s) = \frac{1}{s^2 + \frac{1}{2}s - \frac{1}{2}}$$

The system can also be represented in the state-space form

$$\dot{x} = Ax + bu, \quad y = c^T x \tag{5.68}$$

where

$$x = [x_1 \quad x_2]^T, \quad A = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix}, \quad b = [0 \quad 1]^T, \quad c = [1 \quad 0]^T$$

It is easy to check that the poles of the transfer function $G(s)$ are at $s = -1$ and $s = \frac{1}{2}$, and that these values are also the eigenvalues of the matrix A . Clearly this is an unstable system, with the pole or eigenvalue corresponding to $s = \frac{1}{2}$ located in the right half of the complex plane. In Section 5.6.2 we examine a method of moving this unstable pole to a new location, thus providing a method of overcoming the stability problem.

5.6.2 The pole placement or eigenvalue location technique

We now examine the possibility of introducing **state feedback** into the system. To do this, we use as system input

$$u = k^T x + u_{\text{ext}}$$

where $\mathbf{k} = [k_1 \ k_2]^T$ and u_{ext} is the external input. The state equation in (5.68) then becomes

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} [(k_1 x_1 + k_2 x_2) + u_{\text{ext}}]$$

That is,

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ k_1 + \frac{1}{2} & k_2 - \frac{1}{2} \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_{\text{ext}}$$

Calculating the characteristic equation of the new system matrix, we find that the eigenvalues are given by the roots of

$$\lambda^2 - (k_2 - \frac{1}{2})\lambda - (k_1 + \frac{1}{2}) = 0$$

Suppose that we not only wish to stabilize the system, but also wish to improve the response time. This could be achieved if both eigenvalues were located at (say) $\lambda = -5$, which would require the characteristic equation to be

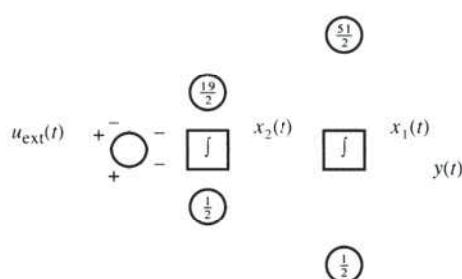
$$\lambda^2 + 10\lambda + 25 = 0$$

In order to make this pole relocation, we should choose

$$-(k_2 - \frac{1}{2}) = 10, \quad -(k_1 + \frac{1}{2}) = 25$$

indicating that we take $k_1 = -\frac{51}{2}$ and $k_2 = -\frac{19}{2}$. Figure 5.45 shows the original system and the additional state-feedback connections as dotted lines. We see that for this example at least, it is possible to locate the system poles or eigenvalues wherever we please in the complex plane, by a suitable choice of the vector \mathbf{k} . This corresponds to the choice of feedback gain, and in practical situations we are of course constrained by the need to specify reasonable values for these. Nevertheless, this technique, referred to as **pole placement**, is a powerful method for system control. There are some questions that remain. For example, can we apply the technique to all systems? Also, can it be extended to systems with more than one input? The following exercises will suggest answers to these questions, and help to prepare the reader for the study of specialist texts.

Figure 5.45 Feedback connections for eigenvalue location.



5.6.3 Exercises

51 An unstable system has Laplace transfer function

$$H(s) = \frac{1}{(s + \frac{1}{2})(s - 1)}$$

Make an appropriate choice of state variables to represent this system in the form

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u, \quad y = \mathbf{c}^T\mathbf{x}$$

where

$$\mathbf{x} = [x_1 \quad x_2]^T, \quad \mathbf{A} = \begin{bmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

$$\mathbf{b} = [0 \quad 1]^T, \quad \mathbf{c} = [1 \quad 0]^T$$

This particular form of the state-space model in which \mathbf{A} takes the companion form and \mathbf{b} has a single 1 in the last row is called the **control canonical form** of the system equations, and pole placement is particularly straightforward in this case.

Find a state-variable feedback control of the form $u = \mathbf{k}^T\mathbf{x}$ that will relocate both system poles at $s = -4$, thus stabilizing the system.

52 Find the control canonical form of the state-space equations for the system characterized by the transfer function

$$G(s) = \frac{2}{(s + 1)(s + \frac{1}{4})}$$

Calculate or (better) simulate the step response of the system, and find a control law that relocates both poles at $s = -5$. Calculate or simulate the step response of the new system. How do the two responses differ?

53 The technique for pole placement can be adapted to multi-input systems in certain cases. Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad y = \mathbf{c}^T\mathbf{x}$$

where

$$\mathbf{x} = [x_1 \quad x_2]^T, \quad \mathbf{u} = [u_1 \quad u_2]^T$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 6 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \quad \mathbf{c} = [1 \quad 0]^T$$

Writing $\mathbf{B}\mathbf{u} = \mathbf{b}_1u_1 + \mathbf{b}_2u_2$, where $\mathbf{b}_1 = [1 \quad 1]^T$ and $\mathbf{b}_2 = [0 \quad 1]^T$, enables us to work with each input separately. As a first step, use only the input u_1 to relocate both the system poles at $s = -5$. Secondly, use input u_2 only to achieve the same result. Note that we can use either or both inputs to obtain any pole locations we choose, subject of course to physical constraints on the size of the feedback gains.

54 The bad news is that it is not always possible to use the procedure described in Exercise 53. In the first place, it assumes that a full knowledge of the state vector $\mathbf{x}(t)$ is available. This may not always be the case; however, in many systems this problem can be overcome by the use of an **observer**. For details, a specialist text on control should be consulted.

There are also circumstances in which the system itself does not permit the use of the technique. Such systems are said to be **uncontrollable**, and the following example, which is more fully discussed in J. G. Reed, *Linear System Fundamentals* (Tokyo, McGraw-Hill, 1983), demonstrates the problem. Consider the system

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & -2 \\ 1 & -3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 2 \\ 1 \end{bmatrix}u$$

with

$$y = [0 \quad 1]\mathbf{x}$$

Find the system poles and attempt to relocate both of them, at, say, $s = -2$. It will be seen that no gain vector \mathbf{k} can be found to achieve this. Calculating the system transfer function gives a clue to the problem, but Exercise 55 shows how the problem could have been seen from the state-space form of the system.

55 In Exercise 46 it was stated that the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u$$

$$y = \mathbf{c}^T\mathbf{x}$$

where \mathbf{A} is an $n \times n$ matrix, is controllable provided that the Kalman matrix

$$\mathbf{M} = [\mathbf{b} \quad \mathbf{Ab} \quad \mathbf{A}^2\mathbf{b} \quad \dots \quad \mathbf{A}^{n-1}\mathbf{b}]$$

is of rank n . This condition must be satisfied if we are to be able to use the procedure for pole placement. Calculate the Kalman controllability matrix for the system in Exercise 54 and confirm that it has rank less than $n = 2$. Verify that the system of Exercise 53 satisfies the controllability condition.

- 56** We have noted that when the system equations are expressed in control canonical form, the calculations for pole relocation are particularly easy. The following technique shows how to transform controllable systems into this form. Given the system

$$\dot{x} = Ax + bu, \quad y = c^T x$$

calculate the Kalman controllability matrix M , defined in Exercise 55, and its inverse M^{-1} . Note that this will only exist for controllable systems. Set v^T as the last row of M^{-1} and form the transformation matrix

$$T = \begin{bmatrix} v^T \\ v^T A \\ \vdots \\ v^T A^{n-1} \end{bmatrix}$$

A transformation of state is now made by introducing the new state vector $z(t) = Tx(t)$, and the resulting system will be in control canonical form. To illustrate the technique, carry out the procedure for the system defined by

$$\dot{x} = \begin{bmatrix} 8 & -2 \\ 35 & -9 \end{bmatrix}x + \begin{bmatrix} 1 \\ 4 \end{bmatrix}u$$

and show that this leads to the system

$$\dot{z} = \begin{bmatrix} 0 & 1 \\ 2 & -1 \end{bmatrix}z + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u$$

Finally, check that the two system matrices have the same eigenvalues, and show that this will always be the case.

5.7 Review exercises (1–18)



Check your answers using MATLAB or MAPLE whenever possible.

- 1** Given that α is a positive constant, use the second shift theorem to

- (i) show that the Laplace transform of $\sin t H(t - \alpha)$ is

$$e^{-\alpha s} \frac{\cos \alpha + s \sin \alpha}{s^2 + 1}$$

- (ii) find the inverse transform of

$$\frac{s e^{-\alpha s}}{s^2 + 2s + 5}$$

- (b) Solve the differential equation

$$\frac{d^2y}{dt^2} + 2\frac{dy}{dt} + 5y = \sin t - \sin t H(t - \pi)$$

given that $y = dy/dt = 0$ when $t = 0$.

- 2** Show that the Laplace transform of the voltage $v(t)$, with period T , defined by

$$v(t) = \begin{cases} 1 & (0 \leq t < \frac{1}{2}T) \\ -1 & (\frac{1}{2}T \leq t < T) \end{cases} \quad v(t+T) = v(t)$$

is

$$V(s) = \frac{1}{s} \frac{1 - e^{-sT/2}}{1 + e^{-sT/2}}$$

This voltage is applied to a capacitor of $100 \mu\text{F}$ and a resistor of 250Ω in series, with no charge initially on the capacitor. Show that the Laplace transform $I(s)$ of the current $i(t)$ flowing, for $t \geq 0$, is

$$I(s) = \frac{1}{250(s+40)} \frac{1 - e^{-sT/2}}{1 + e^{-sT/2}}$$

and give an expression, involving Heaviside step functions, for $i(t)$ where $0 \leq t \leq 2T$. For $T = 10^{-3}$ s, is this a good representation of the steady-state response of the circuit? Briefly give a reason for your answer.

- 3 The response $x(t)$ of a control system to a forcing term $u(t)$ is given by the differential equation

$$\frac{d^2x}{dt^2} + 2\frac{dx}{dt} + 2x = u(t) \quad (t \geq 0)$$

Determine the impulse response of the system, and hence, using the convolution integral, obtain the response of the system to a unit step $u(t) = 1H(t)$ applied at $t = 0$, given that initially the system is in a quiescent state. Check your solution by directly solving the differential equation

$$\frac{d^2x}{dt^2} + 2\frac{dx}{dt} + 2x = 1 \quad (t \geq 0)$$

with $x = dx/dt = 0$ at $t = 0$.

- 4 A light horizontal beam, of length 5 m and constant flexural rigidity EI , built in at the left-hand end $x = 0$, is simply supported at the point $x = 4$ m and carries a distributed load with density function

$$W(x) = \begin{cases} 12 \text{ kNm}^{-1} & (0 < x < 4) \\ 24 \text{ kNm}^{-1} & (4 < x < 5) \end{cases}$$

Write down the fourth-order boundary-value problem satisfied by the deflection $y(x)$. Solve this problem to determine $y(x)$, and write down the resulting expressions for $y(x)$ for the cases $0 \leq x \leq 4$ and $4 \leq x \leq 5$. Calculate the end reaction and moment by evaluating appropriate derivatives of $y(x)$ at $x = 0$. Check that your results satisfy the equation of equilibrium for the beam as a whole.

- 5 (a) Sketch the function defined by

$$f(t) = \begin{cases} 0 & (0 \leq t < 1) \\ 1 & (1 \leq t < 2) \\ 0 & (t > 2) \end{cases}$$

Express $f(t)$ in terms of Heaviside step functions, and use the Laplace transform to solve the differential equation

$$\frac{dx}{dt} + x = f(t)$$

given that $x = 0$ at $t = 0$.

- (b) The Laplace transform $I(s)$ of the current $i(t)$ in a certain circuit is given by

$$I(s) = \frac{E}{s[Ls + R/(1 + Cs)]}$$

where E , L , R and C are positive constants. Determine (i) $\lim_{t \rightarrow 0} i(t)$ and (ii) $\lim_{t \rightarrow \infty} i(t)$.

- 6 Show that the Laplace transform of the half-rectified sine-wave function

$$v(t) = \begin{cases} \sin t & (0 \leq t \leq \pi) \\ 0 & (\pi \leq t \leq 2\pi) \end{cases}$$

of period 2π , is

$$\frac{1}{(1+s^2)(1-e^{-\pi s})}$$

Such a voltage $v(t)$ is applied to a 1Ω resistor and a 1H inductor connected in series. Show that the resulting current, initially zero, is $\sum_{n=0}^{\infty} f(t - n\pi)$, where $f(t) = (\sin t - \cos t + e^{-t})H(t)$. Sketch a graph of the function $f(t)$.

- 7 (a) Find the inverse Laplace transform of $1/s^2(s+1)^2$ by writing the expression in the form $(1/s^2)[1/(s+1)^2]$ and using the convolution theorem.
 (b) Use the convolution theorem to solve the integral equation

$$y(t) = t + 2 \int_0^t y(u) \cos(t-u) du$$

and the integro-differential equation

$$\int_0^t y''(u) y'(t-u) du = y(t)$$

where $y(0) = 0$ and $y'(0) = y_1$. Comment on the solution of the second equation.

- 8 A beam of negligible weight and length $3l$ carries a point load W at a distance l from the left-hand end. Both ends are clamped horizontally at the same level. Determine the equation governing the deflection of the beam. If, in addition, the beam is now subjected to a load per unit length, w , over the shorter part of the beam, what will then be the differential equation determining the deflection?

- 9 (a) Using Laplace transforms, solve the differential equation

$$\frac{d^2x}{dt^2} - 3\frac{dx}{dt} + 3x = H(t-a) \quad (a > 0)$$

where $H(t)$ is the Heaviside unit step function, given that $x = 0$ and $dx/dt = 0$ at $t = 0$.

- (b) The output $x(t)$ from a stable linear control system with input $\sin \omega t$ and transfer function $G(s)$ is determined by the relationship

$$X(s) = G(s)\mathcal{L}\{\sin \omega t\}$$

where $X(s) = \mathcal{L}\{x(t)\}$. Show that, after a long time t , the output approaches $x_s(t)$, where

$$x_s(t) = \operatorname{Re}\left(\frac{e^{j\omega t} G(j\omega)}{j}\right)$$

- 10 Consider the feedback system of Figure 5.46, where K is a constant feedback gain.

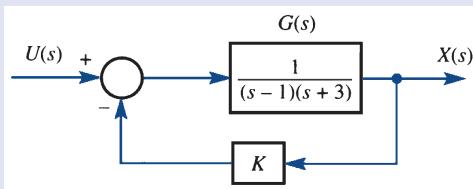


Figure 5.46 Feedback system of Review exercise 10.

- (a) In the absence of feedback (that is, $K=0$) is the system stable?
- (b) Write down the transfer function $G_1(s)$ for the overall feedback system.
- (c) Plot the locus of poles of $G_1(s)$ in the s plane for both positive and negative values of K .
- (d) From the plots in (c), specify for what range of values of K the feedback system is stable.
- (e) Confirm your answer to (d) using the Routh–Hurwitz criterion.

- 11 (a) For the feedback control system of Figure 5.47(a) it is known that the impulse response is $h(t) = 2e^{-2t} \sin t$. Use this to determine the value of the parameter α .
- (b) Consider the control system of Figure 5.47(b), having both proportional and rate feedback. Determine the critical value of the gain K for stability of the closed-loop system.

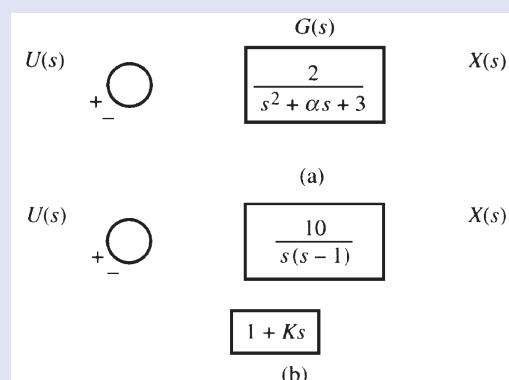


Figure 5.47 Feedback control systems of Review exercise 11.

12

A continuous-time system is specified in state-space form as

$$\begin{aligned}\dot{x}(t) &= \mathbf{A}x(t) + \mathbf{b}u(t) \\ y(t) &= \mathbf{c}^T x(t)\end{aligned}$$

where

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

- (a) Draw a block diagram to represent the system.
- (b) Using Laplace transforms, show that the state transition matrix is given by

$$e^{\mathbf{A}t} = \begin{bmatrix} 3e^{-2t} - 2e^{-3t} & 6e^{-2t} - 6e^{-3t} \\ e^{-3t} - e^{-2t} & 3e^{-3t} - 2e^{-2t} \end{bmatrix}$$

- (c) Calculate the impulse response of the system, and determine the response $y(t)$ of the system to an input $u(t) = 1$ ($t \geq 0$), subject to the initial state $x(0) = [1 \ 0]^T$.

13

A single-input–single-output system is represented in state-space form, using the usual notation, as

$$\begin{aligned}\dot{x}(t) &= \mathbf{A}x(t) + \mathbf{b}u(t) \\ y(t) &= \mathbf{c}^T x(t)\end{aligned}$$

For

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

show that

$$e^{\mathbf{A}t} = \begin{bmatrix} e^{-t}(\cos t - \sin t) & -e^{-t} \sin t \\ 2e^{-t} \sin t & e^{-t}(\cos t + \sin t) \end{bmatrix}$$

and find $x(t)$ given the $x(0) = 0$ and $u(t) = 1$ ($t \geq 0$).

Show that the Laplace transfer function of the system is

$$H(s) = \frac{Y(s)}{U(s)} = \mathbf{c}(sI - \mathbf{A})^{-1}\mathbf{b}$$

and find $H(s)$ for this system. What is the system impulse response?

14

A controllable linear plant that can be influenced by one input $u(t)$ is modelled by the differential equation

$$\dot{x}(t) = \mathbf{A}x(t) + \mathbf{b}u(t)$$

where $\mathbf{x}(t) = [x_1(t) \ x_2(t) \ \dots \ x_n(t)]^T$ is the state vector, \mathbf{A} is a constant matrix with distinct real eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ and $\mathbf{b} = [b_1 \ b_2 \ \dots \ b_n]^T$ is a constant vector.

By the application of the feedback control

$$u(t) = K\mathbf{v}_K^T \mathbf{x}(t)$$

where \mathbf{v}_K is the eigenvector of \mathbf{A}^T corresponding to the eigenvalue λ_K of \mathbf{A}^T (and hence of \mathbf{A}), the eigenvalue λ_K can be changed to a new real value p_K without altering the other eigenvalues. To achieve this, the feedback gain K is chosen as

$$K = \frac{p_K - \lambda_K}{p_K}$$

where $p_K = \mathbf{v}_K^T \mathbf{b}$.

Show that the system represented by

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 1 & 2 & 0 \\ 0 & -1 & 0 \\ -3 & -3 & -2 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u(t)$$

is controllable, and find the eigenvalues and corresponding eigenvectors of the system matrix. Deduce that the system is unstable in the absence of control, and determine a control law that will relocate the eigenvalue corresponding to the unstable mode at the new value -5 .

- 15** A second-order system is modelled by the differential equations

$$\dot{x}_1 + 2x_1 - 4x_2 = u$$

$$\dot{x}_2 - x_2 = u$$

coupled with the output equation

$$y = x_1$$

- (a) Express the model in state-space form.
- (b) Determine the transfer function of the system and show that the system is unstable.
- (c) Show that by using the feedback control law

$$u(t) = r(t) - ky(t)$$

where k is a scalar gain, the system will be stabilized provided $k > \frac{2}{3}$.

- (d) If $r(t) = H(t)$, a unit step function, and $k > \frac{2}{3}$, show that $y(t) \rightarrow 1$ as $t \rightarrow \infty$ if and only if $k = \frac{2}{3}$.

16

(An extended problem) The transient response of a practical control system to a unit step input often exhibits damped oscillations before reaching steady state. The following properties are some of those used to specify the transient response characteristics of an underdamped system:

rise time, the time required for the response to rise from 0 to 100% of its final value;

peak time, the time required for the response to reach the first peak of the overshoot;

settling time, the time required for the response curve to reach and stay within a range about the final value of size specified by an absolute percentage of the final value (usually 2% or 5%);

maximum overshoot, the maximum peak value of the response measured from unity.

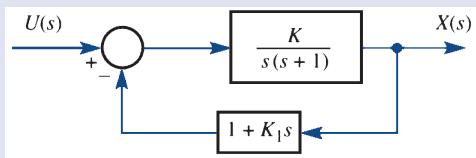


Figure 5.48 Feedback control system of Review exercise 16.

Consider the feedback control system of Figure 5.48 having both proportional and derivative feedback. It is desirable to choose the values of the gains K and K_1 so that the system unit step response has a maximum overshoot of 0.2 and a peak time of 1 s.

- (a) Obtain the overall transfer function of the closed-loop system.
- (b) Show that the unit step response of the system, assuming zero initial conditions, may be written in the form

$$x(t) = 1 - e^{-\omega_n \xi t} \left[\cos \omega_d t + \frac{\xi}{\sqrt{1-\xi^2}} \sin \omega_d t \right] \quad (t \geq 0)$$

where $\omega_d = \omega_n \sqrt{1-\xi^2}$, $\omega_n^2 = K$ and $2\omega_n \xi = 1 + KK_1$.

- (c) Determine the values of the gains K and K_1 so that the desired characteristics are achieved.
- (d) With these values of K and K_1 , determine the rise time and settling time, comparing both the 2% and 5% criteria for the latter.

- 17 (An extended problem) The mass M_1 of the mechanical system of Figure 5.49(a) is subjected to a harmonic forcing term $\sin \omega t$. Determine the steady-state response of the system.

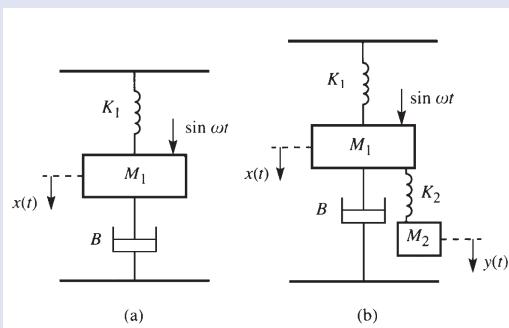


Figure 5.49 Vibration absorber of Review exercise 17.

It is desirable to design a vibration absorber to absorb the steady-state oscillations so that in the steady state $x(t) \equiv 0$. To achieve this, a secondary system is attached as illustrated in Figure 5.49(b).

- Show that, with an appropriate choice of M_2 and K_2 , the desired objective may be achieved.
- What is the corresponding steady-state motion of the mass M_2 ?
- Comment on the practicality of your design.

- 18 (An extended problem) The electronic amplifier of Figure 5.50 has open-loop transfer function $G(s)$ with the following characteristics: a low-frequency gain of 120 dB and simple poles at 1 MHz, 10 MHz and 25 MHz. It may be assumed that the amplifier is ideal, so that $K/(1 + K\beta) \approx 1/\beta$, where β is

the feedback gain and K the steady-state gain associated with $G(s)$.

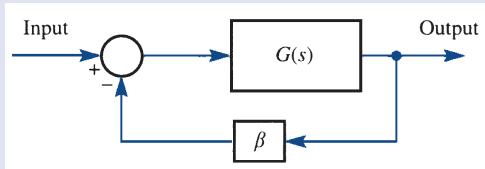


Figure 5.50 Electronic amplifier of Review exercise 18.

- Construct the magnitude versus log frequency and phase versus log frequency plots (Bode plots) for the open-loop system.
- Determine from the Bode plots whether or not the system is stable in the case of unity feedback (that is, $\beta = 1$).
- Determine the value of β for marginal stability, and hence the corresponding value of the closed-loop low-frequency gain.
- Feedback is now applied to the amplifier to reduce the overall closed-loop gain at low frequencies to 100 dB. Determine the gain and phase margin corresponding to this closed-loop configuration.
- Using the given characteristics, express $G(s)$ in the form

$$G(s) = \frac{K}{(1 + s\tau_1)(1 + s\tau_2)(1 + s\tau_3)}$$

and hence obtain the input-output transfer function for the amplifier.

- Write down the characteristic equation for the closed-loop system and, using the Routh-Hurwitz criterion, reconsider parts (b) and (c).



6 The z Transform

Chapter 6 Contents

6.1	Introduction	408
6.2	The z transform	409
6.3	Properties of the z transform	414
6.4	The inverse z transform	420
6.5	Discrete-time systems and difference equations	428
6.6	Discrete linear systems: characterization	435
6.7	The relationship between Laplace and z transforms	455
6.8	Solution of discrete-time state-space equations	456
6.9	Discretization of continuous-time state-space models	464
6.10	Engineering application: design of discrete-time systems	470
6.11	Engineering application: the delta operator and the \mathcal{D} transform	473
6.12	Review exercises (1–18)	480

6.1 Introduction

In this chapter we focus attention on discrete-(time) processes. With the advent of fast and cheap digital computers, there has been renewed emphasis on the analysis and design of digital systems, which represent a major class of engineering systems. The main thrust of this chapter will be in this direction. However, it is a mistake to believe that the mathematical basis of this area of work is of such recent vintage. The first comprehensive text in English dealing with difference equations was *The Treatise of the Calculus of Finite Differences* by George Boole and published in 1860. Much of the early impetus for the **finite calculus** was due to the need to carry out interpolation and to approximate derivatives and integrals. Later, numerical methods for the solution of differential equations were devised, many of which were based on **finite-difference methods**, involving the approximation of the derivative terms to produce a **difference equation**. The underlying idea in each case so far discussed is some form of approximation of an underlying continuous function or continuous-time process. There are situations, however, where it is more appropriate to propose a discrete-time model from the start.

Digital systems operate on digital signals, which are usually generated by **sampling** a continuous-time signal, that is a signal defined for every instant of a possibly infinite time interval. The sampling process generates a **discrete-time signal**, defined only at the instants when sampling takes place so that a digital sequence is generated. After processing by a computer, the output digital signal may be used to construct a new continuous-time signal, perhaps by the use of a **zero-order hold** device, and this in turn might be used to control a plant or process. Digital signal processing devices have made a major impact in many areas of engineering, as well as in the home. For example, compact disc players, which operate using digital technology, offered such a significant improvement in reproduction quality that the 1980s saw them rapidly take over from cassette tape players and vinyl record decks. DVD players have taken over from video players and digital radios are setting the standard for broadcasting. Both of these are based on digital technology.

We have seen in Chapter 5 that the Laplace transform was a valuable aid in the analysis of continuous-time systems, and in this chapter we develop the z transform, which will perform the same task for discrete-time systems. We introduce the transform in connection with the solution of difference equations, and later we show how difference equations arise as discrete-time system models.

The chapter includes two engineering applications. The first is on the design of digital filters, and highlights one of the major applications of transform methods as a design tool. It may be expected that whenever sampling is involved, performance will improve as sampling rate is increased. Engineers have found that this is not the full story, and the second application deals with some of the problems encountered. This leads on to an introduction to the unifying concept of the \mathcal{D} transform, which brings together the theories of the Laplace and z transforms.

6.2 The z transform

Since z transforms relate to sequences, we first review the notation associated with sequences, which were considered in more detail in Chapter 7 of *Modern Engineering Mathematics* (MEM). A finite sequence $\{x_k\}_0^n$ is an ordered set of $n + 1$ real or complex numbers:

$$\{x_k\}_0^n = \{x_0, x_1, x_2, \dots, x_n\}$$

Note that the set of numbers is ordered so that position in the sequence is important. The position is identified by the position index k , where k is an integer. If the number of elements in the set is infinite then this leads to the **infinite sequence**

$$\{x_k\}_0^\infty = \{x_0, x_1, x_2, \dots\}$$

When dealing with sampled functions of time t , it is necessary to have a means of allowing for $t < 0$. To do this, we allow the sequence of numbers to extend to infinity on both sides of the initial position x_0 , and write

$$\{x_k\}_{-\infty}^\infty = \{\dots, x_{-2}, x_{-1}, x_0, x_1, x_2, \dots\}$$

Sequences $\{x_k\}_{-\infty}^\infty$ for which $x_k = 0$ ($k < 0$) are called **causal sequences**, by analogy with continuous-time causal functions $f(t)H(t)$ defined in Section 11.2.1 of MEM as

$$f(t)H(t) = \begin{cases} 0 & (t < 0) \\ f(t) & (t \geq 0) \end{cases}$$

While for some finite sequences it is possible to specify the sequence by listing all the elements of the set, it is normally the case that a sequence is specified by giving a formula for its general element x_k .

6.2.1 Definition and notation

The **z transform** of a sequence $\{x_k\}_{-\infty}^\infty$ is defined in general as

$$\mathcal{Z}\{x_k\}_{-\infty}^\infty = X(z) = \sum_{k=-\infty}^{\infty} \frac{x_k}{z^k} \quad (6.1)$$

whenever the sum exists and where z is a complex variable, as yet undefined.

The process of taking the z transform of a sequence thus produces a function of a complex variable z , whose form depends upon the sequence itself. The symbol \mathcal{Z} denotes the **z -transform operator**; when it operates on a sequence $\{x_k\}$ it transforms the latter into the function $X(z)$ of the complex variable z . It is usual to refer to $\{x_k\}$, $X(z)$ as a **z -transform pair**, which is sometimes written as $\{x_k\} \leftrightarrow X(z)$. Note the similarity to obtaining the Laplace transform of a function in Section 11.2.1 of MEM. We shall return to consider the relationship between Laplace and z transforms in Section 6.7.

For sequences $\{x_k\}_{-\infty}^{\infty}$ that are *causal*, that is

$$x_k = 0 \quad (k < 0)$$

the z transform given in (6.1) reduces to

$$\mathcal{Z}\{x_k\}_0^{\infty} = X(z) = \sum_{k=0}^{\infty} \frac{x_k}{z^k} \quad (6.2)$$

In this chapter we shall be concerned with causal sequences, and so the definition given in (6.2) will be the one that we shall use henceforth. We shall therefore from now on take $\{x_k\}$ to denote $\{x_k\}_0^{\infty}$. Non-causal sequences, however, are of importance, and arise particularly in the field of digital image processing, among others.

Example 6.1

Determine the z transform of the sequence

$$\{x_k\} = \{2^k\} \quad (k \geq 0)$$

Solution

From the definition (6.2),

$$\mathcal{Z}\{2^k\} = \sum_{k=0}^{\infty} \frac{2^k}{z^k} = \sum_{k=0}^{\infty} \left(\frac{2}{z}\right)^k$$

which we recognize as a geometric series, with common ratio $r = 2/z$ between successive terms. The series thus converges for $|z| > 2$, when

$$\sum_{k=0}^{\infty} \left(\frac{2}{z}\right)^k = \lim_{k \rightarrow \infty} \frac{1 - (2/z)^k}{1 - 2/z} = \frac{1}{1 - 2/z}$$

leading to

$$\mathcal{Z}\{2^k\} = \frac{z}{z-2} \quad (|z| > 2) \quad (6.3)$$

so that

$$\left. \begin{aligned} \{x_k\} &= \{2^k\} \\ X(z) &= \frac{z}{z-2} \end{aligned} \right\}$$

is an example of a z -transform pair.

From Example 6.1, we see that the z transform of the sequence $\{2^k\}$ exists provided that we restrict the complex variable z so that it lies outside the circle $|z| = 2$ in the z plane. From another point of view, the function

$$X(z) = \frac{z}{z-2} \quad (|z| > 2)$$

may be thought of as a **generating function** for the sequence $\{2^k\}$, in the sense that the coefficient of z^{-k} in the expansion of $X(z)$ in powers of $1/z$ generates the k th term of the sequence $\{2^k\}$. This can easily be verified, since

$$\frac{z}{z-2} = \frac{1}{1-2/z} = \left(1 - \frac{2}{z}\right)^{-1}$$

and, since $|z| > 2$, we can expand this as

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

and we see that the coefficient of z^{-k} is indeed 2^k , as expected.

We can generalize the result (6.3) in an obvious way to determine $\mathcal{Z}\{a^k\}$, the z -transform of the sequence $\{a^k\}$, where a is a real or complex constant. At once

$$\mathcal{Z}\{a^k\} = \sum_{k=0}^{\infty} \frac{a^k}{z^k} = \frac{1}{1-a/z} \quad (|z| > |a|)$$

so that

$$\mathcal{Z}\{a^k\} = \frac{z}{z-a} \quad (|z| > |a|) \tag{6.4}$$

Example 6.2

Show that

$$\mathcal{Z}\left\{(-\frac{1}{2})^k\right\} = \frac{2z}{2z+1} \quad (|z| > \frac{1}{2})$$

Solution Taking $a = -\frac{1}{2}$ in (6.4), we have

$$\mathcal{Z}\left\{(-\frac{1}{2})^k\right\} = \sum_{k=0}^{\infty} \frac{(-\frac{1}{2})^k}{z^k} = \frac{z}{z - (-\frac{1}{2})} \quad (|z| > \frac{1}{2})$$

so that

$$\mathcal{Z}\left\{(-\frac{1}{2})^k\right\} = \frac{2z}{2z+1} \quad (|z| > \frac{1}{2})$$

Further z -transform pairs can be obtained from (6.4) by formally differentiating with respect to a , which for the moment we regard as a parameter. This gives

$$\frac{d}{da} \mathcal{Z}\{a^k\} = \mathcal{Z}\left\{\frac{da^k}{da}\right\} = \frac{d}{da} \left(\frac{z}{z-a} \right)$$

leading to

$$\mathcal{Z}\{ka^{k-1}\} = \frac{z}{(z-a)^2} \quad (|z| > |a|) \tag{6.5}$$

In the particular case $a = 1$ this gives

$$\mathcal{Z}\{k\} = \frac{z}{(z-1)^2} \quad (|z| > 1) \tag{6.6}$$

Example 6.3 Find the z transform of the sequence

$$\{2k\} = \{0, 2, 4, 6, 8, \dots\}$$

Solution From (6.6),

$$\mathcal{Z}\{k\} = \mathcal{Z}\{0, 1, 2, 3, \dots\} = \sum_{k=0}^{\infty} \frac{k}{z^k} = \frac{z}{(z-1)^2}$$

Using the definition (6.1),

$$\mathcal{Z}\{0, 2, 4, 6, 8, \dots\} = 0 + \frac{2}{z} + \frac{4}{z^2} + \frac{6}{z^3} + \frac{8}{z^4} + \dots = 2 \sum_{k=0}^{\infty} \frac{k}{z^k}$$

so that

$$\mathcal{Z}\{2k\} = 2\mathcal{Z}\{k\} = \frac{2z}{(z-1)^2} \quad (6.7)$$

Example 6.3 demonstrates the ‘linearity’ property of the z transform, which we shall consider further in Section 6.3.1.

A sequence of particular importance is the **unit pulse** or **impulse** sequence

$$\{\delta_k\} = \{1\} = \{1, 0, 0, \dots\}$$

It follows directly from the definition (6.4) that

$$\mathcal{Z}\{\delta_k\} = 1 \quad (6.8)$$



In MATLAB, using the Symbolic Math Toolbox, the z transform of the sequence $\{x_k\}$ is obtained by entering the commands

```
syms k z
ztrans(x_k)
```

As for Laplace transforms (see Section 11.2.2 of MEM), the answer may be simplified using the command `simple(ans)` and reformatted using the `pretty` command. Considering the sequence $\{x_k\} = \{2^k\}$ of Example 6.1, the commands

```
syms k z
ztrans(2^k)
return
ans=1/2*z/(1/2*z-1)
```

Entering the command

```
simple(ans)
returns
```

```
ans=z/(z-2)
```

z transforms can be performed in MAPLE using the `ztrans` function; so the commands:

```

ztrans(2^k,k,z);
simplify(%);

return

```

$$\frac{z}{z-2}$$

6.2.2 Sampling: a first introduction

Sequences are often generated in engineering applications through the sampling of continuous-time signals, described by functions $f(t)$ of a continuous-time variable t . Here we shall not discuss the means by which a signal is sampled, but merely suppose this to be possible in idealized form.

Figure 6.1 Sampling of a continuous-time signal.

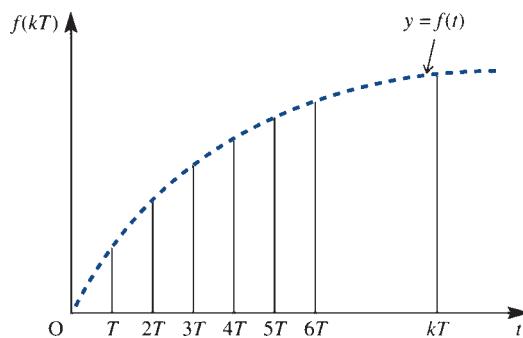


Figure 6.1 illustrates the idealized sampling process in which a continuous-time signal $f(t)$ is sampled instantaneously and perfectly at uniform intervals T , the **sampling interval**. The idealized sampling process generates the sequence

$$\{f(kT)\} = \{f(0), f(T), f(2T), \dots, f(nT), \dots\} \quad (6.9)$$

Using the definition (6.1), we can take the z transform of the sequence (6.9) to give

$$\mathcal{Z}\{f(kT)\} = \sum_{k=0}^{\infty} \frac{f(kT)}{z^k} \quad (6.10)$$

whenever the series converges. This idea is simply demonstrated by an example.

Example 6.4

The signal $f(t) = e^{-t}H(t)$ is sampled at intervals T . What is the z transform of the resulting sequence of samples?

Solution Sampling the causal function $f(t)$ generates the sequence

$$\begin{aligned} \{f(kT)\} &= \{f(0), f(T), f(2T), \dots, f(nT), \dots\} \\ &= \{1, e^{-T}, e^{-2T}, e^{-3T}, \dots, e^{-nT}, \dots\} \end{aligned}$$

Then, using (6.1),

$$\mathcal{Z}\{f(kT)\} = \sum_{k=0}^{\infty} \frac{e^{-kT}}{z^k} = \sum_{k=0}^{\infty} \left(\frac{e^{-T}}{z}\right)^k$$

so that

$$\mathcal{Z}\{e^{-kT}\} = \frac{z}{z - e^{-T}} \quad (|z| > e^{-T}) \quad (6.11)$$

It is important to note in Example 6.4 that the region of convergence depends on the sampling interval T .



In MATLAB the commands

```
syms k T z
ztrans(exp(-k*T));
pretty(simple(ans))
return
ans = z/(z-exp(-T))
```

which confirms (6.11).

In MAPLE the commands:

```
ztrans(exp(-k*T),k,z);
simplify(%);
return
ze^T
ze^T - 1
```

6.2.3 Exercises

- 1 Calculate the z transform of the following sequences, stating the region of convergence in each case:
- (a) $\{\left(\frac{1}{4}\right)^k\}$ (b) $\{3^k\}$ (c) $\{(-2)^k\}$
 (d) $\{-2^k\}$ (e) $\{3k\}$

- 2 The continuous-time signal $f(t) = e^{-2\omega t}$, where ω is a real constant, is sampled when $t \geq 0$ at intervals T . Write down the general term of the sequence of samples, and calculate the z transform of the sequence.

6.3

Properties of the z transform

In this section we establish the basic properties of the z transform that will enable us to develop further z -transform pairs, without having to compute them directly using the definition.

6.3.1 The linearity property

As for Laplace transforms, a fundamental property of the z transform is its linearity, which may be stated as follows.

If $\{x_k\}$ and $\{y_k\}$ are sequences having z transforms $X(z)$ and $Y(z)$ respectively and if α and β are any constants, real or complex, then

$$\mathcal{Z}\{\alpha x_k + \beta y_k\} = \alpha \mathcal{Z}\{x_k\} + \beta \mathcal{Z}\{y_k\} = \alpha X(z) + \beta Y(z) \quad (6.12)$$

As a consequence of this property, we say that the z -transform operator \mathcal{Z} is a **linear operator**. A proof of the property follows readily from the definition (6.4), since

$$\begin{aligned} \mathcal{Z}\{\alpha x_k + \beta y_k\} &= \sum_{k=0}^{\infty} \frac{\alpha x_k + \beta y_k}{z^k} = \alpha \sum_{k=0}^{\infty} \frac{x_k}{z^k} + \beta \sum_{k=0}^{\infty} \frac{y_k}{z^k} \\ &= \alpha X(z) + \beta Y(z) \end{aligned}$$

The region of existence of the z transform, in the z plane, of the linear sum will be the intersection of the regions of existence (that is, the region common to both) of the individual z transforms $X(z)$ and $Y(z)$.

Example 6.5

The continuous-time function $f(t) = \cos \omega t H(t)$, ω a constant, is sampled in the idealized sense at intervals T to generate the sequence $\{\cos k\omega T\}$. Determine the z transform of the sequence.

Solution Using the result $\cos k\omega T = \frac{1}{2}(e^{jk\omega T} + e^{-jk\omega T})$ and the linearity property, we have

$$\mathcal{Z}\{\cos k\omega T\} = \mathcal{Z}\left\{\frac{1}{2}e^{jk\omega T} + \frac{1}{2}e^{-jk\omega T}\right\} = \frac{1}{2}\mathcal{Z}\{e^{jk\omega T}\} + \frac{1}{2}\mathcal{Z}\{e^{-jk\omega T}\}$$

Using (6.7) and noting that $|e^{jk\omega T}| = |e^{-jk\omega T}| = 1$ gives

$$\begin{aligned} \mathcal{Z}\{\cos k\omega T\} &= \frac{1}{2} \frac{z}{z - e^{j\omega T}} + \frac{1}{2} \frac{z}{z - e^{-j\omega T}} \quad (|z| > 1) \\ &= \frac{1}{2} \frac{z(z - e^{-j\omega T}) + z(z - e^{j\omega T})}{z^2 - (e^{j\omega T} + e^{-j\omega T})z + 1} \end{aligned}$$

leading to the z -transform pair

$$\mathcal{Z}\{\cos k\omega T\} = \frac{z(z - \cos \omega T)}{z^2 - 2z \cos \omega T + 1} \quad (|z| > 1) \quad (6.13)$$

In a similar manner to Example 6.5, we can verify the z -transform pair

$$\mathcal{Z}\{\sin k\omega T\} = \frac{z \sin \omega T}{z^2 - 2z \cos \omega T + 1} \quad (|z| > 1) \quad (6.14)$$

and this is left as an exercise for the reader (see Exercise 3).



Check that in MATLAB the commands

```
syms k z w T
ztrans(cos(k*w*T));
pretty(simple(ans))
```

return the transform given in (6.13) and that the MAPLE commands:

```
ztrans(cos(k*w*T),k,z);
simplify(%);
```

do likewise.

6.3.2 The first shift property (delaying)

In this and the next section we introduce two properties relating the z transform of a sequence to the z transform of a shifted version of the same sequence. In this section we consider a delayed version of the sequence $\{x_k\}$, denoted by $\{y_k\}$, with

$$y_k = x_{k-k_0}$$

Here k_0 is the number of steps in the delay; for example, if $k_0 = 2$ then $y_k = x_{k-2}$, so that

$$y_0 = x_{-2}, \quad y_1 = x_{-1}, \quad y_2 = x_0, \quad y_3 = x_1$$

and so on. Thus the sequence $\{y_k\}$ is simply the sequence $\{x_k\}$ moved backward, or delayed, by two steps. From the definition (6.1),

$$\mathcal{Z}\{y_k\} = \sum_{k=0}^{\infty} \frac{y_k}{z^k} = \sum_{k=0}^{\infty} \frac{x_{k-k_0}}{z^k} = \sum_{p=-k_0}^{\infty} \frac{x_p}{z^{p+k_0}}$$

where we have written $p = k - k_0$. If $\{x_k\}$ is a causal sequence, so that $x_p = 0$ ($p < 0$), then

$$\mathcal{Z}\{y_k\} = \sum_{p=0}^{\infty} \frac{x_p}{z^{p+k_0}} = \frac{1}{z^{k_0}} \sum_{p=0}^{\infty} \frac{x_p}{z^p} = \frac{1}{z^{k_0}} X(z)$$

where $X(z)$ is the z transform of $\{x_k\}$.

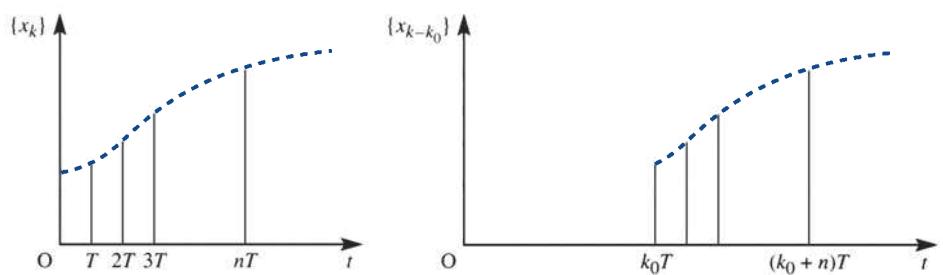
We therefore have the result

$$\mathcal{Z}\{x_{k-k_0}\} = \frac{1}{z^{k_0}} \mathcal{Z}\{x_k\} \tag{6.15}$$

which is referred to as the **first shift property** of z transforms.

If $\{x_k\}$ represents the sampled form, with uniform sampling interval T , of the continuous signal $x(t)$ then $\{x_{k-k_0}\}$ represents the sampled form of the continuous signal $x(t - k_0T)$ which, as illustrated in Figure 6.2, is the signal $x(t)$ delayed by a multiple k_0 of the sampling interval T . The reader will find it of interest to compare this result with the results for the Laplace transforms of integrals [see Section 11.3.2 of MEM, in particular (11.15)].

Figure 6.2
Sequence and its shifted form.



Example 6.6

The causal sequence $\{x_k\}$ is generated by

$$x_k = \left(\frac{1}{2}\right)^k \quad (k \geq 0)$$

Determine the z transform of the shifted sequence $\{x_{k-2}\}$.

Solution By the first shift property,

$$\mathcal{Z}\{x_{k-2}\} = \frac{1}{z^2} \mathcal{Z}\left\{\left(\frac{1}{2}\right)^k\right\}$$

which, on using (6.4), gives

$$\mathcal{Z}\{x_{k-2}\} = \frac{1}{z^2} \frac{z}{z - \frac{1}{2}} \quad (|z| > \frac{1}{2}) = \frac{1}{z^2} \frac{2z}{2z - 1} = \frac{2}{z(2z - 1)} \quad (|z| > \frac{1}{2})$$

We can confirm this result by direct use of the definition (6.1). From this, and the fact that $\{x_k\}$ is a causal sequence,

$$\{x_{k-2}\} = \{x_{-2}, x_{-1}, x_0, x_1, \dots\} = \{0, 0, 1, \frac{1}{2}, \frac{1}{4}, \dots\}$$

Thus,

$$\begin{aligned} \mathcal{Z}\{x_{k-2}\} &= 0 + 0 + \frac{1}{z^2} + \frac{1}{2z^3} + \frac{1}{4z^4} + \dots = \frac{1}{z^2} \left(1 + \frac{1}{2z} + \frac{1}{4z^2} + \dots\right) \\ &= \frac{1}{z^2} \frac{z}{z - \frac{1}{2}} \quad (|z| > \frac{1}{2}) = \frac{z}{z(2z - 1)} \quad (|z| > \frac{1}{2}) \end{aligned}$$

6.3.3 The second shift property (advancing)

In this section we seek a relationship between the z transform of an advanced version of a sequence and that of the original sequence. First we consider a single-step advance. If $\{y_k\}$ is the single-step advanced version of the sequence $\{x_k\}$ then $\{y_k\}$ is generated by

$$y_k = x_{k+1} \quad (k \geq 0)$$

Then

$$\mathcal{Z}\{y_k\} = \sum_{k=0}^{\infty} \frac{y_k}{z^k} = \sum_{k=0}^{\infty} \frac{x_{k+1}}{z^k} = z \sum_{k=0}^{\infty} \frac{x_{k+1}}{z^{k+1}}$$

and putting $p = k + 1$ gives

$$\mathcal{Z}\{y_k\} = z \sum_{p=1}^{\infty} \frac{x_p}{z^p} = z \left(\sum_{p=0}^{\infty} \frac{x_p}{z^p} - x_0 \right) = zX(z) - zx_0$$

where $X(z)$ is the z transform of $\{x_k\}$.

We therefore have the result

$$\mathcal{Z}\{x_{k+1}\} = zX(z) - zx_0 \quad (6.16)$$

In a similar manner it is readily shown that for a two-step advanced sequence $\{x_{k+2}\}$

$$\mathcal{Z}\{x_{k+2}\} = z^2X(z) - z^2x_0 - zx_1 \quad (6.17)$$

Note the similarity in structure between (6.16) and (6.17) on the one hand and those for the Laplace transforms of first and second derivatives (Section 11.3.1 of MEM). In general, it is readily proved by induction that for a k_0 -step advanced sequence $\{x_{k+k_0}\}$

$$\mathcal{Z}\{x_{k+k_0}\} = z^{k_0}X(z) - \sum_{n=0}^{k_0-1} x_n z^{k_0-n} \quad (6.18)$$

In Section 6.5.2 we shall use these results to solve difference equations.

6.3.4 Some further properties

In this section we shall state some further useful properties of the z transform, leaving their verification to the reader as Exercises 9 and 10.

(i) Multiplication by a^k

If $Z\{x_k\} = X(z)$ then for a constant a

$$\mathcal{Z}\{a^k x_k\} = X(a^{-1}z) \quad (6.19)$$

(ii) Multiplication by k^n

If $\mathcal{Z}\{x_k\} = X(z)$ then for a positive integer n

$$\mathcal{Z}\{k^n x_k\} = \left(-z \frac{d}{dz}\right)^n X(z) \quad (6.20)$$

Note that in (6.20) the operator $-z \frac{d}{dz}$ means ‘first differentiate with respect to z and then multiply by $-z$ ’. Raising to the power of n means ‘repeat the operation n times’.

(iii) Initial-value theorem

If $\{x_k\}$ is a sequence with z transform $X(z)$ then the initial-value theorem states that

$$\lim_{z \rightarrow \infty} X(z) = x_0 \quad (6.21)$$

(iv) Final-value theorem

If $\{x_k\}$ is a sequence with z transform $X(z)$ then the final-value theorem states that

$$\lim_{k \rightarrow \infty} x_k = \lim_{z \rightarrow 1} (1 - z^{-1})X(z) \quad (6.22)$$

provided that the poles of $(1 - z^{-1})X(z)$ are inside the unit circle.

6.3.5 Table of z transforms

It is appropriate at this stage to draw together the results proved so far for easy access. This is done in the form of a table in Figure 6.3.

Figure 6.3 A short table of z transforms.

$\{x_k\}$ ($k \geq 0$)	$\mathcal{Z}\{x_k\}$	Region of existence
$x_k = \begin{cases} 1 & (k = 0) \\ 0 & (k > 0) \end{cases}$ (unit pulse sequence)	1	All z
$x_k = 1$ (unit step sequence)	$\frac{z}{z - 1}$	$ z > 1$
$x_k = a^k$ (a constant)	$\frac{z}{z - a}$	$ z > a $
$x_k = k$	$\frac{z}{(z - 1)^2}$	$ z > 1$
$x_k = ka^{k-1}$ (a constant)	$\frac{z}{(z - a)^2}$	$ z > a$
$x_k = e^{-kT}$ (T constant)	$\frac{z}{z - e^{-T}}$	$ z > e^{-T}$
$x_k = \cos k\omega T$ (ωT constants)	$\frac{z(z - \cos \omega T)}{z^2 - 2z \cos \omega T + 1}$	$ z > 1$
$x_k = \sin k\omega T$ (ωT constants)	$\frac{z \sin \omega T}{z^2 - 2z \cos \omega T + 1}$	$ z > 1$

6.3.6 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 3 Use the method of Example 6.5 to confirm (6.14), namely

$$\mathcal{Z}\{\sin k\omega T\} = \frac{z \sin \omega T}{z^2 - 2z \cos \omega T + 1}$$

where ω and T are constants.

- 4 Use the first shift property to calculate the z transform of the sequence $\{y_k\}$, with

$$y_k = \begin{cases} 0 & (k < 3) \\ x_{k-3} & (k \geq 3) \end{cases}$$

where $\{x_k\}$ is causal and $x_k = (\frac{1}{2})^k$. Confirm your result by direct evaluation of $\mathcal{Z}\{y_k\}$ using the definition of the z transform.

- 5 Determine the z transforms of the sequences
 (a) $\{(-\frac{1}{5})^k\}$ (b) $\{\cos k\pi\}$

- 6 Determine $\mathcal{Z}\{(\frac{1}{2})^k\}$. Using (6.6), obtain the z transform of the sequence $\{k(\frac{1}{2})^k\}$.

- 7 Show that for a constant α

$$(a) \mathcal{Z}\{\sinh k\alpha\} = \frac{z \sinh \alpha}{z^2 - 2z \cosh \alpha + 1}$$

$$(b) \mathcal{Z}\{\cosh k\alpha\} = \frac{z^2 - z \cosh \alpha}{z^2 - 2z \cosh \alpha + 1}$$

- 8 Sequences are generated by sampling a causal continuous-time signal $u(t)$ ($t \geq 0$) at uniform intervals T . Write down an expression for u_k , the general term of the sequence, and calculate the corresponding z transform when $u(t)$ is

$$(a) e^{-4t} \quad (b) \sin t \quad (c) \cos 2t$$

- 9 Prove the initial- and final-value theorems given in (6.21) and (6.22).

- 10 Prove the multiplication properties given in (6.19) and (6.20).

6.4

The inverse z transform

In this section we consider the problem of recovering a causal sequence $\{x_k\}$ from knowledge of its z transform $X(z)$. As we shall see, the work on the inversion of Laplace transforms in Section 11.2.7 of MEM will prove a valuable asset for this task.

Formally the symbol $\mathcal{Z}^{-1}[X(z)]$ denotes a causal sequence $\{x_k\}$ whose z transform is $X(z)$; that is,

$$\text{if } \mathcal{Z}\{x_k\} = X(z) \text{ then } \{x_k\} = \mathcal{Z}^{-1}[X(z)]$$

This correspondence between $X(z)$ and $\{x_k\}$ is called the **inverse z transformation**, $\{x_k\}$ being the **inverse transform** of $X(z)$, and \mathcal{Z}^{-1} being referred to as the **inverse z -transform operator**.

As for the Laplace transforms in Section 11.2.8 of MEM, the most obvious way of finding the inverse transform of $X(z)$ is to make use of a table of transforms such as that given in Figure 6.3. Sometimes it is possible to write down the inverse transform directly from the table, but more often than not it is first necessary to carry out some algebraic manipulation on $X(z)$. In particular, we frequently need to determine the inverse transform of a rational expression of the form $P(z)/Q(z)$, where $P(z)$ and $Q(z)$ are polynomials in z . In such cases the procedure, as for Laplace transforms, is first to resolve the expression, or a revised form of the expression, into partial fractions and then to use the table of transforms. We shall now illustrate the approach through some examples.

6.4.1 Inverse techniques

Example 6.7

Find

$$\mathcal{Z}^{-1}\left[\frac{z}{z-2}\right]$$

Solution From Figure 6.3, we see that $z/(z-2)$ is a special case of the transform $z/(z-a)$, with $a=2$. Thus

$$\mathcal{Z}^{-1}\left[\frac{z}{z-2}\right] = \{2^k\}$$

Example 6.8

Find

$$\mathcal{Z}^{-1}\left[\frac{z}{(z-1)(z-2)}\right]$$

Solution Guided by our work on Laplace transforms, we might attempt to resolve

$$Y(z) = \frac{z}{(z-1)(z-2)}$$

into partial fractions. This approach does produce the correct result, as we shall show later. However, we notice that most of the entries in Figure 6.3 contain a factor z in the numerator of the transform. We therefore resolve

$$\frac{Y(z)}{z} = \frac{1}{(z-1)(z-2)}$$

into partial fractions, as

$$\frac{Y(z)}{z} = \frac{1}{z-2} - \frac{1}{z-1}$$

so that

$$Y(z) = \frac{z}{z-2} - \frac{z}{z-1}$$

Then using the result $\mathcal{Z}^{-1}[z/(z-a)] = \{a^k\}$ together with the linearity property, we have

$$\begin{aligned} \mathcal{Z}^{-1}[Y(z)] &= \mathcal{Z}^{-1}\left(\frac{z}{z-2} - \frac{z}{z-1}\right) = \mathcal{Z}^{-1}\left(\frac{z}{z-2}\right) - \mathcal{Z}^{-1}\left(\frac{z}{z-1}\right) \\ &= \{2^k\} - \{1^k\} \quad (k \geq 0) \end{aligned}$$

so that

$$\mathcal{Z}^{-1}\left[\frac{z}{(z-1)(z-2)}\right] = \{2^k - 1\} \quad (k \geq 0) \tag{6.23}$$

Suppose that in Example 6.8 we had not thought so far ahead and we had simply resolved $Y(z)$, rather than $Y(z)/z$, into partial fractions. Would the result be the same? The answer of course is ‘yes’, as we shall now show. Resolving

$$Y(z) = \frac{z}{(z-1)(z-2)}$$

into partial fractions gives

$$Y(z) = \frac{2}{z-2} - \frac{1}{z-1}$$

which may be written as

$$Y(z) = \frac{1}{z} \frac{2z}{z-2} - \frac{1}{z} \frac{z}{z-1}$$

Since

$$\mathcal{Z}^{-1}\left[\frac{2z}{z-2}\right] = 2\mathcal{Z}^{-1}\left(\frac{z}{z-2}\right) = 2\{2^k\}$$

it follows from the first shift property (6.15) that

$$\mathcal{Z}^{-1}\left[\frac{1}{z} \frac{2z}{z-2}\right] = \begin{cases} \{2 \cdot 2^{k-1}\} & (k > 0) \\ 0 & (k = 0) \end{cases}$$

Similarly,

$$\mathcal{Z}^{-1}\left[\frac{1}{z} \frac{z}{z-1}\right] = \begin{cases} \{1^{k-1}\} = \{1\} & (k > 0) \\ 0 & (k = 0) \end{cases}$$

Combining these last two results, we have

$$\begin{aligned} \mathcal{Z}^{-1}[Y(z)] &= \mathcal{Z}^{-1}\left[\frac{1}{z} \frac{2z}{z-2}\right] - \mathcal{Z}^{-1}\left[\frac{1}{z} \frac{z}{z-1}\right] \\ &= \begin{cases} \{2^k - 1\} & (k > 0) \\ 0 & (k = 0) \end{cases} \end{aligned}$$

which, as expected, is in agreement with the answer obtained in Example 6.8.

We can see that adopting this latter approach, while producing the correct result, involved extra effort in the use of a shift theorem. When possible, we avoid this by ‘extracting’ the factor z as in Example 6.8, but of course this is not always possible, and recourse may be made to the shift property, as Example 6.9 illustrates.



The inverse z transform $\{x_k\}$ of $X(z)$ is returned in MATLAB using the command
`iztrans(X(z), k)`

(Note: The command `iztrans(X(z))` by itself returns the inverse transform expressed in terms of n rather than k .)

For the z transform in Example 6.8 the MATLAB command

`iztrans(z / ((z-1) * (z-2)), k)`

returns

$$\text{ans} = -1 + 2^k$$

as required.

The inverse z transform can be performed in MAPLE using the `invztrans` function, so that the command

$$\text{invztrans}(z/(z^2 - 3*z + 2), z, k);$$

also returns the answer

$$2^k - 1$$

Example 6.9

Find

$$\mathcal{Z}^{-1}\left[\frac{2z+1}{(z+1)(z-3)}\right]$$

Solution In this case there is no factor z available in the numerator, and so we must resolve

$$Y(z) = \frac{2z+1}{(z+1)(z-3)}$$

into partial fractions, giving

$$Y(z) = \frac{1}{4} \frac{1}{z+1} + \frac{7}{4} \frac{1}{z-3} = \frac{1}{4} \frac{1}{z} \frac{z}{z+1} + \frac{7}{4} \frac{1}{z} \frac{z}{z-3}$$

Since

$$\mathcal{Z}^{-1}\left[\frac{z}{z+1}\right] = \{(-1)^k\} \quad (k \geq 0)$$

$$\mathcal{Z}^{-1}\left[\frac{z}{z-3}\right] = \{3^k\} \quad (k \geq 0)$$

it follows from the first shift property (6.15) that

$$\mathcal{Z}^{-1}\left[\frac{1}{z} \frac{z}{z+1}\right] = \begin{cases} \{(-1)^{k-1}\} & (k > 0) \\ 0 & (k = 0) \end{cases}$$

$$\mathcal{Z}^{-1}\left[\frac{1}{z} \frac{z}{z-3}\right] = \begin{cases} 3^{k-1} & (k > 0) \\ 0 & (k = 0) \end{cases}$$

Then, from the linearity property,

$$\mathcal{Z}^{-1}[Y(z)] = \frac{1}{4} \mathcal{Z}^{-1}\left[\frac{1}{z} \frac{z}{z+1}\right] + \frac{7}{4} \mathcal{Z}^{-1}\left[\frac{1}{z} \frac{z}{z-3}\right]$$

giving

$$\mathcal{Z}^{-1}\left[\frac{2z+1}{(z+1)(z-3)}\right] = \begin{cases} \left\{\frac{1}{4}(-1)^{k-1} + \frac{7}{4}3^{k-1}\right\} & (k > 0) \\ 0 & (k = 0) \end{cases}$$



In MATLAB the command

```
iztrans((2*z+1)/((z+1)*(z-3)), k)
```

returns

```
ans=-1/3*charfcn[0](k)-1/4*(-1)^k+7/12*3^k
```

(Note: The charfcn function is the characteristic function of the set A , and is defined to be

$$\text{charfcn}[A](k) = \begin{cases} 1 & \text{if } k \text{ is in } A \\ 0 & \text{if } k \text{ is not in } A \end{cases}$$

Thus $\text{charfcn}[0](k) = 1$ if $k = 0$ and 0 otherwise.)

It is left as an exercise to confirm that the answer provided using MATLAB concurs with the calculated answer.

It is often the case that the rational function $P(z)/Q(z)$ to be inverted has a quadratic term in the denominator. Unfortunately, in this case there is nothing resembling the first shift theorem of the Laplace transform which, as we saw in Section 11.2.9 of MEM, proved so useful in similar circumstances. Looking at Figure 6.3, the only two transforms with quadratic terms in the denominator are those associated with the sequences $\{\cos k\omega T\}$ and $\{\sin k\omega T\}$. In practice these prove difficult to apply in the inverse form, and a ‘first principles’ approach is more appropriate. We illustrate this with two examples, demonstrating that all that is really required is the ability to handle complex numbers at the stage of resolution into partial fractions.

Example 6.10

Invert the z transform

$$Y(z) = \frac{z}{z^2 + a^2}$$

where a is a real constant.

Solution

In view of the factor z in the numerator, we resolve $Y(z)/z$ into partial fractions, giving

$$\frac{Y(z)}{z} = \frac{1}{z^2 + a^2} = \frac{1}{(z + ja)(z - ja)} = \frac{1}{j2a} \frac{1}{(z - ja)} - \frac{1}{j2a} \frac{1}{(z + ja)}$$

That is

$$Y(z) = \frac{1}{j2a} \left(\frac{z}{z - ja} - \frac{z}{z + ja} \right)$$

Using the result $\mathcal{Z}^{-1}[z/(z - a)] = \{a^k\}$, we have

$$\mathcal{Z}^{-1}\left[\frac{z}{z - ja}\right] = \{(ja)^k\} = \{j^k a^k\}$$

$$\mathcal{Z}^{-1}\left[\frac{z}{z + ja}\right] = \{(-ja)^k\} = \{(-j)^k a^k\}$$

From the relation $e^{j\theta} = \cos \theta + j \sin \theta$, we have

$$j = e^{j\pi/2}, \quad -j = e^{-j\pi/2}$$

so that

$$\mathcal{Z}^{-1}\left[\frac{z}{z - ja}\right] = \{a^k(e^{j\pi/2})^k\} = \{a^k e^{jk\pi/2}\} = \{a^k(\cos \frac{1}{2}k\pi + j \sin \frac{1}{2}k\pi)\}$$

$$\mathcal{Z}^{-1}\left[\frac{z}{z + ja}\right] = \{a^k(\cos \frac{1}{2}k\pi - j \sin \frac{1}{2}k\pi)\}$$

The linearity property then gives

$$\begin{aligned} \mathcal{Z}^{-1}[Y(z)] &= \left\{ \frac{a^k}{j2a} (\cos \frac{1}{2}k\pi + j \sin \frac{1}{2}k\pi - \cos \frac{1}{2}k\pi + j \sin \frac{1}{2}k\pi) \right\} \\ &= \{a^{k-1} \sin \frac{1}{2}k\pi\} \end{aligned}$$



Whilst MATLAB or MAPLE may be used to obtain the inverse z transform when complex partial fractions are involved, it is difficult to convert results into a simple form, the difficult step being that of expressing complex exponentials in terms of trigonometric functions.

Example 6.11

Invert

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

Solution

The denominator of the transform may be factorized as

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

In exponential form we have $\frac{1}{2} \pm j\frac{1}{2}\sqrt{3} = e^{\pm j\pi/3}$, so the denominator may be written as

$$z^2 - z + 1 = (z - e^{j\pi/3})(z - e^{-j\pi/3})$$

We then have

$$\frac{Y(z)}{z} = \frac{1}{(z - e^{j\pi/3})(z - e^{-j\pi/3})}$$

which can be resolved into partial fractions as

$$\frac{Y(z)}{z} = \frac{1}{e^{j\pi/3} - e^{-j\pi/3}} \frac{1}{z - e^{j\pi/3}} + \frac{1}{e^{-j\pi/3} - e^{j\pi/3}} \frac{1}{z - e^{-j\pi/3}}$$

Noting that $\sin \theta = (e^{j\theta} - e^{-j\theta})/j2$, this reduces to

$$\begin{aligned}\frac{Y(z)}{z} &= \frac{1}{j2 \sin \frac{1}{3}\pi} \frac{z}{z - e^{j\pi/3}} - \frac{1}{j2 \sin \frac{1}{3}\pi} \frac{z}{z - e^{-j\pi/3}} \\ &= \frac{1}{j\sqrt{3}} \frac{z}{z - e^{j\pi/3}} - \frac{1}{j\sqrt{3}} \frac{z}{z - e^{-j\pi/3}}\end{aligned}$$

Using the result $\mathcal{Z}^{-1}[z/(z-a)] = \{a^k\}$, this gives

$$\mathcal{Z}^{-1}[Y(z)] = \frac{1}{j\sqrt{3}} (e^{jk\pi/3} - e^{-jk\pi/3}) = \{2\sqrt{\frac{1}{3}} \sin \frac{1}{3}k\pi\}$$

We conclude this section with two further examples, illustrating the inversion technique applied to frequently occurring transform types.

Example 6.12

Find the sequence whose z transform is

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

Solution $F(z)$ is unlike any z transform treated so far in the examples. However, it is readily expanded in a power series in z^{-1} as

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

Using (6.4), it is then apparent that

$$\mathcal{Z}^{-1}[F(z)] = \{f_k\} = \{1, 2, 0, 1, 0, 0, \dots\}$$



The MATLAB command

```
iztrans((z^3+2*z^2+1)/z^3,k)
```

returns

```
charfcn[0](k)+2*charfcn[1](k)+charfcn[3](k)
```

which corresponds to the sequence

```
{1, 2, 0, 1, 0, 0, ...}
```

Example 6.13

Find $\mathcal{Z}^{-1}[G(z)]$ where

$$G(z) = \frac{z(1 - e^{-aT})}{(z - 1)(z - e^{-aT})}$$

where a and T are positive constants.

Solution Resolving into partial fractions,

$$\frac{G(z)}{z} = \frac{1}{z - 1} - \frac{1}{z - e^{-aT}}$$

giving

$$G(z) = \frac{1}{z-1} - \frac{1}{z-e^{-aT}}$$

Using the result $\mathcal{Z}^{-1}[z/(z-a)] = \{a^k\}$, we have

$$\mathcal{Z}^{-1}[G(z)] = \{(1 - e^{-akT})\} \quad (k \geq 0)$$

In this particular example $G(z)$ is the z transform of a sequence derived by sampling the continuous-time signal

$$f(t) = 1 - e^{-at}$$

at intervals T .



The MATLAB commands

```
syms k z a T
iztrans((z*(1-exp(-a*T))) / ((z-1)*(z-exp(-a*T))), k);
pretty(simple(ans))
return
ans=1-exp(-aT)^k
```

In MAPLE the command

```
invztrans((z*(1-exp(-a*T))) / ((z-1)*(z-exp(-a*T))), z, k);
returns
```

$$-\left(\frac{1}{e^{aT}}\right)^k + 1$$

6.4.2 Exercises



Confirm your answers using MATLAB or MAPLE whenever possible.

11

Invert the following z transforms. Give the general term of the sequence in each case.

(a) $\frac{z}{z-1}$

(b) $\frac{z}{z+1}$

(c) $\frac{z}{z-\frac{1}{2}}$

(d) $\frac{z}{3z+1}$

(e) $\frac{z}{z-j}$

(f) $\frac{z}{z+j\sqrt{2}}$

(g) $\frac{1}{z-1}$

(h) $\frac{z+2}{z+1}$

(f) $\frac{z}{z^2-2\sqrt{3}z+4}$

(g) $\frac{2z^2-7z}{(z-1)^2(z-3)}$

(h) $\frac{z^2}{(z-1)^2(z^2-z+1)}$

13

Find $\mathcal{Z}^{-1}[Y(z)]$ when $Y(z)$ is given by

(a) $\frac{1}{z} + \frac{2}{z^7}$

(b) $1 + \frac{3}{z^2} - \frac{2}{z^9}$

(c) $\frac{3z+z^2+5z^5}{z^5}$

(d) $\frac{1+z}{z^3} + \frac{3z}{3z+1}$

(e) $\frac{2z^3+6z^2+5z+1}{z^2(2z+1)}$

(f) $\frac{2z^2-7z+7}{(z-1)^2(z-2)}$

(g) $\frac{z-3}{z^2-3z+2}$

12

By first resolving $Y(z)/z$ into partial fractions, find

$\mathcal{Z}^{-1}[Y(z)]$ when $Y(z)$ is given by

(a) $\frac{z}{(z-1)(z+2)}$

(b) $\frac{z}{(2z+1)(z-3)}$

(c) $\frac{z^2}{(2z+1)(z-1)}$

(d) $\frac{2z}{2z^2+z-1}$

(e) $\frac{z}{z^2+1}$ [Hint: $z^2+1 = (z+j)(z-j)$]

6.5

Discrete-time systems and difference equations

In Chapter 11 of MEM and Chapter 5 the Laplace transform technique was examined, first as a method for solving differential equations, then as a way of characterizing a continuous-time system. In fact, much could be deduced concerning the behaviour of the system and its properties by examining its transform-domain representation, without looking for specific time-domain responses at all. In this section we shall discuss the idea of a linear discrete-time system and its model, a **difference equation**. Later we shall see that the z transform plays an analogous role to the Laplace transform for such systems, by providing a transform-domain representation of the system.

6.5.1 Difference equations

First we illustrate the motivation for studying difference equations by means of an example.

Suppose that a sequence of observations $\{x_k\}$ is being recorded and we receive observation x_k at (time) step or index k . We might attempt to process (for example, smooth or filter) this sequence of observations $\{x_k\}$ using the discrete-time feedback system illustrated in Figure 6.4. At time step k the observation x_k enters the system as an input, and, after combination with the ‘feedback’ signal at the summing junction S, proceeds to the block labelled D. This block is a unit delay block, and its function is to hold its input signal until the ‘clock’ advances one step, to step $k + 1$. At this time the input signal is passed without alteration to become the signal y_{k+1} , the $(k + 1)$ th member of the output sequence $\{y_k\}$. At the same time this signal is fed back through a scaling block of amplitude α to the summing junction S. This process is instantaneous, and at S the feedback signal is subtracted from the next input observation x_{k+1} to provide the next input to the delay block D. The process then repeats at each ‘clock’ step.

To analyse the system, let $\{r_k\}$ denote the sequence of input signals to D; then, owing to the delay action of D, we have

$$y_{k+1} = r_k$$

Also, owing to the feedback action,

$$r_k = x_k - \alpha y_k$$

where α is the feedback gain. Combining the two expressions gives

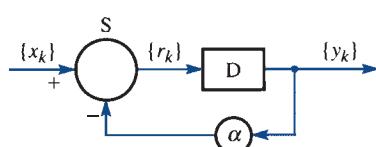
$$y_{k+1} = x_k - \alpha y_k$$

or

$$y_{k+1} + \alpha y_k = x_k \quad (6.24)$$

Equation (6.24) is an example of a first-order difference equation, and it relates adjacent members of the sequence $\{y_k\}$ to each other and to the input sequence $\{x_k\}$.

Figure 6.4 Discrete-time signal processing system.

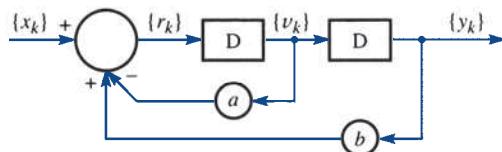


A solution of the difference equation (6.24) is a formula for y_k , the general term of the output sequence $\{y_k\}$, and this will depend on both k and the input sequence $\{x_k\}$ as well as, in this case, the feedback gain α .

Example 6.14

Find a difference equation to represent the system shown in Figure 6.5, having input and output sequences $\{x_k\}$ and $\{y_k\}$ respectively, where D is the unit delay block and a and b are constant feedback gains.

Figure 6.5 The system for Example 6.14.



Solution Introducing intermediate signal sequences $\{r_k\}$ and $\{v_k\}$ as shown in Figure 6.5, at each step the outputs of the delay blocks are

$$y_{k+1} = v_k \quad (6.25)$$

$$v_{k+1} = r_k \quad (6.26)$$

and at the summing junction

$$r_k = x_k - av_k + by_k \quad (6.27)$$

From (6.25),

$$y_{k+2} = v_{k+1}$$

which on using (6.26) gives

$$y_{k+2} = r_k$$

Substituting for r_k from (6.27) then gives

$$y_{k+2} = x_k - av_k + by_k$$

which on using (6.25) becomes

$$y_{k+2} = x_k - ay_{k+1} + by_k$$

Rearranging this gives

$$y_{k+2} + ay_{k+1} - by_k = x_k \quad (6.28)$$

as the difference equation representing the system.

The difference equation (6.28) is an example of a second-order linear constant-coefficient difference equation, and there are strong similarities between this and a second-order linear constant-coefficient differential equation. It is of second order because the term involving the greatest shift of the $\{y_k\}$ sequence is the term in y_{k+2} , implying a shift of two steps. As demonstrated by Example 6.14, the degree of shift, or the order of the equation, is closely related to the number of delay blocks in the block diagram.

6.5.2 The solution of difference equations

Difference equations arise in a variety of ways, sometimes from the direct modelling of systems in discrete time or as an approximation to a differential equation describing the behaviour of a system modelled as a continuous-time system. We do not discuss this further here; rather we restrict ourselves to the technique of solution but examples of applications will be apparent from the exercises. The z -transform method is based upon the second shift property (Section 6.3.3), and it will quickly emerge as a technique almost identical to the Laplace transform method for ordinary differential equations introduced in Section 11.3.3 of MEM. We shall introduce the method by means of an example.

Example 6.15

If in Example 6.14, $a = 1$, $b = 2$ and the input sequence $\{x_k\}$ is the unit step sequence $\{1\}$, solve the resulting difference equation (6.28).

Solution

Substituting for a , b and $\{x_k\}$ in (6.28) leads to the difference equation

$$y_{k+2} + y_{k+1} - 2y_k = 1 \quad (k \geq 0) \quad (6.29)$$

Taking z transforms throughout in (6.29) gives

$$\mathcal{Z}\{y_{k+2} + y_{k+1} - 2y_k\} = \mathcal{Z}\{1, 1, 1, \dots\}$$

which, on using the linearity property and the result $\mathcal{Z}\{1\} = z/(z - 1)$, may be written as

$$\mathcal{Z}\{y_{k+2}\} + \mathcal{Z}\{y_{k+1}\} - 2\mathcal{Z}\{y_k\} = \frac{z}{z - 1}$$

Using (6.16) and (6.17) then gives

$$[z^2Y(z) - z^2y_0 - zy_1] + [zY(z) - zy_0] - 2Y(z) = \frac{z}{z - 1}$$

which on rearranging leads to

$$(z^2 + z - 2)Y(z) = \frac{z}{z - 1} + z^2y_0 + z(y_1 + y_0) \quad (6.30)$$

To proceed, we need some further information, namely the first and second terms y_0 and y_1 of the solution sequence $\{y_k\}$. Without this additional information, we cannot find a unique solution. As we saw in Section 11.3.3 of MEM, this compares with the use of the Laplace transform method to solve second-order differential equations, where the values of the solution and its first derivative at time $t = 0$ are required.

Suppose that we know (or are given) that

$$y_0 = 0, \quad y_1 = 1$$

Then (6.30) becomes

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

or

$$(z + 2)(z - 1)Y(z) = z + \frac{z}{z - 1}$$

and solving for $Y(z)$ gives

$$Y(z) = \frac{z}{(z+2)(z-1)} + \frac{z}{(z+2)(z-1)^2} = \frac{z^2}{(z+2)(z-1)^2} \quad (6.31)$$

To obtain the solution sequence $\{y_k\}$, we must take the inverse transform in (6.31). Proceeding as in Section 6.4, we resolve $Y(z)/z$ into partial fractions as

$$\frac{Y(z)}{z} = \frac{z}{(z+2)(z-1)^2} = \frac{\frac{1}{3}}{(z-1)^2} + \frac{\frac{2}{9}}{z-1} - \frac{\frac{2}{9}}{z+2}$$

and so

$$Y(z) = \frac{\frac{1}{3}}{(z-1)^2} + \frac{\frac{2}{9}}{z-1} - \frac{\frac{2}{9}}{z+2}$$

Using the results $\mathcal{Z}^{-1}[z(z-a)] = \{a^k\}$ and $\mathcal{Z}^{-1}[z/(z-1)^2] = \{k\}$ from Figure 6.3, we obtain

$$\{y_k\} = \left\{ \frac{1}{3}k + \frac{2}{9} - \frac{2}{9}(-2)^k \right\} \quad (k \geq 0)$$

as the solution sequence for the difference equation satisfying the conditions $y_0 = 0$ and $y_1 = 1$.

The method adopted in Example 6.15 is called the ***z*-transform method for solving linear constant-coefficient difference equations**, and is analogous to the Laplace transform method for solving linear constant-coefficient differential equations.

To conclude this section, two further examples are given to help consolidate understanding of the method.



Such difference equations can be solved directly in MAPLE using the `rsolve` command. In the current version of the Symbolic Math Toolbox in MATLAB there appears to be no equivalent command for directly solving a difference equation. However, as we saw in Section 5.2.5, using the `maple` command in MATLAB lets us access MAPLE commands directly. Hence, for the difference equation in Example 6.15, using the command

```
maple('rsolve({y(k+2)+y(k+1)-2*y(k)=1,y(0)=0,y(1)=1},y(k))')
```

in MATLAB returns the calculated answer

```
-2/9 * (-2)^(k+2) / 9 + 1/3 * k
```

In MAPLE difference equations can be solved directly using `rsolve`, so that the command

```
rsolve({y(k+2)+y(k+1)-2*y(k)=1,y(0)=0,y(1)=1},y(k));
```

returns

$$\frac{2}{9} - \frac{2(-2)^k}{9} + \frac{k}{3}$$

Example 6.16

Solve the difference equation

$$8y_{k+2} - 6y_{k+1} + y_k = 9 \quad (k \geq 0)$$

given that $y_0 = 1$ and $y_1 = \frac{3}{2}$.

Solution Taking z transforms

$$8\mathcal{Z}\{y_{k+2}\} - 6\mathcal{Z}\{y_{k+1}\} + \mathcal{Z}\{y_k\} = 9\mathcal{Z}\{1\}$$

Using (6.16) and (6.17) and the result $\mathcal{Z}\{1\} = z/(z - 1)$ gives

$$8[z^2Y(z) - z^2y_0 - zy_1] - 6[zY(z) - zy_0] + Y(z) = \frac{9z}{z-1}$$

which on rearranging leads to

$$(8z^2 - 6z + 1)Y(z) = 8z^2y_0 + 8zy_1 - 6zy_0 + \frac{9z}{z-1}$$

We are given that $y_0 = 1$ and $y_1 = \frac{3}{2}$, so

$$(8z^2 - 6z + 1)Y(z) = 8z^2 + 6z + \frac{9z}{z-1}$$

or

$$\begin{aligned} \frac{Y(z)}{z} &= \frac{8z+6}{(4z-1)(2z-1)} + \frac{9}{(4z-1)(2z-1)(z-1)} \\ &= \frac{z+\frac{3}{4}}{(z-\frac{1}{4})(z-\frac{1}{2})} + \frac{\frac{9}{8}}{(z-\frac{1}{4})(z-\frac{1}{2})(z-1)} \end{aligned}$$

Resolving into partial fractions gives

$$\begin{aligned} \frac{Y(z)}{z} &= \frac{5}{z-\frac{1}{2}} - \frac{4}{z-\frac{1}{4}} + \frac{6}{z-\frac{1}{4}} - \frac{9}{z-\frac{1}{2}} + \frac{3}{z-1} \\ &= \frac{2}{z-\frac{1}{4}} - \frac{4}{z-\frac{1}{2}} + \frac{3}{z-1} \end{aligned}$$

and so

$$Y(z) = \frac{2z}{z-\frac{1}{4}} - \frac{4z}{z-\frac{1}{2}} + \frac{3z}{z-1}$$

Using the result $\mathcal{Z}^{-1}\{z/(z - a)\} = \{a^k\}$ from Figure 6.3, we take inverse transforms, to obtain

$$\{y_k\} = \{2(\frac{1}{4})^k - 4(\frac{1}{2})^k + 3\} \quad (k \geq 0)$$

as the required solution.



Check that in MATLAB the command

```
maple('rsolve({8*y(k+2)-6*y(k+1)+y(k)=9,y(0)=1,
y(1)=3/2},y(k))')
```

returns the calculated answer or alternatively use the command `rsolve` in MAPLE.

Example 6.17

Solve the difference equation

$$y_{k+2} + 2y_k = 0 \quad (k \geq 0)$$

given that $y_0 = 1$ and $y_1 = \sqrt{2}$.

Solution

Taking z transforms, we have

$$[z^2 Y(z) - z^2 y_0 - z y_1] + 2Y(z) = 0$$

and substituting the given values of y_0 and y_1 gives

$$z^2 Y(z) - z^2 - \sqrt{2}z + 2Y(z) = 0$$

or

$$(z^2 + 2)Y(z) = z^2 + \sqrt{2}z$$

Resolving $Y(z)/z$ into partial fractions gives

$$\frac{Y(z)}{z} = \frac{z + \sqrt{2}}{z^2 + 2} = \frac{z + \sqrt{2}}{(z + j\sqrt{2})(z - j\sqrt{2})}$$

Following the approach adopted in Example 6.13, we write

$$j\sqrt{2} = \sqrt{2} e^{j\pi/2}, \quad -j\sqrt{2} = \sqrt{2} e^{-j\pi/2}$$

$$\frac{Y(z)}{z} = \frac{z + \sqrt{2}}{(z - \sqrt{2} e^{j\pi/2})(z - \sqrt{2} e^{-j\pi/2})} = \frac{(1+j)/j2}{z - \sqrt{2} e^{j\pi/2}} - \frac{(1-j)/j2}{z - \sqrt{2} e^{-j\pi/2}}$$

Thus

$$Y(z) = \frac{1}{j2} \left[(1+j) \frac{z}{z - \sqrt{2} e^{j\pi/2}} - (1-j) \frac{z}{z - \sqrt{2} e^{-j\pi/2}} \right]$$

which on taking inverse transforms gives

$$\begin{aligned} \{y_k\} &= \left\{ \frac{2^{k/2}}{j2} \left[(1+j) e^{jk\pi/2} - (1-j) e^{-jk\pi/2} \right] \right\} \\ &= \{2^{k/2} (\cos \frac{1}{2}k\pi + \sin \frac{1}{2}k\pi)\} \quad (k \geq 0) \end{aligned}$$

as the required solution.

The solution in Example 6.17 was found to be a real-valued sequence, and this comes as no surprise because the given difference equation and the ‘starting’ values y_0 and y_1 involved only real numbers. This observation provides a useful check on the algebra when complex partial fractions are involved.



If complex partial fractions are involved then, as was mentioned at the end of Example 6.10, it is difficult to simplify answers when determining inverse z transforms using MATLAB. When such partial fractions arise in the solution of difference equations use of the command `evalc` alongside `rsolve` in MAPLE attempts to express complex exponentials in terms of trigonometric functions, leading in most cases to simplified answers.

Considering the difference equation of Example 6.17, using the command

```
maple('rsolve({y(k+2)+2*y(k)=0,y(0)=1,y(1)
=2^(1/2)},y(k))')
```

in MATLAB returns the answer

$$(1/2+1/2\pi i)*(-i*2^{1/2})^k + (1/2-1/2\pi i)*(i*2^{1/2})^k$$

whilst using the command

```
maple('evalc(rsolve({y(k+2)+2*y(k)=0,y(0)=1,y(1)
=2^(1/2)},y(k)))')
```

returns the answer

$$\exp(1/2\log(2)k)\cos(1/2k\pi) + \exp(1/2\log(2)k)\sin(1/2k\pi)$$

Noting that $e^{\log 2} = 2$ it is readily seen that this corresponds to the calculated answer

$$2^{k/2}(\cos\frac{1}{2}k\pi + \sin\frac{1}{2}k\pi)$$

6.5.3 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

14

Find difference equations representing the discrete-time systems shown in Figure 6.6.

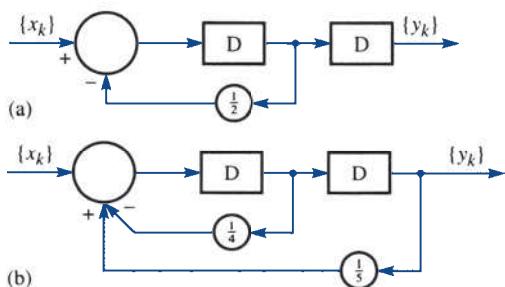


Figure 6.6 The systems for Exercise 14.

15

Using z -transform methods, solve the following difference equations:

(a) $y_{k+2} - 2y_{k+1} + y_k = 0$ subject to $y_0 = 0, y_1 = 1$

(b) $y_{n+2} - 8y_{n+1} - 9y_n = 0$ subject to $y_0 = 2, y_1 = 1$

(c) $y_{k+2} + 4y_k = 0$ subject to $y_0 = 0, y_1 = 1$

(d) $2y_{k+2} - 5y_{k+1} - 3y_k = 0$ subject to $y_0 = 3, y_1 = 2$

16

Using z -transform methods, solve the following difference equations:

(a) $6y_{k+2} + y_{k+1} - y_k = 3$ subject to $y_0 = y_1 = 0$

(b) $y_{k+2} - 5y_{k+1} + 6y_k = 5$ subject to $y_0 = 0, y_1 = 1$

(c) $y_{n+2} - 5y_{n+1} + 6y_n = (\frac{1}{2})^n$ subject to $y_0 = y_1 = 0$

(d) $y_{n+2} - 3y_{n+1} + 3y_n = 1$ subject to $y_0 = 1, y_1 = 0$

(e) $2y_{n+2} - 3y_{n+1} - 2y_n = 6n + 1$ subject to $y_0 = 1, y_1 = 2$

(f) $y_{n+2} - 4y_n = 3n - 5$ subject to $y_0 = y_1 = 0$

- 17 A person's capital at the beginning of, and expenditure during, a given year k are denoted by C_k and E_k respectively, and satisfy the difference equations

$$C_{k+1} = 1.5C_k - E_k$$

$$E_{k+1} = 0.21C_k + 0.5E_k$$

- (a) Show that eventually the person's capital grows at 20% per annum.
- (b) If the capital at the beginning of year 1 is £6000 and the expenditure during year 1 is £3720 then find the year in which the expenditure is a minimum and the capital at the beginning of that year.

- 18 The dynamics of a discrete-time system are determined by the difference equation

$$y_{k+2} - 5y_{k+1} + 6y_k = u_k$$

Determine the response of the system to the unit step input

$$u_k = \begin{cases} 0 & (k < 0) \\ 1 & (k \geq 0) \end{cases}$$

given that $y_0 = y_1 = 1$.

- 19 As a first attempt to model the national economy, it is assumed that the national income I_k at year k is given by

$$I_k = C_k + P_k + G_k$$

where C_k is the consumer expenditure, P_k is private investment and G_k is government expenditure. It is also assumed that the consumer spending is proportional to the national income in the previous year, so that

$$C_k = aI_{k-1} \quad (0 < a < 1)$$

It is further assumed that private investment is proportional to the change in consumer spending over the previous year, so that

$$P_k = b(C_k - C_{k-1}) \quad (0 < b \leq 1)$$

Show that under these assumptions the national income I_k is determined by the difference equation

$$I_{k+2} - a(1+b)I_{k+1} + abI_k = G_{k+2}$$

If $a = \frac{1}{2}$, $b = 1$, government spending is at a constant level (that is, $G_k = G$ for all k) and $I_0 = 2G$, $I_1 = 3G$, show that

$$I_k = 2[1 + (\frac{1}{2})^{k/2} \sin \frac{1}{4} k\pi]G$$

Discuss what happens as $k \rightarrow \infty$.

The difference equation for current in a particular ladder network of N loops is

$$R_1 i_{n+1} + R_2(i_{n+1} - i_n) + R_2(i_{n+1} - i_{n+2}) = 0$$

$$(0 \leq n \leq N-2)$$

where i_n is the current in the $(n+1)$ th loop, and R_1 and R_2 are constant resistors.

- (a) Show that this may be written as

$$i_{n+2} - 2 \cosh \alpha i_{n+1} + i_n = 0 \quad (0 \leq n \leq N-2)$$

where

$$\alpha = \cosh^{-1} \left(1 + \frac{R_1}{2R_2} \right)$$

- (b) By solving the equation in (a), show that

$$i_n = \frac{i_1 \sinh n\alpha - i_0 \sinh(n-1)\alpha}{\sinh \alpha} \quad (2 \leq n \leq N)$$

6.6 Discrete linear systems: characterization

In this section we examine the concept of a discrete-time linear system and its difference equation model. Ideas developed in Chapter 5 for continuous-time system modelling will be seen to carry over to discrete-time systems, and we shall see that the z transform is the key to the understanding of such systems.

6.6.1 z transfer functions

In Section 5.3, when considering continuous-time linear systems modelled by differential equations, we introduced the concept of the system (Laplace) transfer function. This is a powerful tool in the description of such systems, since it contains all the information

on system stability and also provides a method of calculating the response to an arbitrary input signal using a convolution integral. In the same way, we can identify a z transfer function for a discrete-time linear time-invariant system modelled by a difference equation, and we can arrive at results analogous to those of Chapter 5 and Chapter 11 of MEM.

Let us consider the general linear constant-coefficient difference equation model for a linear time-invariant system, with input sequence $\{u_k\}$ and output sequence $\{y_k\}$. Both $\{u_k\}$ and $\{y_k\}$ are causal sequences throughout. Such a difference equation model takes the form

$$\begin{aligned} & a_n y_{k+n} + a_{n-1} y_{k+n-1} + a_{n-2} y_{k+n-2} + \cdots + a_0 y_k \\ & = b_m u_{k+m} + b_{m-1} u_{k+m-1} + b_{m-2} u_{k+m-2} + \cdots + b_0 u_k \end{aligned} \quad (6.32)$$

where $k \geq 0$ and n, m (with $n \geq m$) are positive integers and the a_i and b_j are constants. The difference equation (6.32) differs in one respect from the examples considered in Section 6.5 in that the possibility of delayed terms in the input sequence $\{u_k\}$ is also allowed for. The order of the difference equation is n if $a_n \neq 0$, and for the system to be physically realizable, $n \geq m$.

Assuming the system to be initially in a quiescent state, we take z transforms throughout in (6.32) to give

$$(a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0) Y(z) = (b_m z^m + b_{m-1} z^{m-1} + \cdots + b_0) U(z)$$

where $Y(z) = \mathcal{Z}\{y_k\}$ and $U(z) = \mathcal{Z}\{u_k\}$. The **system discrete or z transfer function** $G(z)$ is defined as

$$G(z) = \frac{Y(z)}{U(z)} = \frac{b_m z^m + b_{m-1} z^{m-1} + \cdots + b_0}{a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0} \quad (6.33)$$

and is normally rearranged (by dividing numerator and denominator by a_n) so that the coefficient of z^n in the denominator is 1. In deriving $G(z)$ in this form, we have assumed that the system was initially in a quiescent state. This assumption is certainly valid for the system (6.32) if

$$y_0 = y_1 = \cdots = y_{n-1} = 0$$

$$u_0 = u_1 = \cdots = u_{m-1} = 0$$

This is not the end of the story, however, and we shall use the term ‘quiescent’ to mean that no non-zero values are stored on the delay elements before the initial time.

On writing

$$P(z) = b_m z^m + b_{m-1} z^{m-1} + \cdots + b_0$$

$$Q(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_0$$

the discrete transfer function may be expressed as

$$G(z) = \frac{P(z)}{Q(z)}$$

As for the continuous model in Section 5.3.1, the equation $Q(z) = 0$ is called the **characteristic equation** of the discrete system, its order, n , determines the **order of the system**, and its roots are referred to as the **poles** of the discrete transfer function. Likewise, the roots of $P(z) = 0$ are referred to as the **zeros** of the discrete transfer function.

Example 6.18

Draw a block diagram to represent the system modelled by the difference equation

$$y_{k+2} + 3y_{k+1} - y_k = u_k \quad (6.34)$$

and find the corresponding z transfer function.

Solution

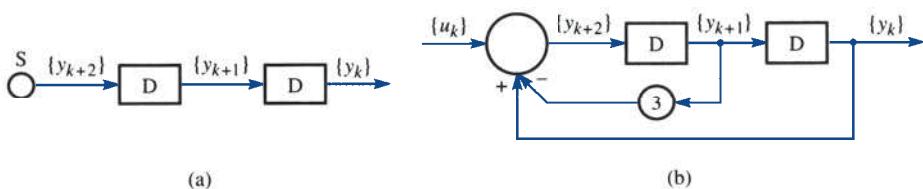
The difference equation may be thought of as a relationship between adjacent members of the solution sequence $\{y_k\}$. Thus at each time step k we have from (6.34)

$$y_{k+2} = -3y_{k+1} + y_k + u_k \quad (6.35)$$

which provides a formula for y_{k+2} involving y_k , y_{k+1} and the input u_k . The structure shown in Figure 6.7(a) illustrates the generation of the sequence $\{y_k\}$ from $\{y_{k+2}\}$ using two delay blocks.

Figure 6.7

(a) The basic second-order block diagram substructure; and
(b) block diagram representation of (6.34) of Example 6.18.



We now use (6.35) as a prescription for generating the sequence $\{y_{k+2}\}$ and arrange for the correct combination of signals to be formed at each step k at the input summing junction S of Figure 6.7(a). This leads to the structure shown in Figure 6.7(b), which is the required block diagram.

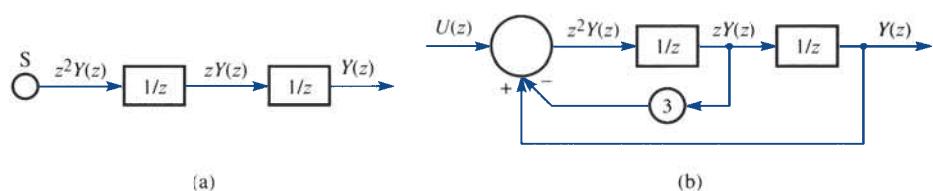
We can of course produce a block diagram in the z -transform domain, using a similar process. Taking the z transform throughout in (6.34), under the assumption of a quiescent initial state, we obtain

$$z^2Y(z) + 3zY(z) - Y(z) = U(z) \quad (6.36)$$

or

$$z^2Y(z) = -3zY(z) + Y(z) + U(z) \quad (6.37)$$

The representation (6.37) is the transform-domain version of (6.35), and the z -transform domain basic structure corresponding to the time-domain structure of Figure 6.7(a) is shown in Figure 6.8(a).

Figure 6.8 (a) The z -transform domain basic second-order block diagram substructure; and (b) the z -transform domain block diagram representation of (6.34) of Example 6.18.

The unit delay blocks, labelled D in Figure 6.7(a), become ' $1/z$ ' elements in the z -transform domain diagram, in line with the first shift property (6.15), where a number k_0 of delay steps involves multiplication by z^{-k_0} .

It is now a simple matter to construct the 'signal' transform $z^2Y(z)$ from (6.37) and arrange for it to be available at the input to the summing junction S in Figure 6.8(a). The resulting block diagram is shown in Figure 6.8(b).

The z transfer function follows at once from (6.36) as

$$G(z) = \frac{Y(z)}{U(z)} = \frac{1}{z^2 + 3z - 1} \quad (6.38)$$

Example 6.19

A system is specified by its z transfer function

$$G(z) = \frac{z - 1}{z^2 + 3z + 2}$$

What is the order n of the system? Can it be implemented using only n delay elements? Illustrate this.

Solution If $\{u_k\}$ and $\{y_k\}$ denote respectively the input and output sequences to the system then

$$G(z) = \frac{Y(z)}{U(z)} = \frac{z - 1}{z^2 + 3z + 2}$$

so that

$$(z^2 + 3z + 2)Y(z) = (z - 1)U(z)$$

Taking inverse transforms, we obtain the corresponding difference equation model assuming the system is initially in a quiescent state

$$y_{k+2} + 3y_{k+1} + 2y_k = u_{k+1} - u_k \quad (6.39)$$

The difference equation (6.39) has a more complex right-hand side than the difference equation (6.34) considered in Example 6.18. This results from the existence of z terms in the numerator of the transfer function. By definition, the order of the difference equation (6.39) is still 2. However, realization of the system with two delay blocks is not immediately apparent, although this can be achieved, as we shall now illustrate.

Introduce a new signal sequence $\{r_k\}$ such that

$$(z^2 + 3z + 2)R(z) = U(z) \quad (6.40)$$

where $R(z) = \mathcal{Z}\{r_k\}$. In other words, $\{r_k\}$ is the output of the system having transfer function $1/(z^2 + 3z + 2)$.

Multiplying both sides of (6.40) by z , we obtain

$$z(z^2 + 3z + 2)R(z) = zU(z)$$

or

$$(z^2 + 3z + 2)zR(z) = zU(z) \quad (6.41)$$

Subtracting (6.40) from (6.41) we have

$$(z^2 + 3z + 2)zR(z) - (z^2 + 3z + 2)R(z) = zU(z) - U(z)$$

giving

$$(z^2 + 3z + 2)[zR(z) - R(z)] = (z - 1)U(z)$$

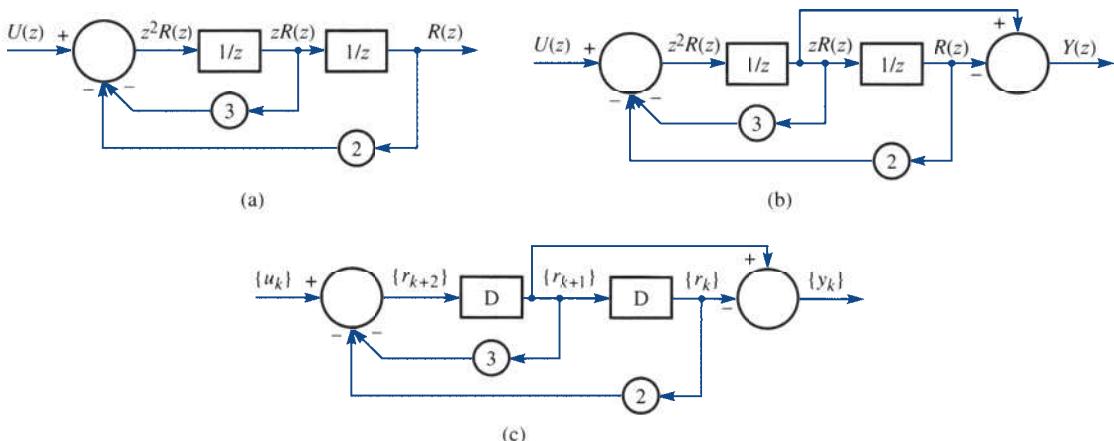


Figure 6.9 The z -transform block diagrams for (a) the system (6.40), (b) the system (6.39), and (c) the time-domain realization of the system in Example 6.19.

Finally, choosing

$$Y(z) = zR(z) - R(z) \quad (6.42)$$

$$(z^2 + 3z + 2)Y(z) = (z - 1)U(z)$$

which is a realization of the given transfer function.

To construct a block diagram realization of the system, we first construct a block diagram representation of (6.40) as in Figure 6.9(a). We now ‘tap off’ appropriate signals to generate $Y(z)$ according to (6.42) to construct a block diagram representation of the specified system. The resulting block diagram is shown in Figure 6.9(b).

In order to implement the system, we must exhibit a physically realizable time-domain structure, that is one containing only D elements. Clearly, since Figure 6.9(b) contains only ‘ $1/z$ ’ blocks, we can immediately produce a realizable time-domain structure as shown in Figure 6.9(c), where, as before, D is the unit delay block.

Example 6.20

A system is specified by its z transfer function

$$G(z) = \frac{z}{z^2 + 0.3z + 0.02}$$

Draw a block diagram to illustrate a time-domain realization of the system. Find a second structure that also implements the system.

Solution

We know that if $\mathcal{Z}\{u_k\} = U(z)$ and $\mathcal{Z}\{y_k\} = Y(z)$ are the z transforms of the input and output sequences respectively then, by definition,

$$G(z) = \frac{Y(z)}{U(z)} = \frac{z}{z^2 + 0.3z + 0.02} \quad (6.43)$$

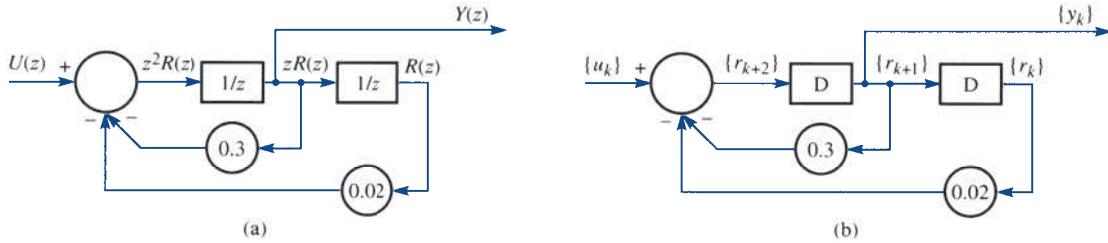


Figure 6.10 (a) The z -transform block diagram for the system of Example 6.20; and (b) the time-domain implementation of (a).

which may be rewritten as

$$(z^2 + 0.3z + 0.02)Y(z) = zU(z)$$

Noting the presence of the factor z on the right-hand side, we follow the procedure of Example 6.19 and consider the system

$$(z^2 + 0.3z + 0.02)R(z) = U(z) \quad (6.44)$$

Multiplying both sides by z , we have

$$(z^2 + 0.3z + 0.02)zR(z) = zU(z)$$

and so, if the output $Y(z) = zR(z)$ is extracted from the block diagram corresponding to (6.44), we have the block diagram representation of the given system (6.43). This is illustrated in Figure 6.10(a), with the corresponding time-domain implementation shown in Figure 6.10(b).

To discover a second form of time-domain implementation, note that

$$G(z) = \frac{z}{z^2 + 0.3z + 0.02} = \frac{2}{z + 0.2} - \frac{1}{z + 0.1}$$

We may therefore write

$$Y(z) = G(z)U(z) = \left(\frac{2}{z + 0.2} - \frac{1}{z + 0.1} \right)U(z)$$

so that

$$Y(z) = R_1(z) - R_2(z)$$

where

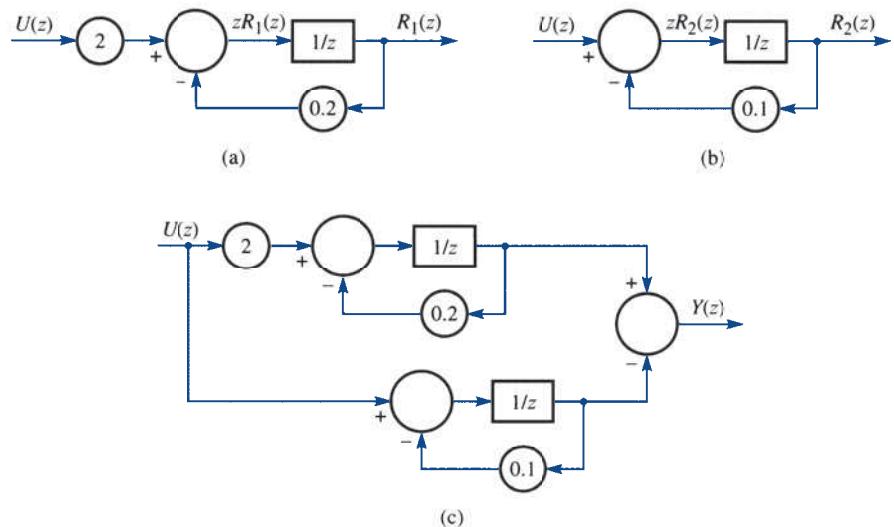
$$R_1(z) = \frac{2}{z + 0.2}U(z) \quad (6.45a)$$

$$R_2(z) = \frac{1}{z + 0.1}U(z) \quad (6.45b)$$

From (6.45a), we have

$$(z + 0.2)R_1(z) = 2U(z)$$

Figure 6.11 The block diagrams for (a) the subsystem (6.45a), (b) the subsystem (6.45b), and (c) an alternative z -transform block diagram for the system of Example 6.20.



which can be represented by the block diagram shown in Figure 6.11(a). Likewise, (6.45b) may be represented by the block diagram shown in Figure 6.11(b).

Recalling that $Y(z) = R_1(z) - R_2(z)$, it is clear that the given system can be represented and then implemented by an obvious coupling of the two subsystems represented by (6.45a, b). The resulting z -transform block diagram is shown in Figure 6.11(c). The time-domain version is readily obtained by replacing the ' $1/z$ ' blocks by D and the transforms $U(z)$ and $Y(z)$ by their corresponding sequences $\{u_k\}$ and $\{y_k\}$ respectively.

6.6.2 The impulse response

In Example 6.20 we saw that two quite different realizations were possible for the same transfer function $G(z)$, and others are possible. Whichever realization of the transfer function is chosen, however, when presented with the same input sequence $\{u_k\}$, the same output sequence $\{y_k\}$ will be produced. Thus we identify the system as characterized by its transfer function as the key concept, rather than any particular implementation. This idea is reinforced when we consider the impulse response sequence for a discrete-time linear time-invariant system, and its role in convolution sums.

Consider the sequence

$$\{\delta_k\} = \{1, 0, 0, \dots\}$$

that is, the sequence consisting of a single ‘pulse’ at $k = 0$, followed by a train of zeros. As we saw in Section 6.2.1, the z transform of this sequence is easily found from the definition (6.1) as

$$\mathcal{Z}\{\delta_k\} = 1 \quad (6.46)$$

The sequence $\{\delta_k\}$ is called the **impulse sequence**, by analogy with the continuous-time counterpart $\delta(t)$, the impulse function. The analogy is perhaps clearer on considering the transformed version (6.46). In continuous-time analysis, using Laplace transform methods, we observed that $\mathcal{L}\{\delta(t)\} = 1$, and (6.46) shows that the ‘entity’

with z transform equal to unity is the sequence $\{\delta_k\}$. It is in fact the property that $\mathcal{Z}\{\delta_k\} = 1$ that makes the impulse sequence of such great importance.

Consider a system with transfer function $G(z)$, so that the z transform $Y(z)$ of the output sequence $\{y_k\}$ corresponding to an input sequence $\{u_k\}$ with z transform $U(z)$ is

$$Y(z) = G(z)U(z) \quad (6.47)$$

If the input sequence $\{y_k\}$ is the impulse sequence $\{\delta_k\}$ and the system is initially quiescent, then the output sequence $\{y_{\delta_k}\}$ is called the impulse response of the system. Hence

$$\mathcal{Z}\{y_{\delta_k}\} = Y_{\delta}(z) = G(z) \quad (6.48)$$

That is, the z transfer function of the system is the z transform of the impulse response. Alternatively, we can say that the impulse response of a system is the inverse z transform of the system transfer function. This compares with the definition of the impulse response for continuous systems given in Section 5.3.3.

Substituting (6.48) into (6.47), we have

$$Y(z) = Y_{\delta}(z)U(z) \quad (6.49)$$

Thus the z transform of the system output in response to any input sequence $\{u_k\}$ is the product of the transform of the input sequence with the transform of the system impulse response. The result (6.49) shows the underlying relationship between the concepts of impulse response and transfer function, and explains why the impulse response (or the transfer function) is thought of as characterizing a system. In simple terms, if either of these is known then we have all the information about the system for any analysis we may wish to do.

Example 6.21

Find the impulse response of the system with z transfer function

$$G(z) = \frac{z}{z^2 + 3z + 2}$$

Solution Using (6.48),

$$Y_{\delta}(z) = \frac{z}{z^2 + 3z + 2} = \frac{z}{(z+2)(z+1)}$$

Resolving $Y_{\delta}(z)/z$ into partial fractions gives

$$\frac{Y_{\delta}(z)}{z} = \frac{1}{(z+2)(z+1)} = \frac{1}{z+1} - \frac{1}{z+2}$$

which on inversion gives the impulse response sequence

$$\begin{aligned} \{Y_{\delta_k}\} &= \mathcal{Z}^{-1}\left[\frac{z}{z+1} - \frac{z}{z+2}\right] \\ &= \{(-1)^k - (-2)^k\} \quad (k \geq 0) \end{aligned}$$



Since the impulse response of a system is the inverse z transform of its transfer function $G(z)$ it can be obtained in MATLAB using the command

```
syms k z
iztrans(G(z), k)
```

so for the $G(z)$ of Example 6.21

```
syms k z
iztrans(z/(z^2+3*z+2), k)
```

returns

$$\text{ans} = (-1)^k - (-2)^k$$

A plot of the impulse response is obtained using the commands

```
z=tf('z', 1);
G=G(z);
impulse(G)
```

Likewise in MAPLE the command

```
invztrans(z/(z^2+3*z+2), z, k);
```

returns the same answer

$$(-1)^k - (-2)^k$$

Example 6.22

A system has the impulse response sequence

$$\{y_{\delta_k}\} = \{a^k - 0.5^k\}$$

where $a > 0$ is a real constant. What is the nature of this response when (a) $a = 0.4$, (b) $a = 1.2$? Find the step response of the system in both cases.

Solution When $a = 0.4$

$$\{y_{\delta_k}\} = \{0.4^k - 0.5^k\}$$

and, since both $0.4^k \rightarrow 0$ as $k \rightarrow \infty$ and $0.5^k \rightarrow 0$ as $k \rightarrow \infty$, we see that the terms of the impulse response sequence go to zero as $k \rightarrow \infty$.

On the other hand, when $a = 1.2$, since $(1.2)^k \rightarrow \infty$ as $k \rightarrow \infty$, we see that in this case the impulse response sequence terms become unbounded, implying that the system ‘blows up’.

In order to calculate the step response, we first determine the system transfer function $G(z)$, using (6.48), as

$$G(z) = Y_{\delta}(z) = \mathcal{Z}\{a^k - 0.5^k\}$$

giving

$$G(z) = \frac{z}{z-a} - \frac{z}{z-0.5}$$

The system step response is the system response to the unit step sequence $\{h_k\} = \{1, 1, 1, \dots\}$ which, from Figure 6.3, has z transform

$$\mathcal{Z}\{h_k\} = \frac{z}{z-1}$$

Hence, from (6.46), the step response is determined by

$$Y(z) = G(z)\mathcal{Z}\{h_k\} = \left(\frac{z}{z-a} - \frac{z}{z-0.5} \right) \frac{z}{z-1}$$

so that

$$\begin{aligned} \frac{Y(z)}{z} &= \frac{z}{(z-a)(z-1)} - \frac{z}{(z-0.5)(z-1)} \\ &= \frac{a}{a-1} \frac{1}{z-a} - \frac{1}{z-0.5} + \left(-2 + \frac{1}{1-a} \right) \frac{1}{z-1} \end{aligned}$$

giving

$$Y(z) = \frac{a}{a-1} \frac{z}{z-a} - \frac{z}{z-0.5} + \left(-2 + \frac{1}{1-a} \right) \frac{z}{z-1}$$

which on taking inverse transforms gives the step response as

$$\{y_k\} = \left\{ \frac{a}{a-1} a^k - (0.5)^k + \left(-2 + \frac{1}{1-a} \right) \right\} \quad (6.50)$$

Considering the output sequence (6.50), we see that when $a = 0.4$, since $(0.4)^k \rightarrow 0$ as $k \rightarrow \infty$ (and $(0.5)^k \rightarrow 0$ as $k \rightarrow \infty$), the output sequence terms tend to the constant value

$$-2 + \frac{1}{1-0.4} = 0.3333$$

In the case of $a = 1.2$, since $(1.2)^k \rightarrow \infty$ as $k \rightarrow \infty$, the output sequence is unbounded, and again the system ‘blows up’.

6.6.3 Stability

Example 6.22 illustrated the concept of system stability for discrete systems. When $a = 0.4$, the impulse response decayed to zero with increasing k , and we observed that the step response remained bounded (in fact, the terms of the sequence approached a constant limiting value). However, when $a = 1.2$, the impulse response became unbounded, and we observed that the step response also increased without limit. In fact, as we saw for continuous systems in Section 5.3.3, a linear constant-coefficient discrete-time system is stable provided that its impulse response goes to zero as $k \rightarrow \infty$. As for the continuous case, we can relate this definition to the poles of the system transfer function

$$G(z) = \frac{P(z)}{Q(z)}$$

As we saw in Section 6.6.1, the system poles are determined as the n roots of its characteristic equation

$$Q(z) = a_n z^n + a_{n-1} z^{n-1} + \dots + a_0 = 0 \quad (6.51)$$

For instance, in Example 6.19 we considered a system with transfer function

$$G(z) = \frac{z-1}{z^2 + 3z + 2}$$

having poles determined by $z^2 + 3z + 2 = 0$, that is poles at $z = -1$ and $z = -2$. Since the impulse response is the inverse transform of $G(z)$, we expect this system to ‘blow up’ or, rather, be unstable, because its impulse response sequence would be expected to contain terms of the form $(-1)^k$ and $(-2)^k$, neither of which goes to zero as $k \rightarrow \infty$. (Note that the term in $(-1)^k$ neither blows up nor goes to zero, simply alternating between +1 and -1; however, $(-2)^k$ certainly becomes unbounded as $k \rightarrow \infty$.) On the other hand, in Example 6.20 we encountered a system with transfer function

$$G(z) = \frac{z}{z^2 + 0.3z + 0.02}$$

having poles determined by

$$Q(z) = z^2 + 0.3z + 0.02 = (z + 0.2)(z + 0.1) = 0$$

that is poles at $z = -0.2$ and $z = -0.1$. Clearly, this system is stable, since its impulse response contains terms in $(-0.2)^k$ and $(-0.1)^k$, both of which go to zero as $k \rightarrow \infty$.

Both of these illustrative examples gave rise to characteristic polynomials $Q(z)$ that were quadratic in form and that had real coefficients. More generally, $Q(z) = 0$ gives rise to a polynomial equation of order n , with real coefficients. From the theory of polynomial equations, we know that $Q(z) = 0$ has n roots α_i ($i = 1, 2, \dots, n$), which may be real or complex (with complex roots occurring in conjugate pairs).

Hence the characteristic equation may be written in the form

$$Q(z) = a_n(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n) = 0 \quad (6.52)$$

The system poles α_i ($i = 1, 2, \dots, n$) determined by (6.52) may be expressed in the polar form

$$\alpha_i = r_i e^{j\theta_i} \quad (i = 1, 2, \dots, n)$$

where $\theta_i = 0$ or π if α_i is real. From the interpretation of the impulse response as the inverse transform of the transfer function $G(z) = P(z)/Q(z)$, it follows that the impulse response sequence of the system will contain terms in

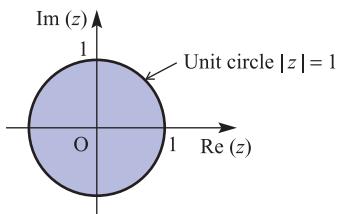
$$r_1^k e^{jk\theta_1}, r_2^k e^{jk\theta_2}, \dots, r_n^k e^{jk\theta_n}$$

Since, for stability, terms in the impulse response sequence must tend to zero as $k \rightarrow \infty$, it follows that a system having characteristic equation $Q(z) = 0$ will be stable provided that

$$r_i < 1 \quad \text{for } i = 1, 2, \dots, n$$

Therefore a linear constant-coefficient discrete-time system with transfer function $G(z)$ is stable if and only if all the poles of $G(z)$ lie within the unit circle $|z| < 1$ in the complex z plane, as illustrated in Figure 6.12. If one or more poles lie outside this unit circle then the system will be unstable. If one or more distinct poles lie on the unit circle $|z| = 1$, with all the other poles inside, then the system is said to be **marginally stable**.

Figure 6.12 Region of stability in the z plane.



Example 6.23

Which of the following systems, specified by their transfer function $G(z)$, are stable?

$$(a) \ G(z) = \frac{1}{z + 0.25} \quad (b) \ G(z) = \frac{z}{z^2 - z + 0.5} \quad (c) \ G(z) = \frac{z^2}{z^3 - 3z^2 + 2.5z - 1}$$

Solution (a) The single pole is at $z = -0.25$, so $r_1 = 0.25 < 1$, and the system is stable.

(b) The system poles are determined by

$$z^2 - z + 0.5 = [z - 0.5(1 + j)][z - 0.5(1 - j)] = 0$$

giving the poles as the conjugate pair $z_1 = 0.5(1 + j)$, $z_2 = 0.5(1 - j)$. The amplitudes $r_1 = r_2 = 0.707 < 1$, and again the system is stable.

(c) The system poles are determined by

$$z^3 - 3z^2 + 2.5z - 1 = (z - 2)[z - 0.5(1 + j)][z - 0.5(1 - j)]$$

giving the poles as $z_1 = 2$, $z_2 = 0.5(1 + j)$, $z_3 = 0.5(1 - j)$, and so their amplitudes are $r_1 = 2$, $r_2 = r_3 = 0.707$. Since $r_1 > 1$, it follows that the system is unstable.

According to our definition, it follows that to prove stability we must show that all the roots of the characteristic equation

$$Q(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0 = 0 \quad (6.53)$$

lie within the unit circle $|z| = 1$ (note that for convenience we have arranged for the coefficient of z^n to be unity in (6.53)). Many mathematical criteria have been developed to test for this property. One such method, widely used in practice, is the **Jury stability criterion** introduced by E. I. Jury in 1963. This procedure gives necessary and sufficient conditions for the polynomial equation (6.53) to have all its roots inside the unit circle $|z| = 1$.

The first step in the procedure is to set up a table as in Figure 6.13 using information from the given polynomial equation (6.53) and where

$$b_k = \begin{vmatrix} 1 & a_k \\ a_0 & a_{n-k} \end{vmatrix}, \quad c_k = \begin{vmatrix} b_0 & b_{n-1-k} \\ b_{n-1} & b_k \end{vmatrix}, \quad d_k = \begin{vmatrix} c_0 & c_{n-2-k} \\ c_{n-2} & c_k \end{vmatrix}, \quad \dots,$$

$$t_0 = \begin{vmatrix} r_0 & r_2 \\ r_2 & r_0 \end{vmatrix}$$

Figure 6.13 Jury stability table for the polynomial equation (6.53).

Row	z^n	z^{n-1}	z^{n-2}	...	z^{n-k}	...	z^2	z^1	z^0
1	1	a_{n-1}	a_{n-2}	...	a_{n-k}	...	a_2	a_1	a_0
2	a_0	a_1	a_2	...	a_k	...	a_{n-2}	a_{n-1}	1
3	$\Delta_1 = b_0$	b_1	b_2	...	b_k	...	b_{n-2}	b_{n-1}	
4	b_{n-1}	b_{n-2}	b_{n-3}	...	b_{n-1-k}	...	b_1	b_0	
5	$\Delta_2 = c_0$	c_1	c_2	...	c_k	...	c_{n-2}		
6	c_{n-2}	c_{n-3}	c_{n-4}	...	c_{n-2-k}	...	c_0		
7	$\Delta_3 = d_0$	d_1	d_2	...	d_k	...			
8	d_{n-3}	d_{n-4}	d_{n-5}	...	d_{n-3-k}	...			
?									
?									
$2n-5$	$\Delta_{n-3} = s_0$	s_1	s_2		s_3				
$2n-4$	s_3	s_2	s_1		s_0				
$2n-3$	$\Delta_{n-2} = r_0$	r_1	r_2						
$2n-2$	r_2	r_1	r_0						
$2n-1$	$\Delta_{n-1} = t_0$								

Note that the elements of row $2j+2$ consist of the elements of row $2j+1$ written in the reverse order for $j = 0, 1, 2, \dots, n$; that is, the elements of the even rows consist of the elements of the odd rows written in reverse order. Necessary and sufficient conditions for the polynomial equation (6.53) to have all its roots inside the unit circle $|z| = 1$ are then given by

- (i) $Q(1) > 0, \quad (-1)^n Q(-1) > 0$ (6.54)
- (ii) $\Delta_1 > 0, \quad \Delta_2 > 0, \quad \Delta_3 > 0, \quad \dots, \quad \Delta_{n-2} > 0, \quad \Delta_{n-1} > 0$

Example 6.24

Show that all the roots of the polynomial equation

$$F(z) = z^3 + \frac{1}{3}z^2 - \frac{1}{4}z - \frac{1}{12} = 0$$

lie within the unit circle $|z| = 1$.

Solution The corresponding Jury stability table is shown in Figure 6.14. In this case

- (i) $F(1) = 1 + \frac{1}{3} - \frac{1}{4} - \frac{1}{12} > 0$
 $(-1)^n F(-1) = (-1)^3 (-1 + \frac{1}{3} + \frac{1}{4} - \frac{1}{12}) > 0$
- (ii) $\Delta_1 = \frac{143}{144} > 0, \quad \Delta_2 = (\frac{143}{144})^2 - \frac{4}{81} > 0$

Thus, by the criteria (6.54), all the roots lie within the unit circle. In this case this is readily confirmed, since the polynomial $F(z)$ may be factorized as

$$F(z) = (z - \frac{1}{2})(z + \frac{1}{2})(z + \frac{1}{3}) = 0$$

So the roots are $z_1 = \frac{1}{2}$, $z_2 = -\frac{1}{2}$ and $z_3 = -\frac{1}{3}$.

Figure 6.14 Jury stability table for Example 6.24.

Row	z^3	z^2	z^1	z^0
1	1	$\frac{1}{3}$	$-\frac{1}{4}$	$-\frac{1}{12}$
2	$-\frac{1}{12}$	$-\frac{1}{4}$	$\frac{1}{3}$	1
3	$\Delta_1 = \begin{vmatrix} 1 & -\frac{1}{12} \\ -\frac{1}{12} & 1 \end{vmatrix} = \frac{143}{144}$	$\begin{vmatrix} 1 & -\frac{1}{4} \\ -\frac{1}{12} & \frac{1}{3} \end{vmatrix} = \frac{5}{16}$	$\begin{vmatrix} 1 & \frac{1}{3} \\ -\frac{1}{12} & -\frac{1}{4} \end{vmatrix} = -\frac{2}{9}$	
4	$-\frac{2}{9}$	$\frac{5}{16}$	$\frac{143}{144}$	
5	$\Delta_2 = \begin{vmatrix} \frac{143}{144} & -\frac{2}{9} \\ -\frac{2}{9} & \frac{143}{144} \end{vmatrix} = 0.93678$			

The Jury stability table may also be used to determine how many roots of the polynomial equation (6.53) lie outside the unit circle. The number of such roots is determined by the number of changes in sign in the sequence

$$1, \quad \Delta_1, \quad \Delta_2, \quad \dots, \quad \Delta_{n-1}$$

Example 6.25

Show that the polynomial equation

$$F(z) = z^3 - 3z^2 - \frac{1}{4}z + \frac{3}{4} = 0$$

has roots that lie outside the unit circle $|z| = 1$. Determine how many such roots there are.

Figure 6.15 Jury stability table for Example 6.25.

Row	z^3	z^2	z^1	z^0
1	1	-3	$-\frac{1}{4}$	$\frac{3}{4}$
2	$\frac{3}{4}$	$-\frac{1}{4}$	-3	1
3	$\Delta_1 = \frac{7}{16}$	$-\frac{45}{16}$	2	
4	2	$-\frac{45}{16}$	$\frac{7}{16}$	
5	$\Delta_2 = -\frac{5}{16}$			

Solution

The corresponding Jury stability table is shown in Figure 6.15. Hence, in this case

$$F(z) = 1 - 3 - \frac{1}{4}z + \frac{3}{4} = -\frac{3}{2}$$

$$(-1)^n F(-1) = (-1)^3 (-1 - 3 + \frac{1}{4}z + \frac{3}{4}) = 3$$

As $F(1) < 0$, it follows from (6.54) that the polynomial equation has roots outside the unit circle $|z| = 1$. From Figure 6.15, the sequence $1, \Delta_1, \Delta_2$ is $1, \frac{7}{16}, -\frac{15}{16}$, and since there is only one sign change in the sequence, it follows that one root lies outside the unit circle. Again this is readily confirmed, since $F(z)$ may be factorized as

$$F(z) = (z - \frac{1}{2})(z + \frac{1}{2})(z - 3) = 0$$

showing that there is indeed one root outside the unit circle at $z = 3$.

Example 6.26

Consider the discrete-time feedback system of Figure 6.16, for which T is the sampling period and $k > 0$ is a constant gain:

- Determine the z transform $G(z)$ corresponding to the Laplace transform $G(s)$.
- Determine the characteristic equation of the system when $T = 1$ and $k = 6$ and show that the discrete-time system is unstable.
- For $T = 1$ show that the system is stable if and only if $0 < k < 4.33$.
- Removing the sampler show that the corresponding continuous-time feedback system is stable for all $k > 0$.

Solution (a) First invert the Laplace transform to give the corresponding time-domain function $f(t)$ and then determine the z transform of $f(t)$:

$$G(s) = \frac{k}{s(s+1)} = \frac{k}{s} - \frac{k}{s+1}$$

$$f(t) = k - ke^{-t}$$

$$G_d(z) = Z\{k\} - Z\{ke^{-t}\} = \frac{kz}{z-1} - \frac{kz}{z-e^{-T}} = \frac{kz(1-e^{-T})}{(z-1)(z-e^{-T})}$$

- (b) With $k = 6$ and $T = 1$

$$G_d(z) = \frac{6(1-e^{-1})z}{(z-1)(z-e^{-1})}$$

The closed-loop transfer function is

$$\frac{G_d(z)}{1 + G_d(z)}$$

giving the characteristic equation

$$1 + G_d(z) = 0 \text{ as } (z-1)(z-e^{-1}) + 6(1-e^{-1})z = 0$$

or

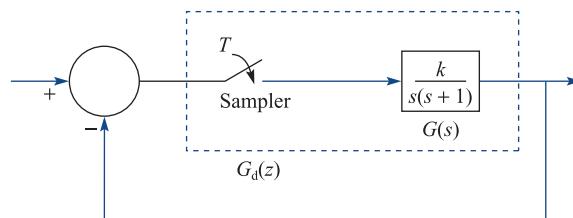
$$z^2 + z[6(1-e^{-1}) - (1+e^{-1})] + e^{-1} = 0$$

which reduces to

$$z^2 + 2.324z + 0.368 = 0$$

The roots of this characteristic equation are $z_1 = -0.171$ and $z_2 = -2.153$. Since one of the roots lies outside the unit circle $|z| = 1$ the system is unstable.

Figure 6.16
Discrete-time system
of Example 6.26.



- (c) For $T = 1$ and general gain $k > 0$ the characteristic equation of the system is

$$F(z) = (z - 1)(z - e^{-1}) + k(1 - e^{-1})z = 0$$

which reduces to

$$F(z) = z^2 + (0.632k - 1.368)z + 0.368 = 0$$

By Jury's procedure conditions for stability are:

$$F(1) = 1 + (0.632k - 1.368) + 0.368 > 0 \text{ since } k > 0$$

$$(-1)^2 F(-1) = 2.736 - 0.632k > 0 \text{ provided } k < \frac{2.736}{0.632} = 4.33$$

$$\Delta_1 = \begin{vmatrix} 1 & 0.368 \\ 0.568 & 1 \end{vmatrix} > 0$$

Thus $F(1) > 0$, $(-1)^2 F(-1) > 0$ and $\Delta_1 > 0$ and system stable if and only if $k < 4.33$.

- (d) In the absence of the sampler the characteristic equation of the continuous-time feedback system is $1 + G(s) = 0$, which reduces to

$$s^2 + s + k = 0$$

All the roots are in the negative half of the s -plane, and the system is stable, for all $k > 0$.

6.6.4 Convolution

Here we shall briefly extend the concept of convolution introduced in Section 5.3.6 to discrete-time systems. From (6.45), for an initially quiescent system with an impulse response sequence $\{y_{\delta_k}\}$ with z transform $Y_\delta(z)$, the z transform $Y(z)$ of the output sequence $\{y_k\}$ in response to an input sequence $\{u_k\}$ with z transform $U(z)$ is given by

$$Y(z) = Y_\delta(z)U(z) \quad (6.49)$$

For the purposes of solving a particular problem, the best approach to determining $\{y_k\}$ for a given $\{u_k\}$ is to invert the right-hand side of (6.49) as an ordinary z transform with no particular thought as to its structure. However, to understand more of the theory of linear systems in discrete time, it is worth exploring the general situation a little further. To do this, we revert to the time domain.

Suppose that a linear discrete-time time-invariant system has impulse response sequence $\{y_{\delta_k}\}$, and suppose that we wish to find the system response $\{y_k\}$ to an input sequence $\{u_k\}$, with the system initially in a quiescent state. First we express the input sequence

$$\{u_k\} = \{u_0, u_1, u_2, \dots, u_n, \dots\} \quad (6.55)$$

as

$$\{u_k\} = u_0\{\delta_k\} + u_1\{\delta_{k-1}\} + u_2\{\delta_{k-2}\} + \dots + u_n\{\delta_{k-n}\} + \dots \quad (6.56)$$

where

$$\delta_{k-j} = \begin{cases} 0 & (k \neq j) \\ 1 & (k = j) \end{cases}$$

In other words, $\{\delta_{k-j}\}$ is simply an impulse sequence with the pulse shifted to $k=j$. Thus, in going from (6.55) to (6.56), we have decomposed the input sequence $\{u_k\}$ into a weighted sum of shifted impulse sequences. Under the assumption of an initially quiescent system, linearity allows us to express the response $\{y_k\}$ to the input sequence $\{u_k\}$ as the appropriately weighted sum of shifted impulse responses. Thus, since the impulse response is $\{y_{\delta_k}\}$, the response to the shifted impulse sequence $\{\delta_{k-j}\}$ will be $\{y_{\delta_{k-j}}\}$, and the response to the weighted impulse sequence $u_j\{\delta_{k-j}\}$ will be simply $u_j\{y_{\delta_{k-j}}\}$. Summing the contributions from all the sequences in (6.56), we obtain

$$\{y_k\} = \sum_{j=0}^{\infty} u_j \{y_{\delta_{k-j}}\} \quad (6.57)$$

as the response of the system to the input sequence $\{u_k\}$. Expanding (6.57), we have

$$\begin{aligned} \{y_k\} &= u_0 \{y_{\delta_k}\} + u_1 \{y_{\delta_{k-1}}\} + \cdots + u_j \{y_{\delta_{k-j}}\} + \cdots \\ &= u_0 \{y_{\delta_0}, y_{\delta_1}, y_{\delta_2}, \dots, y_{\delta_h}, \dots\} \\ &\quad + u_1 \{0, y_{\delta_0}, y_{\delta_1}, \dots, y_{\delta_{h-1}}, \dots\} \\ &\quad + u_2 \{0, 0, y_{\delta_0}, \dots, y_{\delta_{h-2}}, \dots\} \\ &\quad \vdots \\ &\quad + u_h \{0, 0, \dots, 0, y_{\delta_0}, y_{\delta_1}, \dots\} \\ &\quad \uparrow \\ &\quad + \dots \end{aligned}$$

hth position

From this expansion, we find that the h th term of the output sequence is determined by

$$y_h = \sum_{j=0}^h u_j y_{\delta_{h-j}} \quad (6.58)$$

That is,

$$\{y_k\} = \left\{ \sum_{j=0}^k u_j y_{\delta_{k-j}} \right\} \quad (6.59)$$

The expression (6.58) is called the **convolution sum**, and the result (6.59) is analogous to (5.45) for continuous systems.

Example 6.27

A system has z transfer function

$$G(z) = \frac{z}{z + \frac{1}{2}}$$

What is the system step response? Verify the result using (6.59).

Solution From (6.46), the system step response is

$$Y(z) = G(z)\mathcal{Z}\{h_k\}$$

where $\{h_k\} = \{1, 1, 1, \dots\}$. From Figure 6.3, $\mathcal{Z}\{h_k\} = z/(z - 1)$, so

$$Y(z) = \frac{z}{z + \frac{1}{2}} \frac{z}{z - 1}$$

Resolving $Y(z)/z$ into partial fractions gives

$$\frac{Y(z)}{z} = \frac{z}{(z + \frac{1}{2})(z - 1)} = \frac{\frac{2}{3}}{z - 1} + \frac{\frac{1}{3}}{z + \frac{1}{2}}$$

so

$$Y(z) = \frac{\frac{2}{3}}{z - 1} + \frac{\frac{1}{3}}{z + \frac{1}{2}}$$

Taking inverse transforms then gives the step response as

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

Using (6.59), we first have to find the impulse response, which, from (6.48), is given by

$$\{y_{\delta_k}\} = \mathcal{Z}^{-1}[G(z)] = \mathcal{Z}^{-1}\left[\frac{z}{z + \frac{1}{2}}\right]$$

so that

$$\{y_{\delta_k}\} = \left\{ \left(-\frac{1}{2}\right)^k \right\}$$

Taking $\{u_k\}$ to be the unit step sequence $\{h_k\}$, where $h_k = 1$ ($k \geq 0$), the step response may then be determined from (6.59) as

$$\begin{aligned} \{y_k\} &= \left\{ \sum_{j=0}^k u_j y_{\delta_{k-j}} \right\} = \left\{ \sum_{j=0}^k 1 \cdot \left(-\frac{1}{2}\right)^{k-j} \right\} \\ &= \left\{ \left(-\frac{1}{2}\right)^k \sum_{j=0}^k \left(-\frac{1}{2}\right)^{-j} \right\} = \left\{ \left(-\frac{1}{2}\right)^k \sum_{j=0}^k (-2)^j \right\} \end{aligned}$$

Recognizing the sum as the sum to $k + 1$ terms of a geometric series with common ratio -2 , we have

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

which concurs with the sequence obtained by direct evaluation.

Example 6.27 reinforces the remark made earlier that the easiest approach to obtaining the response is by direct inversion of (6.32). However, (6.59), together with the argument leading to it, provides a great deal of insight into the way in which the response sequence $\{y_k\}$ is generated. It also serves as a useful ‘closed form’ for the output of the system, and readers should consult specialist texts on signals and systems

for a full discussion (P. Kranauskas, *Transforms in Signals and Systems*, Wokingham, Addison-Wesley, 1992).

The astute reader will recall that we commenced this section by suggesting that we were about to study the implications of the input–output relationship (6.49), namely

$$Y(z) = Y_\delta(z)U(z)$$

We have in fact explored the time-domain input–output relationship for a linear system, and we now proceed to link this approach with our work in the transform domain. By definition,

$$U(z) = \sum_{k=0}^{\infty} u_k z^{-k} = u_0 + \frac{u_1}{z} + \frac{u_2}{z^2} + \cdots + \frac{u_k}{z^k} + \cdots$$

$$Y_\delta(z) = \sum_{k=0}^{\infty} y_{\delta_k} z^{-k} = y_{\delta_0} + \frac{y_{\delta_1}}{z} + \frac{y_{\delta_2}}{z^2} + \cdots + \frac{y_{\delta_k}}{z^k} + \cdots$$

so

$$Y_\delta(z)U(z) = u_0 y_{\delta_0} + (u_0 y_{\delta_1} + u_1 y_{\delta_0}) \frac{1}{z} + (u_0 y_{\delta_2} + u_1 y_{\delta_1} + u_2 y_{\delta_0}) \frac{1}{z^2} + \cdots \quad (6.60)$$

Considering the k th term of (6.60), we see that the coefficient of z^{-k} is simply

$$\sum_{j=0}^k u_j y_{\delta_{k-j}}$$

However, by definition, since $Y(z) = Y_\delta(z)U(z)$, this is also $y(k)$, the k th term of the output sequence, so that the latter is

$$\{y_k\} = \left\{ \sum_{j=0}^k u_j y_{\delta_{k-j}} \right\}$$

as found in (6.59). We have thus shown that the time-domain and transform-domain approaches are equivalent, and, in passing, we have established the z transform of the convolution sum as

$$\mathcal{Z} \left\{ \sum_{j=0}^k u_j v_{k-j} \right\} = U(z)V(z) \quad (6.61)$$

where

$$\mathcal{Z}\{u_k\} = U(z), \quad \mathcal{Z}\{v_k\} = V(z)$$

Putting $p = k - j$ in (6.61) shows that

$$\sum_{j=0}^k u_j v_{k-j} = \sum_{p=0}^k u_{k-p} v_p \quad (6.62)$$

confirming that the convolution process is commutative.

6.6.5 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 21 Find the transfer functions of each of the following discrete-time systems, given that the system is initially in a quiescent state:

$$\begin{aligned} \text{(a)} \quad & y_{k+2} - 3y_{k+1} + 2y_k = u_k \\ \text{(b)} \quad & y_{k+2} - 3y_{k+1} + 2y_k = u_{k+1} - u_k \\ \text{(c)} \quad & y_{k+3} - y_{k+2} + 2y_{k+1} + y_k = u_k + u_{k-1} \end{aligned}$$

- 22 Draw a block diagram representing the discrete-time system

$$y_{k+2} + 0.5y_{k+1} + 0.25y_k = u_k$$

Hence find a block diagram representation of the system

$$y_{k+2} + 0.5y_{k+1} + 0.25y_k = u_k - 0.6u_{k+1}$$

- 23 Find the impulse response for the systems with z transfer function

$$\begin{array}{ll} \text{(a)} \quad \frac{z}{8z^2 + 6z + 1} & \text{(b)} \quad \frac{z^2}{z^2 - 3z + 3} \\ \text{(c)} \quad \frac{z^2}{z^2 - 0.2z - 0.08} & \text{(d)} \quad \frac{5z^2 - 12z}{z^2 - 6z + 8} \end{array}$$

- 24 Obtain the impulse response for the systems of Exercises 21(a, b).

- 25 Which of the following systems are stable?

$$\begin{array}{l} \text{(a)} \quad 9y_{k+2} + 9y_{k+1} + 2y_k = u_k \\ \text{(b)} \quad 9y_{k+2} - 3y_{k+1} - 2y_k = u_k \\ \text{(c)} \quad 2y_{k+2} - 2y_{k+1} + y_k = u_{k+1} - u_k \\ \text{(d)} \quad 2y_{k+2} + 3y_{k+1} - y_k = u_k \\ \text{(e)} \quad 4y_{k+2} - 3y_{k+1} - y_k = u_{k+1} - 2u_k \end{array}$$

- 26 Use the method of Example 6.27 to calculate the step response of the system with transfer function

$$\frac{z}{z - \frac{1}{2}}$$

Verify the result by direct calculation.

- 27 Following the same procedure as in Example 6.26 show that the closed-loop discrete-time system of Figure 6.17, in which $k > 0$ and $\tau > 0$, is stable if and only if

$$0 < k < 2 \coth\left(\frac{T}{2\tau}\right)$$

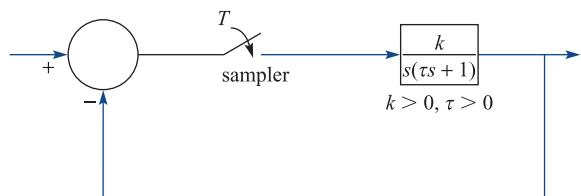


Figure 6.17 Discrete-time system of Exercise 27.

- 28 A sampled data system described by the difference equation

$$y_{n+1} - y_n = u_n$$

is controlled by making the input u_n proportional to the previous error according to

$$u_n = K\left(\frac{1}{2^n} - y_{n-1}\right)$$

where K is a positive gain. Determine the range of values of K for which the system is stable. Taking $K = \frac{2}{9}$, determine the response of the system given $y_0 = y_1 = 0$.

- 29 Show that the system

$$y_{n+2} + 2y_{n+1} + 2y_n = u_{n+1} \quad (n \geq 0)$$

has transfer function

$$D(z) = \frac{z}{z^2 + 2z + 2}$$

Show that the poles of the system are at $z = -1 + j$ and $z = -1 - j$. Hence show that the impulse response of the system is given by

$$h_n = \mathcal{Z}^{-1}D(z) = 2^{n/2} \sin \frac{3}{4} n\pi$$

6.7

The relationship between Laplace and z transforms

Throughout this chapter we have attempted to highlight similarities, where they occur, between results in Laplace transform theory and those for z transforms. In this section we take a closer look at the relationship between the two transforms. In Section 6.2.2 we introduced the idea of sampling a continuous-time signal $f(t)$ instantaneously at uniform intervals T to produce the sequence

$$\{f(nT)\} = \{f(0), f(T), f(2T), \dots, f(nT), \dots\} \quad (6.63)$$

An alternative way of representing the sampled function is to define the continuous-time sampled version of $f(t)$ as $\hat{f}(t)$ where

$$f(t) = \sum_{n=0}^{\infty} f(nT) \delta(t - nT) = \sum_{n=0}^{\infty} f(nT) \delta(t - nT) \quad (6.64)$$

The representation (6.64) may be interpreted as defining a row of impulses located at the sampling points and weighted by the appropriate sampled values (as illustrated in Figure 6.18). Taking the Laplace transform of $\hat{f}(t)$, following the results of Section 5.2.10, we have

$$\begin{aligned} \mathcal{L}\{f(t)\} &= \int_{0^-}^{\infty} \left[\sum_{k=0}^{\infty} f(kT) \delta(t - kT) \right] e^{-st} dt \\ &= \sum_{k=0}^{\infty} f(kT) \int_{0^-}^{\infty} \delta(t - kT) e^{-st} dt \end{aligned}$$

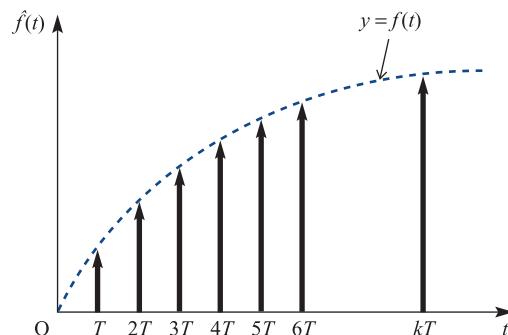
giving

$$\mathcal{L}\{\hat{f}(t)\} = \sum_{k=0}^{\infty} f(kT) e^{-ksT} \quad (6.65)$$

Making the change of variable $z = e^{sT}$ in (6.65) leads to the result

$$\mathcal{L}\{\hat{f}(t)\} = \sum_{k=0}^{\infty} f(kT) z^{-k} = F(z) \quad (6.66)$$

Figure 6.18 Sampled function $\hat{f}(t)$.



where, as in (6.10), $F(z)$ denotes the z transform of the sequence $\{f(kT)\}$. We can therefore view the z transform of a sequence of samples in discrete time as the Laplace transform of the continuous-time sampled function $\hat{f}(t)$ with an appropriate change of variable

$$z = e^{sT} \quad \text{or} \quad s = \frac{1}{T} \ln z$$

In Chapter 4 we saw that under this transformation the left half of the s plane, $\operatorname{Re}(s) < 0$, is mapped onto the region inside the unit circle in the z plane, $|z| < 1$. This is consistent with our stability criteria in the s and z domains.

6.8

Solution of discrete-time state-space equations

The state-space approach to the analysis of continuous-time dynamic systems, developed in Section 5.4, can be extended to the discrete-time case. The discrete form of the state-space representation is quite analogous to the continuous form.

6.8.1 State-space model

Consider the n th-order linear time-invariant discrete-time system modelled by the difference equation

$$y_{k+n} + a_{n-1}y_{k+n-1} + a_{n-2}y_{k+n-2} + \cdots + a_0y_k = b_0u_k \quad (6.67)$$

which corresponds to (6.32), with $b_i = 0$ ($i > 0$). Recall that $\{y_k\}$ is the output sequence, with general term y_k , and $\{u_k\}$ the input sequence, with general term u_k . Following the procedure of Section 1.9.1, we introduce state variables $x_1(k), x_2(k), \dots, x_n(k)$ for the system, defined by

$$x_1(k) = y_k, \quad x_2(k) = y_{k+1}, \quad \dots, \quad x_n(k) = y_{k+n-1} \quad (6.68)$$

Note that we have used the notation $x_i(k)$ rather than the suffix notation $x_{i,k}$ for clarity. When needed, we shall adopt the same convention for the input term and write $u(k)$ for u_k in the interests of consistency. We now define the state vector corresponding to this choice of state variables as $\mathbf{x}(k) = [x_1(k) \quad x_2(k) \quad \dots \quad x_n(k)]^T$. Examining the system of equations (6.68), we see that

$$\begin{aligned} x_1(k+1) &= y_{k+1} = x_2(k) \\ x_2(k+1) &= y_{k+2} = x_3(k) \\ &\vdots \\ x_{n-1}(k+1) &= y_{k+n-1} = x_n(k) \\ x_n(k+1) &= y_{k+n} \\ &= -a_{n-1}y_{k+n-1} - a_{n-2}y_{k+n-2} - \cdots - a_0y_k + b_0u_k \\ &= -a_{n-1}x_n(k) - a_{n-2}x_{n-1}(k) - \cdots - a_0x_1(k) + b_0u(k) \end{aligned}$$

using the alternative notation for u_k .

We can now write the system in the vector-matrix form

$$\mathbf{x}(k+1) = \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ \vdots \\ x_n(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \\ \vdots \\ x_n(k) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ b_0 \end{bmatrix} u(k) \quad (6.69)$$

which corresponds to (1.57) for a continuous-time system. Again, we can write this more concisely as

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{b}u(k) \quad (6.70)$$

where \mathbf{A} and \mathbf{b} are defined as in (6.69). The output of the system is the sequence $\{y_k\}$, and the general term $y_k = x_1(k)$ can be recovered from the state vector $\mathbf{x}(k)$ as

$$y(k) = x_1(k) = [1 \ 0 \ 0 \ \cdots \ 0] \mathbf{x}(k) = \mathbf{c}^T \mathbf{x}(k) \quad (6.71)$$

As in the continuous-time case, it may be that the output of the system is a combination of the state and the input sequence $\{u(k)\}$, in which case (6.71) becomes

$$y(k) = \mathbf{c}^T \mathbf{x}(k) + du(k) \quad (6.72)$$

Equations (6.70) and (6.72) constitute the state-space representation of the system, and we immediately note the similarity with (1.60a, b) derived for continuous-time systems. Likewise, for the multi-input–multi-output case the discrete-time state-space model corresponding to (1.66a, b) is

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k) \quad (6.73a)$$

$$\mathbf{y}(k) = \mathbf{C}\mathbf{x}(k) + \mathbf{D}\mathbf{u}(k) \quad (6.73b)$$

Example 6.28

Determine the state-space representation of the system modelled by the difference equation

$$y_{k+2} + 0.2y_{k+1} + 0.3y_k = u_k \quad (6.74)$$

Solution We choose as state variables

$$x_1(k) = y_k, \quad x_2(k) = y_{k+1}$$

Thus

$$x_1(k+1) = x_2(k)$$

and from (6.74),

$$x_2(k+1) = -0.3x_1(k) - 0.2x_2(k) + u(k)$$

The state-space representation is then

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{b}u(k), \quad \mathbf{y}(k) = \mathbf{c}^T \mathbf{x}(k)$$

with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -0.3 & -0.2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{c}^T = [1 \ 0]$$

We notice, from reference to Section 6.6.1, that the procedure used in Example 6.28 for establishing the state-space form of the system corresponds to labelling the output of each delay block in the system as a state variable. In the absence of any reason for an alternative choice, this is the logical approach. Section 6.6.1 also gives a clue towards a method of obtaining the state-space representation for systems described by the more general form of (6.32) with $m > 0$. Example 6.19 illustrates such a system, with z transfer function

$$G(z) = \frac{z-1}{z^2 + 3z + 2}$$

The block diagram for this system is shown in Figure 6.9(c) and reproduced for convenience in Figure 6.19. We choose as state variables the outputs from each delay block, it being immaterial whether we start from the left- or the right-hand side of the diagram (obviously, different representations will be obtained depending on the choice we make, but the different forms will yield identical information on the system). Choosing to start on the right-hand side (that is, with $x_1(k)$ the output of the right-hand delay block and $x_2(k)$ that of the left-hand block), we obtain

$$x_1(k+1) = x_2(k)$$

$$x_2(k+1) = -3x_2(k) - 2x_1(k) + u(k)$$

with the system output given by

$$y(k) = -x_1(k) + x_2(k)$$

Thus the state-space form corresponding to our choice of state variables is

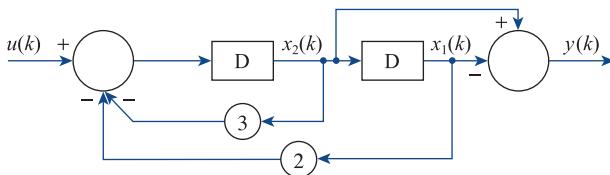
$$\mathbf{x}(k+1) = \mathbf{Ax}(k) + \mathbf{bu}(k), \quad y(k) = \mathbf{c}^T \mathbf{x}(k)$$

with

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{c}^T = [-1 \ 1]$$

We notice that, in contrast with the system of Example 6.28, the row vector $\mathbf{c}^T = [-1 \ 1]$ now combines contributions from both state variables to form the output $y(k)$.

Figure 6.19 Block diagram of system with transfer function $G(z) = (z-1)/(z^2 + 3z + 2)$.



6.8.2 Solution of the discrete-time state equation

As in Section 1.10.1 for continuous-time systems, we first consider the unforced or homogeneous case

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) \quad (6.75)$$

in which the input $\mathbf{u}(k)$ is zero for all time instants k . Taking $k = 0$ in (6.75) gives

$$\mathbf{x}(1) = \mathbf{A}\mathbf{x}(0)$$

Likewise, taking $k = 1$ in (6.75) gives

$$\mathbf{x}(2) = \mathbf{A}\mathbf{x}(1) = \mathbf{A}^2\mathbf{x}(0)$$

and we readily deduce that in general

$$\mathbf{x}(k) = \mathbf{A}^k\mathbf{x}(0) \quad (k \geq 0) \quad (6.76)$$

Equation (6.76) represents the solution of (6.75), and is analogous to (1.77) for the continuous-time case. We define the **transition matrix** $\Phi(k)$ of the discrete-time system (6.75) by

$$\Phi(k) = \mathbf{A}^k$$

and it is the unique matrix satisfying

$$\Phi(k+1) = \mathbf{A}\Phi(k), \quad \Phi(0) = \mathbf{I}$$

where \mathbf{I} is the identity matrix.

Since \mathbf{A} is a constant matrix, the methods discussed in Section 1.7 are applicable for evaluating the transition matrix. From (1.31a),

$$\mathbf{A}^k = \alpha_0(k)\mathbf{I} + \alpha_1(k)\mathbf{A} + \alpha_2(k)\mathbf{A}^2 + \cdots + \alpha_{n-1}(k)\mathbf{A}^{n-1} \quad (6.77)$$

where, using (1.31b), the $\alpha_i(k)$ ($k = 0, \dots, n - 1$) are obtained by solving simultaneously the n equations

$$\lambda_j^k = \alpha_0(k) + \alpha_1(k)\lambda_j + \alpha_2(k)\lambda_j^2 + \cdots + \alpha_{n-1}(k)\lambda_j^{n-1} \quad (6.78)$$

where λ_j ($j = 1, 2, \dots, n$) are the eigenvalues of \mathbf{A} . As in Section 1.7, if \mathbf{A} has repeated eigenvalues then derivatives of λ^k with respect to λ will have to be used. The method for determining \mathbf{A}^k is thus very similar to that used for evaluating $e^{\mathbf{A}t}$ in Section 1.10.3.

Example 6.29

Obtain the response of the second-order unforced discrete-time system

$$\mathbf{x}(k+1) = \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ -1 & \frac{1}{3} \end{bmatrix} \mathbf{x}(k)$$

subject to $\mathbf{x}(0) = [1 \quad 1]^T$.

Solution In this case the system matrix is

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

having eigenvalues $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = \frac{1}{3}$. Since \mathbf{A} is a 2×2 matrix, it follows from (6.77) that

$$\mathbf{A}^k = \alpha_0(k)\mathbf{I} + \alpha_1(k)\mathbf{A}$$

with $\alpha_0(k)$ and $\alpha_1(k)$ given from (6.78),

$$\lambda_j^k = \alpha_0(k) + \alpha_1(k)\lambda_j \quad (j = 1, 2)$$

Solving the resulting two equations

$$(\frac{1}{2})^k = \alpha_0(k) + (\frac{1}{2})\alpha_1(k), \quad (\frac{1}{3})^k = \alpha_0(k) + (\frac{1}{3})\alpha_1(k)$$

for $\alpha_0(k)$ and $\alpha_1(k)$ gives

$$\alpha_0(k) = 3(\frac{1}{3})^k - 2(\frac{1}{2})^k, \quad \alpha_1(k) = 6[(\frac{1}{2})^k - (\frac{1}{3})^k]$$

Thus the transition matrix is

$$\Phi(k) = \mathbf{A}^k = \begin{bmatrix} (\frac{1}{2})^k & 0 \\ 6[(\frac{1}{3})^k - (\frac{1}{2})^k] & (\frac{1}{3})^k \end{bmatrix}$$

Note that $\Phi(0) = \mathbf{I}$, as required.

Then from (6.76) the solution of the unforced system is

$$\mathbf{x}(k+1) = \begin{bmatrix} (\frac{1}{2})^k & 0 \\ 6[(\frac{1}{3})^k - (\frac{1}{2})^k] & (\frac{1}{3})^k \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} (\frac{1}{2})^k \\ 7(\frac{1}{3})^k - 6(\frac{1}{2})^k \end{bmatrix}$$

Having determined the solution of the unforced system, it can be shown that the solution of the state equation (6.73a) for the forced system with input $\mathbf{u}(k)$, analogous to the solution given in (1.78) for the continuous-time system

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$

is

$$\mathbf{x}(k) = \mathbf{A}^k \mathbf{x}(0) + \sum_{j=0}^{k-1} \mathbf{A}^{k-j-1} \mathbf{B} \mathbf{u}(j) \quad (6.79)$$

Having obtained the solution of the state equation, the system output or response $\mathbf{y}(k)$ is obtained from (6.73b) as

$$\mathbf{y}(k) = \mathbf{C}\mathbf{A}^k \mathbf{x}(0) + \mathbf{C} \sum_{j=0}^{k-1} \mathbf{A}^{k-j-1} \mathbf{B} \mathbf{u}(j) + \mathbf{D} \mathbf{u}(k) \quad (6.80)$$

In Section 5.4.1 we saw how the Laplace transform could be used to solve the state-space equations in the case of continuous-time systems. In a similar manner, z transforms can be used to solve the equations for discrete-time systems.

Defining $\mathcal{Z}\{\mathbf{x}(k)\} = \mathbf{X}(z)$ and $\mathcal{Z}\{\mathbf{u}(k)\} = \mathbf{U}(z)$ and taking z transforms throughout in the equation

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{B}\mathbf{u}(k)$$

gives

$$z\mathbf{X}(z) - z\mathbf{x}(0) = \mathbf{A}\mathbf{X}(z) + \mathbf{B}\mathbf{U}(z)$$

which, on rearranging, gives

$$(z\mathbf{I} - \mathbf{A})\mathbf{X}(z) = z\mathbf{x}(0) + \mathbf{B}\mathbf{U}(z)$$

where \mathbf{I} is the identity matrix. Premultiplying by $(z\mathbf{I} - \mathbf{A})^{-1}$ gives

$$\mathbf{X}(z) = z(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{x}(0) + (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(z) \quad (6.81)$$

Taking inverse z transforms gives the response as

$$\mathbf{x}(k) = \mathcal{Z}^{-1}\{\mathbf{X}(z)\} = \mathcal{Z}^{-1}\{z(z\mathbf{I} - \mathbf{A})^{-1}\}\mathbf{x}(0) + \mathcal{Z}^{-1}\{(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(z)\} \quad (6.82)$$

which corresponds to (5.51) in the continuous-time case.

On comparing the solution (6.82) with that given in (6.79), we see that the transition matrix $\Phi(t) = \mathbf{A}^k$ may also be written in the form

$$\Phi(t) = \mathbf{A}^k = \mathcal{Z}^{-1}\{z(z\mathbf{I} - \mathbf{A})^{-1}\}$$

This is readily confirmed from (6.81), since on expanding $z(z\mathbf{I} - \mathbf{A})^{-1}$ by the binomial theorem, we have

$$\begin{aligned} z(z\mathbf{I} - \mathbf{A})^{-1} &= \mathbf{I} + \frac{\mathbf{A}}{z} + \frac{\mathbf{A}^2}{z^2} + \cdots + \frac{\mathbf{A}^r}{z^r} + \cdots \\ &= \sum_{r=0}^{\infty} \frac{\mathbf{A}^r}{z^r} = \mathcal{Z}\{\mathbf{A}^k\} \end{aligned}$$

Example 6.30

Using the z -transform approach, obtain an expression for the state $\mathbf{x}(k)$ of the system characterized by the state equation

$$\mathbf{x}(k+1) = \begin{bmatrix} 2 & 5 \\ -3 & -6 \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mathbf{u}(k) \quad (k \geq 0)$$

when the input is the unit step function

$$\mathbf{u}(k) = \begin{cases} 0 & (k < 0) \\ 1 & (k \geq 0) \end{cases}$$

and subject to the initial condition $\mathbf{x}(0) = [1 \ -1]^T$.

Solution In this case

$$\mathbf{A} = \begin{bmatrix} 2 & 5 \\ -3 & -6 \end{bmatrix} \quad \text{so} \quad z\mathbf{I} - \mathbf{A} = \begin{bmatrix} z-2 & -5 \\ 3 & z+6 \end{bmatrix}$$

giving

$$(z\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{(z+1)(z+3)} \begin{bmatrix} z+6 & 5 \\ -3 & z-2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{5}{2} \frac{z}{z+1} - \frac{3}{2} \frac{z}{z+3} & \frac{5}{2} \frac{z}{z+1} - \frac{5}{2} \frac{z}{z+3} \\ \frac{-3}{2} \frac{z}{z+1} + \frac{3}{2} \frac{z}{z+3} & \frac{-3}{2} \frac{z}{z+1} + \frac{5}{2} \frac{z}{z+3} \end{bmatrix}$$

Then

$$\mathcal{Z}^{-1}\{z(z\mathbf{I} - \mathbf{A})^{-1}\} = \mathcal{Z}^{-1} \begin{bmatrix} \frac{5}{2} \frac{z}{z+1} - \frac{3}{2} \frac{z}{z+3} & \frac{5}{2} \frac{z}{z+1} - \frac{5}{2} \frac{z}{z+3} \\ -\frac{3}{2} \frac{z}{z+1} + \frac{3}{2} \frac{z}{z+3} & -\frac{3}{2} \frac{z}{z+1} + \frac{5}{2} \frac{z}{z+3} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{5}{2}(-1)^k - \frac{3}{2}(-3)^k & \frac{5}{2}(-1)^k - \frac{5}{2}(-3)^k \\ -\frac{3}{2}(-1)^k + \frac{3}{2}(-3)^k & -\frac{3}{2}(-1)^k + \frac{5}{2}(-3)^k \end{bmatrix}$$

so that, with $\mathbf{x}(0) = [1 \ -1]^T$, the first term in the solution (6.82) becomes

$$\mathcal{Z}^{-1}\{z(z\mathbf{I} - \mathbf{A})^{-1}\}\mathbf{x}(0) = \begin{bmatrix} (-3)^k \\ -(-3)^k \end{bmatrix} \tag{6.83}$$

Since $U(z) = \mathcal{Z}\{u(k)\} = z/(z-1)$,

$$(z\mathbf{I} - \mathbf{A})^{-1} \mathbf{B} U(z) = \frac{1}{(z+1)(z+3)} \begin{bmatrix} z+6 & 5 \\ -3 & z-2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \frac{z}{z-1}$$

$$= \frac{z}{(z-1)(z+1)(z+3)} \begin{bmatrix} z+11 \\ z-5 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{3}{2} \frac{z}{z-1} - \frac{5}{2} \frac{z}{z+1} + \frac{z}{z+3} \\ -\frac{1}{2} \frac{z}{z-1} + \frac{3}{2} \frac{z}{z+1} - \frac{z}{z+3} \end{bmatrix}$$

so that the second term in the solution (6.82) becomes

$$\mathcal{Z}^{-1}\{(zI - \mathbf{A})^{-1}\mathbf{B}U(z)\} = \begin{bmatrix} \frac{3}{2} - \frac{5}{2}(-1)^k + (-3)^k \\ -\frac{1}{2} + \frac{3}{2}(-1)^k - (-3)^k \end{bmatrix} \quad (6.84)$$

Combining (6.83) and (6.84), the response $\mathbf{x}(k)$ is given by

$$\mathbf{x}(k) = \begin{bmatrix} \frac{3}{2} - \frac{5}{2}(-1)^k + 2(-3)^k \\ -\frac{1}{2} + \frac{3}{2}(-1)^k - 2(-3)^k \end{bmatrix}$$

Having obtained an expression for a system's state $\mathbf{x}(t)$, its output, or response, $y(t)$ may be obtained from the linear transformation (6.73b).

6.8.3 Exercises



Check your answers using MATLAB or MAPLE whenever possible.

- 30 Use z transforms to determine \mathbf{A}^k for the matrices

$$(a) \begin{bmatrix} 0 & 1 \\ 4 & 0 \end{bmatrix} \quad (b) \begin{bmatrix} -1 & 3 \\ 3 & -1 \end{bmatrix} \quad (c) \begin{bmatrix} -1 & 1 \\ 0 & -1 \end{bmatrix}$$

- 31 Solve the discrete-time system specified by

$$x(k+1) = -7x(k) + 4y(k)$$

$$y(k+1) = -8x(k) + y(k)$$

with $x(0) = 1$ and $y(0) = 2$, by writing it in the form $\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k)$. Use your answer to calculate $\mathbf{x}(1)$ and $\mathbf{x}(2)$, and check your answers by calculating $x(1)$, $y(1)$, $x(2)$, $y(2)$ directly from the given difference equations.

- 32 Using the z -transform approach, obtain an expression for the state $\mathbf{x}(k)$ of the system characterized by the state equation

$$x(k+1) = \begin{bmatrix} 0 & 1 \\ -0.16 & -1 \end{bmatrix}x(k) + \begin{bmatrix} 1 \\ 1 \end{bmatrix}u(k)$$

when the input is the unit step function

$$u(k) = \begin{cases} 0 & (k < 0) \\ 1 & (k \geq 0) \end{cases}$$

and subject to the initial condition $\mathbf{x}(0) = [1 \ -1]^T$.

The difference equation

$$y(k+2) = y(k+1) + y(k)$$

with $y(0) = 0$, and $y(1) = 1$, generates the **Fibonacci sequence** $\{y(k)\}$, which occurs in many practical situations. Taking $x_1(k) = y(k)$ and $x_2(k) = y(k+1)$, express the difference equation in state-space form and hence obtain a general expression for $y(k)$.

Show that as $k \rightarrow \infty$ the ratio $y(k+1)/y(k)$ tends to the constant $\frac{1}{2}(\sqrt{5} + 1)$. This is the so-called **Golden Ratio**, which has intrigued mathematicians for centuries because of its strong influence on art and architecture. The Golden Rectangle, that is one whose two sides are in this ratio, is one of the most visually satisfying of all geometric forms.

6.9

Discretization of continuous-time state-space models

In Sections 1.10 and 5.6 we considered the solutions of the continuous-time state-space model

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad (6.85a)$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \quad (6.85b)$$

If we wish to compute the state $\mathbf{x}(t)$ digitally then we must first approximate the continuous model by a discrete-time state-space model of the form

$$\mathbf{x}[(k+1)T] = \mathbf{G}\mathbf{x}(kT) + \mathbf{H}\mathbf{u}(kT) \quad (6.86a)$$

$$\mathbf{y}(kT) = \mathbf{C}\mathbf{x}(kT) \quad (6.86b)$$

Thus we are interested in determining matrices \mathbf{G} and \mathbf{H} such that the responses to the discrete-time model (6.86) provide a good approximation to sampled-values of the continuous-time model (6.85). We assume that sampling occurs at equally spaced sampling instances $t = kT$, where $T > 0$ is the sampling interval. For clarification we use the notation $\mathbf{x}(kT)$ and $\mathbf{x}[(k+1)T]$ instead of k and $(k+1)$ as in (6.73).

6.9.1 Euler's method

A simple but crude method of determining \mathbf{G} and \mathbf{H} is based on Euler's method considered in Section 10.6 of MEM. Here the derivative of the state is approximated by

$$\dot{\mathbf{x}}(t) \cong \frac{\mathbf{x}(t+T) - \mathbf{x}(T)}{T}$$

which on substituting in (6.85a) gives

$$\frac{\mathbf{x}(t+T) - \mathbf{x}(t)}{T} \cong \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$

which reduces to

$$\mathbf{x}(t+T) \cong (TA + I)\mathbf{x}(t) + TB\mathbf{u}(t) \quad (6.87)$$

Since t is divided into equally spaced sampling intervals of duration T we take $t = kT$, where k is the integer index $k = 0, 1, 2, \dots$, so that (6.87) becomes

$$\mathbf{x}[(k+1)T] \cong (TA + I)\mathbf{x}(kT) + TB\mathbf{u}(kT) \quad (6.88)$$

Defining

$$\mathbf{G} = \mathbf{G}_1 = (TA + I) \text{ and } \mathbf{H} = \mathbf{H}_1 = TB \quad (6.89)$$

(6.86) then becomes the approximating discrete-time model to the continuous-time model (6.85). This approach to discretization is known as **Euler's method** and simply involves a sequential series of calculations.

Example 6.31

Consider the system modelled by the second-order differential equation

$$\ddot{y}(t) + 3\dot{y}(t) + 2y = 2u(t)$$

- (a) Choosing the state-vector $\mathbf{x} = [y \ \dot{y}]^T$ express this in a state-space form.
- (b) Using Euler's method, determine the approximating discrete-time state-space model.
- (c) Illustrate by plotting the responses $y(t)$, for both the exact continuous response and the discretized responses, for a step input $u(t) = 1$ and zero initial conditions, taking $T = 0.2$.

Solution

- (a) Since $x_1 = y$, $x_2 = \dot{y}$ we have that

$$\dot{x}_1 = \dot{y} = x_2$$

$$\dot{x}_2 = \ddot{y} = -2x_1 - 3x_2 + 2u$$

so the state-space model is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} u(t)$$

$$y = [1 \ 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

- (b) From (6.89)

$$\mathbf{G}_1 = T\mathbf{A} + \mathbf{I} = \begin{bmatrix} 1 & T \\ -2T & -3T+1 \end{bmatrix}$$

$$\mathbf{H}_1 = T\mathbf{B} = \begin{bmatrix} 0 \\ 2T \end{bmatrix}$$

so the discretized state-space model is

$$\begin{bmatrix} x_1[(k+1)T] \\ x_2[(k+1)T] \end{bmatrix} = \begin{bmatrix} 1 & T \\ -2T & -3T+1 \end{bmatrix} \begin{bmatrix} x_1(kT) \\ x_2(kT) \end{bmatrix} + \begin{bmatrix} 0 \\ 2T \end{bmatrix} u(kT)$$

$$y(kT) = [1 \ 0] \begin{bmatrix} x_1(kT) \\ x_2(kT) \end{bmatrix}$$



- (c) Using the MATLAB commands:

```
A = [0,1;-2,-3]; B = [0;2]; C = [1,0];
K = 0;
for T = 0.2
k = k + 1;
G1 = [1,T;-2*T,-3*T+1]; H1 = [0;2*T];
```

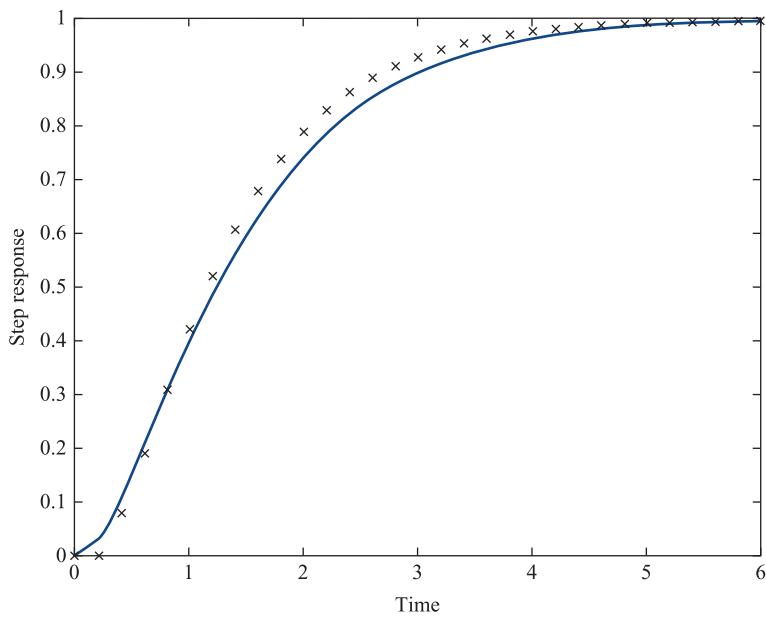
```

T = T*[0:30];
y = step(A,B,C,0,1,t); yd = dstep(G1,H1,C,0,1,31);
plot(t,y,t,yd,'x')
end

```

Step responses for both the continuous model and the Euler discretized model are displayed in Figure 6.20 with ‘ \times ’ denoting the discretized response.

Figure 6.20
The continuous model
and Euler discretized
model of Example 6.31.



6.9.2 Step-invariant method

To determine the matrices \mathbf{G} and \mathbf{H} in the discrete-time model (6.86), use is made of the explicit solution to the state equation (6.85a). From (1.78) the solution of (6.85a) is given by

$$\mathbf{x}(t) = e^{A(t-t_0)} \mathbf{x}(t_0) + \int_{t_0}^t e^{A(t-\tau_1)} \mathbf{B} \mathbf{u}(\tau_1) d\tau_1 \quad (6.90)$$

Taking $t_0 = kT$ and $t = (k+1)T$ in (6.90) gives

$$\mathbf{x}[(k+1)T] = e^{AT} \mathbf{x}(kT) + \int_{kT}^{(k+1)T} e^{A[(k+1)T-\tau_1]} \mathbf{B} \mathbf{u}(\tau_1) d\tau_1$$

Making the substitution $\tau = \tau_1 - kT$ in the integral gives

$$\mathbf{x}[(k+1)T] = e^{AT} \mathbf{x}(kT) + \int_0^T e^{A(T-\tau)} \mathbf{B} \mathbf{u}(kT + \tau) d\tau \quad (6.91)$$

The problem now is: How do we approximate the integral in (6.91)? The simplest approach is to assume that all components of $\mathbf{u}(t)$ are constant over intervals between two consecutive sampling instances so

$$\mathbf{u}(kT + \tau) = \mathbf{u}(kT), \quad 0 \leq \tau \leq T, \quad k = 0, 1, 2, \dots$$

The integral in (6.91) then becomes

$$\left[\int_0^T e^{A(T-\tau)} \mathbf{B} d\tau \right] \mathbf{u}(kT)$$

Defining

$$\mathbf{G} = e^{AT} \quad (6.92a)$$

$$\text{and } \mathbf{H} = \int_0^T e^{(T-\tau)} \mathbf{B} d\tau = \int_0^T e^{At} \mathbf{B} dt, \quad \text{using substitution } t = (T - \tau) \quad (6.92b)$$

then (6.91) becomes the discretized state equation

$$\mathbf{x}[(k+1)T] = \mathbf{Gx}(kT) + \mathbf{Hu}(kT) \quad (6.93)$$

The discretized form (6.93) is frequently referred to as the **step-invariant method**.

Comments

- From Section 5.6.1 we can determine \mathbf{G} using the result

$$e^{At} = \mathcal{L}^{-1}\{(sI - A)^{-1}\} \quad (6.94)$$

- If the state matrix \mathbf{A} is invertible then from (1.34)

$$\mathbf{H} = \int_0^T e^{At} \mathbf{B} d\tau = \mathbf{A}^{-1}(\mathbf{G} - \mathbf{I})\mathbf{B} = (\mathbf{G} - \mathbf{I})\mathbf{A}^{-1}\mathbf{B} \quad (6.95)$$

- Using the power series expansion of e^{At} given in (1.24) we can express \mathbf{G} and \mathbf{H} as the power series

$$\mathbf{G} = \mathbf{I} + T\mathbf{A} + \frac{T^2\mathbf{A}^2}{2!} + \dots = \sum_{r=0}^{\infty} \frac{T^r \mathbf{A}^r}{r!} \quad (6.96)$$

$$\mathbf{H} = (T\mathbf{I} + \frac{T^2\mathbf{A}}{2!} + \dots)\mathbf{B} = \left(\sum_{r=1}^{\infty} \frac{T^r \mathbf{A}^{r-1}}{r!} \right) \mathbf{B} \quad (6.97)$$

We can approximate \mathbf{G} and \mathbf{H} by neglecting higher-order terms in T . In the particular case when we neglect terms of order two or higher in T results (6.97) give

$$\mathbf{G} = \mathbf{I} + T\mathbf{A} \text{ and } \mathbf{H} = T\mathbf{B}$$

which corresponds to Euler's discretization.

Example 6.32

Using the step-invariant method, obtain the discretized form of the state equation for the continuous-time system

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} u(t)$$

considered in Example 6.31. Plot the response $y(kT) = [1 \ 0]\mathbf{x}(kT)$, for a step input $u(t) = 1$ and zero initial conditions, taking $T = 0.2$.

Solution Using (6.93) $\mathbf{G} = e^{AT}$ and $\mathbf{H} = \int_0^T e^{At} \mathbf{B} dt$. From (6.94)

$$\begin{aligned} \mathbf{G} = \mathcal{Z}^{-1}\{(sI - A)^{-1}\} &= \mathcal{Z}^{-1}\left\{\frac{1}{\Delta} \begin{bmatrix} s+3 & 1 \\ -2 & 5 \end{bmatrix}\right\}, \quad \Delta = (s+2)(s+1) \\ &= \mathcal{Z}^{-1} \begin{bmatrix} -\frac{1}{s+2} + \frac{2}{s+1} & -\frac{1}{s+2} + \frac{1}{s+1} \\ \frac{2}{s+2} - \frac{2}{s+1} & \frac{2}{s+2} - \frac{1}{s+1} \end{bmatrix} \end{aligned}$$

so that

$$\mathbf{G} = e^{AT} = \begin{bmatrix} -e^{-2T} + 2e^{-T} & -e^{-2T} + e^{-T} \\ 2e^{-2T} - 2e^{-T} & 2e^{-2T} - e^{-T} \end{bmatrix}$$

and

$$\begin{aligned} \mathbf{H} &= \int_0^T e^{At} \mathbf{B} dt = \left[\begin{bmatrix} \frac{1}{2}e^{-2t} - 2e^{-t} & \frac{1}{2}e^{-2t} - e^{-t} \\ -e^{-2t} + 2e^{-t} & -e^{-2t} + e^{-t} \end{bmatrix} \right]_0^T \begin{bmatrix} 0 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} e^{-2T} - 2e^{-T} + 1 \\ -2e^{-2T} + 2e^{-T} \end{bmatrix} \end{aligned}$$

Thus, the discrete form of the state equation is

$$\mathbf{x}[(k+1)T] = \begin{bmatrix} -e^{-2T} + 2e^{-T} & -e^{-2T} + e^{-T} \\ 2e^{-2T} - 2e^{-T} & 2e^{-2T} - e^{-T} \end{bmatrix} \mathbf{x}(kT) + \begin{bmatrix} e^{-2T} - 2e^{-T} + 1 \\ -2e^{-2T} + e^{-T} \end{bmatrix} u(kT)$$

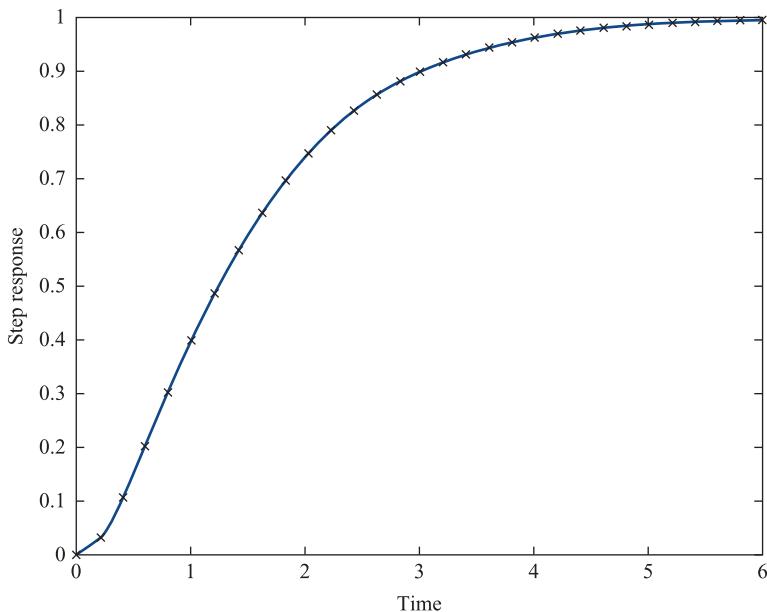
In the particular case $T = 0.2$ the state equation is

$$\mathbf{x}[(k+1)0.2] = \begin{bmatrix} 0.9671 & 0.1484 \\ -0.2968 & 0.5219 \end{bmatrix} \mathbf{x}(k0.2) + \begin{bmatrix} 0.0329 \\ -0.2968 \end{bmatrix} u(k0.2)$$

Using MATLAB step responses for both the continuous-time model and the discretized step-invariant model are displayed in Figure 6.21, with ‘x’ denoting the discretized response.

Figure 6.21

The continuous-time model and discretized step-invariant model of Example 6.32.



For a given value of T the matrices \mathbf{G} and \mathbf{H} may be determined by the step-invariant method using the MATLAB function `c2d` (continuous to discrete). Thus, for the system of Example 6.32 with $T = 0.2$, the commands

```
A = [0, 1; -2, -3];
B = [0; 2];
[G, H] = c2d(A, B, 0.2)
```

return

```
G = 0.9671 0.1484
    -0.2968 0.5219
H = 0.0329
    -0.2968
```

which checks with the answers given in Example 6.32.

6.9.3 Exercises

- 34 Using the step-invariant method obtain the discretized form of the continuous-time state-equation

$$\dot{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} = \begin{bmatrix} 0 & 1 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$

Check your answer using MATLAB for the particular case when the sampling period is $T = 1$.

- 35 An *LCR* circuit, with $L = C = R = 1$, may be modelled by the continuous-time state-space model

$$\dot{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t)$$

$$y = [1 \quad 0] \mathbf{x}$$

- (a) Determine the Euler form of the discretized state-space model.
(b) Determine the discretized state-space model using the step-invariant method.
(Hint: Use (6.95) to determine the \mathbf{H} matrix.)

- (c) Using MATLAB plot, for each of the three models, responses to a unit step input $u(t) = 1$ with zero initial conditions, taking the sampling period $T = 0.1$.

- 36 A linear continuous-time system is characterized by the state matrix

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

- (a) Show that the system is stable.
 (b) Show that the state matrix of the corresponding Euler discrete-time system is

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

- (c) Show that stability of the discretized system requires $T < 1$.

- 37 A simple continuous-time model of a production and inventory control system may be represented by the state-space model

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} \\ &= \begin{bmatrix} -1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} k_1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \end{aligned}$$

where $x_1(t)$ represents the actual production rate and $x_2(t)$ represents the current inventory level; $u_1(t)$

represents the scheduled production rate, $u_2(t)$
 represents the sales rate and k_1 is a constant gain factor.

- (a) Determine, using the step-invariant method, the discretized form of the model. Express the model in the particular case when the sampling period $T = 1$.
 (b) Suppose the production schedule is determined by the feedback policy

$$u_1(kT) = k_c - x_2(kT)$$

where k_c is the desired inventory level. The system is originally in equilibrium with $x_1(0)$ equal to the sales rate and $x_2(0) = k_c$. At time $t = 0$ the sales rate suddenly increases by 10%; that is, $u_2(t) = 1.1x_1(0)$ for $t \geq 0$. Find the resulting discrete-time state model, with sampling rate $T = 1$ and taking $k_1 = \frac{3}{16}$.

- (c) Find the response of the given continuous-time model, subject to the same feedback control policy

$$u_1(t) = k_c - x_2(t)$$

and the same initial conditions.

The exercise may be extended to include simulation studies using MATLAB.

(This exercise is adapted from an illustrative problem in W. L. Brogan, *Modern Control Theory*, second edition, Prentice-Hall, 1985.)

6.10 Engineering application: design of discrete-time systems

An important development in many areas of modern engineering is the replacement of analogue devices by digital ones. Perhaps the most widely known example is the compact disc player, in which mechanical transcription followed by analogue signal processing has been superseded by optical technology and digital signal processing. Also, as stated in the introduction, DVD players and digital radios have set new standards in home entertainment. There are other examples in many fields of engineering, particularly where automatic control is employed.

6.10.1 Analogue filters

At the centre of most signal processing applications are **filters**. These have the effect of changing the spectrum of input signals; that is, attenuating components of signals by an amount depending on the frequency of the component. For example, an analogue **ideal low-pass filter** passes without attenuation all signal components at frequencies less than a critical frequency $\omega = \omega_c$ say. The amplitude of the frequency response $|G(j\omega)|$ (see Section 5.5) of such an ideal filter is shown in Figure 6.22.

One class of analogue filters whose frequency response approximates that of the ideal low-pass filter comprises those known as **Butterworth filters**. As well as having ‘good’ characteristics, these can be implemented using a network as illustrated in Figure 6.23 for the second-order filter.

It can be shown (see M. J. Chapman, D. P. Goodall and N. C. Steele, *Signal Processing in Electronic Communication*, Chichester, Horwood Publishing, 1997) that the transfer function $G_n(s)$ of the n th-order filter is

$$G_n(s) = \frac{1}{B_n(x)} \quad \text{where} \quad B_n(x) = \sum_{k=0}^n a_k x^k$$

with

$$x = \frac{s}{\omega_c}, \quad a_k = \prod_{r=1}^k \frac{\cos((r-1)\alpha)}{\sin r\alpha}, \quad \alpha = \frac{\pi}{2n}$$

Using these relations, it is readily shown that

$$G_2(s) = \frac{\omega_c^2}{s^2 + \sqrt{2}\omega_c s + \omega_c^2} \tag{6.98}$$

$$G_3(s) = \frac{\omega_c^3}{s^3 + 2\omega_c s^2 + 2\omega_c^2 s + \omega_c^3} \tag{6.99}$$

and so on. On sketching the amplitudes of the frequency responses $G_n(j\omega)$, it becomes apparent that increasing n improves the approximation to the response of the ideal low-pass filter of Figure 6.22.

Figure 6.22
Amplitude response
for an ideal low-pass
filter.

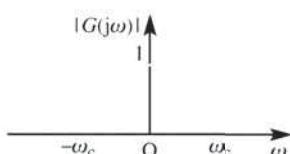
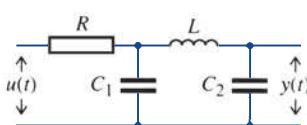


Figure 6.23
LCR network for
implementing a
second-order
Butterworth filter.



6.10.2 Designing a digital replacement filter

Suppose that we now wish to design a discrete-time system, to operate on samples taken from an input signal, that will operate in a similar manner to a Butterworth filter. We shall assume that the input signal $u(t)$ and the output signal $y(t)$ of the analogue filter are both sampled at the same intervals T to generate the input sequence $\{u(kT)\}$ and the output sequence $\{y(kT)\}$ respectively. Clearly, we need to specify what is meant by ‘operate in a similar manner’. In this case, we shall select as our design strategy a method that matches the impulse response sequence of the digital design with a sequence of samples, drawn at the appropriate instants T from the impulse response of an analogue ‘prototype’. We shall select the prototype from one of the Butterworth filters discussed in Section 6.10.1, although there are many other possibilities.

Let us select the first-order filter, with cut-off frequency ω_c , as our prototype. Then the first step is to calculate the impulse response of this filter. The Laplace transfer function of the filter is

$$G(s) = \frac{\omega_c}{s + \omega_c}$$

So, from (5.43), the impulse response is readily obtained as

$$h(t) = \omega_c e^{-\omega_c t} \quad (t \geq 0) \quad (6.100)$$

Next, we sample this response at intervals T to generate the sequence

$$\{h(kT)\} = \{\omega_c e^{-\omega_c kT}\}$$

which on taking the z transform, gives

$$\mathcal{Z}\{h(kT)\} = H(z) = \omega_c \frac{z}{z - e^{-\omega_c T}}$$

Finally, we choose $H(z)$ to be the transfer function of our digital system. This means simply that the input–output relationship for the design of the digital system will be

$$Y(z) = H(z)U(z)$$

where $Y(z)$ and $U(z)$ are the z transforms of the output and input sequences $\{y(kT)\}$ and $\{u(kT)\}$ respectively. Thus we have

$$Y(z) = \omega_c \frac{z}{z - e^{-\omega_c T}} U(z) \quad (6.101)$$

Our digital system is now defined, and we can easily construct the corresponding difference equation model of the system as

$$(z - e^{-\omega_c T})Y(z) = \omega_c z U(z)$$

that is

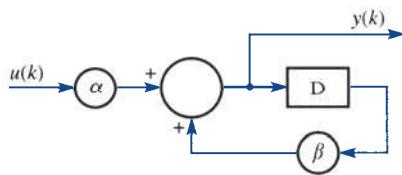
$$zY(z) - e^{-\omega_c T} Y(z) = \omega_c z U(z)$$

Under the assumption of zero initial conditions, we can take inverse transforms to obtain the first-order difference equation model

$$y(k+1) - e^{-\omega_c T} y(k) = \omega_c u(k+1) \quad (6.102)$$

A block diagram implementation of (6.102) is shown in Figure 6.24.

Figure 6.24 Block diagram for the digital replacement filter, $\alpha = k\omega_c$, $\beta = e^{-\alpha_c t}$.



6.10.3 Possible developments

The design method we have considered is called the **impulse invariant technique**, and is only one of many available. The interested reader may develop this study in various ways:

- (1) Write a computer program to evaluate the sequence generated by (6.102) with $\omega_c = 1$, and compare with values obtained at the sampling instants for the impulse response (6.100) of the prototype analogue filter.
- (2) Repeat the design process for the second-order Butterworth filter.
- (3) By setting $s = j\omega$ in the Laplace transfer function of the prototype, and $z = e^{j\omega T}$ in the z transfer function of the digital design, compare the amplitude of the frequency responses in both cases. For an explanation of the results obtained, see Chapter 8.
- (4) An alternative design strategy is to replace s in the Laplace transfer function with

$$\frac{2z - 1}{Tz + 1}$$

(this is a process that makes use of the trapezoidal method of approximate integration). Design alternative digital filters using this technique, which is commonly referred to as the **Tustin** (or **bilinear transform**) method (see Section 6.11.3).

- (5) Show that filters designed using either of these techniques will be stable provided that the prototype design is itself stable.

6.11 Engineering application: the delta operator and the \mathcal{D} transform

6.11.1 Introduction

In recent years, sampling rates for digital systems have increased many-fold, and traditional model formulations based on the z transform have produced unsatisfactory

results in some applications. It is beyond the scope of this text to describe this situation in detail, but it is possible to give a brief introduction to the problem and to suggest an approach to the solution. For further details see R. M. Middleton and G. C. Goodwin, *Digital Control and Estimation, A Unified Approach* (Englewood Cliffs, NJ, Prentice Hall, 1990) or W. Forsythe and R. M. Goodall, *Digital Control* (London, Macmillan, 1991). The contribution of Colin Paterson to the development of this application is gratefully acknowledged.

6.11.2 The q or shift operator and the δ operator

In the time domain we define the shift operator q in terms of its effect on a sequence $\{x_k\}$ as

$$q\{x_k\} = \{x_{k+1}\}$$

That is, the effect of the shift operator is to shift the sequence by one position, so that the k th term of the new sequence is the $(k + 1)$ th term of the original sequence. It is then possible to write the difference equation

$$y_{k+2} + 2y_{k+1} + 5y_k = u_{k+1} - u_k$$

as

$$q^2y_k + 2qy_k + 5y_k = qu_k - u_k$$

or

$$(q^2 + 2q + 5)y_k = (q - 1)u_k \quad (6.103)$$

Note that if we had taken the z transform of the difference equation, with an initially quiescent system, we would have obtained

$$(z^2 + 2z + 5)Y(z) = (z - 1)U(z)$$

We see at once the correspondence between the time-domain q operator and the z -transform operator \mathcal{Z} .

The next step is to introduce the δ operator, defined as

$$\delta = \frac{q - 1}{\Delta}$$

where Δ has the dimensions of time and is often chosen as the sampling period T . Note that

$$\delta y_k = \frac{(q - 1)y_k}{\Delta} = \frac{y_{k+1} - y_k}{\Delta}$$

so that if $\Delta = T$ then, in the limit of rapid sampling,

$$\delta y_k \approx \frac{dy}{dt}$$

Solving for q we see that

$$q = 1 + \Delta\delta$$

The difference equation (6.103) can thus be written as

$$((1 + \Delta\delta)^2 + 2(1 + \Delta\delta) + 5)y_k = [(1 + \Delta\delta) - 1]u_k$$

or

$$[(\Delta\delta)^2 + 4\Delta\delta + 8]y_k = \Delta\delta u_k$$

or, finally, as

$$\left(\delta^2 + \frac{4\delta}{\Delta} + \frac{8}{\Delta^2}\right)y_k = \frac{\delta}{\Delta}u_k$$

6.11.3 Constructing a discrete-time system model

So far, we have simply demonstrated a method of rewriting a difference equation in an alternative form. We now examine the possible advantages of constructing discrete-time system models using the δ operator. To do this, we consider a particular example, in which we obtain two different discrete-time forms of the second-order Butterworth filter, both based on the bilinear transform method, sometimes known as **Tustin's method**. This method has its origins in the trapezoidal approximation to the integration process; full details are given in M. J. Chapman, D. P. Goodall and N. C. Steele, *Signal Processing in Electronic Communication* (Chichester, Horwood Publishing, 1997).

The continuous-time second-order Butterworth filter with cut-off frequency $\omega_c = 1$ is modelled, as indicated by (6.98), by the differential equation

$$\frac{d^2y}{dt^2} + 1.414\ 21\frac{dy}{dt} + y = u(t) \quad (6.104)$$

where $u(t)$ is the input and $y(t)$ the filter response. Taking Laplace transforms throughout on the assumption of quiescent initial conditions, that is $y(0) = (dy/dt)(0) = 0$, we obtain the transformed equation

$$(s^2 + 1.414\ 21s + 1)Y(s) = U(s) \quad (6.105)$$

This represents a stable system, since the system poles, given by

$$s^2 + 1.414\ 21s + 1 = 0$$

are located at $s = -0.707\ 10 \pm j0.707\ 10$ and thus lie in the left half-plane of the complex s plane.

We now seek a discrete-time version of the differential equation (6.104). To do this, we first transform (6.105) into the z domain using the **bilinear transform method**, which involves replacing s by

$$\frac{2z - 1}{Tz + 1}$$

Equation (6.105) then becomes

$$\left[\frac{4}{T^2} \left(\frac{z-1}{z+1}\right)^2 + 1.414\ 21 \frac{2}{T} \left(\frac{z-1}{z+1}\right) + 1\right] Y(z) = U(z)$$

or

$$\begin{aligned} & [(\frac{1}{4}T^2 + 1.41421 \times \frac{1}{2}T + 4)z^2 + (\frac{1}{2}T^2 - 8)z + \frac{1}{4}T^2 - 1.41421 \times \frac{1}{2}T + 4]Y(z) \\ &= \frac{1}{4}T^2(z^2 + 2z + 1)U(z) \end{aligned} \quad (6.106)$$

We can now invert this transformed equation to obtain the time-domain model

$$\begin{aligned} & (\frac{1}{4}T^2 + 1.41421 \times \frac{1}{2}T + 4)y_{k+2} + (\frac{1}{2}T^2 - 8)y_{k+1} + (\frac{1}{4}T^2 - 1.41421 \times \frac{1}{2}T + 4)y_k \\ &= \frac{1}{4}T^2(u_{k+2} + 2u_{k+1} + u_k) \end{aligned} \quad (6.107)$$

For illustrative purposes we set $T = 0.1$ s in (6.107) to obtain

$$4.07321y_{k+2} - 7.99500y_{k+1} + 3.93179y_k = 0.02500(u_{k+2} + 2u_{k+1} + u_k)$$

Note that the roots of the characteristic equation have modulus of about 0.9825, and are thus quite close to the stability boundary.

When $T = 0.01$ s, (6.107) becomes

$$4.00710y_{k+2} - 7.99995y_{k+1} + 3.99295y_k = 0.00003(u_{k+2} + 2u_{k+1} + u_k)$$

In this case the roots have modulus of about 0.9982, and we see that increasing the sampling rate has moved them even closer to the stability boundary, and that *high accuracy in the coefficients is essential*, thus adding to the expense of implementation.

An alternative method of proceeding is to avoid the intermediate stage of obtaining the z -domain model (6.106) and to proceed directly to a discrete-time representation from (6.104), using the transformation

$$s \rightarrow \frac{2q-1}{Tq+1}$$

leading to the same result as in (6.107). Using the δ operator instead of the shift operator q , noting that $q = 1 + \Delta\delta$, we make the transformation

$$s \rightarrow \frac{2}{T} \frac{\Delta\delta}{2 + \Delta\delta}$$

or, if $T = \Delta$, the transformation

$$s \rightarrow \frac{2\delta}{2 + \Delta\delta}$$

in (6.105), which becomes

$$[\delta^2 + 1.41421 \times \frac{1}{2}\delta(2 + \Delta\delta) + \frac{1}{4}(2 + \Delta\delta)^2]y_k = \frac{1}{4}(2 + \Delta\delta)^2u_k$$

Note that in this form it is easy to see that in the limit as $\Delta \rightarrow 0$ (that is, as sampling becomes very fast) we regain the original differential equation model. Rearranging this equation, we have

$$\begin{aligned} & \left[\delta^2 + \frac{(1.41421 + \Delta)}{(1 + 1.41421 \times \frac{1}{2}\Delta + \frac{1}{4}\Delta^2)} \delta + \frac{1}{(1 + 1.41421 \times \frac{1}{2}\Delta + \frac{1}{4}\Delta^2)} \right] y_k \\ &= \frac{(2 + \Delta\delta)^2}{4(1 + 1.41421 \times \frac{1}{2}\Delta + \frac{1}{4}\Delta^2)} u_k \end{aligned} \quad (6.108)$$

In order to assess stability, it is helpful to introduce a transform variable γ associated with the δ operator. This is achieved by defining γ in terms of z as

$$\gamma = \frac{z - 1}{\Delta}$$

The region of stability in the z plane, $|z| < 1$, thus becomes

$$|1 + \Delta\gamma| < 1$$

or

$$\left| \frac{1}{\Delta} + \gamma \right| < \frac{1}{\Delta} \quad (6.109)$$

This corresponds to a circle in the γ domain, centre $(-1/\Delta, 0)$ and radius $1/\Delta$. As $\Delta \rightarrow 0$, we see that this circle expands in such a way that the stability region is the entire open left half-plane, and coincides with the stability region for continuous-time systems.

Let us examine the pole locations for the two cases previously considered, namely $T = 0.1$ and $T = 0.01$. With $\Delta = T = 0.1$, the characteristic equation has the form

$$\gamma^2 + 1.41092\gamma + 0.93178 = 0$$

with roots, corresponding to poles of the system, at $-0.70546 \pm j0.65887$. The centre of the circular stability region is now at $-1/0.1 = -10$, with radius 10, and these roots lie at a radial distance of about 9.3178 from this centre. Note that the distance of the poles from the stability boundary is just less than 0.7. The poles of the original continuous-time model were also at about this distance from the appropriate boundary, and we observe the sharp contrast from our first discretized model, when the discretization process itself moved the pole locations very close to the stability boundary. In that approach the situation became exacerbated when the sampling rate was increased, to $T = 0.01$, and the poles moved nearer to the boundary. Setting $T = 0.01$ in the new formulation, we find that the characteristic equation becomes

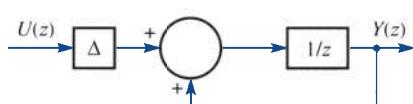
$$\gamma^2 + 1.41413\gamma + 0.99295 = 0$$

with roots at $-0.70706 \pm j0.70214$. The stability circle is now centred at -100 , with radius 100, and the radial distance of the poles is about 99.2954. Thus the distance from the boundary remains at about 0.7. Clearly, in the limit as $\Delta \rightarrow 0$, the pole locations become those of the continuous-time model, with the stability circle enlarging to become the entire left half of the complex γ plane.

6.11.4 Implementing the design

The discussion so far serves to demonstrate the utility of the δ operator formulation, but the problem of implementation of the design remains. It is possible to construct a δ^{-1} block based on delay or $1/z$ blocks, as shown in Figure 6.25. Systems can be realized

Figure 6.25
The δ^{-1} block.



using these structures in cascade or otherwise, and simulation studies have produced successful results. An alternative approach is to make use of the **state-space form** of the system model (see Section 6.18). We demonstrate this approach again for the case $T = 0.01$, when, with $T = \Delta = 0.01$, (6.108) becomes

$$\begin{aligned} & (\delta^2 + 1.414\ 13\delta + 0.992\ 95)y_k \\ &= (0.000\ 02\delta^2 + 0.009\ 30\delta + 0.992\ 95)u_k \end{aligned} \quad (6.110a)$$

Based on (6.110a) we are led to consider the equation

$$(\delta^2 + 1.414\ 13\delta + 0.992\ 95)p_k = u_k \quad (6.110b)$$

Defining the state variables

$$x_{1,k} = p_k, \quad x_{2,k} = \delta p_k$$

equation (6.110b) can be represented by the pair of equations

$$\begin{aligned} \delta x_{1,k} &= x_{2,k} \\ \delta x_{2,k} &= -0.992\ 95x_{1,k} - 1.414\ 13x_{2,k} + u_k \end{aligned}$$

Choosing

$$y_k = 0.992\ 95p_k + 0.009\ 30\delta p_k + 0.000\ 002\delta^2 p_k \quad (6.110c)$$

equations (6.110b) and (6.110c) are equivalent to (6.110a). In terms of the state variables we see that

$$y_k = 0.992\ 93x_{1,k} + 0.009\ 72x_{2,k} + 0.000\ 02u_k$$

Defining the vectors $\mathbf{x}_k = [x_{1,k} \quad x_{2,k}]^T$ and $\delta\mathbf{x}_k = [\delta x_{1,k} \quad \delta x_{2,k}]^T$, equation (6.111a) can be represented in matrix form as

$$\delta\mathbf{x}_k = \begin{bmatrix} 0 & 1 \\ -0.992\ 95 & -1.414\ 13 \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k \quad (6.111a)$$

with

$$y_k = [0.992\ 93 \quad 0.009\ 72] \mathbf{x}_k + 0.000\ 02u_k \quad (6.111b)$$

We now return to the q form to implement the system. Recalling that $\delta = (q - 1)/\Delta$, (6.111a) becomes

$$q\mathbf{x}_k = \mathbf{x}_{k+1} = \mathbf{x}_k + \Delta \left(\begin{bmatrix} 0 & 1 \\ -0.992\ 95 & -1.414\ 13 \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k \right) \quad (6.112)$$

with (6.111b) remaining the same and where $\Delta = 0.01$, in this case. Equations (6.112) and (6.111b) may be expressed in the vector-matrix form

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta[\mathbf{A}(\Delta)\mathbf{x}_k + \mathbf{b}u_k]$$

$$y = \mathbf{c}^T(\Delta)\mathbf{x}_k + d(\Delta)u_k$$

This matrix difference equation can now be implemented without difficulty using standard delay blocks, and has a form similar to the result of applying a simple Euler discretization of the original continuous-time model expressed in state-space form.

6.11.5 The \mathfrak{D} transform

In Section 6.11.3 we introduced a transform variable

$$\gamma = \frac{z - 1}{\Delta}$$

The purpose of this was to enable us to analyse the stability of systems described in the δ form. We now define a transform in terms of the z transform using the notation given by R. M. Middleton and G. C. Goodwin, *Digital Control and Estimation, A Unified Approach* (Englewood Cliffs, NJ, Prentice Hall, 1990). Let the sequence $\{f_k\}$ have z transform $F(z)$; then the new transform is given by

$$\begin{aligned} F'_\Delta(\gamma) &= F(z)|_{z=\Delta\gamma+1} \\ &= \sum_{k=0}^{\infty} \frac{f_k}{(1 + \Delta\gamma)^k} \end{aligned}$$

The \mathfrak{D} transform is formally defined as a slight modification to this form, as

$$\begin{aligned} \mathcal{D}(f_k) &= F_\Delta(\gamma) = \Delta F'_\Delta(\gamma) \\ &= \Delta \sum_{k=0}^{\infty} \frac{f_k}{(1 + \Delta\gamma)^k} \end{aligned}$$

The purpose of this modification is to permit the construction of a *unified theory of transforms* encompassing both continuous- and discrete-time models in the same structure. These developments are beyond the scope of the text, but may be pursued by the interested reader in the reference given above. We conclude the discussion with an example to illustrate the ideas. The ramp sequence $\{u_k\} = \{k\Delta\}$ can be obtained by sampling the continuous-time function $f(t) = t$ at intervals Δ . This sequence has z transform

$$U(z) = \frac{\Delta z}{(z - 1)^2}$$

and the corresponding \mathfrak{D} transform is then

$$\Delta U'_\Delta(\gamma) = \frac{1 + \Delta\gamma}{\gamma^2}$$

Note that on setting $\Delta = 0$ and $\gamma = s$ one recovers the Laplace transform of $f(t)$.

6.11.6 Exercises

- 38 A continuous-time system having input $y(t)$ and output $y(t)$ is defined by its transfer function

$$H(s) = \frac{1}{(s+1)(s+2)}$$

Use the methods described above to find the q and δ form of the discrete-time system model obtained using the transformation

$$s \rightarrow \frac{2}{\Delta} \frac{z-1}{z+1}$$

where Δ is the sampling interval. Examine the stability of the original system and that of the discrete-time systems when $\Delta = 0.1$ and when $\Delta = 0.01$.

- 39 Use the formula in equation (6.99) to obtain the transfer function of the third-order Butterworth filter with $\omega_c = 1$, and obtain the corresponding δ form discrete-time system when $T = \Delta$.

- 40 Make the substitution

$$x_1(t) = y(t)$$

$$x_2(t) = \frac{dy(t)}{dt}$$

in Exercise 38 to obtain the state-space form of the system model,

$$\dot{x}(t) = \mathbf{A}x(t) + \mathbf{b}u(t)$$

$$y(t) = \mathbf{c}^T x(t) + du(t)$$

The **Euler discretization technique** replaces $x(t)$ by

$$\frac{x((k+1)\Delta) - x(k\Delta)}{\Delta}$$

Show that this corresponds to the model obtained above with $\mathbf{A} = \mathbf{A}(0)$, $\mathbf{c} = \mathbf{c}(0)$ and $d = d(0)$.

41

The discretization procedure used in Section 6.11.3 has been based on the bilinear transform method, derived from the trapezoidal approximation to the integration process. An alternative approximation is the Adams–Bashforth procedure, and it can be shown that this means that we should make the transformation

$$s \rightarrow \frac{12}{\Delta} \frac{z^2 - z}{5z^2 + 8z - 1}$$

where Δ is the sampling interval (see W. Forsythe and R. M. Goodall, *Digital Control*, London, Macmillan, 1991). Use this transformation to discretize the system given by

$$H(s) = \frac{s}{s+1}$$

when $\Delta = 0.1$ in

- (a) the z form, and
- (b) the γ form.

6.12 Review exercises (1–18)



Check your answers using MATLAB or MAPLE whenever possible.

- 1 The signal $f(t) = t$ is sampled at intervals T to generate the sequence $\{f(kT)\}$. Show that

$$\mathcal{Z}\{f(kT)\} = \frac{Tz}{(z-1)^2}$$

- 2 Show that

$$\mathcal{Z}\{a^k \sin k\omega\} = \frac{az \sin \omega}{z^2 - 2az \cos \omega + a^2} \quad (a > 0)$$

- 3 Show that

$$\mathcal{Z}\{k^2\} = \frac{z(z+1)}{(z-1)^3}$$

- 4 Find the impulse response for the system with transfer function

$$H(z) = \frac{(3z^2 - z)}{z^2 - 2z + 1}$$

- 5 Calculate the step response for the system with transfer function

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

- 6 A process with Laplace transfer function $H(s) = 1/(s+1)$ is in cascade with a zero-order hold device with Laplace transfer function

$G(s) = (1 - e^{-sT})/s$. The overall transfer function is then

$$\frac{1 - e^{-sT}}{s(s+1)}$$

Write $F(s) = 1/s(s+1)$, and find $f(t) = \mathcal{L}^{-1}\{F(s)\}$. Sample $f(t)$ at intervals T to produce the sequence $\{f(kT)\}$ and find $F(z) = \mathcal{Z}\{f(kT)\}$. Deduce that

$$e^{-sT}F(s) \rightarrow \frac{1}{z}\tilde{F}(z)$$

and hence show that the overall z transfer function for the process and zero-order hold is

$$\frac{1 - e^{-T}}{z - e^{-T}}$$

- 7 A system has Laplace transfer function

$$H(s) = \frac{s+1}{(s+2)(s+3)}$$

Calculate the impulse response, and obtain the z transform of this response when sampled at intervals T .

- 8 It can be established that if $X(z)$ is the z transform of the sequence $\{x_n\}$ then the general term of that sequence is given by

$$x_n = \frac{1}{j2\pi} \oint_C X(z)z^{n-1} dz$$

where C is any closed contour containing all the singularities of $X(z)$. If we assume that all the singularities of $X(z)$ are poles located within a circle of finite radius then it is an easy application of the residue theorem to show that

$$x_n = \sum [\text{residues of } X(z)z^{n-1} \text{ at poles of } X(z)]$$

- (a) Let $X(z) = z/(z-a)(z-b)$, with a and b real. Where are the poles of $X(z)$? Calculate the residues of $z^{n-1}X(z)$, and hence invert the transform to obtain $\{x_n\}$.
- (b) Use the residue method to find

$$(i) \quad \mathcal{Z}^{-1}\left\{\frac{z}{(z-3)^2}\right\} \quad (ii) \quad \mathcal{Z}^{-1}\left\{\frac{z}{z^2-z+1}\right\}$$

- 9 The impulse response of a certain discrete-time system is $\{(-1)^k - 2^k\}$. What is the step response?

10 A discrete-time system has transfer function

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

Find the response to the sequence $\{1, -1, 0, 0, \dots\}$.

- 11 Show that the response of the second-order system with transfer function

$$\frac{z^2}{(z-\alpha)(z-\beta)}$$

to the input $\{1, -(α+β), αβ, 0, 0, 0, \dots\}$ is

$$\{\delta_k\} = \{1, 0, 0, \dots\}$$

Deduce that the response of the system

$$\frac{z}{(z-\alpha)(z-\beta)}$$

to the same input will be

$$\{\delta_{k-1}\} = \{0, 1, 0, 0, \dots\}$$

- 12 A system is specified by its Laplace transfer function

$$H(s) = \frac{s}{(s+1)(s+2)}$$

Calculate the impulse response $y_\delta(t) = \mathcal{L}^{-1}\{H(s)\}$, and show that if this response is sampled at intervals T to generate the sequence $\{y_\delta(nT)\}$ ($n = 0, 1, 2, \dots$) then

$$D(z) = \mathcal{Z}\{y_\delta(nT)\} = \frac{2z}{z-e^{-2T}} - \frac{z}{z-e^{-T}}$$

A discrete-time system is now constructed so that

$$Y(z) = TD(z)X(z)$$

where $X(z)$ is the z transform of the input sequence $\{x_n\}$ and $Y(z)$ that of the output sequence $\{y_n\}$, with $x_n = x(nT)$ and $y_n = y(nT)$. Show that if $T = 0.5$ s then the difference equation governing the system is

$$\begin{aligned} y_{n+2} - 0.9744y_{n+1} + 0.2231y_n \\ = 0.5x_{n+2} - 0.4226x_{n+1} \end{aligned}$$

Sketch a block diagram for the discrete-time system modelled by the difference equation

$$p_{n+2} - 0.9744p_{n+1} + 0.2231p_n = x_n$$

and verify that the signal y_n , as defined above, is generated by taking $y_n = 0.5p_{n+2} - 0.4226p_{n+1}$ as output.

- 13 In a discrete-time position-control system the position y_n satisfies the difference equation

$$y_{n+1} = y_n + av_n \quad (a \text{ constant})$$

where v_n and u_n satisfy the difference equations

$$v_{n+1} = v_n + bu_n \quad (b \text{ constant})$$

$$u_n = k_1(x_n - y_n) - k_2v_n \quad (k_1, k_2 \text{ constants})$$

- (a) Show that if $k_1 = 1/4ab$ and $k_2 = 1/b$ then the z transfer function of the system is

$$\frac{Y(z)}{X(z)} = \frac{1}{(1-2z)^2}$$

where $Y(z) = \mathcal{Z}\{y_n\}$ and $X(z) = \mathcal{Z}\{x_n\}$.

- (b) If also $x_n = A$ (where A is a constant), determine the response sequence $\{y_n\}$ given that $y_0 = y_1 = 0$.

- 14 The step response of a continuous-time system is modelled by the differential equation

$$\frac{d^2y}{dt^2} + 3 \frac{dy}{dt} + 2y = 1 \quad (t \geq 0)$$

with $y(0) = \dot{y}(0) = 0$. Use the backward-difference approximation

$$\frac{dy}{dt} \approx \frac{y_k - y_{k-1}}{T}$$

$$\frac{d^2y}{dt^2} \approx \frac{y_k - 2y_{k-1} + y_{k-2}}{T^2}$$

to show that this differential equation may be approximated by

$$\frac{y_k - 2y_{k-1} + y_{k-2}}{T^2} + 3 \frac{y_k - y_{k-1}}{T} + 2y_k = 1$$

Take the z transform of this difference equation, and show that the system poles are at

$$z = \frac{1}{1+T}, \quad z = \frac{1}{1+2T}$$

Deduce that the general solution is thus

$$y_k = \alpha \left(\frac{1}{1+T}\right)^k + \beta \left(\frac{1}{1+2T}\right)^k + \gamma$$

Show that $\gamma = \frac{1}{2}$ and, noting that the initial conditions $y(0) = 0$ and $\dot{y}(0) = 0$ imply $y_0 = y_{-1} = 0$, deduce that

$$y_k = \frac{1}{2} \left[\left(\frac{1}{1+2T}\right)^k - 2 \left(\frac{1}{1+T}\right)^k + 1 \right]$$

Note that the z -transform method could be used to obtain this result if we redefine $\mathcal{Z}\{y_k\} = \sum_{j=-1}^{\infty} (y_j/z^j)$, with appropriate modifications to the formulae for $\mathcal{Z}\{y_{k+1}\}$ and $\mathcal{Z}\{y_{k+2}\}$.

Explain why the calculation procedure is always stable in theory, but note the pole locations for very small T .

Finally, verify that the solution of the differential equation is

$$y(t) = \frac{1}{2} (e^{-2t} - 2e^{-t} + 1)$$

and plot graphs of the exact and approximate solutions with $T = 0.1$ s and $T = 0.05$ s.

- 15

Again consider the step response of the system modelled by the differential equation

$$\frac{d^2y}{dt^2} + 3 \frac{dy}{dt} + 2y = 1 \quad (t \geq 0)$$

with $y(0) = \dot{y}(0) = 0$. Now discretize using the bilinear transform method; that is, take the Laplace transform and make the transformation

$$s \rightarrow \frac{2z-1}{Tz+1}$$

where T is the sampling interval. Show that the poles of the resulting z transfer function are at

$$z = \frac{1-T}{1+T}, \quad z = \frac{2-T}{2+T}$$

Deduce that the general solution is then

$$y_k = \alpha \left(\frac{1-T}{1+T}\right)^k + \beta \left(\frac{2-T}{2+T}\right)^k + \gamma$$

Deduce that $\gamma = \frac{1}{2}$ and, using the conditions $y_0 = y_{-1} = 0$, show that

$$y_k = \frac{1}{2} \left[(1-T) \left(\frac{1-T}{1+T}\right)^k - (2-T) \left(\frac{2-T}{2+T}\right)^k + 1 \right]$$

Plot graphs to illustrate the exact solution and the approximate solution when $T = 0.1$ s and $T = 0.05$ s.

- 16

Show that the z transform of the sampled version of the signal $f(t) = t^2$ is

$$F(z) = \frac{z(z+1)\Delta^2}{(z-1)^3}$$

where Δ is the sampling interval. Verify that the \mathcal{D} transform is then

$$\frac{(1 + \Delta v)(2 + \Delta v)}{v^3}$$

- 17 Show that the eigenvalues of the matrix

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

are 2, 1 and -1 , and find the corresponding eigenvectors. Write down the modal matrix \mathbf{M} and spectral matrix Λ of \mathbf{A} , and verify that $\mathbf{MA} = \mathbf{AM}$.

Deduce that the system of difference equations

$$\mathbf{x}(k+1) = \mathbf{Ax}(k)$$

where $\mathbf{x}(k) = [x_1(k) \quad x_2(k) \quad x_3(k)]^T$, has a solution

$$\mathbf{x}(k) = \mathbf{My}(k)$$

where $\mathbf{y}(k) = \Lambda^k \mathbf{y}(0)$. Find this solution, given $\mathbf{x}(0) = [1 \quad 0 \quad 0]^T$.

- 18 The system shown in Figure 6.26 is a realization of a discrete-time system. Show that, with state variables $x_1(k)$ and $x_2(k)$ as shown, the system may be represented as

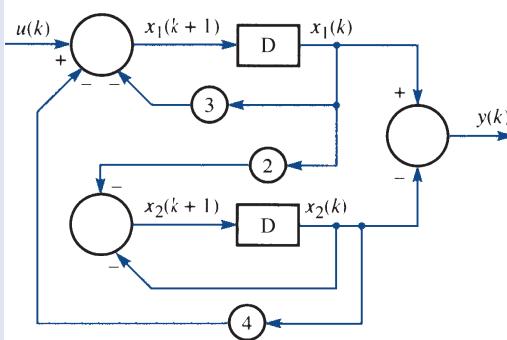


Figure 6.26 Discrete-time system of Review exercise 19.

$$\mathbf{x}(k+1) = \mathbf{Ax}(k) + \mathbf{bu}(k)$$

$$y(k) = \mathbf{c}^T \mathbf{x}(k)$$

where

$$\mathbf{A} = \begin{bmatrix} -3 & -4 \\ -2 & -1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Calculate the z transfer function of the system, $D(z)$, where

$$D(z) = \mathbf{c}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$$

Reduce the system to control canonical form by the following means:

- calculate the controllability matrix \mathbf{M}_c , where $\mathbf{M}_c = [\mathbf{b} \quad \mathbf{Ab}]$ is the matrix with columns \mathbf{b} and \mathbf{Ab} ;
- show that $\text{rank } (\mathbf{M}_c) = 2$, and calculate \mathbf{M}_c^{-1} ;
- write down the vector \mathbf{v}^T corresponding to the last row of \mathbf{M}_c^{-1} ;
- form the matrix $\mathbf{T} = [\mathbf{v}^T \quad \mathbf{v}^T \mathbf{A}]^T$, the matrix with rows \mathbf{v}^T and $\mathbf{v}^T \mathbf{A}$;
- calculate \mathbf{T}^{-1} and using this matrix \mathbf{T} , show that the transformation $\mathbf{z}(k) = \mathbf{T}\mathbf{x}(k)$ produces the system

$$\begin{aligned} \mathbf{z}(k+1) &= \mathbf{T}\mathbf{AT}^{-1}\mathbf{z}(k) + \mathbf{Tb}u(k) \\ &= \mathbf{Cz}(k) + \mathbf{b}_c u(k) \end{aligned}$$

where \mathbf{C} is of the form

$$\begin{bmatrix} 0 & 1 \\ -\alpha & -\beta \end{bmatrix}$$

and $\mathbf{b}_c = [0 \quad 1]^T$. Calculate α and β , and comment on the values obtained in relation to the transfer function $D(z)$.



7

Fourier Series

Chapter 7 Contents

7.1	Introduction	486
7.2	Fourier series of jumps at discontinuities	499
7.3	Engineering application: frequency response and oscillating systems	502
7.4	Complex form of Fourier series	508
7.5	Orthogonal functions	524
7.6	Engineering application: describing functions	532
7.7	Review exercises (1–20)	533

7.1 Introduction

The basics of Fourier series are covered in Chapter 12 of *Modern Engineering Mathematics* (MEM). The reader is referred there to revise this material if required. Here we give only the basic definitions before moving on to more advanced topics.

A **Fourier series** is an expansion of a periodic function $f(t)$ of period $T = 2\pi/\omega$ in which the base set is the set of sine functions, giving an expanded representation of the form

$$f(t) = A_0 + \sum_{n=1}^{\infty} A_n \sin(n\omega t + \phi_n)$$

7.1.1 Periodic functions

A function $f(t)$ is said to be **periodic** if its image values are repeated at regular intervals in its domain. Thus the graph of a periodic function can be divided into ‘vertical strips’ that are replicas of each other, as illustrated in Figure 7.1. The interval between two successive replicas is called the **period** of the function. We therefore say that a function $f(t)$ is periodic with period T if, for all its domain values t ,

$$f(t + mT) = f(t)$$

for any integer m .

To provide a measure of the number of repetitions per unit of t , we define the **frequency** of a periodic function to be the reciprocal of its period, so that

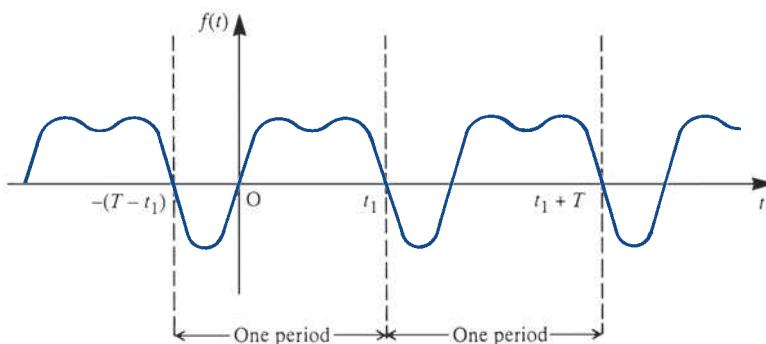
$$\text{frequency} = \frac{1}{\text{period}} = \frac{1}{T}$$

The term **circular frequency** is also used in engineering, and is defined by

$$\text{circular frequency} = 2\pi \times \text{frequency} = \frac{2\pi}{T}$$

and is measured in radians per second. It is common to drop the term ‘circular’ and refer to this simply as the frequency when the context is clear.

Figure 7.1 A periodic function with period T .



7.1.2 Fourier's theorem

This theorem states that a periodic function that satisfies certain conditions can be expressed as the sum of a number of sine functions of different amplitudes, phases and periods. That is, if $f(t)$ is a periodic function with period T then

$$\begin{aligned} f(t) = & A_0 + A_1 \sin(\omega t + \phi_1) + A_2 \sin(2\omega t + \phi_2) + \dots \\ & + A_n \sin(n\omega t + \phi_n) + \dots \end{aligned} \quad (7.1)$$

where the A s and ϕ s are constants and $\omega = 2\pi/T$ is the frequency of $f(t)$. The term $A_1 \sin(\omega t + \phi_1)$ is called the **first harmonic** or the **fundamental mode**, and it has the same frequency ω as the parent function $f(t)$. The term $A_n \sin(n\omega t + \phi_n)$ is called the **n th harmonic**, and it has frequency $n\omega$ which is n times that of the fundamental. A_n denotes the **amplitude** of the n th harmonic and ϕ_n is its **phase angle**, measuring the lag or lead of the n th harmonic with reference to a pure sine wave of the same frequency.

Since

$$\begin{aligned} A_n \sin(n\omega t + \phi_n) &\equiv (A_n \cos \phi_n) \sin n\omega t + (A_n \sin \phi_n) \cos n\omega t \\ &\equiv b_n \sin n\omega t + a_n \cos n\omega t \end{aligned}$$

where

$$b_n = A_n \cos \phi_n, \quad a_n = A_n \sin \phi_n \quad (7.2)$$

the expansion (7.1) may be written as

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t \quad (7.3)$$

where $a_0 = 2A_0$ (we shall see later that taking the first term as $\frac{1}{2}a_0$ rather than a_0 is a convenience that enables us to make a_0 fit a general result). The expansion (7.3) is called the **Fourier series expansion** of the function $f(t)$, and the a s and b s are called the **Fourier coefficients**. In electrical engineering it is common practice to refer to a_n and b_n respectively as the **in-phase** and **phase quadrature components** of the n th harmonic, this terminology arising from the use of the phasor notation $e^{j\omega t} = \cos n\omega t + j \sin n\omega t$. Clearly, (7.1) is an alternative representation of the Fourier series with the amplitude and phase of the n th harmonic being determined from (7.2) as

$$A_n = \sqrt{(a_n^2 + b_n^2)}, \quad \phi_n = \tan^{-1}\left(\frac{a_n}{b_n}\right)$$

with care being taken over choice of quadrant.

The Fourier coefficients are given by

$$a_n = \frac{2}{T} \int_d^{d+T} f(t) \cos n\omega t \, dt \quad (n = 0, 1, 2, \dots) \quad (7.4)$$

$$b_n = \frac{2}{T} \int_d^{d+T} f(t) \sin n\omega t \, dt \quad (n = 1, 2, 3, \dots) \quad (7.5)$$

which are known as **Euler's formulae**.

Further introductory material is found in MEM and Phil Dyke's *Introduction to Laplace Transforms and Fourier series* (second edition, London, Springer, 2014). Here we continue with the calculation of Fourier series where the function exhibits a finite number of finite jumps.

The limits of integration in Euler's formulae may be specified over any period, so that the choice of d is arbitrary, and may be made in such a way as to help in the calculation of a_n and b_n . In practice, it is common to specify $f(t)$ over either the period $-\frac{1}{2}T < t < \frac{1}{2}T$ or the period $0 < t < T$, leading respectively to the limits of integration being $-\frac{1}{2}T$ and $\frac{1}{2}T$ (that is, $d = -\frac{1}{2}T$) or 0 and T (that is, $d = 0$).

It is also worth noting that an alternative approach may simplify the calculation of a_n and b_n . Using the formula

$$e^{jn\omega t} = \cos n\omega t + j \sin n\omega t$$

we have

$$a_n + jb_n = \frac{2}{T} \int_d^{d+T} f(t) e^{jn\omega t} dt \quad (7.6)$$

Evaluating this integral and equating real and imaginary parts on each side gives the values of a_n and b_n . This approach is particularly useful when only the amplitude $|a_n + jb_n|$ of the n th harmonic is required.

7.1.3 Functions of period 2π

If the period T of the periodic function $f(t)$ is taken to be 2π then $\omega = 1$, and the series (7.3) becomes

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos nt + \sum_{n=1}^{\infty} b_n \sin nt \quad (7.7)$$

with the coefficients given by

$$a_n = \frac{1}{\pi} \int_d^{d+2\pi} f(t) \cos nt dt \quad (n = 0, 1, 2, \dots) \quad (7.8)$$

$$b_n = \frac{1}{\pi} \int_d^{d+2\pi} f(t) \sin nt dt \quad (n = 1, 2, \dots) \quad (7.9)$$

Here are several examples of how to calculate a Fourier series based on material in MEM. It should be revision, but readers are urged to consult MEM for more details. In Fourier series, the concepts of even and odd functions are very important. A function f is *even* if, for all x , $f(x) = f(-x)$. Examples include $\cos(x)$ and x^2 . A function is *odd* if, for all x , $f(x) = -f(-x)$. Examples include $\sin(x)$ and x^3 . The key important point is that the Fourier series of an even function has to contain only even functions so must comprise only the constant term and cosine functions, whereas the Fourier series of an odd function can contain only sine terms. The first worked example involves finding the Fourier series of an odd function.

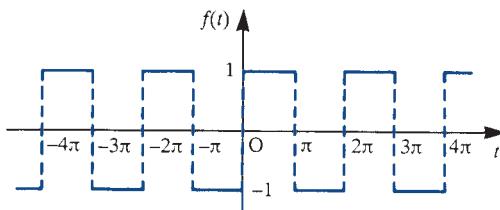
Example 7.1

A periodic function $f(t)$ with period 2π is defined within the period $-\pi < t < \pi$ by

$$f(t) = \begin{cases} -1 & (-\pi < t < 0) \\ 1 & (0 < t < \pi) \end{cases}$$

Find its Fourier series expansion.

Figure 7.2 Square wave of Example 7.1.

**Solution**

A sketch of the function $f(t)$ over the interval $-4\pi < t < 4\pi$ is shown in Figure 7.2. Clearly $f(t)$ is an odd function of t , so that its Fourier series expansion consists of sine terms only. Taking $T = 2\pi$, that is $\omega = 1$ in (7.3), remembering that all the a_n 's are zero, the Fourier series expansion is given by

$$f(t) = \sum_{n=1}^{\infty} b_n \sin nt$$

with

$$\begin{aligned} b_n &= \frac{1}{\pi} \int_0^{2\pi} f(t) \sin nt dt \quad (n = 1, 2, 3, \dots) \\ &= \frac{2}{\pi} \int_0^{\pi} 1 \sin nt dt = \frac{2}{\pi} \left[-\frac{1}{n} \cos nt \right]_0^{\pi} \\ &= \frac{2}{n\pi} (1 - \cos n\pi) = \frac{2}{n\pi} [1 - (-1)^n] \\ &= \begin{cases} 4/n\pi & (\text{odd } n) \\ 0 & (\text{even } n) \end{cases} \end{aligned}$$

Thus the Fourier series expansion of $f(t)$ is

$$f(t) = \frac{4}{\pi} (\sin t + \frac{1}{3} \sin 3t + \frac{1}{5} \sin 5t + \dots) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin (2n-1)t}{2n-1} \quad (7.10)$$

Here is an example showing how to find the Fourier series of an even function.

Example 7.2

Obtain the Fourier series expansion of the rectified sine wave

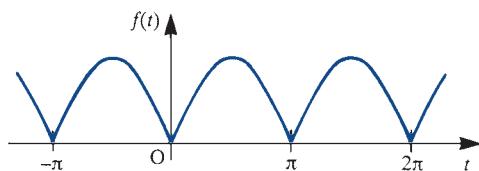
$$f(t) = |\sin t|$$

Solution A sketch of the wave over the interval $-\pi < t < 2\pi$ is shown in Figure 7.3. Clearly, $f(t)$ is periodic with period π . Taking $T = \pi$, that is, $\omega = 2$, in (7.3)–(7.5) the Fourier series expansion is given by

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos 2nt$$

$$a_0 = \frac{2}{\pi} \int_0^\pi \sin t \, dt = \frac{4}{\pi}$$

Figure 7.3 Rectified wave $f(t)$ of Example 7.2.



$$\begin{aligned} a_n &= \frac{2}{\pi} \int_0^\pi \sin t \cos 2nt \, dt \\ &= \frac{1}{\pi} \int_0^\pi [\sin(2n+1)t - \sin(2n-1)t] \, dt \\ &= \frac{1}{\pi} \left[-\frac{\cos 2(n+1)t}{2n+1} + \frac{\cos 2(n-1)t}{2n-1} \right]_0^\pi \\ &= \frac{1}{\pi} \left[\left(\frac{1}{2n+1} - \frac{1}{2n-1} \right) - \left(-\frac{1}{2n+1} + \frac{1}{2n-1} \right) \right] = -\frac{4}{\pi} \frac{1}{4n^2-1} \end{aligned}$$

Thus the Fourier series expansion of $f(t)$ is

$$f(t) = \frac{2}{\pi} - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{4n^2-1} \cos 2nt$$

or, writing out the first few terms,

$$f(t) = \frac{2}{\pi} - \frac{4}{\pi} \left(\frac{1}{3} \cos 2t + \frac{1}{15} \cos 4t + \frac{1}{35} \cos 6t + \dots \right)$$

Finally in this short revision of material from MEM, here is an example of the integration of a Fourier Series term by term.

Example 7.3

Integrate term by term the Fourier series expansion obtained in Example 7.1 for the square wave

$$f(t) = \begin{cases} -1 & (-\pi < t < 0) \\ 1 & (0 < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

illustrated in Figure 7.2.

Solution From (7.10), the Fourier series expansion for $f(t)$ is

$$f(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin(2n-1)t}{2n-1}$$

We now need to integrate between the limits $-\pi$ and t and, owing to the discontinuity in $f(t)$ at $t = 0$, we must consider separately values of t in the intervals $-\pi < t < 0$ and $0 < t < \pi$.

Case (i), interval $-\pi < t < 0$. Integrating (7.10) term by term, we have

$$\int_{-\pi}^t (-1) dt = \frac{4}{\pi} \sum_{n=1}^{\infty} \int_{-\pi}^t \frac{\sin(2n-1)t}{(2n-1)} dt$$

that is,

$$\begin{aligned} -(t + \pi) &= -\frac{4}{\pi} \sum_{n=1}^{\infty} \left[\frac{\cos(2n-1)t}{(2n-1)^2} \right]_{-\pi}^t \\ &= -\frac{4}{\pi} \left[\sum_{n=1}^{\infty} \frac{\cos(2n-1)t}{(2n-1)^2} + \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \right] \end{aligned}$$

It can be shown that

$$\sum_{n=1}^{\infty} \frac{2}{(2n-1)^2} = \frac{1}{8}\pi^2$$

(see Exercise 6), so that the above simplifies to

$$-t = \frac{1}{2}\pi - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos(2n-1)t}{(2n-1)^2} \quad (-\pi < t < 0) \quad (7.11)$$

Case (ii), interval $0 < t < \pi$. Integrating (7.10) term by term, we have

$$\int_{-\pi}^0 (-1) dt + \int_0^t 1 dt = \frac{4}{\pi} \sum_{n=1}^{\infty} \int_{-\pi}^0 \frac{\sin(2n-1)t}{(2n-1)} dt$$

giving

$$t = \frac{1}{2}\pi - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos(2n-1)t}{(2n-1)^2} \quad (0 < t < \pi) \quad (7.12)$$

Taking (7.11) and (7.12) together, we find that the function

$$g(t) = |t| = \begin{cases} -t & (-\pi < t < 0) \\ t & (0 < t < \pi) \end{cases}$$

$$g(t+2\pi) = g(t)$$

has a Fourier series expansion

$$g(t) = |t| = \frac{1}{2}\pi - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos((2n-1)t)}{(2n-1)^2}$$

7.1.4 Functions defined over a finite interval

One of the requirements of Fourier's theorem is that the function to be expanded be periodic. Therefore a function $f(t)$ that is not periodic cannot have a Fourier series representation that converges to it *for all values* of t . However, we can obtain a Fourier series expansion that represents a *non-periodic* function $f(t)$ that is defined only over a finite time interval $0 < t < \tau$. This is a facility that is frequently used to solve problems in practice, particularly boundary-value problems involving partial differential equations, such as the consideration of heat flow along a bar or the vibrations of a string. Various forms of Fourier series representations of $f(t)$, valid only in the interval $0 < t < \tau$, are possible, including series consisting of cosine terms only or series consisting of sine terms only. To obtain these, various periodic extensions of $f(t)$ are formulated.

Full-range series

Suppose the given function $f(t)$ is defined only over the finite time interval $0 < t < \tau$. Then, to obtain a full-range Fourier series representation of $f(t)$ (that is, a series consisting of both cosine and sine terms), we define the **periodic extension** $\phi(t)$ of $f(t)$ by

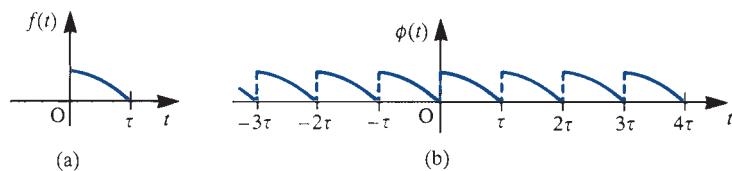
$$\phi(t) = f(t) \quad (0 < t < \tau)$$

$$\phi(t + \tau) = \phi(t)$$

The graphs of a possible $f(t)$ and its periodic extension $\phi(t)$ are shown in Figures 7.4(a) and (b) respectively.

Provided that $f(t)$ satisfies Dirichlet's conditions in the interval $0 < t < \tau$, the new function $\phi(t)$, of period τ , will have a convergent Fourier series expansion. Since, within the particular period $0 < t < \tau$, $\phi(t)$ is identical with $f(t)$, it follows that this Fourier series expansion of $\phi(t)$ will be representative of $f(t)$ within this interval.

Figure 7.4 Graphs of a function defined only over (a) a finite interval $0 < t < \tau$ and (b) its periodic extension.



Example 7.4

Find a full-range Fourier series expansion of $f(t) = t$ valid in the finite interval $0 < t < 4$. Draw graphs of both $f(t)$ and the periodic function represented by the Fourier series obtained.

Solution Define the periodic function $\phi(t)$ by

$$\phi(t) = f(t) = t \quad (0 < t < 4)$$

$$\phi(t+4) = \phi(t)$$

Then the graphs of $f(t)$ and its periodic extension $\phi(t)$ are as shown in Figures 7.5(a) and (b) respectively. Since $\phi(t)$ is a periodic function with period 4, it has a convergent Fourier series expansion. Taking $T = 4$ in (7.4) and (7.5), the Fourier coefficients are determined as

$$a_0 = \frac{1}{2} \int_0^4 f(t) dt = \frac{1}{2} \int_0^4 t dt = 4$$

$$a_n = \frac{1}{2} \int_0^4 f(t) \cos \frac{1}{2}n\pi t dt \quad (n = 1, 2, 3, \dots)$$

$$= \frac{1}{2} \int_0^4 t \cos \frac{1}{2}n\pi t dt = \frac{1}{2} \left[\frac{2t}{n\pi} \sin \frac{1}{2}n\pi t + \frac{4}{(n\pi)^2} \cos \frac{1}{2}n\pi t \right]_0^4 = 0$$

and

$$b_n = \frac{1}{2} \int_0^4 f(t) \sin \frac{1}{2}n\pi t dt \quad (n = 1, 2, 3, \dots)$$

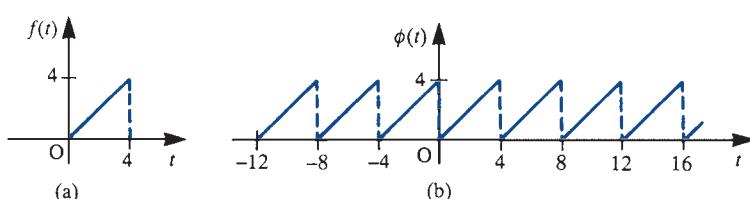
$$= \frac{1}{2} \int_0^4 t \sin \frac{1}{2}n\pi t dt = \frac{1}{2} \left[-\frac{2t}{n\pi} \cos \frac{1}{2}n\pi t + \frac{4}{(n\pi)^2} \sin \frac{1}{2}n\pi t \right]_0^4 = -\frac{4}{n\pi}$$

Thus, by (7.3), the Fourier series expansion of $\phi(t)$ is

$$\phi(t) = 2 - \frac{4}{\pi} (\sin \frac{1}{2}\pi t + \frac{1}{2}\sin \pi t + \frac{1}{3}\sin \frac{3}{2}\pi t + \frac{1}{4}\sin 2t + \frac{1}{5}\sin \frac{5}{2}\pi t + \dots)$$

$$= 2 - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin \frac{1}{2}n\pi t$$

Figure 7.5
The functions $f(t)$ and $\phi(t)$ of Example 7.4.



Since $\phi(t) = f(t)$ for $0 < t < 4$, it follows that this Fourier series is representative of $f(t)$ within this interval, so that

$$f(t) = t = 2 - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin \frac{1}{2} n\pi t \quad (0 < t < 4) \quad (7.13)$$

It is important to appreciate that this series converges to t only within the interval $0 < t < 4$. For values of t outside this interval it converges to the periodic extended function $\phi(t)$. Again convergence is to be interpreted in the sense of Theorem 12.2 of MEM, so that at the end points $t = 0$ and $t = 4$ the series does not converge to t but to the mean of the discontinuity in $\phi(t)$, namely the value 2.

Half-range cosine and sine series

Rather than develop the periodic extension $\phi(t)$ of $f(t)$ as above, it is possible to formulate periodic extensions that are either even or odd functions, so that the resulting Fourier series of the extended periodic functions consist either of cosine terms only or sine terms only.

For a function $f(t)$ defined only over the finite interval $0 \leq t \leq \tau$ its **even periodic extension** $F(t)$ is the even periodic function defined by

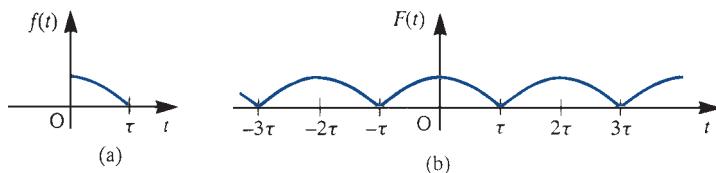
$$F(t) = \begin{cases} f(t) & (0 < t < \tau) \\ f(-t) & (-\tau < t < 0) \end{cases}$$

$$F(t + 2\tau) = f(t)$$

As an illustration, the even periodic extension $F(t)$ of the function $f(t)$ shown in Figure 7.4(a) (redrawn in Figure 7.6(a)) is shown in Figure 7.6(b).

Figure 7.6

- (a) A function $f(t)$;
- (b) its even periodic extension $F(t)$.



Provided that $f(t)$ satisfies Dirichlet's conditions in the interval $0 < t < \tau$, since it is an even function of period 2τ , it follows from the properties of even functions (see Section 12.2.5 of MEM) that the even periodic extension $F(t)$ will have a convergent Fourier series representation consisting of cosine terms only and given by

$$F(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos \frac{n\pi t}{\tau} \quad (7.14)$$

where

$$a_n = \frac{2}{\tau} \int_0^\tau f(t) \cos \frac{n\pi t}{\tau} dt \quad (n = 0, 1, 2, \dots) \quad (7.15)$$

Since, within the particular interval $0 < t < \tau$, $F(t)$ is identical with $f(t)$, it follows that the series (7.14) also converges to $f(t)$ within this interval.

For a function $f(t)$ defined only over the finite interval $0 < t < \tau$, its **odd periodic extension** $G(t)$ is the odd periodic function defined by

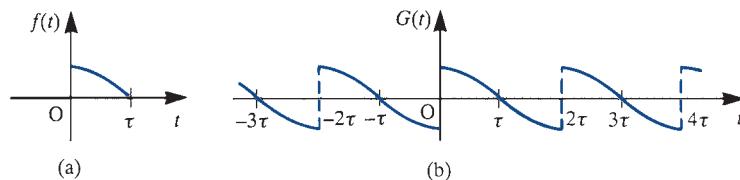
$$G(t) = \begin{cases} f(t) & (0 < t < \tau) \\ -f(-t) & (-\tau < t < 0) \end{cases}$$

$$G(t + 2\tau) = G(t)$$

Again, as an illustration, the odd periodic extension $G(t)$ of the function $f(t)$ shown in Figure 7.4(a) (redrawn in Figure 7.7(a)) is shown in Figure 7.7(b).

Figure 7.7

(a) A function $f(t)$;
(b) its odd periodic extension $G(t)$.



Provided that $f(t)$ satisfies Dirichlet's conditions in the interval $0 < t < \tau$, since it is an odd function of period 2τ , it follows from the properties of odd functions (see Section 12.2.5 of MEM) that the odd periodic extension $G(t)$ will have a convergent Fourier series representation consisting of sine terms only and given by

$$G(t) = \sum_{n=1}^{\infty} b_n \sin \frac{n\pi t}{\tau} \quad (7.16)$$

where

$$b_n = \frac{2}{\tau} \int_0^\tau f(t) \sin \frac{n\pi t}{\tau} dt \quad (n = 1, 2, 3, \dots) \quad (7.17)$$

Again, since, within the particular interval $0 < t < \tau$, $G(t)$ is identical with $f(t)$, it follows that the series (7.16) also converges to $f(t)$ within this interval.

We note that both the even and odd periodic extensions $F(t)$ and $G(t)$ are of period 2τ , which is twice the length of the interval over which $f(t)$ is defined. However, the resulting Fourier series (7.14) and (7.16) are based only on the function $f(t)$, and for this reason are called the **half-range Fourier series expansions** of $f(t)$. In particular, the even half-range expansion $F(t)$, (7.14), is called the **half-range cosine series expansion** of $f(t)$, while the odd half-range expansion $G(t)$, (7.16), is called the **half-range sine series expansion** of $f(t)$.

Example 7.5

For the function $f(t) = t$ defined only in the interval $0 < t < 4$, and considered in Example 7.4, obtain

- (a) a half-range cosine series expansion
- (b) a half-range sine series expansion.

Draw graphs of $f(t)$ and of the periodic functions represented by the two series obtained for $-20 < t < 20$.

Solution (a) *Half-range cosine series.* Define the periodic function $F(t)$ by

$$F(t) = \begin{cases} f(t) = t & (0 < t < 4) \\ f(-t) = -t & (-4 < t < 0) \end{cases}$$

$$F(t + 8) = F(t)$$

Then, since $F(t)$ is an even periodic function with period 8, it has a convergent Fourier series expansion given by (7.14). Taking $\tau = 4$ in (7.15), we have

$$\begin{aligned} a_0 &= \frac{2}{4} \int_0^4 f(t) dt = \frac{1}{2} \int_0^4 t dt = 4 \\ a_n &= \frac{2}{4} \int_0^4 f(t) \cos \frac{1}{4}n\pi t dt \quad (n = 1, 2, 3, \dots) \\ &= \frac{1}{2} \int_0^4 t \cos \frac{1}{4}n\pi t dt = \frac{1}{2} \left[\frac{4t}{n\pi} \sin \frac{1}{4}n\pi t + \frac{16}{(n\pi)^2} \cos \frac{1}{4}n\pi t \right]_0^4 \\ &= \frac{8}{(n\pi)^2} (\cos n\pi - 1) = \begin{cases} 0 & (\text{even } n) \\ -16/(n\pi)^2 & (\text{odd } n) \end{cases} \end{aligned}$$

Then, by (7.14), the Fourier series expansion of $F(t)$ is

$$F(t) = 2 - \frac{16}{\pi^2} (\cos \frac{1}{4}\pi t + \frac{1}{3^2} \cos \frac{3}{4}\pi t + \frac{1}{5^2} \cos \frac{5}{4}\pi t + \dots)$$

or

$$F(t) = 2 - \frac{16}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos \frac{1}{4}(2n-1)\pi t$$

Since $F(t) = f(t)$ for $0 < t < 4$, it follows that this Fourier series is representative of $f(t)$ within this interval. Thus the half-range cosine series expansion of $f(t)$ is

$$f(t) = t = 2 - \frac{16}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos \frac{1}{4}(2n-1)\pi t \quad (0 < t < 4) \tag{7.18}$$

(b) *Half-range sine series.* Define the periodic function $G(t)$ by

$$G(t) = \begin{cases} f(t) = t & (0 < t < 4) \\ -f(-t) = t & (-4 < t < 0) \end{cases}$$

$$G(t + 8) = G(t)$$

Then, since $G(t)$ is an odd periodic function with period 8, it has a convergent Fourier series expansion given by (7.16). Taking $\tau=4$ in (7.17), we have

$$\begin{aligned} b_n &= \frac{2}{4} \int_0^4 f(t) \sin \frac{1}{4}n\pi t dt \quad (n = 1, 2, 3, \dots) \\ &= \frac{1}{2} \int_0^4 t \sin \frac{1}{4}n\pi t dt = \frac{1}{2} \left[-\frac{4t}{n\pi} \cos \frac{1}{4}n\pi t + \frac{16}{(n\pi)^2} \sin \frac{1}{4}n\pi t \right]_0^4 \\ &= -\frac{8}{n\pi} \cos n\pi = -\frac{8}{n\pi}(-1)^n \end{aligned}$$

Thus, by (7.16), the Fourier series expansion of $G(t)$ is

$$G(t) = \frac{8}{\pi} \left(\sin \frac{1}{4}\pi t - \frac{1}{2} \sin \frac{1}{2}\pi t + \frac{1}{3} \sin \frac{3}{4}\pi t - \dots \right)$$

or

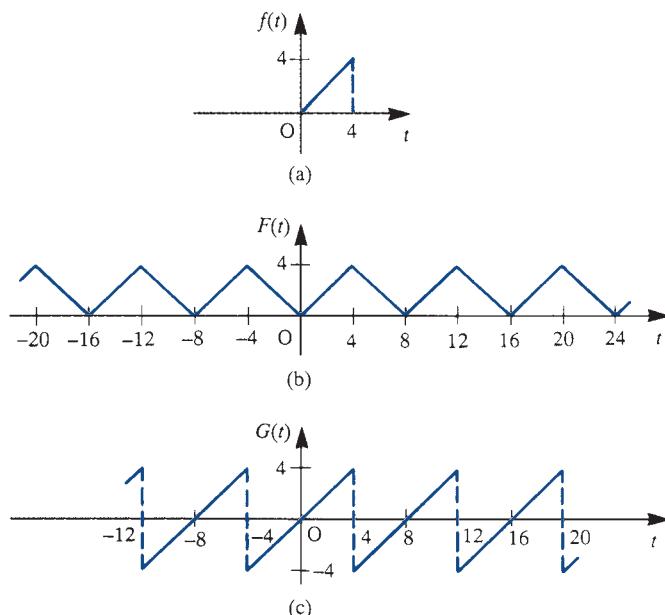
$$G(t) = \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin \frac{1}{4}n\pi t$$

Since $G(t)=f(t)$ for $0 < t < 4$, it follows that this Fourier series is representative of $f(t)$ within this interval. Thus the half-range sine series expansion of $f(t)$ is

$$f(t) = t = \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin \frac{1}{4}n\pi t \quad (0 < t < 4) \quad (7.19)$$

Graphs of the given function $f(t)$ and of the even and odd periodic expansions $F(t)$ and $G(t)$ are given in Figures 7.8(a), (b) and (c) respectively.

Figure 7.8
The functions $f(t)$,
 $F(t)$ and $G(t)$ of
Example 7.5.



It is important to realize that the three different Fourier series representations (7.13), (7.18) and (7.19) are representative of the function $f(t) = t$ only within the defined interval $0 < t < 4$. Outside this interval the three Fourier series converge to the three different functions $\phi(t)$, $F(t)$ and $G(t)$, illustrated in Figures 7.5(b), 7.8(b) and 7.8(c) respectively.

7.1.5 Exercises

- 1 Show that the half-range Fourier sine series expansion of the function $f(t) = 1$, valid for $0 < t < \pi$, is

$$f(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin(2n-1)t}{2n-1} \quad (0 < t < \pi)$$

Sketch the graphs of both $f(t)$ and the periodic function represented by the series expansion for $-3\pi < t < 3\pi$.

- 2 Determine the half-range cosine series expansion of the function $f(t) = 2t - 1$, valid for $0 < t < 1$. Sketch the graphs of both $f(t)$ and the periodic function represented by the series expansion for $-2 < t < 2$.

- 3 The function $f(t) = 1 - t^2$ is to be represented by a Fourier series expansion over the finite interval $0 < t < 1$. Obtain a suitable

- (a) full-range series expansion,
- (b) half-range sine series expansion,
- (c) half-range cosine series expansion.

Draw graphs of $f(t)$ and of the periodic functions represented by each of the three series for $-4 < t < 4$.

- 4 A function $f(t)$ is defined by

$$f(t) = \pi t - t^2 \quad (0 \leq t \leq \pi)$$

and is to be represented by either a half-range Fourier sine series or a half-range Fourier cosine series. Find both of these series and sketch the graphs of the functions represented by them for $-2\pi < t < 2\pi$.

- 5 A tightly stretched flexible uniform string has its ends fixed at the points $x = 0$ and $x = l$. The midpoint of the string is displaced a distance a , as shown in Figure 7.9. If $f(x)$ denotes the displaced profile of

the string, express $f(x)$ as a Fourier series expansion consisting only of sine terms.

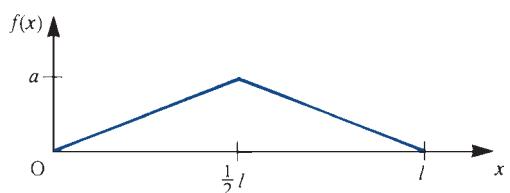


Figure 7.9 Displaced string of Exercise 5.

- 6 Repeat Exercise 5 for the case where the displaced profile of the string is as shown in Figure 7.10.

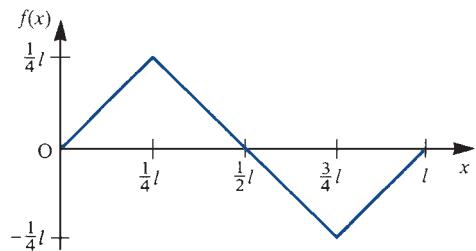


Figure 7.10 Displaced string of Exercise 6.

- 7 A function $f(t)$ is defined on $0 \leq t \leq \pi$ by

$$f(t) = \begin{cases} \sin t & (0 \leq t \leq \frac{1}{2}\pi) \\ 0 & (\frac{1}{2}\pi \leq t \leq \pi) \end{cases}$$

Find a half-range Fourier series expansion of $f(t)$ on this interval. Sketch a graph of the function represented by the series for $-2\pi \leq t \leq 2\pi$.

- 8 A function $f(t)$ is defined on the interval $-l \leq x \leq l$ by

$$f(x) = \frac{A}{l}(|x| - l)$$

Obtain a Fourier series expansion of $f(x)$ and sketch a graph of the function represented by the series for $-3l \leq x \leq 3l$.

9

The temperature distribution $T(x)$ at a distance x , measured from one end, along a bar of length L is given by

$$T(x) = Kx(L-x) \quad (0 \leq x \leq L), \quad K = \text{constant}$$

Express $T(x)$ as a Fourier series expansion consisting of sine terms only.

10

Find the Fourier series expansion of the function $f(t)$ valid for $-1 < t < 1$, where

$$f(t) = \begin{cases} 1 & (-1 < t < 0) \\ \cos \pi t & (0 < t < 1) \end{cases}$$

To what value does this series converge when $t = 1$?

7.2

Fourier series of jumps at discontinuities

For periodic functions that, within a period, are piecewise polynomials and exhibit jump discontinuities, the Fourier coefficients may be determined in terms of the magnitude of the jumps and those of derived functions. This method is useful for determining describing functions (see Section 7.6) for nonlinear characteristics in control engineering, where only the fundamental component of the Fourier series is important; this applies particularly to the case of multivalued nonlinearities.

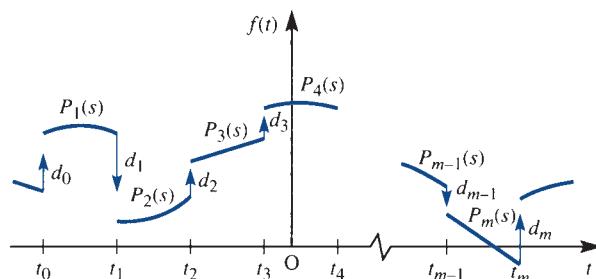
Consider a periodic function $f(t)$, of period T , having within the time interval $-\frac{1}{2}T \leq t \leq \frac{1}{2}T$ a finite number $(m+1)$ of jump discontinuities d_0, d_1, \dots, d_m at times t_0, t_1, \dots, t_m , with $t_0 = -\frac{1}{2}T$ and $t_m = \frac{1}{2}T$. Furthermore, within the interval $t_{s-1} < t < t_s$ ($s = 1, 2, \dots, m$) let $f(t)$ be represented by polynomial functions $P_s(t)$ ($s = 1, 2, \dots, m$), as illustrated in Figure 7.11. If $f(t)$ is to be represented in terms of the Fourier series

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t$$

then, from (7.4),

$$a_n = \frac{2}{T} \sum_{s=1}^m \int_{t_{s-1}}^{t_s} P_s(t) \cos n\omega t \, dt$$

Figure 7.11 Piecewise polynomial periodic function exhibiting jump discontinuities.



Defining the magnitude of the jump discontinuities as in Section 5.2.11, namely

$$d_i = f(t_i + 0) - f(t_i - 0)$$

and noting that $t_0 = -\frac{1}{2}T$ and $t_m = \frac{1}{2}T$, integration by parts and summation gives

$$a_n = -\frac{1}{n\pi} \sum_{s=1}^m \left[d_s \sin n\omega t_s + \int_{t_{s-1}}^{t_s} P_s^{(1)}(t) \sin n\omega t \, dt \right] \quad (7.20)$$

where $P_s^{(1)}(t)$ denotes the piecewise components of the derivative $f^{(1)}(t) \equiv f'(t)$ in the generalized sense of (5.21). In a similar manner the integral terms of (7.20) may be expressed as

$$\sum_{s=1}^m \int_{t_{s-1}}^{t_s} P_s^{(1)} \sin n\omega t \, dt = \frac{1}{n\omega} \sum_{s=1}^m \left[d_s^{(1)} \cos n\omega t_s + \int_{t_{s-1}}^{t_s} P_s^{(2)}(t) \cos n\omega t \, dt \right]$$

where $d_s^{(1)}$ ($s = 1, 2, \dots, m$) denotes the magnitude of the jump discontinuities in the derivative $f^{(1)}(t)$.

Continuing in this fashion, integrals involving higher derivatives may be obtained. However, since all $P_s(t)$ ($s = 1, 2, \dots, m$) are polynomials, a stage is reached when all the integrals vanish. If the $\deg(P_s(t)) \leq N$ for $s = 1, 2, \dots, m$ then

$$a_n = \frac{1}{n\pi} \sum_{s=1}^m \sum_{r=0}^N (-1)^{r+1} (n\omega)^{-2r} [d_s^{(2r)} \sin n\omega t_s + (n\omega)^{-1} d_s^{(2r+1)} \cos n\omega t_s] \quad (n \neq 0) \quad (7.21)$$

where $d_s^{(r)}$ denotes the magnitudes of the jump discontinuities in the r th derivative of $f(t)$ according to (5.21). Similarly, it may be shown that

$$b_n = \frac{1}{n\pi} \sum_{s=1}^m \sum_{r=0}^N (-1)^r (n\omega)^{-2r} [d_s^{(2r)} \cos n\omega t_s - (n\omega)^{-1} d_s^{(2r+1)} \sin n\omega t_s] \quad (7.22)$$

and the coefficient a_0 is found by direct integration of the corresponding Euler formula

$$a_0 = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \, dt \quad (7.23)$$

Example 7.6

Using (7.21)–(7.23), obtain the Fourier series expansion of the periodic function $f(t)$ defined by

$$f(t) = \begin{cases} t^2 & (-\pi < t < 0) \\ -2 & (0 < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

Solution In this case $N = 2$, and the graphs of $f(t)$ together with those of its first two derivatives are shown in Figure 7.12.

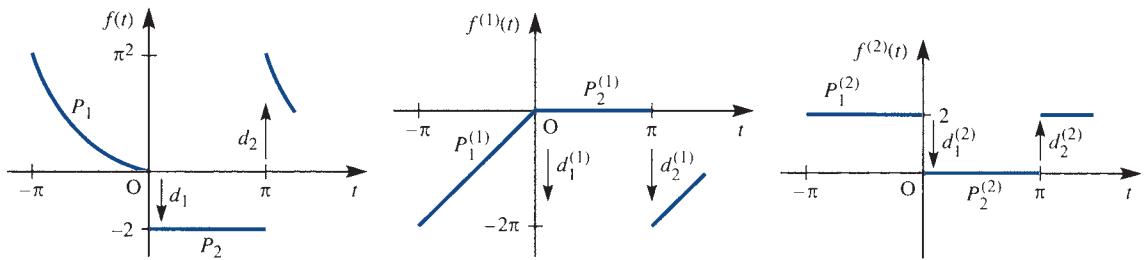


Figure 7.12 The functions $f(t)$, $f^{(1)}(t)$, $f^{(2)}(t)$ of Example 7.6.

Jump discontinuities occur at $t = -\pi$, 0 and π , and so $m = 2$. The piecewise polynomials involved and the corresponding jump discontinuities are

$$(a) \quad P_1(t) = t^2, \quad P_2(t) = -2 \\ d_1 = -2, \quad d_2 = \pi^2 + 2$$

$$(b) \quad P_1^{(1)}(t) = 2t, \quad P_2^{(1)}(t) = 0 \\ d_1^{(1)} = 0, \quad d_2^{(1)} = -2\pi$$

$$(c) \quad P_1^{(2)}(t) = 2, \quad P_2^{(2)}(t) = 0 \\ d_1^{(2)} = -2, \quad d_2^{(2)} = 2$$

with $d_1^{(r)} = d_2^{(r)} = 0$ for $r > 2$. Taking $\omega = 1$ (since $T = 2\pi$) in (7.21) gives

$$a_n = \frac{1}{n\pi} \left(- \sum_{s=1}^2 d_s \sin nt_s - \frac{1}{n} \sum_{s=1}^2 d_s^{(1)} \cos nt_s + \frac{1}{n^2} \sum_{s=1}^2 d_s^{(2)} \sin nt_s \right)$$

Since $t_1 = 0$, $t_2 = \pi$, $\sin 0 = \sin n\pi = 0$, $\cos 0 = 1$ and $\cos n\pi = (-1)^n$, we have

$$a_n = \frac{2}{n^2} (-1)^n \quad (n = 1, 2, 3, \dots)$$

Likewise, from (7.22),

$$\begin{aligned} b_n &= \frac{1}{n\pi} \left(\sum_{s=1}^2 d_s \cos nt_s - \frac{1}{n} \sum_{s=1}^2 d_s^{(1)} \sin nt_s - \frac{1}{n^2} \sum_{s=1}^2 d_s^{(2)} \cos nt_s \right) \\ &= \frac{1}{n\pi} \left\{ -2 + (\pi^2 + 2)(-1)^n - \frac{1}{n^2} [-2 + 2(-1)^n] \right\} \text{ (by writing out each series)} \\ &= \frac{1}{n\pi} \left\{ \left(\frac{2}{n^2} - 2 \right) [1 - (-1)^n] + \pi^2 (-1)^n \right\} \quad (n = 1, 2, 3, \dots) \end{aligned}$$

and, from (7.23),

$$a_0 = \frac{1}{\pi} \left[\int_{-\pi}^0 t^2 dt + \int_0^\pi (-2) dt \right] = \frac{1}{3}\pi^2 - 2$$

Thus the Fourier expansion for $f(t)$ is

$$f(t) = (\frac{1}{6}\pi^2 - 1) + \sum_{n=1}^{\infty} \frac{2}{n^2} (-1)^n \cos nt + \sum_{n=1}^{\infty} \frac{1}{n\pi} \left\{ \left(\frac{2}{n^2} - 2 \right) [1 - (-1)^n] + \pi^2 (-1)^n \right\} \sin nt$$

7.2.1 Exercises

- 11 Consider the periodic function

$$f(t) = \begin{cases} 0 & (-\pi < t < -\frac{1}{2}\pi) \\ \pi + 2t & (-\frac{1}{2}\pi < t < 0) \\ \pi - 2t & (0 < t < \frac{1}{2}\pi) \\ 0 & (\frac{1}{2}\pi < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

- (a) Sketch a graph of the function for $-4\pi < t < 4\pi$.
 (b) Use (7.21)–(7.23) to obtain the Fourier series expansion

$$f(t) = \frac{1}{4}\pi - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{n^2} (\cos \frac{1}{2}n\pi - 1) \cos nt$$

and write out the first 10 terms of this series.

(Note: Although the function $f(t)$ itself has no jump discontinuities, the method may be used since the derivative does have jump discontinuities.)

- 12 Use the method of Section 7.2 to obtain the Fourier series expansions for the following periodic functions:

$$(a) f(t) = \begin{cases} 0 & (-\pi < t < 0) \\ t^2 & (0 < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

$$(b) f(t) = \begin{cases} 2 & (-\pi < t < -\frac{1}{2}\pi) \\ t^3 & (-\frac{1}{2}\pi < t < \frac{1}{2}\pi) \\ -2 & (\frac{1}{2}\pi < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

$$(c) f(t) = \begin{cases} t & (0 < t < 1) \\ 1-t & (1 < t < 2) \end{cases}$$

$$f(t + 2) = f(t)$$

$$(d) f(t) = \begin{cases} \frac{1}{2} + t & (-\frac{1}{2} < t < 0) \\ \frac{1}{2} - t & (0 < t < \frac{1}{2}) \end{cases}$$

$$f(t + 1) = f(t)$$

7.3 Engineering application: frequency response and oscillating systems

7.3.1 Response to periodic input

In Section 5.5 we showed that the frequency response, defined as the steady-state response to a sinusoidal input $A \sin \omega t$, of a stable linear system having a transfer function $G(s)$ is given by (5.63) as

$$x_{ss}(t) = A |G(j\omega)| \sin [\omega t + \arg G(j\omega)] \quad (7.24)$$

By employing a Fourier series expansion, we can use this result to determine the steady-state response of a stable linear system to a non-sinusoidal periodic input. For a stable linear system having a transfer function $G(s)$, let the input be a periodic function $P(t)$ of period $2T$ (that is, one having frequency $\omega = \pi/T$ in rad s⁻¹). $P(t)$ may be expressed in the form of the Fourier series expansion

$$P(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} A_n \sin(n\omega t + \phi_n) \quad (7.25)$$

where A_n and ϕ_n are defined as in Section 7.1.1. The steady-state response to each term in the series expansion (7.25) may be obtained using (7.24). Since the system is linear, the principle of superposition holds, so that the steady-state response to the periodic input $P(t)$ may be obtained as the sum of the steady-state responses to the individual sinusoids comprising the sum in (7.25). Thus the steady-state response to the input $P(t)$ is

$$x_{ss}(t) = \frac{1}{2}a_0 G(0) + \sum_{n=1}^{\infty} A_n |G(jn\omega)| \sin[n\omega t + \phi_n + \arg G(jn\omega)] \quad (7.26)$$

There are two issues related to this steady-state response that are worthy of note.

- (a) For practical systems $|G(j\omega)| \rightarrow 0$ as $\omega \rightarrow \infty$, so that $|G(jn\omega)| \rightarrow 0$ as $n \rightarrow \infty$ in (7.26). As a consequence, the Fourier series representation of the steady-state response $x_{ss}(t)$ converges more rapidly than the Fourier series representation of the periodic input $P(t)$. From a practical point of view, this is not surprising, since it is a consequence of the smoothing action of the system (that is, integration is a ‘smoothing’ operation).
- (b) There is a significant difference between the steady-state response (7.26) to a non-sinusoidal periodic input of frequency ω and the steady-state response (7.23) to a pure sinusoid at the same frequency. As indicated in (7.24), in the case of a sinusoidal input at frequency ω the steady-state response is also a sinusoid at the same frequency ω . However, for a non-sinusoidal periodic input $P(t)$ at frequency ω the steady-state response (7.26) is no longer at the same frequency; rather it comprises an infinite sum of sinusoids having frequencies $n\omega$ that are integer multiples of the input frequency ω . This clearly has important practical implications, particularly when considering the responses of oscillating or vibrating systems. If the frequency $n\omega$ of one of the harmonics in (7.26) is close to the natural oscillating frequency of an underdamped system then the phenomenon of **resonance** will arise.

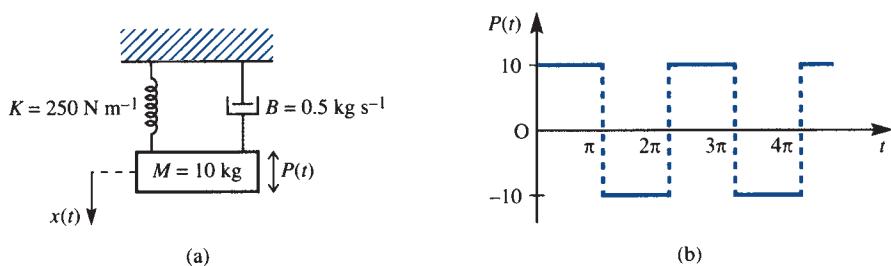
To someone unfamiliar with the theory, it may seem surprising that a practical system may resonate at a frequency much higher than that of the input. The phenomenon of resonance is important in practice, and it is therefore important that engineers

have some knowledge of the theory associated with Fourier series, so that the possible dominance of a system response by one of the higher harmonics, rather than the fundamental, may be properly interpreted.

Example 7.7

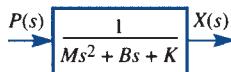
The mass–spring–damper system of Figure 7.13(a) is initially at rest in a position of equilibrium. Determine the steady-state response of the system when the mass is subjected to an externally applied periodic force $P(t)$ having the form of the square wave shown in Figure 7.13(b).

Figure 7.13 (a) System and (b) input for Example 7.7.



Solution From Newton's law, the displacement $x(t)$ of the mass at time t is given by

$$M \frac{d^2x}{dt^2} + B \frac{dx}{dt} + Kx = P(t) \quad (7.27)$$



so that the system may be represented by the block diagram of Figure 7.14. Thus the system transfer function is

$$G(s) = \frac{1}{Ms^2 + Bs + K} \quad (7.28)$$

From Example 7.1, the Fourier series expansion for the square wave $P(t)$ is

$$P(t) = \frac{40}{\pi} \left[\sin t + \frac{\sin 3t}{3} + \frac{\sin 5t}{5} + \dots + \frac{\sin(2n-1)t}{2n-1} + \dots \right]$$

that is,

$$P(t) = u_1(t) + u_2(t) + u_3(t) + \dots + u_n(t) + \dots \quad (7.29)$$

where

$$u_n(t) = \frac{40}{\pi} \frac{\sin(2n-1)t}{2n-1} \quad (7.30)$$

Substituting the given values for M , B and K , the transfer function (7.18) becomes

$$G(s) = \frac{1}{10s^2 + 0.5s + 250}$$

Thus

$$G(j\omega) = \frac{1}{-10\omega^2 + 0.5j\omega + 250} = \frac{250 - 10\omega^2}{D} - j \frac{0.5\omega}{D}$$

where $D = (250 - 10\omega^2)^2 + 0.25\omega^2$, so that

$$\begin{aligned} |G(j\omega)| &= \sqrt{\left[\frac{(250 - 10\omega^2)^2 + 0.25\omega^2}{D^2} \right]} \\ &= \frac{1}{\sqrt{D}} = \frac{1}{\sqrt{[(250 - 10\omega^2)^2 + 0.25\omega^2]}} \end{aligned} \quad (7.31)$$

$$\arg G(j\omega) = -\tan^{-1}\left(\frac{0.5\omega}{250 - 10\omega^2}\right) \quad (7.32)$$

Using (7.24), the steady-state response of the system to the n th harmonic $u_n(t)$ given by (7.30) is

$$x_{ssn}(t) = \frac{40}{\pi(2n-1)} |G(j(2n-1))| \sin[(2n-1)t + \arg G(j(2n-1))] \quad (7.33)$$

where $|G(j\omega)|$ and $\arg G(j\omega)$ are given by (7.31) and (7.32) respectively. The steady-state response $x_{ss}(t)$ of the system to the square-wave input $P(t)$ is then determined as the sum of the steady-state responses due to the individual harmonics in (7.29); that is,

$$x_{ss}(t) = \sum_{n=1}^{\infty} x_{ssn}(t) \quad (7.34)$$

where $x_{ssn}(t)$ is given by (7.33).

Evaluating the first few terms of the response (7.34), we have

$$x_{ss1}(t) = \frac{40}{\pi} \frac{1}{\sqrt{[(250 - 10)^2 + 0.25]}} \sin\left[t - \tan^{-1}\left(\frac{0.5}{240}\right)\right] = 0.053 \sin(t - 0.003)$$

$$x_{ss2}(t) = \frac{40}{3\pi} \frac{1}{\sqrt{[(250 - 90)^2 + 2.25]}} \sin \left[3t - \tan^{-1} \left(\frac{1.5}{160} \right) \right] = 0.027 \sin(3t - 0.009)$$

$$x_{ss3}(t) = \frac{40}{5\pi} \frac{1}{\sqrt{(6.25)}} \sin \left[5t - \tan^{-1} \left(\frac{2.5}{0} \right) \right] = 1.02 \sin(5t - \frac{1}{2}\pi)$$

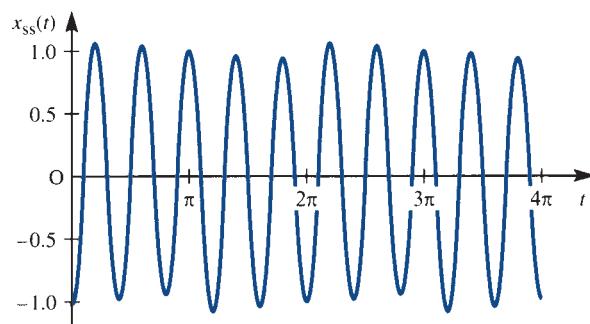
$$x_{ss4}(t) = \frac{40}{7\pi} \frac{1}{\sqrt{[(250 - 490)^2 + 12.25]}} \sin \left[7t - \tan^{-1} \left(\frac{3.5}{-240} \right) \right] = 0.0076 \sin(7t - 3.127)$$

Thus a good approximation to the steady-state response (7.34) is

$$\begin{aligned} x_{ss}(t) \simeq & 0.053 \sin(t - 0.003) + 0.027 \sin(3t - 0.54) + 1.02 \sin(5t - \frac{1}{2}\pi) \\ & + 0.0076 \sin(7t - 3.127) \end{aligned} \quad (7.35)$$

The graph of this displacement is shown in Figure 7.15, and it appears from this that the response has a frequency about five times that of the input. This is because the term $1.02 \sin(5t - \frac{1}{2}\pi)$ dominates in the response (7.35); this is a consequence of the fact that the natural frequency of oscillation of the system is $\sqrt{(K/M)} = 5 \text{ rad s}^{-1}$, so that it is in resonance with this particular harmonic.

Figure 7.15
Steady-state response
of system of
Figure 7.13.



In conclusion, it should be noted that it was not essential to introduce transfer functions to solve this problem. Alternatively, by determining the particular integral of the differential equation (7.27), the steady-state response to an input $A \sin \omega t$ is determined as

$$x_{ss}(t) = \frac{A \sin(\omega t - \alpha)}{\sqrt{[(K - M\omega^2)^2 + B^2\omega^2]}}, \quad \tan \alpha = \frac{\omega B}{K - M\omega^2}$$

giving $x_{ssn}(t)$ as in (7.34). The solution then proceeds as before.

7.3.2 Exercises

13

Determine the steady-state current in the circuit of Figure 7.16(a) as a result of the applied periodic voltage shown in Figure 7.16(b).

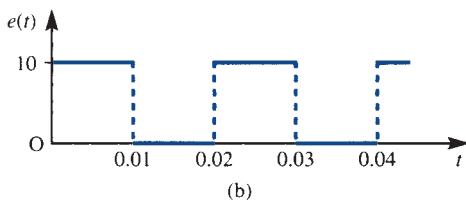
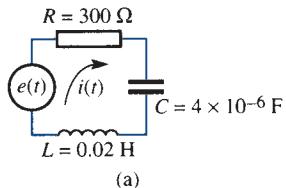


Figure 7.16 (a) Circuit of Exercise 13; (b) applied voltage.

14

Determine the steady-state response of the mass–spring–damper system of Figure 7.17(a) when the mass is subjected to the externally applied periodic force $f(t)$ shown in Figure 7.17(b).

What frequency dominates the response, and why?

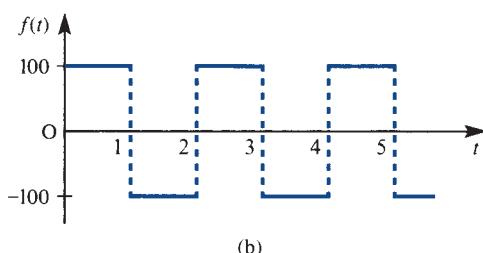
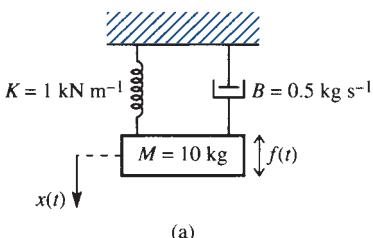


Figure 7.17 (a) Mass–spring–damper system of Exercise 14; (b) applied force.

15

Determine the steady-state motion of the mass of Figure 7.18(a) when it is subjected to the externally applied force of Figure 7.18(b).

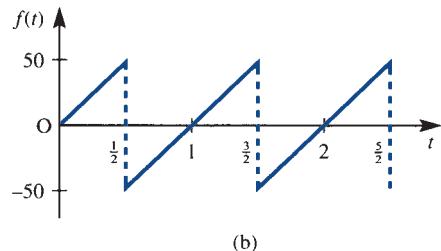
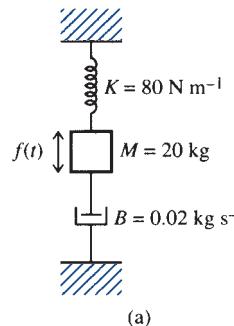


Figure 7.18 (a) Mass–spring–damper system of Exercise 15; (b) applied force.

16

Determine the steady-state current in the circuit shown in Figure 7.19(a) when the applied voltage is of the form shown in Figure 7.19(b).

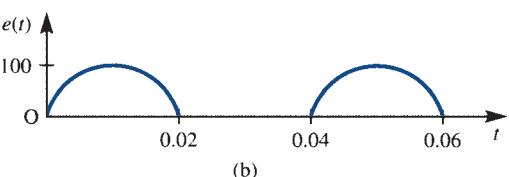
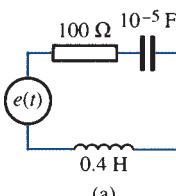


Figure 7.19 (a) Circuit of Exercise 16; (b) applied voltage.

7.4

Complex form of Fourier series

An alternative to the trigonometric form of the Fourier series considered so far is the complex or exponential form. As a result of the properties of the exponential function, this form is easily manipulated mathematically. It is widely used by engineers in practice, particularly in work involving signal analysis, and provides a smoother transition from the consideration of Fourier series for dealing with periodic signals to the consideration of Fourier transforms for dealing with aperiodic signals, which will be dealt with in Chapter 8.

7.4.1 Complex representation

To develop the complex form of the Fourier series

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t \quad (7.36)$$

representing a periodic function $f(t)$ of period T , we proceed as follows. Substituting the results

$$\sin n\omega t = \frac{1}{2j}(e^{jn\omega t} - e^{-jn\omega t})$$

$$\cos n\omega t = \frac{1}{2}(e^{jn\omega t} + e^{-jn\omega t})$$

into (7.36) gives

$$\begin{aligned} f(t) &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \frac{e^{jn\omega t} + e^{-jn\omega t}}{2} + \sum_{n=1}^{\infty} b_n \frac{e^{jn\omega t} - e^{-jn\omega t}}{2j} \\ &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} \frac{1}{2}a_n(e^{jn\omega t} + e^{-jn\omega t}) + \sum_{n=1}^{\infty} -\frac{1}{2}jb_n(e^{jn\omega t} - e^{-jn\omega t}) \\ &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} [\frac{1}{2}(a_n - jb_n)e^{jn\omega t} + \frac{1}{2}(a_n + jb_n)e^{-jn\omega t}] \end{aligned} \quad (7.37)$$

Writing

$$c_0 = \frac{1}{2}a_0, \quad c_n = \frac{1}{2}(a_n - jb_n), \quad c_{-n} = c_n^* = \frac{1}{2}(a_n + jb_n) \quad (7.38)$$

(7.37) becomes

$$\begin{aligned} f(t) &= c_0 + \sum_{n=1}^{\infty} c_n e^{jn\omega t} + \sum_{n=1}^{\infty} c_{-n} e^{-jn\omega t} = c_0 + \sum_{n=1}^{\infty} c_n e^{jn\omega t} + \sum_{n=-1}^{-\infty} c_n e^{jn\omega t} \\ &= \sum_{n=-\infty}^{\infty} c_n e^{jn\omega t}, \quad \text{since } c_0 e^0 = c_0 \end{aligned}$$

Thus the Fourier series (7.36) becomes simply

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega t} \quad (7.39)$$

which is referred to as the **complex** or **exponential form** of the Fourier series expansion of the function $f(t)$.

In order that we can apply this result directly, it is necessary to obtain a formula for calculating the complex coefficients c_n . To do this, we incorporate the definitions of the Fourier coefficients into the definitions given in (7.38), leading to

$$c_0 = \frac{1}{2}a_0 = \frac{1}{T} \int_d^{d+T} f(t) dt \quad (7.40)$$

$$\begin{aligned} c_n &= \frac{1}{2}(a_n - jb_n) = \frac{1}{T} \left[\int_d^{d+T} f(t) \cos n\omega t dt - j \int_d^{d+T} f(t) \sin n\omega t dt \right] \\ &= \frac{1}{T} \int_d^{d+T} f(t) (\cos n\omega t - j \sin n\omega t) dt \\ &= \frac{1}{T} \int_d^{d+T} f(t) e^{-jn\omega t} dt \end{aligned} \quad (7.41)$$

$$\begin{aligned} c_{-n} &= \frac{1}{2}(a_n + jb_n) = \frac{1}{T} \int_d^{d+T} f(t) (\cos n\omega t + j \sin n\omega t) dt \\ &= \frac{1}{T} \int_d^{d+T} f(t) e^{jn\omega t} dt \end{aligned} \quad (7.42)$$

From (7.40)–(7.42), it is readily seen that for all values of n

$$c_n = \frac{1}{T} \int_d^{d+T} f(t) e^{-jn\omega t} dt \quad (7.43)$$

Summary

In summary, the complex form of the Fourier series expansion of a periodic function $f(t)$, of period T , is

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn\omega t} \quad (7.39)$$

where

$$c_n = \frac{1}{T} \int_d^{d+T} f(t) e^{-jn\omega t} dt \quad (n = 0, \pm 1, \pm 2, \dots) \quad (7.40)$$

In general the coefficients c_n ($n = 0, \pm 1, \pm 2, \dots$) are complex, and may be expressed in the form

$$c_n = |c_n| e^{j\phi_n}$$

where $|c_n|$, the magnitude of c_n , is given from the definitions (7.38) by

$$|c_n| = \sqrt{\left(\frac{1}{2}a_n\right)^2 + \left(\frac{1}{2}b_n\right)^2} = \frac{1}{2}\sqrt{(a_n^2 + b_n^2)}$$

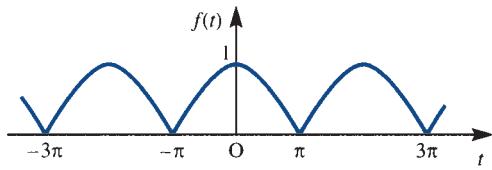
so that $2|c_n|$ is the amplitude of the n th harmonic. The argument ϕ_n of c_n is related to the phase of the n th harmonic.

Example 7.8

Find the complex form of the Fourier series expansion of the periodic function $f(t)$ defined by

$$f(t) = \cos \frac{1}{2}t \quad (-\pi < t < \pi), \quad f(t + 2\pi) = f(t)$$

Figure 7.20 Function $f(t)$ of Example 7.8.



Solution

A graph of the function $f(t)$ over the interval $-3\pi \leq t \leq 3\pi$ is shown in Figure 7.20. Here the period T is 2π , so from (7.43) the complex coefficients c_n are given by

$$\begin{aligned} c_n &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos \frac{1}{2}t e^{-jnt} dt = \frac{1}{4\pi} \int_{-\pi}^{\pi} (e^{jt/2} + e^{-jt/2}) e^{-jnt} dt \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} (e^{-j(n-1/2)t} + e^{-j(n+1/2)t}) dt \\ &= \frac{1}{4\pi} \left[\frac{-2e^{-j(2n-1)t/2}}{j(2n-1)} - \frac{2e^{-j(2n+1)t/2}}{j(2n+1)} \right]_{-\pi}^{\pi} \\ &\quad \frac{j}{2\pi} \left[\left(\frac{e^{-jn\pi} e^{j\pi/2}}{2n-1} + \frac{e^{-jn\pi} e^{-j\pi/2}}{2n+1} \right) - \left(\frac{e^{jn\pi} e^{-j\pi/2}}{2n-1} + \frac{e^{jn\pi} e^{j\pi/2}}{2n+1} \right) \right] \end{aligned}$$

Now $e^{jn\pi/2} = \cos \frac{1}{2}\pi + j \sin \frac{1}{2}\pi = j$, $e^{-jn\pi/2} = -j$ and $e^{jn\pi} = e^{-jn\pi} = \cos n\pi = (-1)^n$, so that

$$\begin{aligned} c_n &= \frac{j}{2\pi} \left(\frac{j}{2n-1} - \frac{j}{2n+1} + \frac{j}{2n-1} - \frac{j}{2n+1} \right) (-1)^n \\ &= \frac{(-1)^n}{\pi} \left(\frac{1}{2n+1} - \frac{1}{2n-1} \right) = \frac{-2(-1)^n}{(4n^2-1)\pi} \end{aligned}$$

Note that in this case c_n is real, which is as expected, since the function $f(t)$ is an even function of t .

From (7.39), the complex Fourier series expansion for $f(t)$ is

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{2(-1)^{n+1}}{(4n^2-1)\pi} e^{jnt}$$

This may readily be converted back to the trigonometric form, as by definitions (7.38),

$$a_0 = 2c_0, \quad a_n = c_n + c_n^*, \quad b_n = j(c_n - c_n^*)$$

so that in this particular case

$$a_0 = \frac{4}{\pi}, \quad a_n = 2 \left[\frac{2}{\pi} \frac{(-2)^{n+1}}{4n^2 + 1} \right] = \frac{4}{\pi} \frac{(-1)^{n+1}}{4n^2 - 1}, \quad b_n = 0$$

Thus the trigonometric form of the Fourier series is

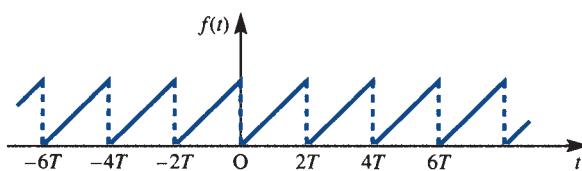
$$f(t) = \frac{2}{\pi} + \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{4n^2 - 1} \cos nt$$

Example 7.9

Obtain the complex form of the Fourier series of the sawtooth function $f(t)$ defined by

$$f(t) = \frac{2t}{T} \quad (0 < t < 2T), \quad f(t + 2T) = f(t)$$

Figure 7.21 Function $f(t)$ of Example 7.9.



Solution

A graph of the function $f(t)$ over the interval $-6T < t < 6T$ is shown in Figure 7.21. Here the period is $2T$, that is $\omega = \pi/T$, so from (7.43) the complex coefficients c_n are given by

$$\begin{aligned} c_n &= \frac{1}{2T} \int_0^{2T} f(t) e^{-jn\pi t/T} dt = \frac{1}{2T} \int_0^{2T} \frac{2}{T} t e^{-jn\pi t/T} dt \\ &= \frac{1}{T^2} \left[\frac{Tt}{-jn\pi} e^{-jn\pi t/T} - \frac{T^2}{(jn\pi)^2} e^{-jn\pi t/T} \right]_0^{2T} \quad (n \neq 0) \end{aligned}$$

Now $e^{-jn2\pi} = e^{-j0} = 1$, so

$$c_n = \frac{1}{T^2} \left[\frac{2T^2}{-jn\pi} + \frac{T^2}{(n\pi)^2} - \frac{T^2}{(n\pi)^2} \right] = \frac{j2}{n\pi} \quad (n \neq 0)$$

In the particular case $n = 0$

$$c_0 = \frac{1}{2T} \int_0^{2T} f(t) dt = \frac{1}{2T} \int_0^{2T} \frac{2t}{T} dt = \frac{1}{T^2} \left[\frac{1}{2} t^2 \right]_0^{2T} = 2$$

Thus from (7.39) the complex form of the Fourier series expansion of $f(t)$ is

$$f(t) = 2 + \sum_{n=-\infty}^{-1} \frac{j2}{n\pi} e^{jn\pi t/T} + \sum_{n=1}^{\infty} \frac{j2}{n\pi} e^{jn\pi t/T} = 2 + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{j2}{n\pi} e^{jn\pi t/T}$$

Noting that $j = e^{j\pi/2}$, this result may also be written in the form

$$f(t) = 2 + \frac{2}{\pi} \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{1}{n} e^{j(n\pi t/T + \pi/2)}$$

As in Example 7.8, the Euler coefficients in the corresponding trigonometric series are

$$a_0 = 2c_0 = 4, \quad a_n = c_n + c_n^* = 0, \quad b_n = j(c_n + c_n^*) = j\left(\frac{2j}{n\pi} + \frac{2j}{n\pi}\right) = -\frac{4}{n\pi}$$

so that the corresponding trigonometric Fourier series expansion of $f(t)$ is

$$f(t) = 2 - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin \frac{n\pi t}{T}$$

7.4.2 The multiplication theorem and Parseval's theorem

Two useful results, particularly in the application of Fourier series to signal analysis, are the **multiplication theorem** and **Parseval's theorem**. The multiplication theorem enables us to write down the mean value of the product of two periodic functions over a period in terms of the coefficients of their Fourier series expansions, while Parseval's theorem enables us to write down the mean square value of a periodic function, which, as we will see in Section 7.4.4, determines the power spectrum of the function.

Theorem 7.1

The multiplication theorem

If $f(t)$ and $g(t)$ are two periodic functions having the same period T then

$$\frac{1}{T} \int_c^{c+T} f(t)g(t) dt = \sum_{n=-\infty}^{\infty} c_n d_n^* \quad (7.44)$$

where the c_n and d_n are the coefficients in the complex Fourier series expansions of $f(t)$ and $g(t)$ respectively.

Proof Let $f(t)$ and $g(t)$ have complex Fourier series given by

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{jn2\pi t/T} \quad (7.45a)$$

with

$$c_n = \frac{1}{T} \int_c^{c+T} f(t) e^{-jn2\pi t/T} dt \quad (7.45b)$$

and

$$g(t) = \sum_{n=-\infty}^{\infty} d_n e^{jn2\pi t/T} \quad (7.46a)$$

with

$$d_n = \frac{1}{T} \int_c^{c+T} g(t) e^{-jn2\pi t/T} dt \quad (7.46b)$$

Then

$$\begin{aligned} \frac{1}{T} \int_c^{c+T} f(t) g(t) dt &= \frac{1}{T} \int_c^{c+T} \left(\sum_{n=-\infty}^{\infty} c_n e^{jn2\pi t/T} \right) g(t) dt && \text{using (7.45a)} \\ &= \sum_{n=-\infty}^{\infty} c_n \left[\frac{1}{T} \int_c^{c+T} g(t) e^{jn2\pi t/T} dt \right] && \text{assuming term-by-term} \\ &= \sum_{n=-\infty}^{\infty} c_n d_{-n} && \text{integration is possible} \\ &&& \text{using (7.45b)} \end{aligned}$$

Since $d_{-n} = d_n^*$, the complex conjugate of d_n , this reduces to the required result:

$$\frac{1}{T} \int_c^{c+T} f(t) g(t) dt = \sum_{n=-\infty}^{\infty} c_n d_n^*$$

end of theorem

In terms of the real coefficients a_n , b_n and α_n , β_n of the corresponding trigonometric Fourier series expansions of $f(t)$ and $g(t)$,

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n2\pi t}{T}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n2\pi t}{T}\right)$$

$$g(t) = \frac{1}{2}\alpha_0 + \sum_{n=1}^{\infty} \alpha_n \cos\left(\frac{n2\pi t}{T}\right) + \sum_{n=1}^{\infty} \beta_n \sin\left(\frac{n2\pi t}{T}\right)$$

and using the definitions (7.38), the multiplication theorem result (7.44) reduces to

$$\begin{aligned} \frac{1}{T} \int_c^{c+T} f(t) g(t) dt &= \sum_{n=1}^{\infty} c_{-n} d_n + c_0 d_0 + \sum_{n=1}^{\infty} c_n d_{-n} \\ &= \frac{1}{4} \alpha_0 a_0 + \frac{1}{4} \sum_{n=1}^{\infty} [(a_n - jb_n)(\alpha_n + jb_n) + (a_n + jb_n)(\alpha_n - jb_n)] \end{aligned}$$

giving

$$\frac{1}{T} \int_c^{c+T} f(t) g(t) dt = \frac{1}{4} \alpha_0 a_0 + \frac{1}{2} \sum_{n=1}^{\infty} (a_n \alpha_n + b_n \beta_n)$$

Theorem 7.2 Parseval's theorem

If $f(t)$ is a periodic function with period T then

$$\frac{1}{T} \int_c^{c+T} [f(t)]^2 dt = \sum_{n=-\infty}^{\infty} c_n c_n^* = \sum_{n=-\infty}^{\infty} |c_n|^2 \quad (7.47)$$

where the c_n are the coefficients in the complex Fourier series expansion of $f(t)$.

Proof This result follows from the multiplication theorem, since, taking $g(t) = f(t)$ in (7.44), we obtain

$$\frac{1}{T} \int_c^{c+T} [f(t)]^2 dt = \sum_{n=-\infty}^{\infty} c_n c_n^* = \sum_{n=-\infty}^{\infty} |c_n|^2$$

[end of theorem](#)

Using (7.46), Parseval's theorem may be written in terms of the real coefficients a_n and b_n of the trigonometric Fourier series expansion of the function $f(t)$ as

$$\frac{1}{T} \int_c^{c+T} [f(t)]^2 dt = \frac{1}{4} a_0^2 + \frac{1}{2} \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \quad (7.48)$$

The **root mean square (RMS)** value f_{RMS} of a periodic function $f(t)$ of period T , defined by

$$f_{\text{RMS}}^2 = \frac{1}{T} \int_c^{c+T} [f(t)]^2 dt$$

may therefore be expressed in terms of the Fourier coefficients using (7.47) or (7.48).

Example 7.10

By applying Parseval's theorem to the function

$$f(t) = \frac{2t}{T} \quad (0 < t < T), \quad f(t + 2T) = f(t)$$

considered in Example 7.9, show that

$$\frac{1}{6}\pi^2 = \sum_{n=1}^{\infty} \frac{1}{n^2}$$

Solution From Example 7.9, the coefficients of the complex Fourier series expansion of $f(t)$ are

$$c_0 = 2, c_n = \frac{j2}{n\pi} \quad (n \neq 0)$$

Thus, applying the Parseval's theorem result (7.47), noting that the period in this case is $2T$, we obtain

$$\frac{1}{2T} \int_0^{2T} [f(t)]^2 dt = c_0^2 + \sum_{n=-\infty}^{-1} |c_n|^2 + \sum_{n=1}^{\infty} |c_n|^2$$

giving

$$\frac{1}{2T} \int_0^{2T} \frac{4t^2}{T^2} dt = 4 + 2 \sum_{n=1}^{\infty} \left(\frac{2}{n\pi} \right)^2$$

which reduces to

$$\frac{16}{3} = 4 + \sum_{n=1}^{\infty} \frac{8}{n^2 \pi^2}$$

and leads to the required result

$$\frac{1}{6}\pi^2 = \sum_{n=1}^{\infty} \frac{1}{n^2}$$

7.4.3 Discrete frequency spectra

In expressing a periodic function $f(t)$ by its Fourier series expansion, we are decomposing the function into its **harmonic** or **frequency components**. We have seen that if $f(t)$ is of period T then it has frequency components at frequencies

$$\omega_n = \frac{2n\pi}{T} = n\omega_0 \quad (n = 1, 2, 3, \dots) \quad (7.49)$$

where ω_0 is the frequency of the parent function $f(t)$. (All frequencies here are measured in rad s⁻¹.)

A Fourier series may therefore be interpreted as constituting a **frequency spectrum** of the periodic function $f(t)$, and provides an alternative representation of the function to its time-domain waveform. This frequency spectrum is often displayed by plotting graphs of both the amplitudes and phases of the various harmonic components against angular frequency ω_n . A plot of amplitude against angular frequency is called the **amplitude spectrum**, while that of phase against angular frequency is called the **phase spectrum**. For a periodic function $f(t)$, of period T , harmonic components only occur at discrete frequencies ω_n , given by (7.45), so that these spectra are referred to as **discrete frequency spectra** or **line spectra**. In Chapter 8 Fourier transforms will be used to define continuous spectra for aperiodic functions. With the growing ability to process signals digitally, the representation of signals by their corresponding spectra is an approach widely used in almost all branches of engineering, especially electrical engineering, when considering topics such as filtering and modulation. An example of the use of a discrete spectral representation of a periodic function is in distortion measurements on amplifiers, where the harmonic content of the output, measured digitally, to a sinusoidal input provides a measure of the distortion.

If the Fourier series expansion of a periodic function $f(t)$, with period T , has been obtained in the trigonometric form

$$f(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2n\pi t}{T}\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2n\pi t}{T}\right)$$

then this may be expressed in terms of the various harmonic components as

$$f(t) = A_0 + \sum_{n=1}^{\infty} A_n \sin\left(\frac{2n\pi t}{T} + \phi_n\right) \quad (7.50)$$

where

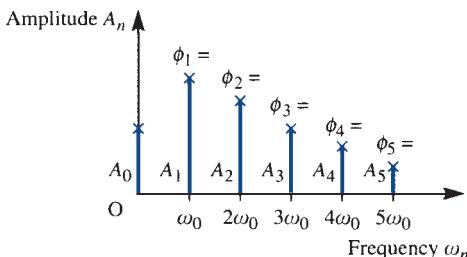
$$A_0 = \frac{1}{2}a_0, \quad A_n = \sqrt{(a_n^2 + b_n^2)}$$

and the ϕ_n are determined by

$$\sin \phi_n = \frac{b_n}{A_n}, \quad \cos \phi_n = \frac{a_n}{A_n}$$

In this case a plot of A_n against angular frequency ω_n will constitute the amplitude spectrum and that of ϕ_n against ω_n the phase spectrum. These may be incorporated in the same graph by indicating the various phases on the amplitude spectrum as illustrated in Figure 7.22. It can be seen that the amplitude spectrum consists of a series of equally spaced vertical lines whose lengths are proportional to the amplitudes of the various harmonic components making up the function $f(t)$. Clearly the trigonometric form of the Fourier series does not in general lend itself to the plotting of the discrete frequency spectrum, and the amplitudes A_n and phases ϕ_n must first be determined from the values of a_n and b_n previously determined.

Figure 7.22 Real discrete frequency spectrum.

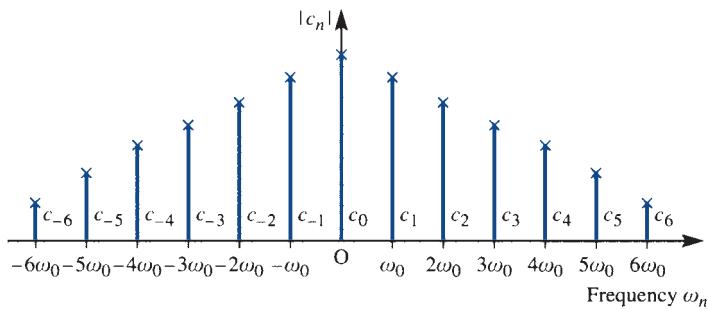


In work on signal analysis it is much more common to use the complex form of the Fourier series. For a periodic function $f(t)$, of period T , this is given by (7.39), with the complex coefficients being given by

$$c_n = |c_n| e^{j\phi_n} \quad (n = 0, \pm 1, \pm 2, \dots)$$

in which $|c_n|$ and ϕ_n denote the magnitude and argument of c_n respectively. Since in general c_n is a complex quantity, we need two line spectra to determine the discrete frequency spectrum; the amplitude spectrum being a plot of $|c_n|$ against ω_n and the phase spectrum that of ϕ_n against ω_n . In cases where c_n is real a single spectrum may be used to represent the function $f(t)$. Since $|c_{-n}| = |c_n^*| = |c_n|$, the amplitude spectrum will be symmetrical about the vertical axis, as illustrated in Figure 7.23.

Figure 7.23 Complex form of the amplitude spectrum.



Note that in the complex form of the discrete frequency spectrum we have components at the discrete frequencies $0, \pm\omega_0, \pm 2\omega_0, \pm 3\omega_0, \dots$; that is, both positive and negative discrete frequencies are involved. Clearly signals having negative frequencies are not physically realizable, and have been introduced for mathematical convenience. At frequency $n\omega_0$ we have the component $e^{jn\omega_0 t}$, which in itself is not a physical signal; to obtain a physical signal, we must consider this alongside the corresponding component $e^{-jn\omega_0 t}$ at the frequency $-n\omega_0$, since then we have

$$e^{jn\omega_0 t} + e^{-jn\omega_0 t} = 2 \cos n\omega_0 t \quad (7.51)$$

Example 7.11

Plot the discrete amplitude and phase spectra for the periodic function

$$f(t) = \frac{2t}{T} \quad (0 < t < 2T), \quad f(t + 2T) = f(t)$$

of Example 7.9. Consider both complex and real forms.

Solution

In Example 7.9 the complex coefficients were determined as

$$c_0 = 2, \quad c_n = \frac{j2}{n\pi} \quad (n = \pm 1, \pm 2, \pm 3, \dots)$$

Thus

$$|c_n| = \begin{cases} 2/n\pi & (n = 1, 2, 3, \dots) \\ -2/n\pi & (n = -1, -2, -3, \dots) \end{cases}$$

$$\phi_n = \arg c_n = \begin{cases} \frac{1}{2}\pi & (n = 1, 2, 3, \dots) \\ -\frac{1}{2}\pi & (n = -1, -2, -3, \dots) \end{cases}$$

The corresponding amplitude and phase spectra are shown in Figures 7.24(a) and (b) respectively.

In Example 7.9 we saw that the coefficients in the trigonometric form of the Fourier series expansion of $f(t)$ are

$$a_0 = 4, \quad a_n = 0, \quad b_n = -\frac{4}{n\pi}$$

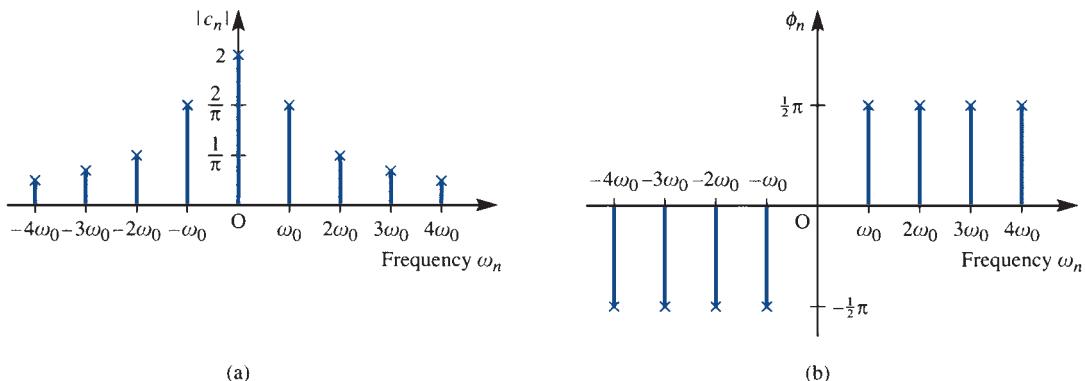
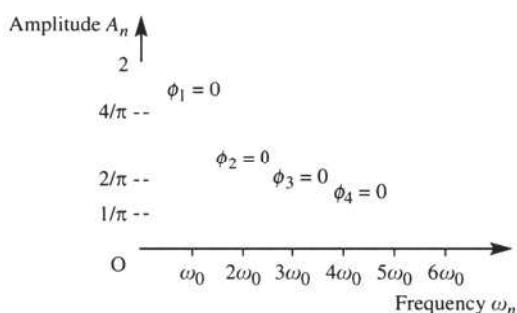


Figure 7.24 Complex discrete frequency spectra for Example 7.11, with $\omega_0 = \pi/T$: (a) amplitude spectrum; (b) phase spectrum.

Figure 7.25 Real discrete frequency spectrum for Example 7.11 (corresponding to sinusoidal expansion).



so that the amplitude coefficients in (7.49) are

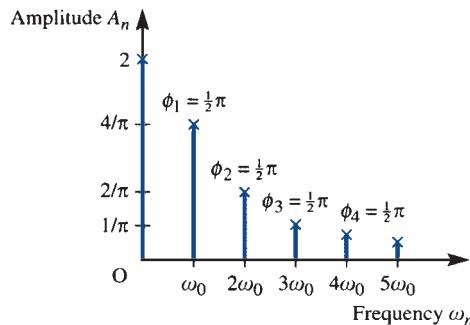
$$A_0 = 2, \quad A_n = \frac{4}{n\pi} \quad (n = 1, 2, 3, \dots)$$

leading to the real discrete frequency spectrum of Figure 7.25.

Since $|c_n| = \frac{1}{2}\sqrt{(a_n^2 + b_n^2)} = \frac{1}{2}A_n$, the amplitude spectrum lines in the complex form (Figure 7.24) are, as expected, halved in amplitude relative to those in the real representation (Figure 7.25), the other half-value being allocated to the corresponding negative frequency. In the complex representation the phases at negative frequencies (Figure 7.24b) are the negatives of those at the corresponding positive frequencies. In our particular representation (7.50) of the real form the phases at positive frequencies differ by $\frac{1}{2}\pi$ between the real and complex form. Again this is not surprising, since from (7.51) we see that combining positive and negative frequencies in the complex form leads to a cosinusoid at that frequency rather than a sinusoid. In order to maintain equality of the phases at positive frequencies between the complex and real representations, a cosinusoidal expansion

$$f(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos\left(\frac{n\pi t}{T} + \phi_n\right) \quad (7.52)$$

Figure 7.26 Real discrete frequency spectrum for Example 7.11 (corresponding to cosinusoidal expansion).



of the real Fourier series is frequently adopted as an alternative to the sinusoidal series expansion (7.50). Taking (7.52), the phase spectrum will be determined by

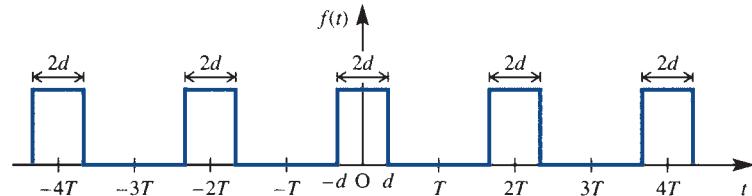
$$\sin \phi_n = -\frac{b_n}{A_n}, \quad \cos \phi_n = \frac{a_n}{A_n}$$

showing a phase shift of $\frac{1}{2}\pi$ from that of (7.50). Adopting the real representation (7.52), the corresponding real discrete frequency spectrum for the function $f(t)$ of Example 7.11 is as illustrated in Figure 7.26.

Example 7.12

Determine the complex form of the Fourier series expansion of the periodic (period $2T$) infinite train of identical rectangular pulses of magnitude A and duration $2d$ illustrated in Figure 7.27. Draw the discrete frequency spectrum in the particular case when $d = \frac{1}{10}$ and $T = \frac{1}{2}$.

Figure 7.27
Infinite train of rectangular pulses of Example 7.12.



Solution Over one period $-T < t < T$ the function $f(t)$ representing the train is expressed as

$$f(t) = \begin{cases} 0 & (-T < t < -d) \\ A & (-d < t < d) \\ 0 & (d < t < T) \end{cases}$$

From (7.43), the complex coefficients c_n are given by

$$\begin{aligned} c_n &= \frac{1}{2T} \int_{-T}^T f(t) e^{-jn\pi t/T} dt = \frac{1}{2T} \int_{-d}^d A e^{-jn\pi t/T} dt = \frac{A}{2T} \left[\frac{-T}{jn\pi} e^{-jn\pi t/T} \right]_{-d}^d \quad (n \neq 0) \\ &= \frac{A}{n\pi} \frac{e^{jn\pi d/T} - e^{-jn\pi d/T}}{j2} = \frac{A}{n\pi} \sin\left(\frac{n\pi d}{T}\right) = \frac{Ad}{T} \frac{\sin(n\pi d/T)}{n\pi d/T} \quad (n = \pm 1, \pm 2, \dots) \end{aligned}$$

In the particular case when $n = 0$

$$c_0 = \frac{1}{2T} \int_{-T}^T f(t) dt = \frac{1}{2T} \int_{-d}^d A dt = \frac{Ad}{T}$$

so that

$$c_n = \frac{Ad}{T} \operatorname{sinc}\left(\frac{n\pi d}{T}\right) \quad (n = 0, \pm 1, \pm 2, \dots)$$

where the **sinc function** is defined by

$$\operatorname{sinc} t = \begin{cases} \frac{\sin t}{t} & (t \neq 0) \\ 1 & (t = 0) \end{cases}$$

Thus from (7.39) the complex Fourier series expansion for the infinite train of pulses $f(t)$ is

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{Ad}{T} \operatorname{sinc}\left(\frac{n\pi d}{T}\right) e^{j n \pi t / T}$$

As expected, since $f(t)$ is an even function, c_n is real, so we need only plot the discrete amplitude spectrum to represent $f(t)$. Since the amplitude spectrum is a plot of $|c_n|$ against frequency $n\omega_0$, with $\omega_0 = \pi/T$, it will only take values at the discrete frequency values

$$0, \pm \frac{\pi}{T}, \pm \frac{2\pi}{T}, \pm \frac{3\pi}{T}, \dots$$

In the particular case $d = \frac{1}{10}$, $T = \frac{1}{2}$, $\omega_0 = 2\pi$ the amplitude spectrum will only exist at frequency values

$$0, \pm 2\pi, \pm 4\pi, \dots$$

Since in this case

$$c_n = \frac{1}{5} A \operatorname{sinc} \frac{1}{5} n\pi \quad (n = 0, \pm 1, \pm 2, \dots)$$

noting that $\operatorname{sinc} \frac{1}{5} n\pi = 0$ when $\frac{1}{5} n\pi = m\pi$ or $n = 5m$ ($m = \pm 1, \pm 2, \dots$), the spectrum is as shown in Figure 7.28.

Figure 7.28 Discrete amplitude spectrum for an infinite train of pulses when $d = \frac{1}{10}$ and $T = \frac{1}{2}$.

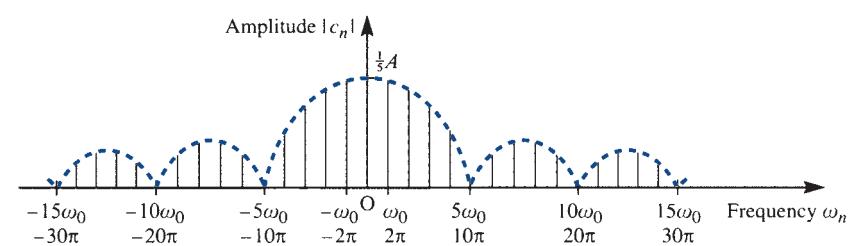
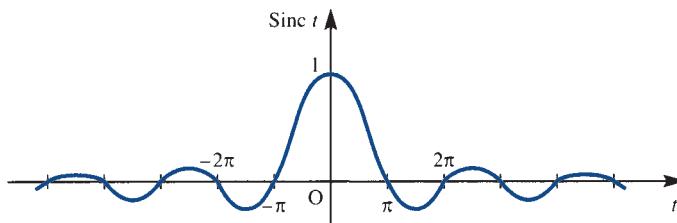


Figure 7.29
Graph of $\text{sinc } t$.



As we will see in Chapter 8, the sinc function $\text{sinc } t = (\sin t)/t$ plays an important role in signal analysis, and it is sometimes referred to as the **sampling function**. A graph of $\text{sinc } t$ is shown in Figure 7.29, and it is clear that the function oscillates over intervals of length 2π and decreases in amplitude with increasing t . Note also that the function has zeros at $t = \pm n\pi$ ($n = 1, 2, 3, \dots$).

7.4.4 Power spectrum

The **average power** P associated with a periodic signal $f(t)$, of period T , is defined as the mean square value; that is,

$$P = \frac{1}{T} \int_d^{d+T} [f(t)]^2 dt \quad (7.53)$$

For example, if $f(t)$ represents a voltage waveform applied to a resistor then P represents the average power, measured in watts, dissipated by a 1Ω resistor.

By Parseval's theorem (Theorem 7.2),

$$P = \frac{1}{4}a_0^2 + \frac{1}{2} \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \quad (7.54)$$

Since

$$\frac{1}{2}a_n^2 = \frac{1}{T} \int_d^{d+T} \left[a_n \cos\left(\frac{2n\pi t}{T}\right) \right]^2 dt, \quad \frac{1}{2}b_n^2 = \frac{1}{T} \int_d^{d+T} \left[b_n \sin\left(\frac{2n\pi t}{T}\right) \right]^2 dt$$

the power in the n th harmonic is

$$P_n = \frac{1}{2}(a_n^2 + b_n^2) \quad (7.55)$$

and it follows from (7.54) that the power of the periodic function $f(t)$ is the sum of the power of the individual harmonic components contained in $f(t)$.

In terms of the complex Fourier coefficients, Parseval's theorem gives

$$P = \sum_{n=-\infty}^{\infty} |c_n|^2 \quad (7.56)$$

As discussed in Section 7.4.3, the component $e^{jn\omega_0 t}$ at frequency $\omega_n = n\omega_0$, $\omega_0 = 2\pi/T$, must be considered alongside the component $e^{-jn\omega_0 t}$ at the corresponding negative frequency $-\omega_n$ in order to form the actual n th harmonic component of the function $f(t)$. Since $|c_{-n}|^2 = |c_n^*|^2 = |c_n|^2$, it follows that the power associated with the n th harmonic is the sum of the power associated with $e^{jn\omega_0 t}$ and $e^{-jn\omega_0 t}$; that is,

$$P_n = 2|c_n|^2 \quad (7.57)$$

which, since $|c_n| = \frac{1}{2}\sqrt{(a_n^2 + b_n^2)}$, corresponds to (7.55). Thus in the complex form half the power of the n th harmonic is associated with the positive frequency and half with the negative frequency.

Since the total power of a periodic signal is the sum of the power associated with each of the harmonics of which the signal is composed, it is again useful to consider a spectral representation, and a plot of $|c_n|^2$ against angular frequency ω_n is called the **power spectrum** of the function $f(t)$. Clearly such a spectrum is readily deduced from the discrete amplitude spectrum of $|c_n|$ against angular frequency ω_n .

Example 7.13

For the spectrum of the infinite train of rectangular pulses shown in Figure 7.27, determine the percentage of the total power contained within the frequency band up to the first zero value (called the **zero crossing** of the spectrum) at $10\pi \text{ rad s}^{-1}$.

Solution From (7.53), the total power associated with the infinite train of rectangular pulses $f(t)$ is

$$P = \frac{1}{2T} \int_{-T}^T [f(t)]^2 dt = \frac{1}{2T} \int_{-d}^d A^2 dt$$

which in the particular case when $d = \frac{1}{10}$ and $T = \frac{1}{2}$ becomes

$$P = \int_{-1/10}^{1/10} A^2 dt = \frac{1}{5} A^2$$

The power contained in the frequency band up to the first zero crossing at $10\pi \text{ rad s}^{-1}$ is

$$P_1 = c_0^2 + 2(c_1^2 + c_2^2 + c_3^2 + c_4^2)$$

where

$$c_n = \frac{1}{5} A \operatorname{sinc} \frac{1}{5} n \pi$$

That is,

$$\begin{aligned} P_1 &= \frac{1}{25} A^2 + \frac{2}{25} A^2 (\operatorname{sinc}^2 \frac{1}{5} \pi + \operatorname{sinc}^2 \frac{2}{5} \pi + \operatorname{sinc}^2 \frac{3}{5} \pi + \operatorname{sinc}^2 \frac{4}{5} \pi) \\ &= \frac{1}{25} A^2 [1 + 2(0.875 + 0.756 + 0.255 + 0.055)] = \frac{1}{5} A^2 (0.976) \end{aligned}$$

Thus $P_1 = 0.976P$, so that approximately 97.6% of the total power associated with $f(t)$ is contained in the frequency band up to the first zero crossing at $10\pi \text{ rad s}^{-1}$.

Suppose that a periodic voltage $v(t)$, of period T , applied to a linear circuit, results in a corresponding current $i(t)$, having the same period T . Then, given the Fourier series representation of both the voltage and current at a pair of terminals, we can use the multiplication theorem (Theorem 7.1) to obtain an expression for the average power P at the terminals. Thus, given

$$v(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi nt/T}, \quad i(t) = \sum_{n=-\infty}^{\infty} d_n e^{j2\pi nt/T}$$

the instantaneous power at the terminals is vi and the average power is

$$P = \frac{1}{T} \int_d^{d+T} vi \, dt = \sum_{n=-\infty}^{\infty} c_n d_n^*$$

or, in terms of the corresponding trigonometric Fourier series coefficients a_n , b_n and α_n , β_n ,

$$P = \frac{1}{4} \alpha_0 \beta_0 + \frac{1}{2} \sum_{n=1}^{\infty} (a_n \alpha_n + b_n \beta_n)$$

7.4.5 Exercises

- 17 Show that the complex form of the Fourier series expansion of the periodic function

$$f(t) = t^2 \quad (-\pi < t < \pi)$$

$$f(t + 2\pi) = f(t)$$

is

$$f(t) = \frac{\pi^2}{6} + \sum_{n=0}^{\infty} \frac{2}{n^2} (-1)^n e^{jnt}$$

Using (7.28), obtain the corresponding trigonometric series.

- 18 Obtain the complex form of the Fourier series expansion of the square wave

$$f(t) = \begin{cases} 0 & (-2 < t < 0) \\ 1 & (0 < t < 2) \end{cases}$$

$$f(t + 4) = f(t)$$

Using (7.28), obtain the corresponding trigonometric series.

- 19 Obtain the complex form of the Fourier series expansion of the following periodic functions.

$$(a) f(t) = \begin{cases} \pi & (-\pi < t < 0) \\ t & (0 < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

$$(b) f(t) = \begin{cases} a \sin \omega t & (0 < t < \frac{1}{2}T) \\ 0 & (\frac{1}{2}T < t < T) \end{cases}$$

$$f(t + T) = f(t), \quad T = 2\pi/\omega$$

$$(c) f(t) = \begin{cases} 2 & (-\pi < t < 0) \\ 1 & (0 < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

$$(d) f(t) = |\sin t| \quad (-\pi < t < \pi)$$

$$f(t + 2\pi) = f(t)$$

20 A periodic function $f(t)$, of period 2π , is defined within the period $-\pi < t < \pi$ by

$$f(t) = \begin{cases} 0 & (-\pi < t < 0) \\ 1 & (0 < t < \pi) \end{cases}$$

Using the Fourier coefficients of $f(t)$, together with Parseval's theorem, show that

$$\sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} = \frac{1}{8}\pi^2$$

(Note: The Fourier coefficients may be deduced from Exercise 18.)

- 21 (a) Show that the Fourier series expansion of the periodic function

$$f(t) = 500\pi t \quad (0 < t < \frac{1}{50})$$

$$f(t + \frac{1}{50}) = f(t)$$

may be expressed as

$$f(t) = 5\pi - 10 \sum_{n=1}^{\infty} \frac{1}{n} \sin 100n\pi t$$

- (b) Using (7.40), estimate the RMS value of $f(t)$ by

- (i) using the first four terms of the Fourier series;
- (ii) using the first eight terms of the Fourier series.

- (c) Obtain the true RMS value of $f(t)$, and hence determine the percentage errors in the estimated values obtained in (b).

- 22 A periodic voltage $v(t)$ (in V) of period 5 ms and specified by

$$v(t) = \begin{cases} 60 & (0 < t < 1.25 \text{ ms}) \\ 0 & (1.25 \text{ ms} < t < 5 \text{ ms}) \end{cases}$$

$$v(t + 5 \text{ ms}) = v(t)$$

is applied across the terminals of a 15Ω resistor.

- (a) Obtain expressions for the coefficients c_n of the complex Fourier series representation of $v(t)$, and write down the values of the first five non-zero terms.
 (b) Calculate the power associated with each of the first five non-zero terms of the Fourier expansion.
 (c) Calculate the total power delivered to the 15Ω resistor.
 (d) What is the percentage of the total power delivered to the resistor by the first five non-zero terms of the Fourier series?

7.5

Orthogonal functions

As was noted in the introduction, the fact that the set of functions $\{1, \cos \omega t, \sin \omega t, \dots, \cos n\omega t, \sin n\omega t, \dots\}$ is an orthogonal set of functions on the interval $d \leq t \leq d + T$ was crucial in the evaluation of the coefficients in the Fourier series expansion of a function $f(t)$. It is natural to ask whether it is possible to express the function $f(t)$ as a series expansion in other sets of functions. In the case of periodic functions $f(t)$ there is no natural alternative, but if we are concerned with representing a function $f(t)$ only in a finite interval $t_1 \leq t \leq t_2$ then a variety of other possibilities exist. These possibilities are drawn from a class of functions called **orthogonal functions**, of which the trigonometric set $\{1, \cos \omega t, \sin \omega t, \dots, \cos n\omega t, \sin n\omega t\}$ is a particular example.

7.5.1 Definitions

Two real functions $f(t)$ and $g(t)$ that are piecewise-continuous in the interval $t_1 \leq t \leq t_2$ are said to be **orthogonal** in this interval if

$$\int_{t_1}^{t_2} f(t)g(t) dt = 0$$

A set of real functions $\phi_1(t), \phi_2(t), \dots \equiv \{\phi_n(t)\}$, each of which is piecewise-continuous on $t_1 \leq t \leq t_2$, is said to be an **orthogonal set** on this interval if $\phi_n(t)$ and $\phi_m(t)$ are orthogonal for each pair of distinct indices n, m ; that is, if

$$\int_{t_1}^{t_2} \phi_n(t)\phi_m(t) dt = 0 \quad (n \neq m) \tag{7.58}$$

We shall also assume that no member of the set $\{\phi_n(t)\}$ is identically zero except at a finite number of points, so that

$$\int_{t_1}^{t_2} \phi_m^2(t) dt = \gamma_m \quad (m = 1, 2, 3, \dots) \tag{7.59}$$

where γ_m ($m = 1, 2, \dots$) are all non-zero constants.

An orthogonal set $\{\phi_n(t)\}$ is said to be **orthonormal** if each of its components is also normalized; that is, $\gamma_m = 1$ ($m = 1, 2, 3, \dots$). We note that any orthogonal set $\{\phi_n(t)\}$ can be converted into an orthonormal set by dividing each member $\phi_m(t)$ of the set by $\sqrt{\gamma_m}$.

Example 7.14

We know already that

$$\{1, \cos t, \sin t, \cos 2t, \sin 2t, \dots, \cos nt, \sin nt\}$$

is an orthogonal set on the interval $d \leq t \leq d + 2\pi$, and so the set

$$\left\{ \frac{1}{\sqrt{(2\pi)}}, \frac{\cos t}{\sqrt{\pi}}, \frac{\sin t}{\sqrt{\pi}}, \dots, \frac{\cos nt}{\sqrt{\pi}}, \frac{\sin nt}{\sqrt{\pi}} \right\}$$

forms an orthonormal set on the same interval. The latter follows since

$$\begin{aligned} \int_d^{d+2\pi} \left[\frac{1}{\sqrt{(2\pi)}} \right]^2 dt &= 1 \\ \int_d^{d+2\pi} \left(\frac{\cos nt}{\sqrt{\pi}} \right)^2 dt &= \int_d^{d+2\pi} \left(\frac{\sin nt}{\sqrt{\pi}} \right)^2 dt = 1 \quad (n = 1, 2, 3, \dots) \end{aligned}$$

The definition of orthogonality considered so far applies to real functions, and has to be amended if members of the set $\{\phi_n(t)\}$ are complex functions of the real variable t . In such a case the set $\{\phi_n(t)\}$ is said to be an orthogonal set on the interval $t_1 \leq t \leq t_2$ if

$$\int_{t_1}^{t_2} \phi_n(t) \phi_m^*(t) dt = \begin{cases} 0 & (n \neq m) \\ \gamma & (n = m) \end{cases} \quad (7.60)$$

where $\phi_m^*(t)$ denotes the complex conjugate of $\phi_m(t)$.

Example 7.15

Verify that the set of complex exponential functions

$$\{e^{jnt\pi/T}\} \quad (n = 0, \pm 1, \pm 2, \pm 3, \dots)$$

used in the complex representation of the Fourier series is an orthogonal set on the interval $0 \leq t \leq 2T$.

Solution

First,

$$\int_0^{2T} e^{jnt\pi/T} 1 dt = \left[\frac{T}{jn\pi} e^{jnt\pi/T} \right]_0^{2T} = 0 \quad (n \neq 0)$$

since $e^{j2n\pi} = e^0 = 1$. Secondly,

$$\int_0^{2T} e^{jnt\pi/T} (e^{jm\pi t/T})^* dt = \int_0^{2T} e^{j(n-m)\pi t/T} dt = \left[\frac{T}{j(n-m)\pi} e^{j(n-m)\pi t/T} \right]_0^{2T} = 0 \quad (n \neq m)$$

and, when $n = m$,

$$\int_0^{2T} e^{j\pi nt/T} (e^{j\pi nt/T})^* dt = \int_0^{2T} 1 dt = 2T$$

Thus

$$\begin{aligned} \int_0^{2T} e^{j\pi nt/T} 1 dt &= 0 \quad (n \neq 0) \\ \int_0^{2T} e^{j\pi nt/T} (e^{j\pi nt/T})^* dt &= \begin{cases} 0 & (n \neq m) \\ 2T & (n = m) \end{cases} \end{aligned}$$

and, from (7.60), the set is an orthogonal set on the interval $0 \leq t \leq 2T$.

The trigonometric and exponential sets are examples of orthogonal sets that we have already used in developing the work on Fourier series. Examples of other sets of orthogonal functions that are widely used in practice are Legendre polynomials, Bessel functions, Hermite polynomials, Laguerre polynomials, Jacobi polynomials, Chebyshev polynomials, Zernike polynomials and Walsh functions. Over recent years wavelets are another set of orthogonal functions that have been widely used, particularly in applications such as signal processing and data compression.

7.5.2 Generalized Fourier series

Let $\{\phi_n(t)\}$ be an orthogonal set on the interval $t_1 \leq t \leq t_2$ and suppose that we wish to represent the piecewise-continuous function $f(t)$ in terms of this set within this interval. Following the Fourier series development, suppose that it is possible to express $f(t)$ as a series expansion of the form

$$f(t) = \sum_{n=1}^{\infty} c_n \phi_n(t) \tag{7.61}$$

We now wish to determine the coefficients c_n , and to do so we again follow the Fourier series development. Multiplying (7.61) throughout by $\phi_m(t)$ and integrating term by term, we obtain

$$\int_{t_1}^{t_2} f(t) \phi_m(t) dt = \sum_{n=1}^{\infty} c_n \int_{t_1}^{t_2} \phi_m(t) \phi_n(t) dt$$

which, on using (7.58) and (7.59), reduces to

$$\int_{t_1}^{t_2} f(t) \phi_n(t) dt = c_n \gamma_n$$

giving

$$c_n = \frac{1}{\gamma_n} \int_{t_1}^{t_2} f(t) \phi_n(t) dt \quad (n = 1, 2, 3, \dots) \tag{7.62}$$

Summary

Summarizing, if $f(t)$ is a piecewise-continuous function on the interval $t_1 \leq t \leq t_2$ and $\{\phi_n(t)\}$ is an orthogonal set on this interval then the series

$$f(t) = \sum_{n=1}^{\infty} c_n \phi_n(t)$$

is called the **generalized Fourier series** of $f(t)$ with respect to the basis set $\{\phi_n(t)\}$, and the coefficients c_n , given by (7.62), are called the **generalized Fourier coefficients** with respect to the same basis set.

A parallel can be drawn between a generalized Fourier series expansion of a function $f(t)$ with respect to an orthogonal basis set of functions $\{\phi_n(t)\}$ and the representation of a vector \mathbf{f} in terms of an orthogonal basis set of vectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ as

$$\mathbf{f} = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n$$

where

$$\alpha_i = \frac{\mathbf{f} \cdot \mathbf{v}_i}{\mathbf{v}_i \cdot \mathbf{v}_i} = \frac{\mathbf{f} \cdot \mathbf{v}_i}{|\mathbf{v}_i|^2}$$

There is clearly a similarity between this pair of results and the pair (7.61)–(7.62).

7.5.3 Convergence of generalized Fourier series

As in the case of a Fourier series expansion, partial sums of the form

$$F_N(t) = \sum_{n=1}^N c_n \phi_n(t) \quad (7.63)$$

can be considered, and we wish this representation to be, in some sense, a ‘close approximation’ to the parent function $f(t)$. The question arises when considering such a partial sum as to whether choosing the coefficients c_n as the generalized Fourier coefficients (7.62) leads to the ‘best’ approximation. Defining the **mean square error** E_N between the actual value of $f(t)$ and the approximation $F_N(t)$ as

$$E_N = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} [f(t) - F_N(t)]^2 dt$$

it can be shown that E_N is minimized, for all N , when the coefficients c_n are chosen according to (7.62). Thus in this sense the finite generalized Fourier series gives the best approximation. To verify this result, assume, for convenience, that the set $\{\phi_n(t)\}$ is orthonormal, and consider the N th partial sum

$$F_N(t) = \sum_{n=1}^N \tilde{c}_n \phi_n(t)$$

where the \tilde{c}_n are to be chosen in order to minimize the mean square error E_N . Now

$$\begin{aligned}(t_2 - t_1)E_N &= \int_{t_1}^{t_2} \left[f(t) - \sum_{n=1}^N \tilde{c}_n \phi_n(t) \right]^2 dt \\ &= \int_{t_1}^{t_2} f^2(t) dt - 2 \sum_{n=1}^N \tilde{c}_n \int_{t_1}^{t_2} f(t) \phi_n(t) dt + \sum_{n=1}^N \tilde{c}_n^2 \int_{t_1}^{t_2} \phi_n^2(t) dt \\ &= \int_{t_1}^{t_2} f^2(t) dt - 2 \sum_{n=1}^N \tilde{c}_n c_n + \sum_{n=1}^N \tilde{c}_n^2\end{aligned}$$

as $\{\phi_n(t)\}$ is an orthonormal set. That is,

$$(t_2 - t_1)E_N = \int_{t_1}^{t_2} f^2(t) dt - \sum_{n=1}^N c_n^2 + \sum_{n=1}^N (\tilde{c}_n - c_n)^2 \quad (7.64)$$

which is clearly minimized when $\tilde{c}_n = c_n$.

Taking $\tilde{c}_n = c_n$ in (7.64), the mean square error E_N in approximating $f(t)$ by $F_N(t)$ of (7.59) is given by

$$E_N = \frac{1}{t_2 - t_1} \left[\int_{t_1}^{t_2} f^2(t) dt - \sum_{n=1}^N c_n^2 \right]$$

if the set $\{\phi_n(t)\}$ is orthonormal, and is given by

$$E_N = \frac{1}{t_2 - t_1} \left[\int_{t_1}^{t_2} f^2(t) dt - \sum_{n=1}^N \gamma_n c_n^2 \right] \quad (7.65)$$

if the set $\{\phi_n(t)\}$ is orthogonal. Since E_N is non-negative, it follows from (7.65) that

$$\int_{t_1}^{t_2} f^2(t) dt \geq \sum_{n=1}^N \gamma_n c_n^2 \quad (7.66)$$

a result known as **Bessel's inequality**. The question that arises in practice is whether or not $E_N \rightarrow 0$ as $N \rightarrow \infty$, indicating that the sum

$$\sum_{n=1}^N c_n \phi_n(t)$$

converges to the function $f(t)$. If this were the case then, from (7.63),

$$\int_{t_1}^{t_2} f^2(t) dt = \sum_{n=1}^{\infty} \gamma_n c_n^2 \quad (7.67)$$

which is the **generalized form of Parseval's theorem**, and the set $\{\phi_n(t)\}$ is said to be complete. Strictly speaking, the fact that Parseval's theorem holds ensures that the partial sum $F_N(t)$ converges in the mean to the parent function $f(t)$ as $N \rightarrow \infty$, and this does not necessarily guarantee convergence at any particular point. In engineering applications, however, this distinction may be overlooked, since for the functions

met in practice convergence in the mean also ensures pointwise convergence at points where $f(t)$ is convergent, and convergence to the mean of the discontinuity at points where $f(t)$ is discontinuous.

Example 7.16

The set $\{1, \cos t, \sin t, \dots, \cos nt, \sin nt\}$ is a complete orthogonal set in the interval $d \leq t \leq d + 2\pi$. Following the same argument as above, it is readily shown that for a function $f(t)$ that is piecewise-continuous on $d \leq t \leq d + 2\pi$ the mean square error between $f(t)$ and the finite Fourier series

$$F_N(t) = \frac{1}{2}\tilde{a}_0 + \sum_{n=1}^N \tilde{a}_n \cos nt + \sum_{n=1}^N \tilde{b}_n \sin nt$$

is minimized when \tilde{a}_0, \tilde{a}_n and \tilde{b}_n ($n = 1, 2, 3, \dots$) are equal to the corresponding Fourier coefficients a_0, a_n and b_n ($n = 1, 2, 3, \dots$) determined using (7.4) and (7.5). In this case the mean square error E_N is given by

$$E_N = \frac{1}{2\pi} \left[\int_d^{d+2\pi} f^2(t) dt - \pi \left[\frac{1}{2} a_0^2 + \sum_{n=1}^N (a_n^2 + b_n^2) \right] \right]$$

Bessel's inequality (7.66) becomes

$$\int_d^{d+2\pi} f^2(t) dt \geq \pi \left[\frac{1}{2} a_0^2 + \sum_{n=1}^N (a_n^2 + b_n^2) \right]$$

and Parseval's theorem (7.67) reduces to

$$\frac{1}{2\pi} \int_d^{d+2\pi} f^2(t) dt = \frac{1}{4} a_0^2 + \frac{1}{2} \sum_{n=1}^{\infty} (a_n^2 + b_n^2)$$

which conforms with (7.48). Since, in this case, the basis set is complete, Parseval's theorem holds, and the Fourier series converges to $f(t)$ in the sense discussed above.

7.5.4 Exercises

- 23 The Fourier series expansion for the periodic square wave

$$f(t) = \begin{cases} -1 & (-\pi < t < 0) \\ 1 & (0 < t < \pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

is

$$f(t) = \sum_{n=1}^{\infty} \frac{4}{\pi(2n-1)} \sin(2n-1)t$$

Determine the mean square error corresponding to approximations to $f(t)$ based on the use of one term, two terms and three terms respectively in the series expansion.

24

The Legendre polynomials $P_n(t)$ are generated by the formula

$$P_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} (t^2 - 1)^n \quad (n = 0, 1, 2, \dots)$$

and satisfy the recurrence relationship

$$nP_n(t) = (2n-1)tP_{n-1}(t) - (n-1)P_{n-2}(t)$$

- (a) Deduce that

$$\begin{aligned} P_0(t) &= 1, \quad P_1(t) = t \\ P_2(t) &= \frac{1}{2}(3t^2 - 1), \quad P_3(t) = \frac{1}{2}(5t^3 - 3t) \end{aligned}$$

- (b) Show that the polynomials form an orthogonal set on the interval $(-1, 1)$ and, in particular, that

$$\int_{-1}^1 P_m(t)P_n(t) dt = \begin{cases} 0 & (n \neq m) \\ 2/(2n+1) & (n = m; m = 0, 1, 2, \dots) \end{cases}$$

- (c) Given that the function

$$f(t) = \begin{cases} -1 & (-1 < t < 0) \\ 0 & (t = 0) \\ 1 & (0 < t < 1) \end{cases}$$

is expressed as a Fourier–Legendre series expansion

$$f(t) = \sum_{r=0}^{\infty} c_r P_r(t)$$

determine the values of c_0 , c_1 , c_2 and c_3 .

- (d) Plot graphs to illustrate convergence of the series obtained in (c), and compare the mean square error with that of the corresponding Fourier series expansion.

25 Repeat parts (c) and (d) of Exercise 24 for the function

$$f(x) = \begin{cases} 0 & (-1 < x < 0) \\ x & (0 < x < 0) \end{cases}$$

26 Laguerre polynomials $L_n(t)$ are generated by the formula

$$L_n(t) = e^t \frac{d^n}{dt^n} (t^n e^{-t}) \quad (n = 0, 1, 2, \dots)$$

and satisfy the recurrence relation

$$L_n(t) = (2n - 1 - t)L_{n-1}(t) - (n - 1)^2 L_{n-2}(t) \quad (n = 2, 3, \dots)$$

These polynomials are orthogonal on the interval $0 \leq t < \infty$ with respect to the weighting function e^{-t} , so that

$$\int_0^\infty e^{-t} L_n(t) L_m(t) dt = \begin{cases} 0 & (n \neq m) \\ (n!)^2 & (n = m) \end{cases}$$

- (a) Deduce that

$$\begin{aligned} L_0(t) &= 1, \quad L_1(t) = 1 - t \\ L_2(t) &= 2 - 4t + t^2 \\ L_3(t) &= 6 - 18t + 9t^2 - t^3 \end{aligned}$$

- (b) Confirm the above orthogonality result in the case of L_0 , L_1 , L_2 and L_3 .

- (c) Given that the function $f(t)$ is to be approximated over the interval $0 \leq t < \infty$ by

$$f(t) = \sum_{r=0}^{\infty} c_r L_r(t)$$

show that

$$c_r = \frac{1}{(r!)^2} \int_0^\infty f(t) e^{-t} L_r(t) dt \quad (r = 0, 1, 2, \dots)$$

(Note: Laguerre polynomials are of particular importance to engineers, since they can be generated as the impulse responses of relatively simple networks.)

27

Hermite polynomials $H_n(t)$ are generated by the formula

$$H_n(t) = (-1)^n e^{t^2/2} \frac{d^n}{dt^n} e^{-t^2/2}$$

$$(n = 0, 1, 2, \dots)$$

and satisfy the recurrence relationship

$$H_n(t) = t H_{n-1}(t) - (n - 1) H_{n-2}(t)$$

$$(n = 2, 3, \dots)$$

The polynomials are orthogonal on the interval $-\infty < t < \infty$ with respect to the weighting function $e^{-t^2/2}$, so that

$$\int_{-\infty}^{\infty} e^{-t^2/2} H_n(t) H_m(t) dt = \begin{cases} 0 & (n \neq m) \\ \sqrt{(2\pi)n!} & (n = m) \end{cases}$$

- (a) Deduce that

$$\begin{aligned} H_0(t) &= 1, \quad H_1(t) = t \\ H_2(t) &= t^2 - 1, \quad H_3(t) = t^3 - 3t \\ H_4(t) &= t^4 - 6t^2 + 3 \end{aligned}$$

- (b) Confirm the above orthogonality result for H_0 , H_1 , H_2 and H_3 .

- (c) Given that the function $f(t)$ is to be approximated over the interval $-\infty < t < \infty$ by

$$f(t) = \sum_{r=0}^{\infty} c_r H_r(t)$$

show that

$$c_r = \frac{1}{r! \sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} f(t) H_r(t) dt \quad (r = 0, 1, \dots)$$

- 28 Chebyshev polynomials $T_n(t)$ are generated by the formula

$$T_n(t) = \cos(n \cos^{-1} t) \quad (n = 0, 1, 2, \dots)$$

or

$$T_n(t) = \sum_{r=0}^{[n/2]} (-1)^r \frac{n!}{(2r)!(n-2r)!} (1-t^2)^r t^{n-2r} \quad (n = 0, 1, 2, \dots)$$

where

$$[n/2] = \begin{cases} n/2 & (\text{even } n) \\ (n-1)/2 & (\text{odd } n) \end{cases}$$

They also satisfy the recurrence relationship

$$T_n(t) = 2tT_{n-1}(t) - T_{n-2}(t) \quad (n = 2, 3, \dots)$$

and are orthogonal on the interval $-1 \leq t \leq 1$ with respect to the weighting function $1/\sqrt{1-t^2}$, so that

$$\int_{-1}^1 \frac{T_n(t)T_m(t)}{\sqrt{1-t^2}} dt = \begin{cases} 0 & (m \neq n) \\ \frac{1}{2}\pi & (m = n \neq 0) \\ \pi & (m = n = 0) \end{cases}$$

- (a) Deduce that

$$T_0(t) = 1, \quad T_1(t) = t$$

$$T_2(t) = 2t^2 - 1, \quad T_3(t) = 4t^3 - 3t$$

$$T_4(t) = 8t^4 - 8t^2 + 1$$

$$T_5(t) = 16t^5 - 20t^3 + 5t$$

- (b) Confirm the above orthogonality result for T_0, T_1, T_2 and T_3 .

- (c) Given that the function $f(t)$ is to be approximated over the interval $-1 \leq t \leq 1$ by

$$f(t) = \sum_{r=0}^{\infty} c_r T_r(t)$$

show that

$$c_0 = \frac{1}{\pi} \int_{-1}^1 \frac{f(t)T_0(t)}{\sqrt{1-t^2}} dt$$

$$c_r = \frac{2}{\pi} \int_{-1}^1 \frac{f(t)T_r(t)}{\sqrt{1-t^2}} dt \quad (r = 1, 2, \dots)$$

With developments in digital techniques, Walsh functions $W_n(t)$ have become of considerable importance in practice, since they are so easily generated by digital logic circuitry. The first four Walsh functions may be defined on the interval $0 \leq t \leq T$ by

$$W_0(t) = \frac{1}{\sqrt{T}} \quad (0 \leq t \leq T)$$

$$W_1(t) = \begin{cases} 1/\sqrt{T} & (0 \leq t < \frac{1}{2}T) \\ -1/\sqrt{T} & (\frac{1}{2}T < t \leq T) \end{cases}$$

$$W_2(t) = \begin{cases} 1/\sqrt{T} & (0 \leq t < \frac{1}{4}T, \frac{3}{4}T < t \leq T) \\ -1/\sqrt{T} & (\frac{1}{4}T < t < \frac{3}{4}T) \end{cases}$$

$$W_3(t) =$$

$$\begin{cases} 1/\sqrt{T} & (0 \leq t < \frac{1}{8}T, \frac{3}{8}T < t < \frac{5}{8}T, \frac{7}{8}T < t \leq T) \\ -1/\sqrt{T} & (\frac{1}{8}T < t < \frac{3}{8}T, \frac{5}{8}T < t < \frac{7}{8}T) \end{cases}$$

- (a) Plot graphs of the functions $W_0(t), W_1(t), W_2(t)$ and $W_3(t)$, and show that they are orthonormal on the interval $0 \leq t \leq T$. Write down an expression for $W_n(t)$.

- (b) The Walsh functions may be used to obtain a Fourier–Walsh series expansion for a function $f(t)$, over the interval $0 \leq t \leq T$, in the form

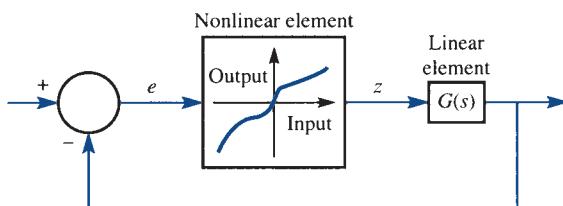
$$f(t) = \sum_{r=0}^{\infty} c_r W_r(t)$$

Illustrate this for the square wave of Exercise 23. What is the corresponding mean square error? Comment on your answer.

7.6 Engineering application: describing functions

Many control systems containing a nonlinear element may be represented by the block diagram of Figure 7.30. In practice, describing function techniques are used to analyse and design such control systems. Essentially the method involves replacing the nonlinearity by an equivalent gain N and then using the techniques developed for linear systems, such as the frequency response methods of Section 5.5. If the nonlinear element is subjected to a sinusoidal input $e(t) = X \sin \omega t$ then its output $z(t)$ may be represented by the Fourier series expansion

Figure 7.30 Nonlinear control system.



$$\begin{aligned} z(t) &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t \\ &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} A_n \sin(n\omega t + \phi_n) \end{aligned}$$

with $A_n = \sqrt{(a_n^2 + b_n^2)}$ and $\phi_n = \tan^{-1}(a_n/b_n)$.

The **describing function** $N(X)$ of the nonlinear element is then defined to be the complex ratio of the fundamental component of the output to the input; that is,

$$N(X) = \frac{A_1}{X} e^{j\phi_1}$$

with $N(X)$ being independent of the input frequency ω if the nonlinear element is memory-free.

Having determined the describing function, the behaviour of the closed-loop system is then determined by the characteristic equation

$$1 + N(X)G(j\omega) = 0$$

If a combination of X and ω can be found to satisfy this equation then the system is capable of sustained oscillations at that frequency and magnitude; that is, the system exhibits **limit-cycle behaviour**. In general, more than one combination can be found, and the resulting oscillations can be a stable or unstable limit cycle.

Normally the characteristic equation is investigated graphically by plotting $G(j\omega)$ and $-1/N(X)$, for all values of X , on the same polar diagram. Limit cycles then occur at frequencies and amplitudes corresponding to points of intersection of the curves. Sometimes plotting can be avoided by calculating the maximum value of $N(X)$ and hence the value of the gain associated with $G(s)$ that will just cause limit cycling to occur.

Using this background information, the following investigation is left as an exercise for the reader to develop.

Figure 7.31 (a) Relay; (b) relay with dead zone.

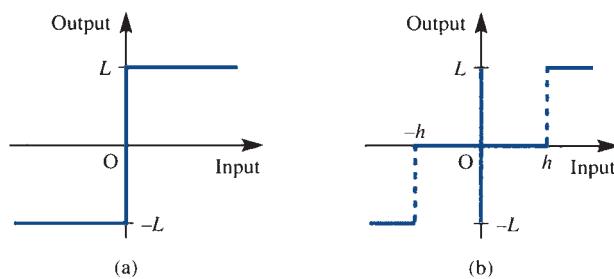
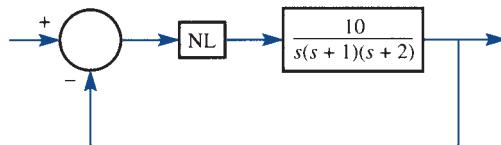


Figure 7.32 Nonlinear system of exercise.



- (a) Show that the describing functions $N_1(X)$ and $N_2(X)$ corresponding respectively to the relay (on-off nonlinearity) of Figure 7.31(a) and the relay with dead zone of Figure 7.31(b) are

$$N_1(X) = \frac{4L}{\pi X}, \quad N_2(X) = \frac{4L}{\pi X} \sqrt{\left[1 - \left(\frac{h}{X}\right)^2\right]}$$

- (b) For the system of Figure 7.32 show that a limit cycle exists when the nonlinearity is the relay of Figure 7.31(a) with $L = 1$. Determine the amplitude and frequency of this limit cycle.

In an attempt to eliminate the limit-cycle oscillation, the relay is replaced by the relay with dead zone illustrated in Figure 7.31(b), again with $L = 1$. Show that this allows our objective to be achieved provided that $h > 10/3\pi$.

7.7 Review exercises (1–20)

- 1 A periodic function $f(t)$ is defined by

$$f(t) = \begin{cases} t^2 & (0 \leq t < \pi) \\ 0 & (\pi < t \leq 2\pi) \end{cases}$$

$$f(t + 2\pi) = f(t)$$

Obtain a Fourier series expansion of $f(t)$ and deduce that

$$\frac{1}{6}\pi^2 = \sum_{r=1}^{\infty} \frac{1}{r^2}$$

- 2 Determine the full-range Fourier series expansion of the even function $f(t)$ of period 2π defined by

$$f(t) = \begin{cases} \frac{2}{3}t & (0 \leq t \leq \frac{1}{3}\pi) \\ \frac{1}{3}(\pi - t) & (\frac{1}{3}\pi \leq t \leq \pi) \end{cases}$$

To what value does the series converge at $t = \frac{1}{3}\pi$?

- 3 A function $f(t)$ is defined for $0 \leq t \leq \frac{1}{2}T$ by

$$f(t) = \begin{cases} t & (0 \leq t \leq \frac{1}{4}T) \\ \frac{1}{2}T - t & (\frac{1}{4}T \leq t \leq \frac{1}{2}T) \end{cases}$$

Sketch odd and even functions that have a period T and are equal to $f(t)$ for $0 \leq t \leq \frac{1}{2}T$.

- Find the half-range Fourier sine series of $f(t)$.
- To what value will the series converge for $t = -\frac{1}{4}T$?
- What is the sum of the following series?

$$S = \sum_{r=1}^{\infty} \frac{1}{(2r-1)^2}$$

- 4** Prove that if $g(x)$ is an odd function and $f(x)$ an even function of x , the product $g(x)[c + f(x)]$ is an odd function if c is a constant.

A periodic function with period 2π is defined by

$$F(\theta) = \frac{1}{12} \theta(\pi^2 - \theta^2)$$

in the interval $-\pi \leq \theta \leq \pi$. Show that the Fourier series representation of the function is

$$F(\theta) = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^3} \sin n\theta$$

- 5** A repeating waveform of period 2π is described by

$$f(t) = \begin{cases} \pi + t & (-\pi \leq t \leq -\frac{1}{2}\pi) \\ -t & (-\frac{1}{2}\pi \leq t \leq \frac{1}{2}\pi) \\ t - \pi & (\frac{1}{2}\pi \leq t \leq \pi) \end{cases}$$

Sketch the waveform over the range $t = -2\pi$ to $t = 2\pi$ and find the Fourier series representation of $f(t)$, making use of any properties of the waveform that you can identify before any integration is performed.

- 6** A function $f(x)$ is defined in the interval $-1 \leq x \leq 1$ by

$$f(x) = \begin{cases} 1/2\varepsilon & (-\varepsilon < x < \varepsilon) \\ 0 & (-1 \leq x < -\varepsilon; \varepsilon < x \leq 1) \end{cases}$$

Sketch a graph of $f(x)$ and show that a Fourier series expansion of $f(x)$ valid in the interval $-1 \leq x \leq 1$ is given by

$$f(x) = \frac{1}{2} + \sum_{n=1}^{\infty} \frac{\sin n\pi\varepsilon}{n\pi\varepsilon} \cos n\pi x$$

- 7** Show that the half-range Fourier sine series for the function

$$f(t) = \left(1 - \frac{t}{\pi}\right)^2 \quad (0 \leq t \leq \pi)$$

is

$$f(t) = \sum_{n=1}^{\infty} \frac{2}{n\pi} \left\{ 1 - \frac{2}{n^2\pi^2} [1 - (-1)^n] \right\} \sin nt$$

8 Find a half-range Fourier sine and Fourier cosine series for $f(x)$ valid in the interval $0 < x < \pi$ when $f(x)$ is defined by

$$f(x) = \begin{cases} x & (0 \leq x \leq \frac{1}{2}\pi) \\ \pi - x & (\frac{1}{2}\pi \leq x \leq \pi) \end{cases}$$

Sketch the graph of the Fourier series obtained for $-2\pi < x \leq 2\pi$.

9

A function $f(x)$ is periodic of period 2π and is defined by $f(x) = e^x$ ($-\pi < x < \pi$). Sketch the graph of $f(x)$ from $x = -2\pi$ to $x = 2\pi$ and prove that

$$f(x) = \frac{2 \sinh \pi}{\pi} \left[\frac{1}{2} + \sum_{n=1}^{\infty} \frac{(-1)^n}{1+n^2} (\cos nx - n \sin nx) \right]$$

10

A function $f(t)$ is defined on $0 < t < \pi$ by

$$f(t) = \pi - t$$

Find

- a half-range Fourier sine series, and
- a half-range Fourier cosine series for $f(t)$ valid for $0 < t < \pi$.

Sketch the graphs of the functions represented by each series for $-2\pi < t < 2\pi$.

11

Show that the Fourier series

$$\frac{1}{2}\pi - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos(2n-1)t}{(2n-1)^2}$$

represents the function $f(t)$, of period 2π , given by

$$f(t) = \begin{cases} t & (0 \leq t \leq \pi) \\ -t & (-\pi \leq t \leq 0) \end{cases}$$

Deduce that, apart from a transient component (that is, a complementary function that dies away as $t \rightarrow \infty$), the differential equation

$$\frac{dx}{dt} + x = f(t)$$

has the solution

$$x = \frac{1}{2}\pi - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\cos(2n-1)t + (2n-1)\sin(2n-1)t}{(2n-1)^2[1+(2n-1)^2]}$$

- 12 Show that if $f(t)$ is a periodic function of period 2π and

$$f(t) = \begin{cases} t/\pi & (0 < t < \pi) \\ (2\pi-t)/\pi & (\pi < t < 2\pi) \end{cases}$$

then

$$f(t) = \frac{1}{2} - \frac{4}{\pi^2} \sum_{n=0}^{\infty} \frac{\cos(2n+1)t}{(2n+1)^2}$$

Show also that, when ω is not an integer,

$$\begin{aligned} y &= \frac{1}{2\omega^2}(1 - \cos \omega t) \\ &\quad - \frac{4}{\pi^2} \sum_{n=0}^{\infty} \frac{\cos(2n+1)t - \cos \omega t}{(2n+1)^2[\omega^2 - (2n+1)^2]} \end{aligned}$$

satisfies the differential equation

$$\frac{d^2y}{dt^2} + \omega^2 y = f(t)$$

subject to the initial conditions $y = dy/dt = 0$ at $t = 0$.

- 13 (a) A periodic function $f(t)$, of period 2π , is defined in $-\pi \leq t \leq \pi$ by

$$f(t) = \begin{cases} -t & (-\pi < t < 0) \\ t & (0 < t < \pi) \end{cases}$$

Obtain a Fourier series expansion for $f(t)$, and from it, using Parseval's theorem, deduce that

$$\frac{1}{96}\pi^4 = \sum_{n=0}^{\infty} \frac{1}{(2n-1)^4}$$

- (b) By formally differentiating the series obtained in (a), obtain the Fourier series expansion of the periodic square wave

$$g(t) = \begin{cases} -1 & (-\pi < t < 0) \\ 0 & (t = 0) \\ 1 & (0 < t < \pi) \end{cases}$$

$$g(t+2\pi) = g(t)$$

Check the validity of your result by determining directly the Fourier series expansion of $g(t)$.

14

A periodic function $f(t)$, of period 2π , is defined in the range $-\pi < t < \pi$ by

$$f(t) = \sin \frac{1}{2}t$$

Show that the complex form of the Fourier series expansion for $f(t)$ is

$$f(t) = \sum_{n=-\infty}^{\infty} \frac{j4n(-1)^n}{\pi(4n^2-1)} e^{int}$$

15

- (a) Find the Fourier series expansion of the voltage $v(t)$ represented by the half-wave rectified sine wave

$$v(t) = \begin{cases} 10 \sin(2\pi t/T) & (0 < t < \frac{1}{2}T) \\ 0 & (\frac{1}{2}T < t < T) \end{cases}$$

$$v(t+T) = v(t)$$

- (b) If the voltage $v(t)$ in (a) is applied to a 10Ω resistor, what is the total average power delivered to the resistor? What percentage of the total power is carried by the second-harmonic component of the voltage?

The periodic waveform $f(t)$ shown in Figure 7.33 may be written as

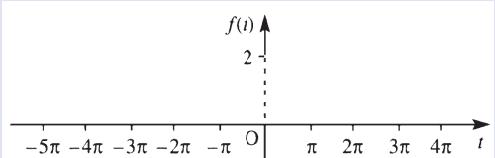


Figure 7.33 Waveform $f(t)$ of Review exercise 16.

$$f(t) = 1 + g(t)$$

where $g(t)$ represents an odd function.

- (a) Sketch the graph of $g(t)$.
(b) Obtain the Fourier series expansion for $g(t)$, and hence write down the Fourier series expansion for $f(t)$.

17

Show that the complex Fourier series expansion for the periodic function

$$f(t) = t \quad (0 < t < 2\pi)$$

$$f(t+2\pi) = f(t)$$

is

$$f(t) = \pi + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{j e^{jnt}}{n}$$

- 18 (a) A square-wave voltage $v(t)$ of period T is defined by

$$v(t) = \begin{cases} -1 & (-\frac{1}{2}T < t < 0) \\ 1 & (0 < t < \frac{1}{2}T) \end{cases}$$

$$v(t+T) = v(t)$$

Show that its Fourier series expansion is given by

$$v(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin[(4n-2)\pi t/T]}{2n-1}$$

- (b) Find the steady-state response of the circuit shown in Figure 7.34 to the sinusoidal input voltage

$$v_{\omega}(t) = \sin \omega t$$

and hence write down the Fourier series expansion of the circuit's steady-state response to the square-wave voltage $v(t)$ in (a).

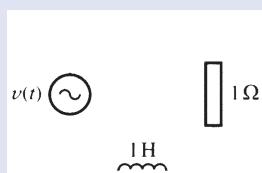


Figure 7.34 Circuit of Review exercise 18.

- 19 (a) Defining the n th Chebyshev polynomial by

$$T_n(t) = \cos(n \cos^{-1} t)$$

use Euler's formula $\cos \theta = \frac{1}{2}(e^{j\theta} + e^{-j\theta})$ to obtain the expansions of t^{2k} and t^{2k+1} in Chebyshev polynomials, where k is a positive integer.

- (b) Establish the recurrence relation

$$T_n(t) = 2tT_{n-1}(t) - T_{n-2}(t)$$

- (c) Write down the values of $T_0(t)$ and $T_1(t)$ from the definition, and then use (b) to find $T_2(t)$ and $T_3(t)$.

- (d) Express $t^5 - 5t^4 + 7t^3 + 6t - 8$ in Chebyshev polynomials.

- (e) Find the cubic polynomial that approximates to

$$t^5 - 5t^4 + 7t^3 + 6t - 8$$

over the interval $(-1, 1)$ with the smallest maximum error. Give an upper bound for this error. Is there a value of t for which this upper bound is attained?

20

The relationship between the input and output of a relay with a dead zone Δ and no hysteresis is shown in Figure 7.35. Show that the describing function is

$$N(x_i) = \frac{4M}{\pi x_i} \left[1 - \left(\frac{\Delta}{2x_i} \right)^2 \right]^{1/2}$$

for an input amplitude x_i .

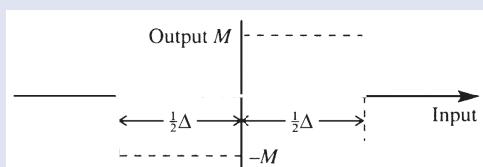


Figure 7.35 Relay with dead zone of Review exercise 20.

If this relay is used in the forward path of the on-off positional control system shown in Figure 7.36, where the transfer function

$$\frac{K}{s(T_1 s + 1)(T_2 s + 1)}$$

characterizes the time constant of the servo-motor, and the inertia and viscous damping of the load, show that a limit-cycle oscillation will not occur provided that the dead zone in the relay is such that

$$\Delta > \frac{4MK}{\pi} \frac{T_1 T_2}{T_1 + T_2}$$

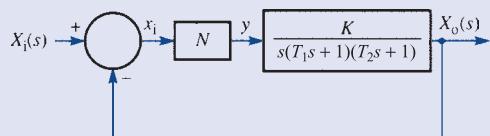


Figure 7.36 Positional control system of Review exercise 20.



8 The Fourier Transform

Chapter 8 Contents

8.1	Introduction	538
8.2	The Fourier transform	539
8.3	Properties of the Fourier transform	552
8.4	The frequency response	558
8.5	Transforms of the step and impulse functions	563
8.6	The Fourier transform in discrete time	575
8.7	Engineering application: the design of analogue filters	599
8.8	Engineering application: direct design of digital filters and windows	602
8.9	Review exercises (1–25)	611

8.1

Introduction

In Chapter 7 [and Chapter 12 of *Modern Engineering Mathematics* (MEM)] we saw how Fourier series provided an ideal framework for analysing the steady-state response of systems to a periodic input signal. In this chapter we extend the ideas of Fourier analysis to deal with non-periodic functions. We do this via the introduction of the Fourier transform. As the theory develops, we shall see how the complex exponential form of the Fourier series representation of a periodic function emerges as a special case of the Fourier transform. Similarities between the transform and the Laplace transform, discussed in Chapter 11 of MEM and Chapter 5 in this text, will also be highlighted.

While Fourier transforms first found most application in the solution of partial differential equations, it is probably true to say that today Fourier transform methods are most heavily used in the analysis of signals and systems. This chapter is therefore developed with such applications in mind, and its main aim is to develop an understanding of the underlying mathematics as a preparation for a specialist study of application areas in various branches of engineering.

While much early work on signal analysis was implemented using analogue devices, the bulk of modern equipment exploits digital technology. In Chapter 11 of MEM and Chapter 5 in this text we developed the Laplace transform as an aid to the analysis and design of continuous-time systems while in Chapter 6 we introduced the z and \mathcal{D} transforms to assist with the analysis and design of discrete-time systems. In this chapter the frequency-domain analysis introduced in Chapter 11 of MEM and Chapter 5 in this text for continuous-time systems is consolidated and then extended to provide a framework for the frequency-domain description of discrete-time systems through the introduction of discrete Fourier transforms. These discrete transforms provide one of the most advanced methods for discrete signal analysis, and are widely used in such fields as communications theory and speech and image processing. In practice, the computational aspects of the work assume great importance, and the use of appropriate computational algorithms for the calculation of the discrete Fourier transform is essential. For this reason we have included an introduction to the fast Fourier transform algorithm, based on the pioneering work of J. W. Cooley and J. W. Tukey published in 1965, which it is hoped will serve the reader with the necessary understanding for progression to the understanding of specialist engineering applications.

In the engineering application (Section 8.8) we discuss the discrete-time Fourier transform to provide the means of describing the so-called direct design method for digital filters which is based on the use of the desired frequency response, without using an analogue prototype design. This naturally leads to considering ‘windowing’ and a brief introduction to this topic is included.

Wavelets were developed in the 1980s and 1990s and they provide a technique for analysing pulses. Their application to medical signals has grown since 2010 but, as yet, they have found little use in mainstream engineering. This will change. However, for now, the study of wavelets and the introduction of the wavelet transform is postponed. For a brief introduction see Phil Dyke’s *An Introduction to Laplace Transforms and Fourier Series* (second edition, London, Springer, 2014).

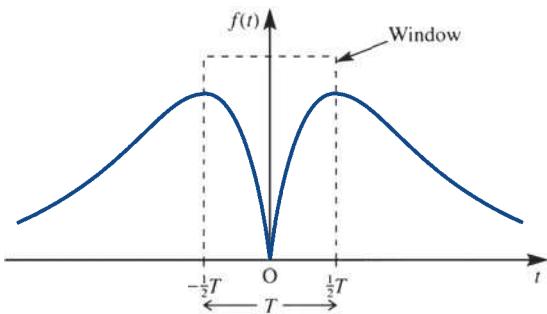


Figure 8.1 The view of $f(t)$ through a window of length T .

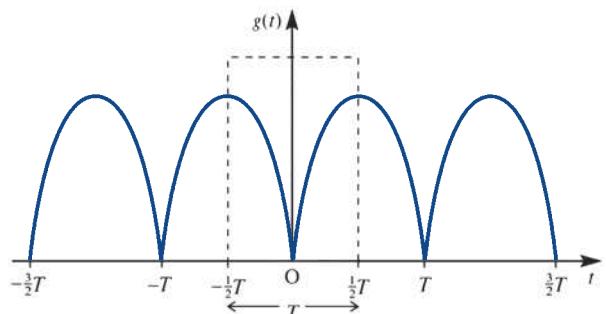


Figure 8.2 The periodic function $g(t)$ based on the ‘windowed’ view of $f(t)$.

8.2 The Fourier transform

8.2.1 The Fourier integral

In Chapter 7 (and Chapter 12 of MEM) we saw how Fourier series methods provided a technique for the frequency-domain representation of periodic functions. As indicated in Section 7.4.3, in expressing a function as its Fourier series expansion we are decomposing the function into its harmonic or frequency components. Thus a periodic function $f(t)$, of period T' , has frequency components at discrete frequencies

$$\omega_n = \frac{2\pi n}{T'} = n\omega_0 \quad (n = 0, 1, 2, 3, \dots)$$

where ω_0 is the fundamental frequency, that is the frequency of the parent function $f(t)$. Consequently, we were able to interpret a Fourier series as constituting a **discrete frequency spectrum** of the periodic function $f(t)$, thus providing an alternative frequency-domain representation of the function to its time-domain waveform. However, not all functions are periodic and so we need to develop an approach that will give a similar representation for non-periodic functions, defined on $-\infty < t < \infty$. One way of achieving this is to look at a portion of a non-periodic function $f(t)$ over an interval T , by imagining that we are looking at a graph of $f(t)$ through a ‘window’ of length T , and then to consider what happens as T gets larger.

Figure 8.1 depicts this situation, with the window placed symmetrically about the origin. We could now concentrate only on the ‘view through the window’ and carry out a Fourier series development based on that portion of $f(t)$ alone. Whatever the behaviour of $f(t)$ outside the window, the Fourier series thus generated would represent the periodic function defined by

$$g(t) = \begin{cases} f(t) & (|t| < \frac{1}{2}T) \\ f(t - nT) & \left(\frac{1}{2}(2n-1)T < |t| < \frac{1}{2}(2n+1)T\right) \end{cases}$$

Figure 8.2 illustrates $g(t)$, and we can see that the graphs of $f(t)$ and $g(t)$ agree on the interval $(-\frac{1}{2}T, \frac{1}{2}T)$.

Using the complex or exponential form of the Fourier series expansion, we have from (7.39) that

$$g(t) = \sum_{n=-\infty}^{\infty} G_n e^{jn\omega_0 t} \quad (8.1)$$

with

$$G_n = \frac{1}{T} \int_{-T/2}^{T/2} g(t) e^{-jn\omega_0 t} dt \quad (8.2)$$

and where

$$\omega_0 = 2\pi/T \quad (8.3)$$

Equation (8.2) in effect *transforms* the time-domain function $g(t)$ into the associated frequency-domain components G_n , where n is *any* integer. Equation (8.1) can also be viewed as transforming the discrete components G_n in the frequency-domain representation to the time-domain form $g(t)$. Substituting for G_n in (8.1), using (8.2), we obtain

$$g(t) = \sum_{n=-\infty}^{\infty} \left[\frac{1}{T} \int_{-T/2}^{T/2} g(\tau) e^{-jn\omega_0 \tau} d\tau \right] e^{jn\omega_0 t} \quad (8.4)$$

The frequency of the general term in the expansion (8.4) is

$$\frac{2\pi n}{T} = n\omega_0 = \omega_n$$

and so the difference in frequency between successive terms is

$$\frac{2\pi}{T}[(n+1) - n] = \frac{2\pi}{T} = \Delta\omega$$

Since $\Delta\omega = \omega_0$, we can express (8.4) as

$$g(t) = \sum_{n=-\infty}^{\infty} \left[\frac{1}{2\pi} \int_{-T/2}^{T/2} g(\tau) e^{-j\omega_n \tau} d\tau \right] e^{j\omega_n t} \Delta\omega \quad (8.5)$$

Defining $G(j\omega)$ as

$$G(j\omega) = \int_{-T/2}^{T/2} g(\tau) e^{-j\omega\tau} d\tau \quad (8.6)$$

we have

$$g(t) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-j\omega_n t} G(j\omega_n) \Delta\omega \quad (8.7)$$

As $T \rightarrow \infty$, our window widens, so that $g(t) = f(t)$ everywhere and $\Delta\omega \rightarrow 0$. Since we also have

$$\lim_{\Delta\omega \rightarrow 0} \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{j\omega_n t} G(j\omega_n) \Delta\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega t} G(j\omega) d\omega$$

it follows from (8.7) and (8.6) that

$$f(t) = \int_{-\infty}^{\infty} \left[\frac{1}{2\pi} e^{j\omega t} \int_{-\infty}^{\infty} f(\tau) e^{-j\omega\tau} d\tau \right] d\omega \quad (8.8)$$

The result (8.8) is known as the **Fourier integral representation** of $f(t)$. A set of conditions that are sufficient for the existence of the Fourier integral is a revised form of Dirichlet's conditions for Fourier series convergence (see Theorem 12.2 of MEM for further details). These conditions may be stated in the form of Theorem 8.1.

Theorem 8.1

Dirichlet's conditions for the Fourier integral

If the function $f(t)$ is such that

- (a) it is absolutely integrable, so that

$$\int_{-\infty}^{\infty} |f(t)| dt < \infty$$

(that is, the integral is finite), and

- (b) it has at most a finite number of maxima and minima and a finite number of discontinuities in any finite interval

then the Fourier integral representation of $f(t)$, given in (8.8), converges to $f(t)$ at all points where $f(t)$ is continuous and to the average of the right- and left-hand limits of $f(t)$ where $f(t)$ is discontinuous (that is, to the mean of the discontinuity).

[end of theorem](#)

As was indicated in the introduction in Chapter 7, the use of the equality sign in (8.8) must be interpreted carefully because of the non-convergence to $f(t)$ at points of discontinuity. Again the symbol \sim (read as 'behaves as' or 'represented by') rather than $=$ is frequently used.

Condition (a) of Theorem 8.1 implies that the absolute area under the graph of $y = f(t)$ is finite. Clearly this is so if $f(t)$ decays sufficiently fast with time. However, in general the condition seems to imply a very tight constraint on the nature of $f(t)$, since clearly functions of the form $f(t) = \text{constant}$, $f(t) = e^{at}$, $f(t) = e^{-at}$, $f(t) = \sin \omega t$, and so on, defined for $-\infty < t < \infty$, do not meet the requirement. In practice, however, signals are usually causal and only exist for a finite time. Also, in practice no signal amplitude goes to infinity, so consequently no **practical signal** $f(t)$ can have an infinite area under its graph $y = f(t)$. Thus for practical signals the integral in (8.8) exists.

To obtain the trigonometric (or real) form of the Fourier integral, we substitute

$$e^{-j\omega(\tau-t)} = \cos \omega(\tau-t) - j \sin \omega(\tau-t)$$

in (8.8) to give

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\tau) [\cos \omega(\tau-t) - j \sin \omega(\tau-t)] d\tau d\omega$$

Since $\sin \omega(\tau-t)$ is an odd function of ω , this reduces to

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\tau) \cos \omega(\tau-t) d\tau d\omega$$

which, on noting that the integrand is an even function of ω reduces further to

$$f(t) = \frac{1}{\pi} \int_0^{\infty} d\omega \int_{-\infty}^{\infty} f(\tau) \cos \omega(\tau-t) d\tau \quad (8.9)$$

The representation (8.9) is then the required trigonometric form of the Fourier integral.

If $f(t)$ is either an odd function or an even function then further simplifications of (8.9) are possible. Detailed calculations are left as an exercise for the reader, and we shall simply quote the results.

(a) If $f(t)$ is an even function then (8.9) reduces to

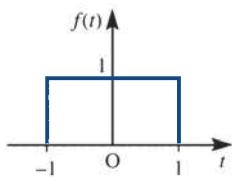
$$f(t) = \frac{2}{\pi} \int_0^\infty \int_0^\infty f(\tau) \cos \omega \tau \cos \omega t d\tau d\omega \quad (8.10)$$

which is referred to as the **Fourier cosine integral**.

(b) If $f(t)$ is an odd function then (8.9) reduces to

$$f(t) = \frac{2}{\pi} \int_0^\infty \int_0^\infty f(\tau) \sin \omega \tau \sin \omega t d\tau d\omega \quad (8.11)$$

which is referred to as the **Fourier sine integral**.



In the case of the Fourier series representation of a periodic function it was a matter of some interest to determine how well the first few terms of the expansion represented the function. The corresponding problem in the non-periodic case is to investigate how well the Fourier integral represents a function when only the components in the lower part of the (continuous) frequency range are taken into account. To illustrate, consider the rectangular pulse of Figure 8.3 given by

Figure 8.3 Rectangular pulse

$$f(t) = \begin{cases} 1 & (|t| \leq 1) \\ 0 & (|t| > 1) \end{cases}$$

$$f(t) = \begin{cases} 1 & (|t| \leq 1) \\ 0 & (|t| > 1) \end{cases}$$

This is clearly an even function, so from (8.10) its Fourier integral is

$$f(t) = \frac{2}{\pi} \int_0^\infty \int_0^1 1 \cos \omega \tau \cos \omega t d\tau d\omega = \frac{2}{\pi} \int_0^\infty \frac{\cos \omega t \sin \omega}{\omega} d\omega$$

An elementary evaluation of this integral is not possible, so we consider frequencies $\omega < \omega_0$, when

$$\begin{aligned} f(t) &\simeq \frac{2}{\pi} \int_0^{\omega_0} \frac{\cos \omega t \sin \omega}{\omega} d\omega = \frac{1}{\pi} \int_0^{\omega_0} \frac{\sin \omega(t+1)}{\omega} d\omega - \frac{1}{\pi} \int_0^{\omega_0} \frac{\sin \omega(t-1)}{\omega} d\omega \\ &= \frac{1}{\pi} \int_0^{\omega_0(t+1)} \frac{\sin u}{u} du - \frac{1}{\pi} \int_0^{\omega_0(t-1)} \frac{\sin u}{u} du \end{aligned}$$

The integral

$$\text{Si}(x) = \int_0^x \frac{\sin u}{u} du \quad (x \geq 0)$$

occurs frequently, and it can be shown that

$$\text{Si}(x) = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)(2n+1)!}$$

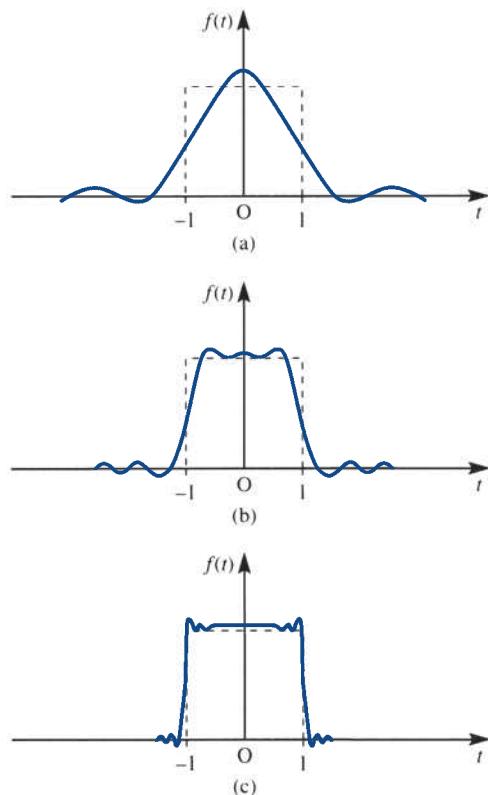
Its values have been tabulated and are freely available on the world wide web (for example, search for ‘values of the Si(x) function’). Thus

$$f(t) \simeq \text{Si}(\omega_0(t+1)) - \text{Si}(\omega_0(t-1)) \quad (8.12)$$

This has been plotted for $\omega_0 = 4, 8$ and 16 , and the responses are shown in Figures 8.4(a), (b) and (c) respectively. Physically, these responses describe the output of an ideal low-pass filter, cutting out all frequencies $\omega > \omega_0$, when the input signal is the rectangular pulse of Figure 8.3.

Figure 8.4

Plot of (8.12):
 (a) $\omega_0 = 4$; (b) $\omega_0 = 8$;
 (c) $\omega_0 = 16$.



8.2.2 The Fourier transform pair

We note from (8.6) and (8.7) that the Fourier integral (8.8) may be written in the form of the pair of equations

$$F(j\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \quad (8.13)$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega \quad (8.14)$$

$F(j\omega)$ as defined by (8.13) is called the **Fourier transform** of $f(t)$, and it provides a frequency-domain representation of the non-periodic function $f(t)$, whenever the integral in (8.13) exists. Note that we have used the notation $F(j\omega)$ for the Fourier transform of $f(t)$ rather than the alternative $F(\omega)$, which is also in common use. The reason for this choice is a consequence of the relationship between the Fourier and Laplace transforms, which will emerge later in Section 8.4.1. We stress that this is a *choice* that we have made, but the reader should have no difficulty in using either form, provided that once the choice has been made it is then adhered to. Equation (8.14) then provides us with a way of reconstructing $f(t)$ if we know its Fourier transform $F(j\omega)$.

A word of caution is in order here regarding the scaling factor $1/2\pi$ in (8.14). Although the convention that we have adopted here is fairly standard, some authors associate the factor $1/2\pi$ with (8.13) rather than (8.14), while others associate a factor of $(2\pi)^{-1/2}$ with each of (8.13) and (8.14). In all cases the pair combine to give the Fourier integral (8.8). We could overcome this possible confusion by measuring the frequency in cycles per second or hertz rather than in radians per second, this being achieved using the substitution $f = \omega/2\pi$, where f is in hertz and ω is in radians per second. We have not adopted this approach, since ω is so widely used by engineers.

In line with our notation for Laplace transforms in Chapter 5, we introduce the symbol \mathcal{F} to denote the Fourier transform operator. Then from (8.13) the Fourier transform $\mathcal{F}\{f(t)\}$ of a function $f(t)$ is defined by

$$\mathcal{F}\{f(t)\} = F(j\omega) = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \quad (8.15)$$

whenever the integral exists. Similarly, using (8.14), we define the inverse Fourier transform of $G(j\omega)$ as

$$\mathcal{F}^{-1}\{G(j\omega)\} = g(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(j\omega) e^{j\omega t} d\omega \quad (8.16)$$

whenever the integral exists. The relations (8.15) and (8.16) together constitute the **Fourier transform pair**, and they provide a pathway between the time- and frequency-domain representations of a function. Equation (8.15) expresses $f(t)$ in the frequency domain, and is analogous to resolving it into harmonic components with a continuously varying frequency ω . This contrasts with a Fourier series representation of a periodic function, where the resolved frequencies take discrete values.

The conditions for the existence of the Fourier transform $F(j\omega)$ of the function $f(t)$ are Dirichlet's conditions (Theorem 8.1). Corresponding trigonometric forms of the Fourier transform pair may be readily written down from (8.9), (8.10) and (8.11).

Example 8.1

Does the function

$$f(t) = 1 \quad (-\infty < t < \infty)$$

have a Fourier transform representation?

Solution

Since the area under the curve of $y = f(t)$ ($-\infty < t < \infty$) is infinite, it follows that $\int_{-\infty}^{\infty} |f(t)| dt$ is unbounded, so the conditions of Theorem 8.1 are not satisfied. We can confirm that the Fourier transform does not exist from the definition (8.15). We have

$$\begin{aligned} \int_{-\infty}^{\infty} 1 e^{-j\omega t} dt &= \lim_{\alpha \rightarrow \infty} \int_{-\alpha}^{\alpha} e^{-j\omega t} dt \\ &= \lim_{\alpha \rightarrow \infty} \left[-\frac{1}{j\omega} (e^{-j\omega\alpha} - e^{j\omega\alpha}) \right] \\ &= \lim_{\alpha \rightarrow \infty} \frac{2 \sin \omega\alpha}{\omega} \end{aligned}$$

Since this last limit does not exist, we conclude that $f(t) = 1$ ($-\infty < t < \infty$) does not have a Fourier transform representation.

It is clear, using integration by parts, that $f(t) = t$ ($-\infty < t < \infty$) does not have a Fourier transform, nor indeed does $f(t) = t^n$ ($n > 1$, an integer; $-\infty < t < \infty$). While neither e^{at} nor e^{-at} ($a > 0$) has a Fourier transform, when we consider the causal signal $f(t) = H(t) e^{-at}$ ($a > 0$), we do obtain a transform.

Example 8.2

Find the Fourier transform of the one-sided exponential function

$$f(t) = H(t) e^{-at} \quad (a > 0)$$

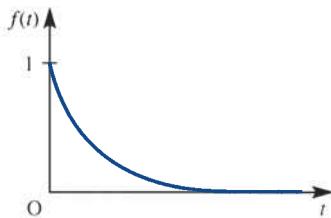
where $f(t)$ is the Heaviside unit step function.

Solution

The graph of $f(t)$ is shown in Figure 8.5, and we can show that the area under the graph is bounded. Hence, by Theorem 8.1, a Fourier transform exists. Using the definition (8.15), we have

$$\begin{aligned} \mathcal{F}\{f(t)\} &= \int_{-\infty}^{\infty} H(t) e^{-at} e^{-j\omega t} dt \quad (a > 0) \\ &= \int_0^{\infty} e^{-(a+j\omega)t} dt = \left[-\frac{e^{-(a+j\omega)t}}{a + j\omega} \right]_0^{\infty} \end{aligned}$$

Figure 8.5
The ‘one-sided’ exponential function of Example 8.2.



so that

$$\mathcal{F}\{H(t)e^{-at}\} = \frac{1}{a + j\omega} \quad (8.17)$$

Example 8.3

Calculate the Fourier transform of the rectangular pulse

$$f(t) = \begin{cases} A & (|t| \leq T) \\ 0 & (|t| > T) \end{cases}$$

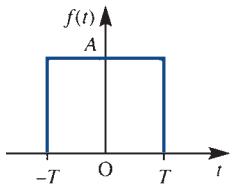
Solution

The graph of $f(t)$ is shown in Figure 8.6, and since the area under it is finite, a Fourier transform exists. From the definition (8.15), we have

$$\begin{aligned} \mathcal{F}\{f(t)\} &= \int_{-T}^T A e^{-j\omega t} dt = \begin{cases} \left[-\frac{A}{j\omega} e^{-j\omega t} \right]_{-T}^T & \omega \neq 0 \\ 2A & \omega = 0 \end{cases} \\ &= 2AT \operatorname{sinc} \omega T \end{aligned}$$

Figure 8.6 The rectangular pulse of Example 8.3.

Figure 8.7
A brief table of Fourier transforms.



where $\operatorname{sinc} x$ is defined, as in Example 7.12, by

$$\operatorname{sinc} x = \begin{cases} \frac{\sin x}{x} & (x \neq 0) \\ 1 & (x = 0) \end{cases}$$

$f(t)$	$\mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt$
$e^{-at} H(t) \quad (a > 0)$	$\frac{1}{a + j\omega}$
$t e^{-at} H(t) \quad (a > 0)$	$\frac{1}{(a + j\omega)^2}$
$\begin{cases} A & (t \leq T) \\ 0 & (t > T) \end{cases}$	$2AT \operatorname{sinc} \omega T$
$e^{-a t } (a > 0)$	$\frac{2a}{a^2 + \omega^2}$

By direct use of the definition (8.15), we can, as in Examples 8.2 and 8.3, determine the Fourier transforms of some standard functions. A brief table of transforms is given in Figure 8.7.



In MATLAB, incorporating the Symbolic Math Toolbox, the Fourier transform $F(j\omega)$ of $f(t)$ is obtained using the commands

```
syms w t
F=fourier(f(t),t,w)
```

whilst the inverse Fourier transform $f(t)$ of $F(j\omega)$ is obtained using the command

```
f=ifourier(F(jw),w,t)
```

Returning to Example 8.2, and considering the particular case of $a = 2$, the commands

```
syms w t
H=sym('heaviside(t)');
F=fourier(H*exp(-2*t))
```

in MATLAB return

```
F=1/(2+w^2)
```

as expected. In MATLAB there is an `assume` command (as in MAPLE) to enable us to specify that $a > 0$. However, since $\text{abs}(a) = a$ for $a > 0$, the following commands in MATLAB can be used to deal with the general case

```
syms w t a
H=sym('heaviside(t)');
F=fourier(H*exp(-abs(a)*t),t,w)
```

As another illustration, consider the function $f(t) = e^{-a|t|}$, $a > 0$, given in the table of Figure 8.7. Considering the particular case $a = 2$ then the MATLAB commands

```
syms w t
F=fourier(exp(-2*abs(t)),t,w)
```

return

```
F=4/(4+w^2)
```

as specified in the table. It is left as an exercise to consider the general case of a . To illustrate the use, in MATLAB, of the `ifourier` command this transform can be inverted using the commands

```
syms w t
f=ifourier(4/(w^2+4),w,t)
```

which return

```
f= exp(-2*abs(t))
```

As another illustration consider the Fourier transform $F(\omega) = 1/(a + j\omega)^2$ given in the second entry of the table in Figure 8.7. The MATLAB commands

```
syms w t a
f=ifourier(1/(a+i*w)^2,w,t)
f=simplify(f)
return
f=(t*exp(-a*t)*(sign(real(a))+sign(t)))/2
```

corresponding to the table.

Considering the rectangular pulse $f(t)$ of Example 8.3, we first express the pulse in terms of Heaviside functions as

$$f(t) = A(H(t + T) - H(t - T))$$

and then use the MATLAB commands

```
syms w t T A
H=sym('heaviside(t+T)-heaviside(t-T)');
F=fourier(A*H,t,w);
F=simplify(F)
```

which return

$$F=2*A*sin(T*w)/w$$

8.2.3 The continuous Fourier spectra

From Figure 8.7, it is clear that Fourier transforms are generally complex-valued functions of the real frequency variable ω . If $\mathcal{F}\{f(t)\} = F(j\omega)$ is the Fourier transform of the signal $f(t)$ then $F(j\omega)$ is also known as the **(complex) frequency spectrum** of $f(t)$. Writing $F(j\omega)$ in the exponential form

$$F(j\omega) = |F(j\omega)| e^{j \arg F(j\omega)}$$

plots of $|F(j\omega)|$ and $\arg F(j\omega)$, which are both real-valued functions of ω , are called the **amplitude** and **phase spectra** respectively of the signal $f(t)$. These two spectra represent the **frequency-domain portrait** of the signal $f(t)$. In contrast to the situation when $f(t)$ was periodic, where (as shown in Section 7.4.3) the amplitude and phase spectra were defined only at discrete values of ω , we now see that both spectra are defined for all values of the continuous variable ω .

Example 8.4

Determine the amplitude and phase spectra of the causal signal

$$f(t) = e^{-at} H(t) \quad (a > 0)$$

and plot their graphs.

Solution From (8.17),

$$\mathcal{F}\{f(t)\} = F(j\omega) = \frac{1}{a + j\omega}$$

Thus the amplitude and argument of $F(j\omega)$ are

$$|F(j\omega)| = \frac{1}{\sqrt{(a^2 + \omega^2)}} \quad (8.18)$$

$$\arg F(j\omega) = \tan^{-1}(1) - \tan^{-1}\left(\frac{\omega}{a}\right) = -\tan^{-1}\left(\frac{\omega}{a}\right) \quad (8.19)$$

These are the amplitude and phase spectra of $f(t)$, and are plotted in Figure 8.8.

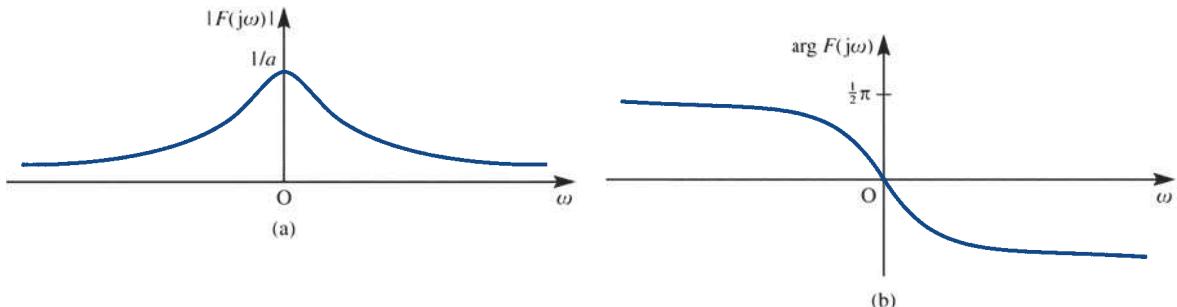


Figure 8.8 (a) Amplitude and (b) phase spectra of the one-sided exponential function $f(t)$ of Example 8.4.

Generally, as we have observed, the Fourier transform and thus the frequency spectrum are complex-valued quantities. In some cases, as for instance in Example 8.3, the spectrum is purely real. In Example 8.3 we found that the transform of the pulse illustrated in Figure 8.6 was

$$F(j\omega) = 2AT \operatorname{sinc} \omega T$$

where

$$\operatorname{sinc} \omega T = \begin{cases} \frac{\sin \omega T}{\omega T} & (\omega \neq 0) \\ 1 & (\omega = 0) \end{cases}$$

is an even function of ω , taking both positive and negative values. In this case the amplitude and phase spectra are given by

$$|F(j\omega)| = 2AT |\operatorname{sinc} \omega T| \quad (8.20)$$

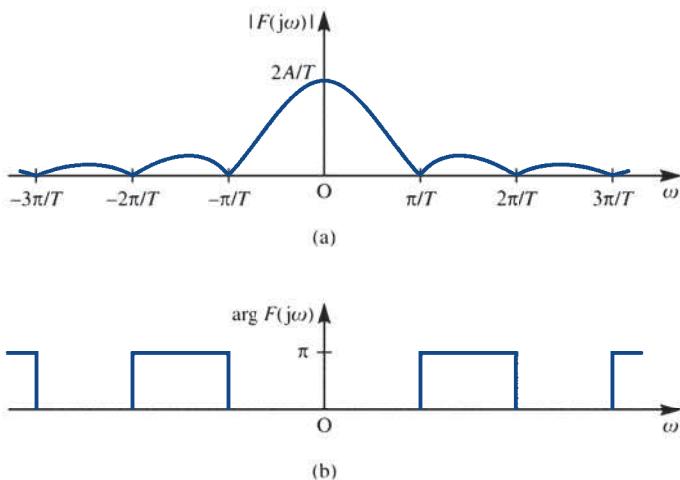
$$\arg F(j\omega) = \begin{cases} 0 & (\operatorname{sinc} \omega T \geq 0) \\ \pi & (\operatorname{sinc} \omega T < 0) \end{cases} \quad (8.21)$$

with corresponding graphs shown in Figure 8.9.

Figure 8.9

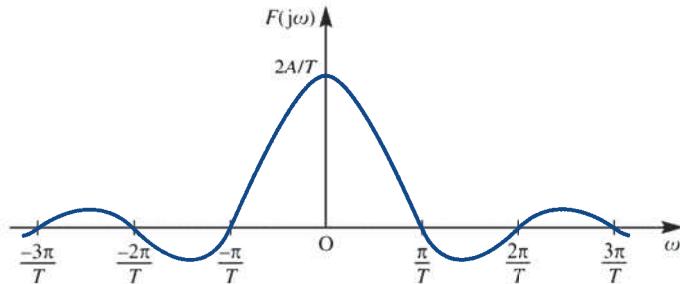
(a) Amplitude and
(b) spectra of the pulse

$$f(t) = \begin{cases} A & (|t| \leq T) \\ 0 & (|t| > T). \end{cases}$$

**Figure 8.10**

Frequency spectrum
(real-valued) of the pulse

$$f(t) = \begin{cases} A & (|t| \leq T) \\ 0 & (|t| > T). \end{cases}$$



In fact, when the Fourier transform is a purely real-valued function, we can plot all the information on a single frequency spectrum of $F(j\omega)$ versus ω . For the rectangular pulse of Figure 8.6 the resulting graph is shown in Figure 8.10.

From Figure 8.7, we can see that the Fourier transforms discussed so far have two properties in common. First, the amplitude spectra are even functions of the frequency variable ω . This is always the case when the time signal $f(t)$ is real; that is, loosely speaking, a consequence of the fact that we have decomposed, or analysed $f(t)$, relative to complex exponentials rather than real-valued sines and cosines. The second common feature is that all the amplitude spectra decrease rapidly as ω increases. This means that most of the information concerning the ‘shape’ of the signal $f(t)$ is contained in a fairly small interval of the frequency axis around $\omega = 0$. From another point of view, we see that a device capable of passing signals of frequencies up to about $\omega = 3\pi/T$ would pass a reasonably accurate version of the rectangular pulse of Example 8.3.

8.2.4 Exercises



Whenever possible check your answers using MATLAB or MAPLE.

- 1 Calculate the Fourier transform of the two-sided exponential pulse given by

$$f(t) = \begin{cases} e^{at} & (t \leq 0) \\ e^{-at} & (t > 0) \end{cases} \quad (a > 0)$$

- 2 Determine the Fourier transform of the ‘on–off’ pulse shown in Figure 8.11.

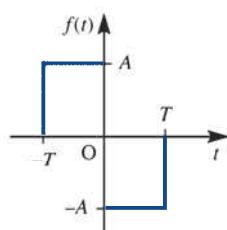


Figure 8.11 The ‘on–off’ pulse of Exercise 2.

- 3 A triangular pulse is defined by

$$f(t) = \begin{cases} (A/T)t + A & (-T \leq t \leq 0) \\ (-A/T)t + A & (0 < t \leq T) \end{cases}$$

Sketch $f(t)$ and determine its Fourier transform. What is the relationship between this pulse and that of Exercise 2?

- 4 Determine the Fourier transforms of

$$f(t) = \begin{cases} 2K & (|t| \leq 2) \\ 0 & (|t| > 2) \end{cases}$$

$$g(t) = \begin{cases} K & (|t| \leq 1) \\ 0 & (|t| > 1) \end{cases}$$

Sketch the function $h(t) = f(t) - g(t)$ and determine its Fourier transform.

- 5 Calculate the Fourier transform of the ‘off–on–off’ pulse $f(t)$ defined by

$$f(t) = \begin{cases} 0 & (t < -2) \\ -1 & (-2 \leq t < -1) \\ 1 & (-1 \leq t \leq 1) \\ -1 & (1 < t \leq 2) \\ 0 & (t > 2) \end{cases}$$

- 6 Show that the Fourier transform of

$$f(t) = \begin{cases} \sin at & (|t| \leq \pi/a) \\ 0 & (|t| > \pi/a) \end{cases}$$

is

$$\frac{j2a \sin(\pi\omega/a)}{\omega^2 - a^2}$$

- 7 Calculate the Fourier transform of

$$f(t) = e^{-at} \sin \omega_0 t H(t)$$

- 8 Based on (8.10) and (8.11), define the **Fourier sine transform** as

$$F_s(x) = \int_0^\infty f(t) \sin xt dt$$

and the **Fourier cosine transform** as

$$F_c(x) = \int_0^\infty f(t) \cos xt dt$$

Show that

$$f(t) = \begin{cases} 0 & (t < 0) \\ \cos at & (0 \leq t \leq a) \\ 0 & (t > a) \end{cases}$$

has Fourier cosine transform

$$\frac{1}{2} \left[\frac{\sin(1+x)a}{1+x} + \frac{\sin(1-x)a}{1-x} \right]$$

- 9 Show that the Fourier sine and cosine transforms of

$$f(t) = \begin{cases} 0 & (t < 0) \\ 1 & (0 \leq t \leq a) \\ 0 & (t > a) \end{cases}$$

are

$$\frac{1 - \cos xa}{x}, \quad \frac{\sin xa}{x}$$

respectively.

- 10 Find the sine and cosine transforms of $f(t) = e^{-at} H(t)$ ($a > 0$).

8.3

Properties of the Fourier transform

In this section we establish some of the properties of the Fourier transform that allow its use as a practical tool in system analysis and design.

8.3.1 The linearity property

Linearity, as with the Laplace transform, is a fundamental property of the Fourier transform, and may be stated as follows.

If $f(t)$ and $g(t)$ are functions having Fourier transforms $F(j\omega)$ and $G(j\omega)$ respectively, and if α and β are constants, then

$$\mathcal{F}\{\alpha f(t) + \beta g(t)\} = \alpha \mathcal{F}\{f(t)\} + \beta \mathcal{F}\{g(t)\} = \alpha F(j\omega) + \beta G(j\omega) \quad (8.22)$$

As a consequence of this, we say that the Fourier transform operator \mathcal{F} is a **linear operator**. The proof of this property follows readily from the definition (8.15), since

$$\begin{aligned} \mathcal{F}\{\alpha f(t) + \beta g(t)\} &= \int_{-\infty}^{\infty} [\alpha f(t) + \beta g(t)] e^{-j\omega t} dt \\ &= \alpha \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt + \beta \int_{-\infty}^{\infty} g(t) e^{-j\omega t} dt \\ &= \alpha F(j\omega) + \beta G(j\omega) \end{aligned}$$

Clearly the linearity property also applies to the inverse transform operator \mathcal{F}^{-1} .

8.3.2 Time-differentiation property

If the function $f(t)$ has a Fourier transform $F(j\omega)$ then, by (8.16),

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega$$

Differentiating with respect to t gives

$$\frac{df}{dt} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\partial}{\partial t} [F(j\omega) e^{j\omega t}] d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} (j\omega) F(j\omega) e^{j\omega t} d\omega$$

implying that the time signal df/dt is the inverse Fourier transform of $(j\omega)F(j\omega)$. In other words

$$\mathcal{F}\left\{\frac{df}{dt}\right\} = (j\omega)F(j\omega)$$

Repeating the argument n times, it follows that

$$\mathcal{F}\left\{\frac{d^n f}{dt^n}\right\} = (j\omega)^n F(j\omega) \quad (8.23)$$

The result (8.23) is referred to as the **time-differentiation property**, and may be used to obtain frequency-domain representations of differential equations.

Example 8.5

Show that if the time signals $y(t)$ and $u(t)$ have Fourier transforms $Y(j\omega)$ and $U(j\omega)$ respectively, and if

$$\frac{d^2y(t)}{dt^2} + 3\frac{dy(t)}{dt} + 7y(t) = 3\frac{du(t)}{dt} + 2u(t) \quad (8.24)$$

then $Y(j\omega) = G(j\omega)U(j\omega)$ for some function $G(j\omega)$.

Solution

Taking Fourier transforms throughout in (8.24), we have

$$\mathcal{F}\left\{\frac{d^2y(t)}{dt^2} + 3\frac{dy(t)}{dt} + 7y(t)\right\} = \mathcal{F}\left\{3\frac{du(t)}{dt} + 2u(t)\right\}$$

which, on using the linearity property (8.22), reduces to

$$\mathcal{F}\left\{\frac{d^2y(t)}{dt^2}\right\} + 3\mathcal{F}\left\{\frac{dy(t)}{dt}\right\} + 7\mathcal{F}\{y(t)\} = 3\mathcal{F}\left\{\frac{du(t)}{dt}\right\} + 2\mathcal{F}\{u(t)\}$$

Then, from (8.23),

$$(j\omega)^2 Y(j\omega) + 3(j\omega) Y(j\omega) + 7Y(j\omega) = 3(j\omega) U(j\omega) + 2U(j\omega)$$

that is,

$$(-\omega^2 + j3\omega + 7)Y(j\omega) = (j3\omega + 2)U(j\omega)$$

giving

$$Y(j\omega) = G(j\omega)U(j\omega)$$

where

$$G(j\omega) = \frac{2 + j3\omega}{7 - \omega^2 + j3\omega}$$

The reader may at this stage be fearing that we are about to propose yet *another* method for solving differential equations. This is not the idea! Rather, we shall show that the Fourier transform provides an essential tool for the analysis (and synthesis) of linear systems from the viewpoint of the frequency domain.

8.3.3 Time-shift property

If a function $f(t)$ has Fourier transform $F(j\omega)$ then what is the Fourier transform of the shifted version $g(t) = f(t - \tau)$, where τ is a constant? From the definition (8.15),

$$\mathcal{F}\{g(t)\} = \int_{-\infty}^{\infty} g(t) e^{-j\omega t} dt = \int_{-\infty}^{\infty} f(t - \tau) e^{-j\omega t} dt$$

Making the substitution $x = t - \tau$, we have

$$\mathcal{F}\{g(t)\} = \int_{-\infty}^{\infty} f(x) e^{-j\omega(x+\tau)} dx = e^{-j\omega\tau} \int_{-\infty}^{\infty} f(x) e^{-j\omega x} dx = e^{-j\omega\tau} F(j\omega)$$

that is,

$$\mathcal{F}\{f(t - \tau)\} = e^{-j\omega\tau} F(j\omega) \quad (8.25)$$

The result (8.25) is known as the **time-shift property**, and implies that delaying a signal by a time τ causes its Fourier transform to be multiplied by $e^{-j\omega\tau}$. Note the similarity between this and the first-shift theorem in Laplace transforms, Theorem 5.1.

Since

$$|e^{-j\omega\tau}| = |\cos \omega\tau - j \sin \omega\tau| = \sqrt{(\cos^2 \omega\tau + \sin^2 \omega\tau)} = 1$$

we have

$$|e^{-j\omega\tau} F(j\omega)| = |F(j\omega)|$$

indicating that the amplitude spectrum of $f(t - \tau)$ is identical with that of $f(t)$. However,

$$\arg[e^{-j\omega\tau} F(j\omega)] = \arg F(j\omega) - \arg e^{j\omega\tau} = \arg F(j\omega) - \omega\tau$$

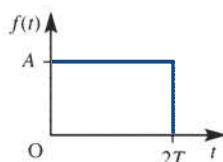
indicating that each frequency component is shifted by an amount proportional to its frequency ω

Example 8.6

Determine the Fourier transform of the rectangular pulse $f(t)$ shown in Figure 8.12.

Figure 8.12

Rectangular pulse of Example 8.6.



Solution

This is just the pulse of Example 8.3 (shown in Figure 8.6), delayed by T . The pulse of Example 8.3 had a Fourier transform $2AT \operatorname{sinc} \omega T$, and so, using the shift property (8.25) with $\tau = T$, we have

$$\mathcal{F}\{f(t)\} = F(j\omega) = e^{-j\omega T} 2AT \operatorname{sinc} \omega T = 2AT e^{-j\omega T} \operatorname{sinc} \omega T$$

8.3.4 Frequency-shift property

Suppose that a function $f(t)$ has Fourier transform $F(j\omega)$. Then, from the definition (8.15), the Fourier transform of the related function $g(t) = e^{j\omega_0 t} f(t)$ is

$$\begin{aligned}
 \mathcal{F}\{g(t)\} &= \int_{-\infty}^{\infty} e^{j\omega_0 t} f(t) e^{-j\omega t} dt = \int_{-\infty}^{\infty} f(t) e^{-j(\omega - \omega_0)t} dt \\
 &= \int_{-\infty}^{\infty} f(t) e^{-j\tilde{\omega}t} dt, \quad \text{where } \tilde{\omega} = \omega - \omega_0 \\
 &= F(j\tilde{\omega}), \quad \text{by definition}
 \end{aligned}$$

Thus

$$\mathcal{F}\{e^{j\omega_0 t} f(t)\} = F(j(\omega - \omega_0)) \quad (8.26)$$

The result (8.26) is known as the **frequency-shift property**, and indicates that multiplication by $e^{j\omega_0 t}$ simply shifts the spectrum of $f(t)$ so that it is centred on the point $\omega = \omega_0$ in the frequency domain. This phenomenon is the mathematical foundation for the process of **modulation** in communication theory, illustrated in Example 8.7.

Example 8.7

Determine the frequency spectrum of the signal $g(t) = f(t) \cos \omega_c t$.

Solution Since $\cos \omega_c t = \frac{1}{2}(e^{j\omega_c t} + e^{-j\omega_c t})$, it follows, using the linearity property (8.22), that

$$\begin{aligned}
 \mathcal{F}\{g(t)\} &= \mathcal{F}\left\{\frac{1}{2}f(t)(e^{j\omega_c t} + e^{-j\omega_c t})\right\} \\
 &= \frac{1}{2}\mathcal{F}\{f(t)e^{j\omega_c t}\} + \frac{1}{2}\mathcal{F}\{f(t)e^{-j\omega_c t}\}
 \end{aligned}$$

If $\mathcal{F}\{f(t)\} = F(j\omega)$ then, using (8.26),

$$\mathcal{F}\{f(t) \cos \omega_c t\} = \mathcal{F}\{g(t)\} = \frac{1}{2}F(j(\omega - \omega_c)) + \frac{1}{2}F(j(\omega + \omega_c))$$

The effect of multiplying the signal $f(t)$ by the **carrier signal** $\cos \omega_c t$ is thus to produce a signal whose spectrum consists of two (scaled) versions of $F(j\omega)$, the spectrum of $f(t)$: one centred on $\omega = \omega_c$ and the other on $\omega = -\omega_c$. The carrier signal $\cos \omega_c t$ is said to be modulated by the signal $f(t)$.

Demodulation is considered in Review exercise 5 (Section 8.9).

8.3.5 The symmetry property

From the definition of the transform pair (8.15) and (8.16) it is apparent that there is some symmetry of structure in relation to the variables t and ω . We can establish the exact form of this symmetry as follows. From (8.16),

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega$$

or, equivalently, by changing the ‘dummy’ variable in the integration,

$$2\pi f(t) = \int_{-\infty}^{\infty} F(jy) e^{jyt} dy$$

so that

$$2\pi f(-t) = \int_{-\infty}^{\infty} F(jy) e^{-jyt} dy$$

or, on replacing t by ω

$$2\pi f(-\omega) = \int_{-\infty}^{\infty} F(jy) e^{-jy\omega} dy \quad (8.27)$$

The right-hand side of (8.27) is simply the definition (8.15) of the Fourier transform of $F(jt)$, with the integration variable t replaced by y . We therefore conclude that

$$\mathcal{F}\{F(jt)\} = 2\pi f(-\omega) \quad (8.28a)$$

given that

$$\mathcal{F}\{f(t)\} = F(j\omega) \quad (8.28b)$$

This (8.28) tells us that if $f(t)$ and $F(j\omega)$ form a Fourier transform pair then $F(jt)$ and $2\pi f(-\omega)$ also form a Fourier transform pair. This property is referred to as the **symmetry property of Fourier transforms or duality property**.

Example 8.8

Determine the Fourier transform of the signal

$$g(t) = C \operatorname{sinc} at = \begin{cases} \frac{C \sin at}{at} & (t \neq 0) \\ C & (t = 0) \end{cases} \quad (8.29)$$

Solution From Example 8.3, we know that if

$$f(t) = \begin{cases} A & (|t| \leq T) \\ 0 & (|t| > T) \end{cases} \quad (8.30)$$

then

$$\mathcal{F}\{f(t)\} = F(j\omega) = 2AT \operatorname{sinc} \omega T$$

Thus, by the symmetry property (8.28), $F(jt)$ and $2\pi f(-\omega)$ are also a Fourier transform pair. In this case

$$F(jt) = 2AT \operatorname{sinc} tT$$

and so, choosing $T = a$ and $A = C/2a$ to correspond to (8.29), we see that

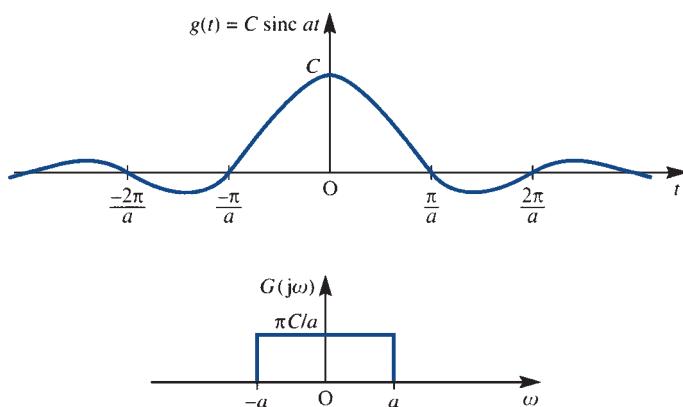
$$F(jt) = C \operatorname{sinc} at = g(t)$$

has Fourier transform $2\pi f(-\omega)$. Rewriting (8.30), we find that, since $|\omega| = |-\omega|$,

$$\mathcal{F}\{C \operatorname{sinc} at\} = \begin{cases} 2\pi C/2a & (|\omega| \leq a) \\ 0 & (|\omega| > a) \end{cases} = \begin{cases} \pi C/a & (|\omega| \leq a) \\ 0 & (|\omega| > a) \end{cases}$$

A graph of $g(t)$ and its Fourier transform $G(j\omega) = 2\pi f(-\omega)$ is shown in Figure 8.13.

Figure 8.13
The Fourier
transform pair
 $g(t)$ and $G(j\omega)$ of
Example 8.8.



Using the MATLAB commands

```
syms w t a C
F=fourier(C*sin(a*t)/(a*t),t,w);
F=simplify(F)
```

returns

$$F = (C \cdot \pi \cdot (\text{heaviside}(a+w) + \text{heaviside}(a-w) - 1)) / a$$

which is the answer given in the solution expressed in terms of Heaviside functions.

8.3.6 Exercises



Whenever possible check your answers using MATLAB or MAPLE.

- 11 Use the linearity property to verify the result in Exercise 4.

- 12 If $y(t)$ and $u(t)$ are signals with Fourier transforms $Y(j\omega)$ and $U(j\omega)$ respectively, and

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

show that $Y(j\omega) = H(j\omega)U(j\omega)$ for some function $H(j\omega)$. What is $H(j\omega)$?

- 13 Use the time-shift property to calculate the Fourier transform of the double pulse defined by

$$f(t) = \begin{cases} 1 & (1 \leq |t| \leq 2) \\ 0 & (\text{otherwise}) \end{cases}$$

- 14 Calculate the Fourier transform of the windowed cosine function

$$f(t) = \cos \omega_0 t [H\left(t + \frac{1}{2}T\right) - H\left(t - \frac{1}{2}T\right)]$$

- 15 Find the Fourier transform of the shifted form of the windowed cosine function

$$g(t) = \cos \omega_0 t [H(t) - H(t - T)]$$

- 16 Calculate the Fourier transform of the windowed sine function

$$f(t) = \sin 2t [H(t + 1) - H(t - 1)]$$

8.4 The frequency response

In this section we first consider the relationship between the Fourier and Laplace transforms, and then proceed to consider the frequency response in terms of the Fourier transform.

8.4.1 Relationship between Fourier and Laplace transforms

The differences between the Fourier and Laplace transforms are quite subtle. At first glance it appears that to obtain the Fourier transform from the Laplace transform we merely write $j\omega$ for s , and that the difference ends there. This is true in some cases, but not in all. Strictly, the Fourier and Laplace transforms are distinct, and neither is a generalization of the other.

Writing down the defining integrals, we have

The Fourier transform

$$\mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt \quad (8.31)$$

The bilateral Laplace transform

$$\mathcal{L}_B\{f(t)\} = \int_{-\infty}^{\infty} f(t) e^{-st} dt \quad (8.32)$$

The unilateral Laplace transform

$$\mathcal{L}\{f(t)\} = \int_{0^-}^{\infty} f(t) e^{-st} dt \quad (8.33)$$

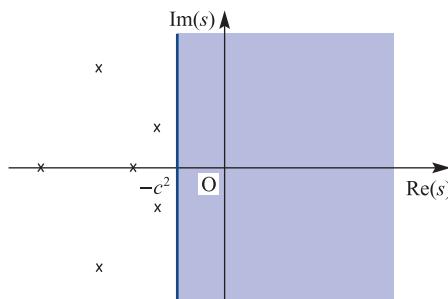
There is an obvious structural similarity between (8.31) and (8.32), while the connection with (8.33) is not so clear in view of the lower limit of integration. In the Laplace transform definitions recall that s is a complex variable, and may be written as

$$s = \sigma + j\omega \quad (8.34)$$

where σ and ω are real variables. We can then interpret (8.31), the Fourier transform of $f(t)$, as a special case of (8.32), when $\sigma = 0$, provided that the Laplace transform exists when $\sigma = 0$, or equivalently when $s = j\omega$ (that is, s describes the imaginary axis in the s plane). If we restrict our attention to causal functions, that is functions (or signals) that are zero whenever $t < 0$, the bilateral Laplace transform (8.32) is identical with the unilateral Laplace transform (8.33). The Fourier transform can thus be regarded as a special case of the unilateral Laplace transform for causal functions, provided again that the unilateral Laplace transform exists on the imaginary axis $s = j\omega$.

The next part of the story is concerned with a class of time signals $f(t)$ whose Laplace transforms do exist on the imaginary axis $s = j\omega$. Recall from (5.43) that a causal linear time-invariant system with Laplace transfer function $G(s)$ has an impulse response $h(t)$ given by

Figure 8.14
Pole locations for $G(s)$ and the region of existence of $G(s)$.



$$h(t) = \mathcal{L}^{-1}\{G(s)\} = g(t)H(t), \text{ say} \quad (8.35)$$

Furthermore, if the system is stable then all the poles of $G(s)$ are in the left half-plane, implying that $g(t)H(t) \rightarrow 0$ as $t \rightarrow \infty$. Let the pole locations of $G(s)$ be

$$p_1, p_2, \dots, p_n$$

where

$$p_k = -a_k^2 + jb_k$$

in which a_k, b_k are real and $a_k \neq 0$ for $k = 1, 2, \dots, n$. Examples of such poles are illustrated in Figure 8.14, where we have assumed that $G(s)$ is the transfer function of a real system so that poles that do not lie on the real axis occur in conjugate pairs. The Laplace transfer function $G(s)$ will exist in the shaded region of Figure 8.14 defined by

$$\operatorname{Re}(s) > -c^2$$

where $-c^2$ is the abscissa of convergence and is such that

$$0 < c^2 < \min a_k^2$$

The important conclusion is that for such systems $G(s)$ always exists on the imaginary axis $s = j\omega$, and so $h(t) = g(t)H(t)$ always has a Fourier transform. In other words, we have demonstrated that the impulse response function $h(t)$ of a *stable causal*, linear time-invariant system always has a Fourier transform. Moreover, we have shown that this can be found by evaluating the Laplace transform on the imaginary axis; that is, by putting $s = j\omega$ in the Laplace transform. We have thus established that Fourier transforms exist for a significant class of useful signals; this knowledge will be used in Section 8.4.2.

Example 8.9

Which of the following causal time-invariant systems have impulse responses that possess Fourier transforms? Find the latter when they exist.

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

$$(b) \frac{d^2y(t)}{dt^2} + \omega^2y(t) = u(t)$$

$$(c) \frac{d^2y(t)}{dt^2} + \frac{dy(t)}{dt} + y(t) = 2u(t) + \frac{du(t)}{dt}$$

Solution Assuming that the systems are initially in a quiescent state when $t < 0$, taking Laplace transforms gives

$$(a) \quad Y(s) = \frac{1}{s^2 + 3s + 2} U(s) = G_1(s)U(s)$$

$$(b) \quad Y(s) = \frac{1}{s^2 + \omega^2} U(s) = G_2(s)U(s)$$

$$(c) \quad Y(s) = \frac{s+2}{s^2+s+1} U(s) = G_3(s)U(s)$$

In case (a) the poles of $G_1(s)$ are at $s = -1$ and $s = -2$, so the system is stable and the impulse response has a Fourier transform given by

$$\begin{aligned} G_1(j\omega) &= \left. \frac{1}{s^2 + 3s + 2} \right|_{s=j\omega} = \frac{1}{2 - \omega^2 + j3\omega} \\ &= \frac{2 - \omega^2 - j3\omega}{(2 - \omega^2)^2 + 9\omega^2} = \frac{(2 - \omega^2) - j3\omega}{\omega^4 + 5\omega^2 + 4} \end{aligned}$$

In case (b) we find that the poles of $G_2(s)$ are at $s = j\omega$ and $s = -j\omega$; that is, on the imaginary axis. The system is not stable (notice that the impulse response does not decay to zero), and the impulse response does not possess a Fourier transform.

In case (c) the poles of $G_3(s)$ are at $s = -\frac{1}{2} + j\frac{1}{2}\sqrt{3}$ and $s = -\frac{1}{2} - j\frac{1}{2}\sqrt{3}$. Since these are in the left half-plane, $\text{Re}(s) < 0$, we conclude that the system is stable. The Fourier transform of the impulse response is then

$$G_3(j\omega) = \frac{2 + j\omega}{1 - \omega^2 + j\omega}$$

8.4.2 The frequency response

For a linear time-invariant system, initially in a quiescent state, having a Laplace transfer function $G(s)$, the response $y(t)$ to an input $u(t)$ is given in (5.28) as

$$Y(s) = G(s)U(s) \tag{8.36}$$

where $Y(s)$ and $U(s)$ are the Laplace transforms of $y(t)$ and $u(t)$ respectively. In Section 5.5 we saw that, subject to the system being stable, the steady-state response $y_{ss}(t)$ to a sinusoidal input $u(t) = A \sin \omega t$ is given by (5.63) as

$$y_{ss}(t) = A |G(j\omega)| \sin [\omega t + \arg G(j\omega)] \tag{8.37}$$

That is, the steady-state response is also sinusoidal, with the same frequency as the input signal but having an amplitude gain $|G(j\omega)|$ and a phase shift $\arg G(j\omega)$.

More generally, we could have taken the input to be the complex sinusoidal signal

$$u(t) = A e^{j\omega t}$$

and, subject to the stability requirement, showed that the steady-state response is

$$y_{ss}(t) = A G(j\omega) e^{j\omega t} \tag{8.38}$$

or

$$y_{ss}(t) = A |G(j\omega)| e^{j[\omega t + \arg G(j\omega)]} \tag{8.39}$$

As before, $|G(j\omega)|$ and $\arg G(j\omega)$ are called the amplitude gain and phase shift respectively. Both are functions of the real frequency variable ω , and their plots versus ω constitute the **system frequency response**, which, as we saw in Section 5.5, characterizes the behaviour of the system. Note that taking imaginary parts throughout in (8.39) leads to the sinusoidal response (8.37).

We note that the steady-state response (8.38) is simply the input signal $Ae^{j\omega t}$ multiplied by the Fourier transform $G(j\omega)$ of the system's impulse response. Consequently $G(j\omega)$ is called the **frequency transfer function** of the system. Therefore if the system represented in (8.36) is stable, so that $G(j\omega)$ exists as the Fourier transform of its impulse response, and the input $u(t) = \mathcal{L}^{-1}\{U(s)\}$ has a Fourier transform $U(j\omega)$, then we may represent the system in terms of the frequency transfer function as

$$Y(j\omega) = G(j\omega)U(j\omega) \quad (8.40)$$

Equation (8.40) thus determines the Fourier transform of the system output, and can be used to determine the frequency spectrum of the output from that of the input. This means that both the amplitude and phase spectra of the output are available, since

$$|Y(j\omega)| = |G(j\omega)| |U(j\omega)| \quad (8.41a)$$

$$\arg Y(j\omega) = \arg G(j\omega) + \arg U(j\omega) \quad (8.41b)$$

We shall now consider an example that will draw together both these and some earlier ideas which serve to illustrate the relevance of this material in the communications industry.

Example 8.10

A signal $f(t)$ consists of two components:

- (a) a symmetric rectangular pulse of duration 2π (see Example 8.3) and
- (b) a second pulse, also of duration 2π (that is, a copy of (a)), modulating a signal with carrier frequency $\omega_0 = 3$ (the process of modulation was introduced in Section 8.3.4).

Write down an expression for $f(t)$ and illustrate its amplitude spectrum. Describe the amplitude spectrum of the output signal if $f(t)$ is applied to a stable causal system with a Laplace transfer function

$$G(s) = \frac{1}{s^2 + \sqrt{2}s + 1}$$

Solution

Denoting the pulse of Example 8.3, with $T = \pi$, by $P_\pi(t)$, and noting the use of the term 'carrier signal' in Example 8.7, we have

$$f(t) = P_\pi(t) + (\cos 3t)P_\pi(t)$$

From Example 8.3,

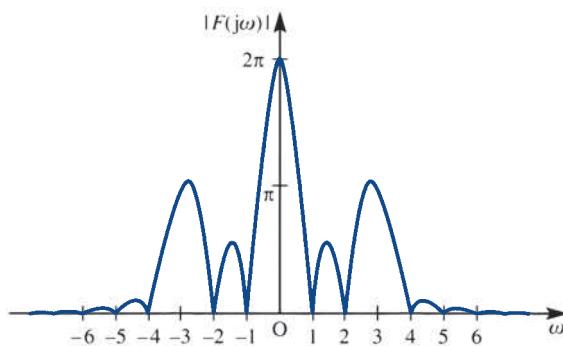
$$\mathcal{F}\{P_\pi(t)\} = 2\pi \operatorname{sinc} \omega\pi$$

so, using the result of Example 8.7, we have

$$\mathcal{F}\{f(t)\} = F(j\omega) = 2\pi \operatorname{sinc} \omega\pi + \frac{1}{2}[2\pi \operatorname{sinc}(\omega - 3)\pi + 2\pi \operatorname{sinc}(\omega + 3)\pi]$$

The corresponding amplitude spectrum obtained by plotting $|F(j\omega)|$ versus ω is illustrated in Figure 8.15.

Figure 8.15
Amplitude spectrum
of the signal of
Example 8.10.



Since the system with transfer function

$$G(s) = \frac{1}{s^2 + \sqrt{2}s + 1}$$

is stable and causal, it has a frequency transfer function

$$G(j\omega) = \frac{1}{1 - \omega^2 + j\sqrt{2}\omega}$$

so that its amplitude gain is

$$|G(j\omega)| = \frac{1}{\sqrt{(\omega^4 + 1)}}$$

The amplitude spectrum of the output signal $|Y(j\omega)|$ when the input is $f(t)$ is then obtained from (8.41a) as the product of $|F(j\omega)|$ and $|G(j\omega)|$. Plots of both the amplitude gain spectrum $|G(j\omega)|$ and the output amplitude spectrum $|Y(j\omega)|$ are shown in Figures 8.16(a) and (b) respectively. Note from Figure 8.16(b) that we have a reasonably good copy of the amplitude spectrum of $P_\pi(t)$ (see Figure 8.9 with $A = \pi$, $T = 1$). However, the second element of $f(t)$ has effectively vanished. Our system has ‘filtered out’ this latter component while ‘passing’ an almost intact version of the first. Examination of the time-domain response would show that the first component does in fact experience some ‘smoothing’, which, roughly speaking, consists of rounding of the sharp edges. The system considered here is a second-order ‘low-pass’ Butterworth filter (introduced in Section 6.10.1).

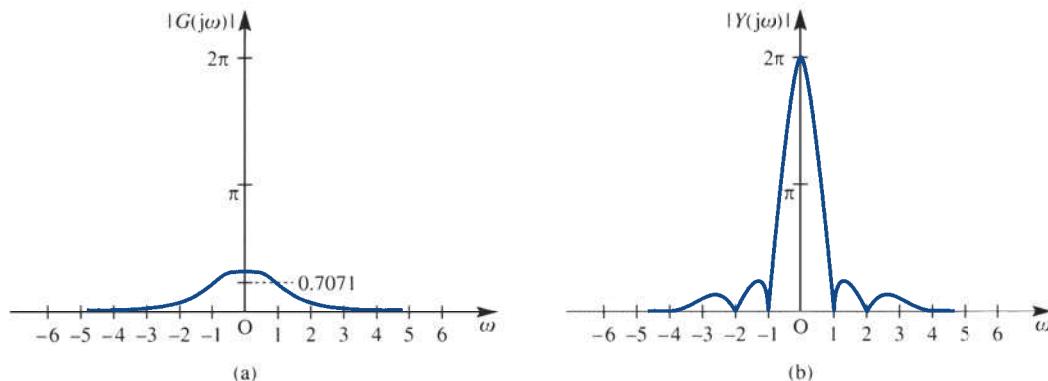


Figure 8.16 (a) Amplitude gain spectrum of the system with $G(s) = 1/(s^2 + \sqrt{2}s + 1)$; (b) amplitude spectrum of the output signal $|Y(j\omega)|$ of Example 8.10.

8.4.3 Exercises

- 17 Find the impulse response of systems (a) and (c) of Example 8.9. Calculate the Fourier transform of each using the definition (8.15), and verify the results given in Example 8.9.
- 18 Use the time-shift property to calculate the Fourier transform of the double rectangular pulse $f(t)$ illustrated in Figure 8.17.

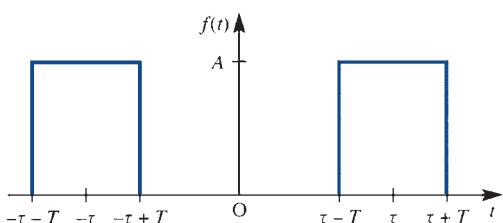


Figure 8.17 The double rectangular pulse of Exercise 18.

- 19 The system with transfer function

$$G(s) = \frac{1}{s^2 + \sqrt{2s+1}}$$

was discussed in Example 8.10. Make a transformation

$$s \rightarrow \frac{1}{s'}$$

and write down $G(s')$. Examine the frequency response of a system with transfer function $G(s')$ and in particular find the amplitude response when $\omega = 0$ and as $\omega \rightarrow \infty$. How would you describe such a system?

- 20 Use the symmetry property, and the result of Exercise 1, to calculate the Fourier transform of

$$f(t) = \frac{1}{a^2 + t^2}$$

Sketch $f(t)$ and its transform (which is real).

- 21 Using the results of Examples 8.3 and 8.7, calculate the Fourier transform of the pulse-modulated signal

$$f(t) = P_T(t) \cos \omega_0 t$$

where

$$P_T(t) = \begin{cases} 1 & (|t| \leq T) \\ 0 & (|t| > T) \end{cases}$$

is the pulse of duration $2T$.

8.5 Transforms of the step and impulse functions

In this section we consider the application of Fourier transforms to the concepts of energy, power and convolution. In so doing, we shall introduce the Fourier transform of the Heaviside unit step function $H(t)$ and the impulse function $\delta(t)$.

8.5.1 Energy and power

In Section 7.4.4 we introduced the concept of the power spectrum of a periodic signal and found that it enabled us to deduce useful information relating to the latter. In this section we define two quantities associated with time signals $f(t)$, defined for $-\infty < t < \infty$, namely signal energy and signal power. Not only are these important quantities in themselves, but, as we shall see, they play an important role in characterizing signal types.

The total **energy** associated with the signal $f(t)$ is defined as

$$E = \int_{-\infty}^{\infty} [f(t)]^2 dt \quad (8.42)$$

If $f(t)$ has a Fourier transform $F(j\omega)$, so that, from (8.16),

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega$$

then (8.42) may be expressed as

$$E = \int_{-\infty}^{\infty} f(t)f(t) dt = \int_{-\infty}^{\infty} f(t) \left[\frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega \right] dt$$

On changing the order of integration, this becomes

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) \left[\int_{-\infty}^{\infty} f(t) e^{j\omega t} dt \right] d\omega \quad (8.43)$$

From the defining integral (8.15) for $F(j\omega)$, we recognize the part of the integrand within the square brackets as $F(-j\omega)$, which, if $f(t)$ is real, is such that $F(-j\omega) = F^*(j\omega)$, where $F^*(j\omega)$ is the complex conjugate of $F(j\omega)$. Thus (8.43) becomes

$$E = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) F^*(j\omega) d\omega$$

so that

$$E = \int_{-\infty}^{\infty} [f(t)]^2 dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(j\omega)|^2 d\omega \quad (8.44)$$

Equation (8.44) relates the total energy of the signal $f(t)$ to the integral over all frequencies of $|F(j\omega)|^2$. For this reason, $|F(j\omega)|^2$ is called the **energy spectral density**, and a plot of $|F(j\omega)|^2$ versus ω is called the **energy spectrum** of the signal $f(t)$. The result (8.44) is called **Parseval's theorem**, and is an extension of the result contained in Theorem 7.2 for periodic signals.

Example 8.11

Determine the energy spectral densities of

- (a) the one-sided exponential function $f(t) = e^{-at} H(t)$ ($a > 0$),
- (b) the rectangular pulse of Figure 8.6.

Solution (a) From (8.17), the Fourier transform of $f(t)$ is

$$F(j\omega) = \frac{a - j\omega}{a^2 + \omega^2}$$

The energy spectral density of the function is therefore

$$|F(j\omega)|^2 = F(j\omega)F^*(j\omega) = \frac{a - j\omega}{a^2 + \omega^2} \frac{a + j\omega}{a^2 + \omega^2}$$

that is,

$$|F(j\omega)|^2 = \frac{1}{a^2 + \omega^2}$$

- (b) From Example 8.3, the Fourier transform $F(j\omega)$ of the rectangular pulse is

$$F(j\omega) = 2AT \operatorname{sinc} \omega T$$

Thus the energy spectral density of the pulse is

$$|F(j\omega)|^2 = 4A^2T^2 \operatorname{sinc}^2 \omega T$$

There are important signals $f(t)$, defined in general for $-\infty < t < \infty$, for which the integral $\int_{-\infty}^{\infty} [f(t)]^2 dt$ in (8.42) either is unbounded (that is, it becomes infinite) or does not converge to a finite limit; for example, $\sin t$. For such signals, instead of considering energy, we consider the average power P , frequently referred to as the **power** of the signal. This is defined by

$$P = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} [f(t)]^2 dt \quad (8.45)$$

Note that for signals that satisfy the Dirichlet conditions (Theorem 8.1) the integral in (8.42) exists and, since in (8.45) we divide by the signal duration, it follows that such signals have zero power associated with them.

We now pose the question: ‘Are there other signals which possess Fourier transforms?’ As you may expect, the answer is ‘Yes’, although the manner of obtaining the transforms will be different from our procedure so far. We shall see that the transforms so obtained, on using the inversion integral (8.16), yield some very ‘ordinary’ signals so far excluded from our discussion. We begin by considering the Fourier transform of the generalized function $\delta(t)$, the Dirac delta function introduced in Section 5.2.8. Recall from (5.10) that $\delta(t)$ satisfies the sifting property; that is, for a continuous function $g(t)$,

$$\int_a^b g(t) \delta(t - c) dt = \begin{cases} g(c) & (a < c < b) \\ 0 & \text{otherwise} \end{cases}$$

Using the defining integral (8.15), we readily obtain the following two Fourier transforms:

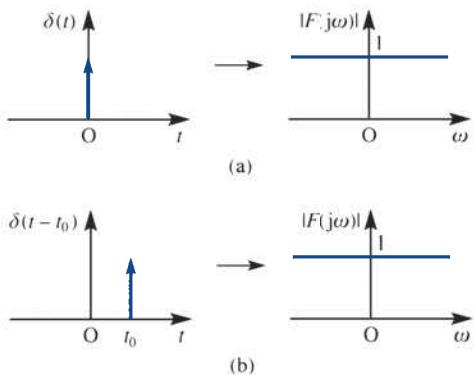
$$\mathcal{F}\{\delta(t)\} = \int_{-\infty}^{\infty} \delta(t) e^{-j\omega t} dt = 1 \quad (8.46)$$

$$\mathcal{F}\{\delta(t - t_0)\} = \int_{-\infty}^{\infty} \delta(t - t_0) e^{-j\omega t} dt = e^{-j\omega t_0} \quad (8.47)$$

These two transforms are, by now, unremarkable, and, noting that $|e^{-j\omega t_0}| = 1$, we illustrate the signals and their spectra in Figure 8.18.

Figure 8.18

(a) $\delta(t)$ and its amplitude spectrum;
 (b) $\delta(t - t_0)$ and its amplitude spectrum.



These results may be confirmed in MATLAB. Using the commands

```
syms w t
D=sym('dirac(t)');
F=fourier(D,t,w)
```

returns

```
F=1
```

in agreement with (8.46); whilst the commands

```
syms w t T
D1=sym('dirac(t-T)');
F1=fourier(D1,t,w)
```

return

```
F1=exp(-T*w*1i)
```

which confirms (8.47), with T replacing t_0 .

Likewise in MAPLE the commands

```
with(inttrans):
fourier(Dirac(t),t,w);
```

return the answer 1.

We now depart from the definition of the Fourier transform given in (8.15) and seek new transform pairs based on (8.46) and (8.47). Using the symmetry (duality) property of Section 8.3.5, we deduce from (8.46) that

$$1 \quad \text{and} \quad 2\pi\delta(-\omega) = 2\pi\delta(\omega) \quad (8.48)$$

is another Fourier transform pair. Likewise, from (8.47), we deduce that

$$e^{-j t_0 t} \quad \text{and} \quad 2\pi\delta(-\omega - t_0)$$

is also a Fourier transform pair. Substituting $t_0 = -\omega_0$ into the latter, we have

$$e^{j\omega_0 t} \quad \text{and} \quad 2\pi\delta(\omega_0 - \omega) = 2\pi\delta(\omega - \omega_0) \quad (8.49)$$

as another Fourier transform pair.

We are thus claiming that in (8.48) and (8.49) that $f_1(t) = 1$ and $f_2(t) = e^{j\omega_0 t}$, which do not have ‘ordinary’ Fourier transforms as defined by (8.15), actually do have ‘generalized’ Fourier transforms given by

$$F_1(j\omega) = 2\pi\delta(\omega) \quad (8.50)$$

$$F_2(j\omega) = 2\pi\delta(\omega - \omega_0) \quad (8.51)$$

respectively. These are the ones given in MATLAB in these cases.

The term ‘generalized’ has been used because the two transforms contain the generalized functions $\delta(\omega)$ and $\delta(\omega - \omega_0)$. Let us now test our conjecture that (8.50) and (8.51) are Fourier transforms of $f_1(t)$ and $f_2(t)$ respectively. If (8.50) and (8.51) really are Fourier transforms then their time-domain images $f_1(t)$ and $f_2(t)$ respectively should reappear via the inverse transform (8.16). Substituting $F_1(j\omega)$ from (8.50) into (8.16), we have

$$\mathcal{F}^{-1}\{F_1(j\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_1(j\omega) e^{j\omega t} d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} 2\pi\delta(\omega) e^{j\omega t} d\omega = 1$$

so $f_1(t) = 1$ is recovered.

Similarly, using (8.51), we have

$$\mathcal{F}^{-1}\{F_2(j\omega)\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} F_2(j\omega) e^{j\omega t} d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} 2\pi\delta(\omega - \omega_0) e^{j\omega t} d\omega = e^{j\omega_0 t}$$

so that $f_2(t) = e^{j\omega_0 t}$ is also recovered.

Our approach has therefore been successful, and we do indeed have a way of generating new pairs of transforms. We shall therefore use the approach to find generalized Fourier transforms for the signals

$$f_3(t) = \cos \omega_0 t, \quad f_4(t) = \sin \omega_0 t$$

Since

$$f_3(t) = \cos \omega_0 t = \frac{1}{2}(e^{j\omega_0 t} + e^{-j\omega_0 t})$$

the linearity property (8.22) gives

$$\mathcal{F}\{f_3(t)\} = \frac{1}{2}\mathcal{F}\{e^{j\omega_0 t}\} + \frac{1}{2}\mathcal{F}\{e^{-j\omega_0 t}\}$$

which, on using (8.49), leads to the generalized Fourier transform pair

$$\mathcal{F}\{\cos \omega_0 t\} = \pi[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)] \quad (8.52)$$

Likewise, we deduce the generalized Fourier transform pair

$$\mathcal{F}\{\sin \omega_0 t\} = j\pi[\delta(\omega + \omega_0) - \delta(\omega - \omega_0)] \quad (8.53)$$

The development of (8.53) and the verification that both (8.52) and (8.53) invert correctly using the inverse transform (8.16) is left as an exercise for the reader.

It is worth noting at this stage that defining the Fourier transform $\mathcal{F}\{f(t)\}$ of $f(t)$ in (8.15) as

$$\mathcal{F}\{f(t)\} = \int_{-\infty}^{\infty} f(t) e^{-j\omega t} dt$$

whenever the integral exists does not preclude the existence of other Fourier transforms, such as the generalized one just introduced, defined by other means.

It is clear that the total energy

$$E = \int_{-\infty}^{\infty} \cos^2 \omega_0 t dt$$

associated with the signal $f_3(t) = \cos \omega_0 t$ is unbounded. However, from (8.45), we can calculate the power associated with the signal as

$$P = \lim_{T \rightarrow \infty} \frac{1}{T} \int_{-T/2}^{T/2} \cos^2 \omega_0 t \, dt = \lim_{T \rightarrow \infty} \frac{1}{T} \left[t + \frac{1}{2\omega_0} \sin 2\omega_0 t \right]_{-T/2}^{T/2} = \frac{1}{2}$$

Thus, while the signal $f_3(t) = \cos \omega_0 t$ has unbounded energy associated with it, its power content is $\frac{1}{2}$. Signals whose associated energy is finite, for example $f(t) = e^{-at} H(t)$ ($a > 0$), are sometimes called **energy signals**, while those whose associated energy is unbounded but whose total power is finite are known as **power signals**. The concepts of power signals and power spectral density are important in the analysis of random signals, and the interested reader should consult specialized texts.

Example 8.12

Suppose that a periodic function $f(t)$, defined on $-\infty < t < \infty$, may be expanded in a Fourier series having exponential form

$$f(t) = \sum_{n=-\infty}^{\infty} F_n e^{jn\omega_0 t}$$

What is the (generalized) Fourier transform of $f(t)$?

Solution From the definition,

$$\mathcal{F}\{f(t)\} = \mathcal{F}\left\{ \sum_{n=-\infty}^{\infty} F_n e^{jn\omega_0 t} \right\} = \sum_{n=-\infty}^{\infty} F_n \mathcal{F}\{e^{jn\omega_0 t}\}$$

which, on using (8.49), gives

$$\mathcal{F}\{f(t)\} = \sum_{n=-\infty}^{\infty} F_n 2\pi \delta(\omega - n\omega_0)$$

That is,

$$\mathcal{F}\{f(t)\} = 2\pi \sum_{n=-\infty}^{\infty} F_n \delta(\omega - n\omega_0)$$

where F_n ($-\infty < n < \infty$) are the coefficients of the exponential form of the Fourier series representation of $f(t)$.

Example 8.13

Use the result of Example 8.12 to verify the Fourier transform of $f(t) = \cos \omega_0 t$ given in (8.52).

Solution Since

$$f(t) = \cos \omega_0 t = \frac{1}{2} e^{j\omega_0 t} + \frac{1}{2} e^{-j\omega_0 t}$$

the F_n of Example 8.12 are

$$F_{-1} = F_1 = \frac{1}{2}$$

$$F_n = 0 \quad (n \neq \pm 1)$$

Thus, using the result

$$\mathcal{F}\{f(t)\} = 2\pi \sum_{n=-\infty}^{\infty} F_n \delta(\omega - \omega_0)$$

we have

$$\begin{aligned}\mathcal{F}\{\cos \omega_0 t\} &= 2\pi F_{-1} \delta(\omega + \omega_0) + 2\pi F_1 \delta(\omega - \omega_0) \\ &= \pi[\delta(\omega + \omega_0) + \delta(\omega - \omega_0)]\end{aligned}$$

in agreement with (8.52).



Confirm this answer using the MATLAB commands

```
syms w t a
F=fourier(cos(a*t),t,w)
```

where a has been used to represent ω_0 .

Example 8.14

Determine the (generalized) Fourier transform of the periodic ‘sawtooth’ function, defined by

$$f(t) = \frac{2t}{T} \quad (0 < t < 2T)$$

$$f(t + 2T) = f(t)$$

Solution In Example 7.9 we saw that the exponential form of the Fourier series representation of $f(t)$ is

$$f(t) = \sum_{n=-\infty}^{\infty} F_n e^{jn\omega_0 t}$$

with

$$\omega_0 = \frac{2\pi}{2T} = \frac{\pi}{T}, F_0 = 2, F_n = \frac{j2}{n\pi} \quad (n \neq 0)$$

It follows from Example 8.12 that the Fourier transform $\mathcal{F}\{f(t)\}$ is

$$\begin{aligned}\mathcal{F}\{f(t)\} &= F(j\omega) = 4\pi\delta(\omega) + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} j\frac{4}{n} \delta(\omega - n\omega_0) \\ &= 4\pi\delta(\omega) + j4 \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{1}{n} \delta\left(\omega - \frac{n\pi}{T}\right)\end{aligned}$$

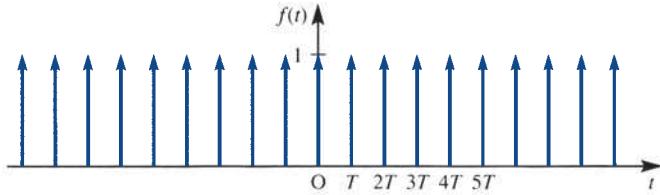
Thus we see that the amplitude spectrum simply consists of pulses located at integer multiples of the fundamental frequency $\omega_0 = \pi/T$. The discrete line spectra obtained via the exponential form of the Fourier series for this periodic function is thus reproduced, now with a scaling factor of 2π .

Example 8.15

Determine the (generalized) Fourier transform of the unit impulse train $f(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT)$, sometimes called the Shah function, shown symbolically in Figure 8.19.

Figure 8.19

Unit impulse train $f(t)$ of Example 8.15.

**Solution**

Although $f(t)$ is a generalized function, and not a function in the ordinary sense, it follows that since

$$\begin{aligned} f(t + kT) &= \sum_{n=-\infty}^{\infty} \delta(t + (k - n)T) \quad (k \text{ an integer}) \\ &= \sum_{m=-\infty}^{\infty} \delta(t - mT) \quad (m = n - k) \\ &= f(t) \end{aligned}$$

it is periodic, with period T . Moreover, we can formally expand $f(t)$ as a Fourier series

$$f(t) = \sum_{n=-\infty}^{\infty} F_n e^{jn\omega_0 t} \quad \left(\omega_0 = \frac{2\pi}{T} \right)$$

with

$$F_n = \frac{1}{T} \int_{-T/2}^{T/2} f(t) e^{-jn\omega_0 t} dt = \frac{1}{T} \int_{-T/2}^{T/2} \delta(t) e^{-jn\omega_0 t} dt = \frac{1}{T} \quad \text{for all } n$$

It follows from Example 8.12 that

$$\mathcal{F}\{f(t)\} = 2\pi \sum_{n=-\infty}^{\infty} \frac{1}{T} \delta(\omega - n\omega_0) = \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0)$$

Thus we have shown that

$$\mathcal{F}\left\{ \sum_{n=-\infty}^{\infty} \delta(t - nT) \right\} = \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \quad (8.54)$$

where $\omega_0 = 2\pi/T$. That is, the time-domain impulse train has another impulse train as its transform. We shall see in Section 8.6.4 that this result is of particular importance in dealing with sampled time signals.

Following our successful hunt for generalized Fourier transforms, we are led to consider the possibility that the Heaviside unit step function $H(t)$ defined in Section 5.2.1 may have a transform in this sense. Recall from (5.18) that if

$$f(t) = H(t)$$

then

$$\frac{df(t)}{dt} = \delta(t)$$

From the time-differentiation property (8.23), we might expect that if

$$\mathcal{F}\{H(t)\} = \bar{H}(j\omega)$$

then

$$(j\omega)\bar{H}(j\omega) = \mathcal{F}\{\delta(t)\} = 1 \quad (8.55)$$

Equation (8.55) suggests that a candidate for $\bar{H}(j\omega)$ might be $1/j\omega$, but this is not the case, since inversion using (8.16) does not give $H(t)$ back. Using (8.16) and complex variable techniques, it can be shown that

$$\mathcal{F}^{-1}\left\{\frac{1}{j\omega}\right\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{j\omega t}}{j\omega} d\omega = \begin{cases} \frac{1}{2} & (t > 0) \\ 0 & (t = 0) \\ -\frac{1}{2} & (t < 0) \end{cases} = \frac{1}{2} \operatorname{sgn}(t)$$

where $\operatorname{sgn}(t)$ is the **signum (or sign) function**, defined by

$$\operatorname{sgn}(t) = \begin{cases} 1 & (t > 0) \\ 0 & (t = 0) \\ -1 & (t < 0) \end{cases}$$



Note: This last result may be obtained in terms of Heaviside functions using the MATLAB commands

```
syms w t
f=ifourier(1/(i*w))
```

or using the MAPLE commands

```
with(inttrans):
invfourier(1(I*w),w,t);
```

However, we note that (8.55) is also satisfied by

$$\bar{H}(j\omega) = \frac{1}{j\omega} + c\delta(\omega) \quad (8.56)$$

where c is a constant. This follows from the equivalence property (see Definition 5.2, Section 5.2.11) $f(\omega)\delta(\omega) = f(0)\delta(\omega)$ with $f(\omega) = j\omega$, which gives

$$(j\omega)\bar{H}(j\omega) = 1 + (j\omega)c\delta(\omega) = 1$$

Inverting (8.56) using (8.16), we have

$$\begin{aligned} g(t) &= \mathcal{F}^{-1}\left\{\frac{1}{j\omega} + c\delta(\omega)\right\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[\frac{1}{j\omega} + c\delta(\omega) \right] e^{j\omega t} d\omega \\ &= \begin{cases} c/2\pi + \frac{1}{2} & (t > 0) \\ c/2\pi & (t = 0) \\ c/2\pi - \frac{1}{2} & (t < 0) \end{cases} \end{aligned}$$

and, choosing $c = \pi$, we have

$$g(t) = \begin{cases} 1 & (t > 0) \\ \frac{1}{2} & (t = 0) \\ 0 & (t < 0) \end{cases}$$

Thus we have (almost) recovered the step function $H(t)$. Here $g(t)$ takes the value $\frac{1}{2}$ at $t = 0$, but this is not surprising in view of the convergence of the Fourier integral at points of discontinuity as given in Theorem 8.1. With this proviso, we have shown that

$$\bar{H}(j\omega) = \mathcal{F}\{H(t)\} = \frac{1}{j\omega} + \pi\delta(\omega) \quad (8.57)$$

We must confess to having made an informed guess as to what additional term to add in (8.56) to produce the Fourier transform (8.57). We could instead have chosen $c\delta(k\omega)$ with k a constant as an additional term. While it is possible to show that this would not lead to a different result, proving uniqueness is not trivial and is beyond the scope of this book.



Using the MATLAB commands

```
syms w t
H=sym('heaviside(t)');
F=fourier(H,t,w)
```

returns

```
F=pi*dirac(w)-1i/w
```

which, noting that $-i = 1/i$, confirms result (8.57).

Likewise the MATLAB commands

```
syms w t T
H=sym('heaviside(t-T)');
F=fourier(H,t,w)
```

return

```
F=exp(-T*w*1i)*(pi*dirac(w)-1i/w)
```

which gives us another Fourier transform

$$\mathcal{F}\{H(t-T)\} = e^{-j\omega T}(\pi\delta(\omega) + 1/j\omega)$$

8.5.2 Convolution

In Section 5.3.6 we saw that the convolution integral, in conjunction with the Laplace transform, provided a useful tool for *discussing* the nature of the solution of a differential equation, although it was not perhaps the most efficient way of evaluating the solution to a particular problem. As the reader may now have come to expect, in view of the duality between time and frequency domains, there are two convolution results involving the Fourier transform.

Convolution in time

Suppose that

$$\mathcal{F}\{u(t)\} = U(j\omega) = \int_{-\infty}^{\infty} u(t) e^{-j\omega t} dt$$

$$\mathcal{F}\{v(t)\} = V(j\omega) = \int_{-\infty}^{\infty} v(t) e^{-j\omega t} dt$$

then the Fourier transform of the convolution

$$y(t) = \int_{-\infty}^{\infty} u(\tau)v(t - \tau) d\tau = u(t) * v(t) \quad (8.58)$$

is

$$\begin{aligned} \mathcal{F}\{y(t)\} &= Y(j\omega) = \int_{-\infty}^{\infty} e^{-j\omega t} \left[\int_{-\infty}^{\infty} u(\tau)v(t - \tau) d\tau \right] dt \\ &= \int_{-\infty}^{\infty} u(\tau) \left[\int_{-\infty}^{\infty} e^{-j\omega t} v(t - \tau) dt \right] d\tau \end{aligned}$$

Introducing the change of variables $z \rightarrow t - \tau$, $\tau \rightarrow \tau$ and following the procedure for change of variable from Section 5.3.6, the transform can be expressed as

$$\begin{aligned} Y(j\omega) &= \int_{-\infty}^{\infty} u(\tau) \left[\int_{-\infty}^{\infty} v(z) e^{-j\omega(z+\tau)} dz \right] d\tau \\ &= \int_{-\infty}^{\infty} u(\tau) e^{-j\omega\tau} d\tau \int_{-\infty}^{\infty} v(z) e^{-j\omega z} dz \end{aligned}$$

so that

$$Y(j\omega) = U(j\omega)V(j\omega) \quad (8.59)$$

That is,

$$\mathcal{F}\{u(t) * v(t)\} = \mathcal{F}\{v(t) * u(t)\} = U(j\omega)V(j\omega) \quad (8.60)$$

indicating that a convolution in the time domain is transformed into a product in the frequency domain.

Convolution in frequency

If

$$\mathcal{F}\{u(t)\} = U(j\omega), \quad \text{with } u(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} U(j\omega) e^{j\omega t} d\omega$$

$$\mathcal{F}\{v(t)\} = V(j\omega), \quad \text{with } v(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} V(j\omega) e^{j\omega t} d\omega$$

then the inverse transform of the convolution

$$U(j\omega) * V(j\omega) = \int_{-\infty}^{\infty} U(jy)V(j(\omega - y)) dy$$

is given by

$$\begin{aligned}\mathcal{F}^{-1}\{U(j\omega) * V(j\omega)\} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{j\omega t} \left[\int_{-\infty}^{\infty} U(jy)V(j(\omega-y)) dy \right] d\omega \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} U(jy) \left[\int_{-\infty}^{\infty} V(j(\omega-y)) e^{j\omega t} d\omega \right] dy\end{aligned}$$

A change of variable $z \rightarrow \omega - y$, $\omega \rightarrow \omega$ leads to

$$\begin{aligned}\mathcal{F}^{-1}\{U(j\omega) * V(j\omega)\} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} U(jy) \left[\int_{-\infty}^{\infty} V(jz) e^{j(z+y)t} dz \right] dy \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} U(jy) e^{jyt} dy \int_{-\infty}^{\infty} V(jz) e^{jzt} dz \\ &= 2\pi u(t)v(t)\end{aligned}$$

That is,

$$\mathcal{F}\{u(t)v(t)\} = \frac{1}{2\pi} U(j\omega) * V(j\omega) \quad (8.61)$$

and thus multiplication in the time domain corresponds to convolution in the frequency domain (subject to the scaling factor $1/(2\pi)$).

8.5.3 Exercises

22 Verify that $\mathcal{F}^{-1}\{\pi[\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]\} = \cos \omega_0 t$.



$$\int_{-\infty}^{\infty} f(t)g(t) dt = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega)G(-j\omega) d\omega$$

23 Show that $\mathcal{F}\{\sin \omega_0 t\} = j\pi[\delta(\omega + \omega_0) - \delta(\omega - \omega_0)]$. Use (8.16) to verify that



$$\mathcal{F}^{-1}\{j\pi[\delta(\omega + \omega_0) - \delta(\omega - \omega_0)]\} = \sin \omega_0 t$$

24 Suppose that $f(t)$ and $g(t)$ have Fourier transforms $F(j\omega)$ and $G(j\omega)$ respectively, defined in the ‘ordinary’ sense (that is, using (8.15)), and show that

$$\int_{-\infty}^{\infty} f(t)G(jt) dt = \int_{-\infty}^{\infty} F(jt)g(t) dt$$

This result is known as **Parseval’s formula**.

25 Use the results of Exercise 24 and the symmetry property to show that

26 Use the convolution result in the frequency domain to obtain $\mathcal{F}\{H(t) \sin \omega_0 t\}$.

27 Calculate the exponential form of the Fourier series for the periodic pulse train shown in Figure 8.20. Hence show that

$$\mathcal{F}\{f(t)\} = \frac{2\pi A d}{T} \sum_{n=-\infty}^{\infty} \text{sinc}\left(\frac{n\pi d}{T}\right) \delta(\omega - n\omega_0)$$

($\omega_0 = 2\pi/T$), and A is the height of the pulse.

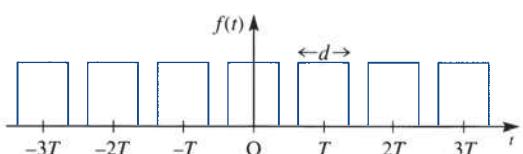


Figure 8.20 Periodic pulse train of Exercise 27.

8.6 The Fourier transform in discrete time

8.6.1 Introduction

The earlier sections of this chapter have discussed the Fourier transform of signals defined as functions of the continuous-time variable t . We have seen that a major area of application is in the analysis of signals in the frequency domain, leading to the concept of the frequency response of a linear system. In Chapter 7 we considered signals defined at discrete-time instants, together with linear systems modelled by difference equations. There we found that in system analysis the z transform plays a role similar to that of the Laplace transform for continuous-time systems. We now attempt to develop a theory of Fourier analysis to complement that for continuous-time systems, and then consider the problem of estimating the continuous-time Fourier transform in a form suitable for computer execution.

8.6.2 A Fourier transform for sequences

First we return to our work on Fourier series and write down the exponential form of the Fourier series representation for the periodic function $F(e^{j\theta})$ of period 2π . Writing $\theta = \omega t$, we infer from (7.39) that

$$F(e^{j\theta}) = \sum_{n=-\infty}^{\infty} f_n e^{jn\theta} \quad (8.62)$$

where

$$f_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(e^{j\theta}) e^{-jn\theta} d\theta \quad (8.63)$$

Thus the operation has generated a sequence of numbers $\{f_n\}$ from the periodic function $F(e^{j\theta})$ of the continuous variable θ . Let us reverse the process and imagine that we start with a sequence $\{g_k\}$ and use (8.62) to define a periodic function $\tilde{G}'(e^{j\theta})$ such that

$$\tilde{G}'(e^{j\theta}) = \sum_{n=-\infty}^{\infty} g_n e^{jn\theta} \quad (8.64)$$

We have thus defined a transformation from the sequence $\{g_k\}$ to $\tilde{G}'(e^{j\theta})$. This transformation can be inverted, since, from (8.63),

$$g_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{G}'(e^{j\theta}) e^{-jk\theta} d\theta \quad (8.65)$$

and we recover the terms of the sequence $\{g_k\}$ from $\tilde{G}'(e^{j\theta})$.

It is convenient for our later work if we modify the definition slightly, defining the Fourier transform of the sequence $\{g_k\}$ as

$$\mathcal{F}\{g_k\} = G(e^{j\theta}) = \sum_{n=-\infty}^{\infty} g_n e^{-jn\theta} \quad (8.66)$$

whenever the series converges. The inverse transform is then given from (8.65), by

$$g_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(e^{j\theta}) e^{jk\theta} d\theta \quad (8.67)$$

The results (8.66) and (8.67) thus constitute the Fourier transform pair for the sequence $\{g_k\}$. Note that $G(e^{j\theta})$ is a function of the continuous variable θ , and since it is a function of $e^{j\theta}$ it is periodic (with a period of at most 2π), irrespective of whether or not the sequence $\{g_k\}$ is periodic.

Note that we have adopted the notation $G(e^{j\theta})$ rather than $G(\theta)$ for the Fourier transform, similar to our use of $F(j\omega)$ rather than $F(\omega)$ in the case of continuous-time signals. In the present case we shall be concerned with the relationship with the z transform of Chapter 6, where $z = r e^{j\theta}$, and the significance of our choice will soon emerge.

Example 8.16

Find the transform of the sequence $\{g_k\}_{-\infty}^{\infty}$, where $g_0 = 2$, $g_2 = g_{-2} = 1$ and $g_k = g_{-k} = 0$ for $k \neq 0$ or 2.

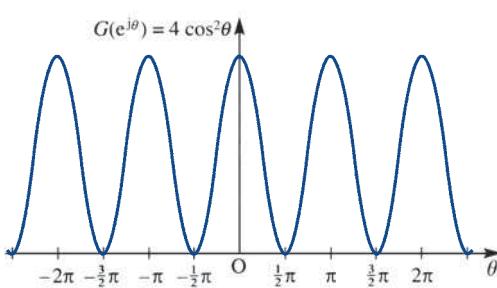
Solution

From the definition (8.66),

$$\begin{aligned} \mathcal{F}\{g_k\} = G(e^{j\theta}) &= \sum_{n=-\infty}^{\infty} g_n e^{-jn\theta} \\ &= g_{-2} e^{j2\theta} + g_0 1 + g_2 e^{-j2\theta} = e^{j2\theta} + 2 + e^{-j2\theta} \\ &= 2(1 + \cos 2\theta) = 4 \cos^2 \theta \end{aligned}$$

In this particular case the transform is periodic of period π , rather than 2π . This is because $g_1 = g_{-1} = 0$, so that $\cos \theta$ does not appear in the transform. Since $G(e^{j\theta})$ is purely real, we may plot the transform as in Figure 8.21.

Figure 8.21
Transform of the sequence of Example 8.16.



Having defined a Fourier transform for sequences, we now wish to link it to the frequency response of discrete-time systems. In Section 8.4.2 the link between frequency responses and the Fourier transforms of continuous-time systems was established using the Laplace transform. We suspect therefore that the z transform should yield the necessary link for discrete-time systems. Indeed, the argument follows closely that of Section 8.4.2.

For a causal linear time-invariant discrete-time system with z transfer function $G(z)$ the relationship between the input sequence $\{u_k\}$ and output sequence $\{y_k\}$ in the transform domain is given from Section 6.6.1 by

$$Y(z) = G(z)U(z) \quad (8.68)$$

where $U(z) = \mathcal{Z}\{u_k\}$ and $Y(z) = \mathcal{Z}\{y_k\}$.

To investigate the system frequency response, we seek the output sequence corresponding to an input sequence

$$\{u_k\} = \{Ae^{j\omega kT}\} = \{Ae^{jk\theta}\}, \quad \theta = \omega T \quad (8.69)$$

which represents samples drawn, at equal intervals T , from the continuous-time complex sinusoidal signal $e^{j\omega t}$.

The frequency response of the discrete-time system is then its steady-state response to the sequence $\{u_k\}$ given in (8.69). As for the continuous-time case (Section 8.4.2), the complex form $e^{j\omega t}$ is used in order to simplify the algebra, and the steady-state sinusoidal response is easily recovered by taking imaginary parts, if necessary.

From Figure 6.3, we saw that

$$\mathcal{Z}\{Ae^{jk\theta}\} = \mathcal{Z}\{A(e^{j\theta})^k\} = \frac{Az}{z - e^{j\theta}}$$

so, from (8.68), the response of the system to the input sequence (8.69) is determined by

$$Y(z) = G(z) \frac{Az}{z - e^{j\theta}} \quad (8.70)$$

Taking the system to be of order n , and under the assumption that the n poles p_r ($r = 1, 2, \dots, n$) of $G(z)$ are distinct and none is equal to $e^{j\theta}$, we can expand $Y(z)/z$ in terms of partial fractions to give

$$\frac{Y(z)}{z} = \frac{c}{z - e^{j\theta}} + \sum_{r=1}^n \frac{c_r}{z - p_r} \quad (8.71)$$

where, in general, the constants c_r ($r = 1, 2, \dots, n$) are complex. Taking inverse z transforms throughout in (8.71) then gives the response sequence as

$$\{y_k\} = \mathcal{Z}^{-1}\{\frac{Y(z)}{z}\} = \mathcal{Z}^{-1}\left\{\frac{zc}{z - e^{j\theta}}\right\} + \sum_{r=1}^n \mathcal{Z}^{-1}\left\{\frac{zc_r}{z - p_r}\right\}$$

that is,

$$\{y_k\} = c\{e^{jk\theta}\} + \sum_{r=1}^n c_r\{p_r^k\} \quad (8.72)$$

If the transfer function $G(z)$ corresponds to a stable discrete-time system then all its poles p_r ($r = 1, 2, \dots, n$) lie within the unit circle $|z| < 1$, so that all the terms under the summation sign in (8.72) tend to zero as $k \rightarrow \infty$. This is clearly seen by expressing p_r in the form $p_r = |p_r|e^{j\phi_r}$ and noting that if $|p_r| < 1$ then $|p_r|^k \rightarrow 0$ as $k \rightarrow \infty$. Consequently, for stable systems the steady-state response corresponding to (8.72) is

$$\{y_{k_{ss}}\} = c\{e^{jk\theta}\}$$

Using the ‘cover-up’ rule for partial fractions, the constant c is readily determined from (8.70) as

$$c = AG(e^{j\theta})$$

so that the steady-state response becomes

$$\{y_{k_{ss}}\} = AG(e^{j\theta})\{e^{jk\theta}\} \quad (8.73)$$

We have assumed that the poles of $G(z)$ are distinct in order to simplify the algebra. Extending the development to accommodate multiple poles is readily accomplished, leading to the same steady-state response as given in (8.73).

The result (8.73) corresponds to (8.38) for continuous-time systems, and indicates that the steady-state response sequence is simply the input sequence with each term multiplied by $G(e^{j\theta})$. Consequently $G(e^{j\theta})$ is called the **frequency transfer function** of the discrete-time system and, as for the continuous case, it characterizes the system's frequency response. Clearly $G(e^{j\theta})$ is simply $G(z)$, the z transfer function, with $z = e^{j\theta}$, and so we are simply evaluating the z transfer function around the unit circle $|z| = 1$. The z transfer function $G(z)$ will exist on $|z| = 1$ if and only if the system is stable, and thus the result is the exact analogue of the result for continuous-time systems in Section 8.4.2, where the Laplace transfer function was evaluated along the imaginary axis to yield the frequency response of a stable linear continuous-time system.

To complete the analogy with continuous-time systems, we need one further result. From Section 6.6.2, the impulse response of the linear causal discrete-time system with z transfer function $G(z)$ is

$$\{y_{k_\delta}\} = \mathcal{Z}^{-1}\{G(z)\} = \{g_k\}_{k=0}^{\infty}, \text{ say}$$

Taking inverse transforms then gives

$$G(z) = \sum_{k=0}^{\infty} g_k z^{-k} = \sum_{k=-\infty}^{\infty} g_k z^{-k}$$

since $g_k = 0$ ($k < 0$) for a causal system. Thus

$$G(e^{j\theta}) = \sum_{k=-\infty}^{\infty} g_k e^{-jk\theta}$$

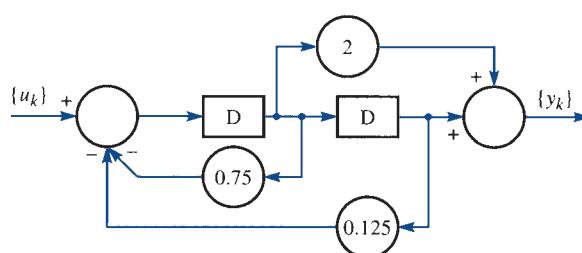
and we conclude from (8.66) that $G(e^{j\theta})$ is simply the Fourier transform of the sequence $\{g_k\}$. Therefore the discrete-time frequency transfer function $G(e^{j\theta})$ is the Fourier transform of the impulse response sequence.

Example 8.17

Determine the frequency transfer function of the causal discrete-time system shown in Figure 8.22 and plot its amplitude spectrum.

Figure 8.22

Discrete-time system of Example 8.17.



Solution Using the methods of Section 6.6.1, we readily obtain the z transfer function as

$$G(z) = \frac{2z + 1}{z^2 + 0.75z + 0.125}$$

Next we check for system stability. Since $z^2 + 0.75z + 0.125 = (z + 0.5)(z + 0.25)$, the poles of $G(z)$ are at $p_1 = -0.5$ and $p_2 = -0.25$, and since both are inside the unit circle $|z| = 1$, the system is stable. The frequency transfer function may then be obtained as $G(e^{j\theta})$, where

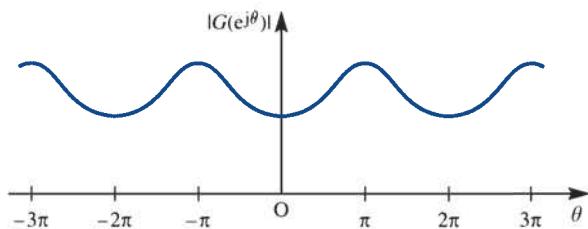
$$G(e^{j\theta}) = \frac{2 e^{j\theta} + 1}{e^{j2\theta} + 0.75 e^{j\theta} + 0.125}$$

To determine the amplitude spectrum, we evaluate $|G(e^{j\theta})|$ as

$$\begin{aligned}|G(e^{j\theta})| &= \frac{|2 e^{j\theta} + 1|}{|e^{j2\theta} + 0.75 e^{j\theta} + 0.125|} \\&= \frac{\sqrt{(5 + 4 \cos \theta)}}{\sqrt{(1.578 + 1.688 \cos \theta + 0.25 \cos 2\theta)}}\end{aligned}$$

A plot of $|G(e^{j\theta})|$ versus θ then leads to the amplitude spectrum of Figure 8.23.

Figure 8.23
Amplitude spectrum
of the system of
Example 8.17.



In Example 8.17 we note the periodic behaviour of the amplitude spectrum, which is inescapable when discrete-time signals and systems are concerned. Note, however, that the periodicity is in the variable $\theta = \omega T$ and that we may have control over the choice of T , the time between samples of our input signal.

8.6.3 The discrete Fourier transform

The Fourier transform of sequences discussed in Section 8.6.2 transforms a sequence $\{g_k\}$ into a continuous function $G(e^{j\theta})$ of a frequency variable θ , where $\theta = \omega T$ and T is the time between signal samples. In this section, with an eye to computer requirements, we look at the implications of sampling $G(e^{j\theta})$. The overall operation will have commenced with samples of a time signal $\{g_k\}$ and proceeded via a Fourier transformation process, finally producing a sequence $\{G_k\}$ of samples drawn from the frequency-domain image $G(e^{j\theta})$ of $\{g_k\}$.

Suppose that we have a sequence $\{g_k\}$ of N samples drawn from a continuous-time signal $g(t)$, at equal intervals T ; that is,

$$\{g_k\} = \{g(kT)\}_{k=0}^{N-1}$$

Using (8.66), the Fourier transform of this sequence is

$$\mathcal{F}\{g_k\} = G(e^{j\theta}) = \sum_{n=-\infty}^{\infty} g_n e^{-jn\theta} \quad (8.74)$$

where $g_k = 0$ ($k \notin [0, N-1]$). Then, with $\theta = \omega T$, we may write (8.74) as

$$G(e^{j\omega T}) = \sum_{n=0}^{N-1} g_n e^{-jn\omega T} \quad (8.75)$$

We now sample this transform $G(e^{j\omega T})$ at intervals $\Delta\omega$ in such a way as to create N samples spread equally over the interval $0 \leq \theta \leq 2\pi$; that is, over one period of the essentially periodic function $G(e^{j\theta})$. We then have

$$N\Delta\theta = 2\pi$$

where $\Delta\theta$ is the normalized frequency spacing. Since $\theta = \omega T$ and T is a constant such that $\Delta\theta = T\Delta\omega$, we deduce that

$$\Delta\omega = \frac{2\pi}{NT} \quad (8.76)$$

Sampling (8.75) at intervals $\Delta\omega$ produces the sequence

$$\{G_k\}_{k=0}^{N-1}, \quad \text{where} \quad G_k = \sum_{n=0}^{N-1} g_n e^{-jn k \Delta\omega T} \quad (8.77)$$

Since

$$\begin{aligned} G_{k+N} &= \sum_{n=0}^{N-1} g_n e^{-jn(k+N)\Delta\omega T} \\ &= \sum_{n=0}^{N-1} g_n e^{-jn k \Delta\omega T} e^{-jn 2\pi}, \quad \text{using (8.76)} \\ &= \sum_{n=0}^{N-1} g_n e^{-jn k \Delta\omega T} = G_k \end{aligned}$$

it follows that the sequence $\{G_k\}_{-\infty}^{\infty}$ is periodic, with period N . We have therefore generated a sequence of samples in the frequency domain that in some sense represents the spectrum of the underlying continuous-time signal. We shall postpone the question of the exact nature of this representation for the moment, but as the reader will have guessed, it is crucial to the purpose of this section. First, we consider the question of whether, from knowledge of the sequence $\{G_k\}_{k=0}^{N-1}$ of (8.77), we can recover the original sequence $\{g_n\}_{n=0}^{N-1}$. To see how this can be achieved, consider a sum of the form

$$S_r = \sum_{k=0}^{N-1} G_k e^{-jkr\Delta\omega T}, \quad (N-1) \leq r \leq 0 \quad (8.78)$$

Substituting for G_k from (8.77), we have

$$S_r = \sum_{k=0}^{N-1} \left(\sum_{m=0}^{N-1} g_m e^{-jm k \Delta\omega T} \right) e^{-jkr\Delta\omega T} = \sum_{k=0}^{N-1} \sum_{m=0}^{N-1} g_m e^{-jk\Delta\omega(m+r)T}$$

That is, on interchanging the order of integration,

$$S_r = \sum_{m=0}^{N-1} g_m \sum_{k=0}^{N-1} e^{-jk\Delta\omega(m+r)T} \quad (8.79)$$

Now

$$\sum_{k=0}^{N-1} e^{-jk\Delta\omega(m+r)T}$$

is a geometric progression with first term $e^0 = 1$ and common ratio $e^{-j\Delta\omega(m+r)T}$, and so the sum to N terms is thus

$$\sum_{k=0}^{N-1} e^{-jk\Delta\omega(m+r)T} = \frac{1 - e^{-j\Delta\omega(m+r)NT}}{1 - e^{-j\Delta\omega(m+r)T}} = \frac{1 - e^{-j(m+r)2\pi}}{1 - e^{-j\Delta\omega(m+r)T}} = 0 \quad (m \neq -r + nN)$$

When $m = -r$

$$\sum_{k=0}^{N-1} e^{-jk\Delta\omega(m+r)T} = \sum_{k=0}^{N-1} 1 = N$$

Thus

$$\sum_{k=0}^{N-1} e^{-jk\Delta\omega(m+r)T} = N\delta_{m,-r} \quad (8.80)$$

where δ_{ij} is the Kronecker delta defined by

$$\delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

Substituting (8.80) into (8.79), we have

$$S_r = N \sum_{m=0}^{N-1} g_m \delta_{m,-r} = Ng_{-r}$$

Returning to (8.78) and substituting for S_r we see that

$$g_{-r} = \frac{1}{N} \sum_{k=0}^{N-1} G_k e^{-jkr\Delta\omega T}$$

which on taking $n = -r$ gives

$$g_n = \frac{1}{N} \sum_{k=0}^{N-1} G_k e^{jkn\Delta\omega T} \quad (8.81)$$

Thus (8.81) allows us to determine the members of the sequence

$$\{g_n\}_{n=0}^{N-1}$$

that is, it enables us to recover the time-domain samples from the frequency-domain samples *exactly*.

The relations

$$G_k = \sum_{n=0}^{N-1} g_n e^{-jn k \Delta \omega T} \quad (8.77)$$

$$g_n = \frac{1}{N} \sum_{k=0}^{N-1} G_k e^{jn k \Delta \omega T} \quad (8.81)$$

with $\Delta \omega = 2\pi/NT$, between the time- and frequency-domain sequences $\{g_n\}_{n=0}^{N-1}$ and $\{G_k\}_{k=0}^{N-1}$ define the **discrete Fourier transform (DFT)** pair. The pair provide pathways between time and frequency domains for discrete-time signals in exactly the same sense that (8.15) and (8.16) defined similar pathways for continuous-time signals. It should be stressed again that, whatever the properties of the sequences $\{g_n\}$ and $\{G_k\}$ on the right-hand sides of (8.77) and (8.81), the sequences generated on the left-hand sides will be periodic, with period N .

Example 8.18

The sequence $\{g_k\}_{k=0}^2 = \{1, 2, 1\}$ is generated by sampling a time signal $g(t)$ at intervals with $T = 1$. Determine the discrete Fourier transform of the sequence, and verify that the sequence can be recovered exactly from its transform.

Solution From (8.77), the discrete Fourier transform sequence $\{G_k\}_{k=0}^2$ is generated by

$$G_k = \sum_{n=0}^2 g_n e^{-jn k \Delta \omega T} \quad (k = 0, 1, 2)$$

In this case $T = 1$ and, with $N = 3$, (8.76) gives

$$\Delta \omega = \frac{2\pi}{3 \times 1} = \frac{2}{3}\pi$$

Thus

$$G_0 = \sum_{n=0}^2 g_n e^{-jn \times 0 \times 2\pi/3} = \sum_{n=0}^2 g_n = g_0 + g_1 + g_2 = 1 + 2 + 1 = 4$$

$$\begin{aligned} G_1 &= \sum_{n=0}^2 g_n e^{-jn \times 1 \times 2\pi/3} = g_0 e^0 + g_1 e^{-j2\pi/3} + g_2 e^{-j4\pi/3} = 1 + 2 e^{-j2\pi/3} + 1 e^{-j4\pi/3} \\ &= e^{-j2\pi/3} (e^{j2\pi/3} + 2 + e^{-j2\pi/3}) = 2 e^{-j2\pi/3} (1 + \cos \frac{2}{3}\pi) = e^{-j2\pi/3} \end{aligned}$$

$$\begin{aligned} G_2 &= \sum_{n=0}^2 g_n e^{-jn \times 2 \times 2\pi/3} = \sum_{n=0}^2 g_n e^{-jn4\pi/3} = g_0 e^0 + g_1 e^{-j4\pi/3} + g_2 e^{-j8\pi/3} \\ &= e^{-j4\pi/3} [e^{j4\pi/3} + 2 + e^{-j4\pi/3}] = 2 e^{-j4\pi/3} (1 + \cos \frac{4}{3}\pi) = e^{-j4\pi/3} \end{aligned}$$

Thus

$$\{G_k\}_{k=0}^2 = \{4, e^{-j2\pi/3}, e^{-j4\pi/3}\}$$

We must now show that use of (8.81) will recover the original sequence $\{g_k\}_{k=0}^2$. From (8.81), the inverse transform of $\{G_k\}_{k=0}^2$ is given by

$$\tilde{g}_n = \frac{1}{N} \sum_{k=0}^{N-1} G_k e^{jkn\Delta\omega T}$$

again with $T = 1$, $\Delta\omega = \frac{2}{3}\pi$ and $N = 3$. Thus

$$\begin{aligned}\tilde{g}_0 &= \frac{1}{3} \sum_{k=0}^2 G_k e^{jk \times 0 \times 2\pi/3} = \frac{1}{3} \sum_{k=0}^2 G_k = \frac{1}{3} (4 + e^{-j2\pi/3} + e^{-j4\pi/3}) \\ &= \frac{1}{3} [4 + e^{-j\pi}(e^{j\pi/3} + e^{-j\pi/3})] = \frac{1}{3} (4 - 2 \cos \frac{1}{3}\pi) = 1 \\ \tilde{g}_1 &= \frac{1}{3} \sum_{k=0}^2 G_k e^{jk \times 1 \times 2\pi/3} = \frac{1}{3} (G_0 + G_1 e^{j2\pi/3} + G_2 e^{j4\pi/3}) \\ &= \frac{1}{3} (4 + 1 + 1) = 2 \\ \tilde{g}_2 &= \frac{1}{3} \sum_{k=0}^2 G_k e^{jk \times 2 \times 2\pi/3} = \frac{1}{3} (G_0 + G_1 e^{j4\pi/3} + G_2 e^{j8\pi/3}) \\ &= \frac{1}{3} [4 + e^{j\pi}(e^{j\pi/3} + e^{-j\pi/3})] = \frac{1}{3} (4 - 2 \cos \frac{1}{3}\pi) = 1\end{aligned}$$

That is

$$\{\tilde{g}_n\}_{n=0}^2 = \{1, 2, 1\} = \{g_k\}_{k=0}^2$$

and thus the original sequence has been recovered exactly from its transform.

We see from Example 8.18 that the operation of calculating N terms of the transformed sequence involved $N \times N = N^2$ multiplications and $N(N - 1)$ summations, all of which are operations involving complex numbers in general. The computation of the discrete Fourier transform in this direct manner is thus said to be a computation of complexity N^2 . Such computations rapidly become impossible as N increases, owing to the time required for this execution.

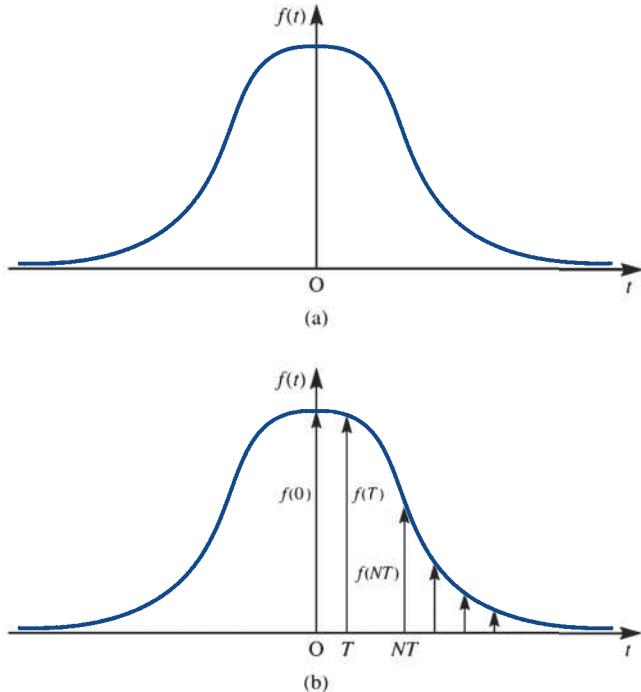
8.6.4 Estimation of the continuous Fourier transform

We saw in Section 8.4.2 that the continuous Fourier transform provides a means of examining the frequency response of a stable linear time-invariant *continuous-time* system. Similarly, we saw in Section 8.6.2 how a discrete-time Fourier transform could be developed that allows examination of the frequency response of a stable linear time-invariant *discrete-time* system. By sampling this latter transform, we developed the discrete Fourier transform itself. Why did we do this? First we have found a way (at least in theory) of involving the computer in our efforts. Secondly, as we shall now show, we can use the discrete Fourier transform to estimate the continuous Fourier transform of a continuous-time signal. To see how this is done, let us first examine what happens when we sample a continuous-time signal.

Suppose that $f(t)$ is a non-periodic continuous-time signal, a portion of which is shown in Figure 8.24(a). Let us sample the signal at equal intervals T , to generate the sequence

Figure 8.24

(a) Continuous-time signal $f(t)$;
 (b) samples drawn from $f(t)$.



$$\{f(0), f(T), \dots, f(nT), \dots\}$$

as shown in Figure 8.24(b). Imagine now that each of these samples is presented in turn, at the appropriate instant, as the input to a continuous linear time-invariant system with impulse response $h(t)$. The output would then be, from Section 5.3.6,

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} h(t - \tau) f(0) \delta(\tau) d\tau + \int_{-\infty}^{\infty} h(t - \tau) f(\tau) \delta(\tau - T) d\tau \\ &\quad + \dots + \int_{-\infty}^{\infty} h(t - \tau) f(nT) \delta(\tau - nT) d\tau + \dots \\ &= \int_{-\infty}^{\infty} h(t - \tau) \sum_{k=0}^{\infty} f(kT) \delta(\tau - kT) d\tau \end{aligned}$$

Thus

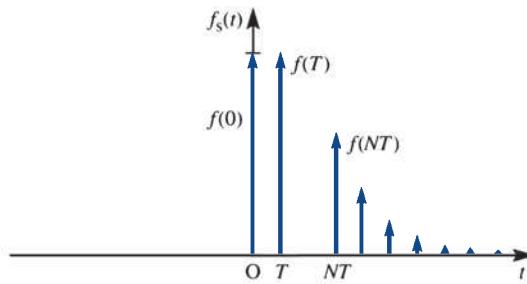
$$y(t) = \int_{-\infty}^{\infty} h(t - \tau) f_s(\tau) d\tau \tag{8.82}$$

where

$$f_s(t) = \sum_{k=0}^{\infty} f(kT) \delta(t - kT) = f(t) \sum_{k=0}^{\infty} \delta(t - kT) \tag{8.83}$$

which we identify as a ‘continuous-time’ representation of the sampled version of $f(t)$. We are thus led to picture $f_s(t)$ as in Figure 8.25.

Figure 8.25
Visualization of $f_s(t)$ defined in (8.83).



In order to admit the possibility of signals that are non-zero for $t < 0$, we can generalize (8.83) slightly by allowing in general that

$$f_s(t) = f(t) \sum_{k=-\infty}^{\infty} \delta(t - kT) \quad (8.84)$$

We can now use convolution to find the Fourier transform $F_s(j\omega)$ of $f_s(t)$. Using the representation (8.84) for $f_s(t)$, we have

$$F_s(j\omega) = \mathcal{F}\{f_s(t)\} = \mathcal{F}\left\{f(t) \sum_{k=-\infty}^{\infty} \delta(t - kT)\right\}$$

which, on using (8.61), leads to

$$F_s(j\omega) = \frac{1}{2\pi} F(j\omega) * \mathcal{F}\left\{\sum_{k=-\infty}^{\infty} \delta(t - kT)\right\} \quad (8.85)$$

where

$$\mathcal{F}\{f(t)\} = F(j\omega)$$

From (8.54),

$$\mathcal{F}\left\{\sum_{k=-\infty}^{\infty} \delta(t - kT)\right\} = \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi k}{T}\right)$$

so that, assuming the interchange of the order of integration and summation to be possible, (8.85) becomes

$$\begin{aligned} F_s(j\omega) &= \frac{1}{2\pi} F(j\omega) * \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi k}{T}\right) \\ &= \frac{1}{T} \int_{-\infty}^{\infty} F(j[\omega - \omega']) \sum_{k=-\infty}^{\infty} \delta\left(\omega' - \frac{2\pi k}{T}\right) d\omega' \\ &= \frac{1}{T} \sum_{k=-\infty}^{\infty} \int_{-\infty}^{\infty} F(j[\omega - \omega']) \delta\left(\omega' - \frac{2\pi k}{T}\right) d\omega' \\ &= \frac{1}{T} \sum_{k=-\infty}^{\infty} F\left(j\left(\omega - \frac{2\pi k}{T}\right)\right) \end{aligned}$$

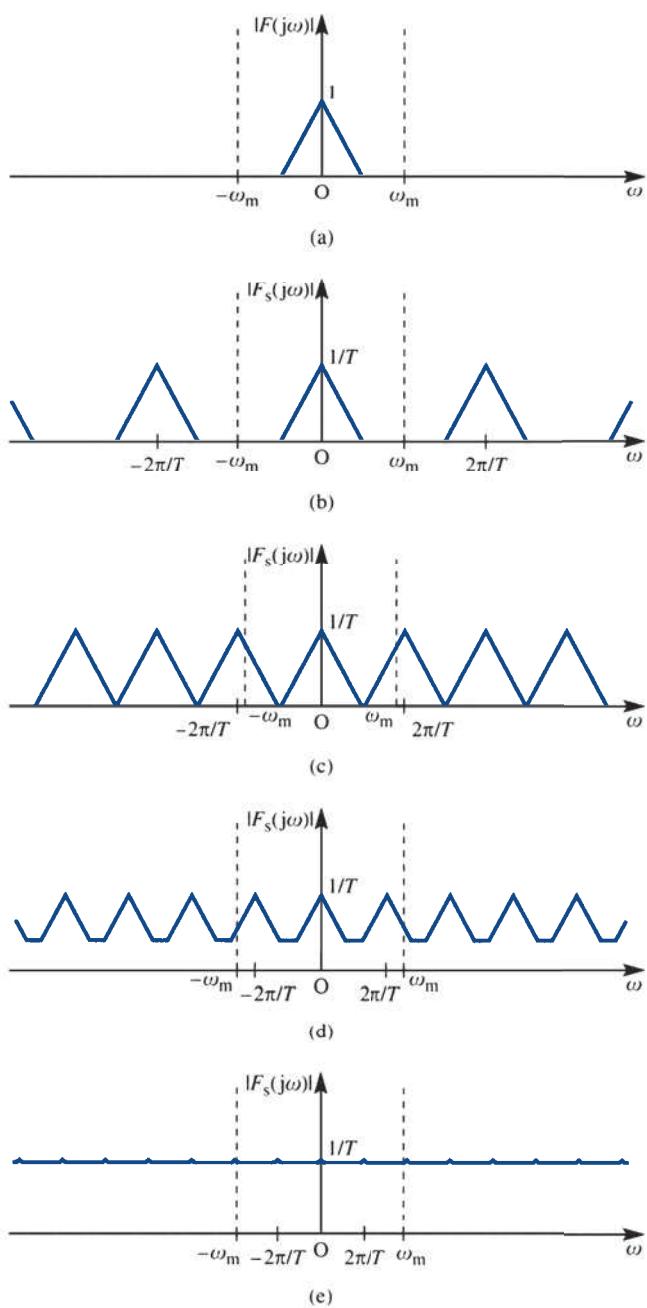
Thus

$$F_s(j\omega) = \frac{1}{T} \sum_{k=-\infty}^{\infty} F(j[\omega - k\omega_0]), \quad \omega_0 = \frac{2\pi}{T} \quad (8.86)$$

Examining (8.86), we see that the spectrum $F_s(j\omega)$ of the sampled version $f_s(t)$ of $f(t)$ consists of repeats of the spectrum $F(j\omega)$ of $f(t)$ scaled by a factor $1/T$, these repeats being spaced at intervals $\omega_0 = 2\pi/T$ apart. Figure 8.26(a) shows the amplitude spectrum

Figure 8.26

(a) Amplitude spectrum of a band-limited signal $f(t)$; (b)–(e) amplitude spectrum $|F_s(j\omega)|$ of $f_s(t)$, showing periodic repetition of $|F_s(j\omega)|$ and interaction effects as T increases.



$|F(j\omega)|$ of a band-limited signal $f(t)$; that is, a signal whose spectrum is zero for $|\omega| > \omega_m$. Figures 8.26(b–e) show the amplitude spectrum $|F_s(j\omega)|$ of the sampled version for increasing values of the sampling interval T . Clearly, as T increases, the spectrum of $F(j\omega)$, as observed using $|F_s(j\omega)|$ in $-\omega_m < \omega < \omega_m$, becomes more and more misleading because of ‘interaction’ from neighbouring copies.

As we saw in Section 8.6.2, the periodicity in the amplitude spectrum $|F_s(j\omega)|$ of $f_s(t)$ is inevitable as a consequence of the sampling process, and ways have to be found to minimize the problems it causes. The interaction observed in Figure 8.26 between the periodic repeats is known as **aliasing error**, and it is clearly essential to minimize this effect. This can be achieved in an obvious way if the original unsampled signal $f(t)$ is band-limited as in Figure 8.26(a). It is apparent that we must arrange that the periodic repeats of $|F(j\omega)|$ be far enough apart to prevent interaction between the copies. This implies that we have

$$\omega_0 \geq 2\omega_m$$

at an absolute (and impractical!) minimum. Since $\omega_0 = 2\pi/T$, the constraint implies that

$$T \leq \pi/\omega_m$$

where T is the interval between samples. The minimum time interval allowed is

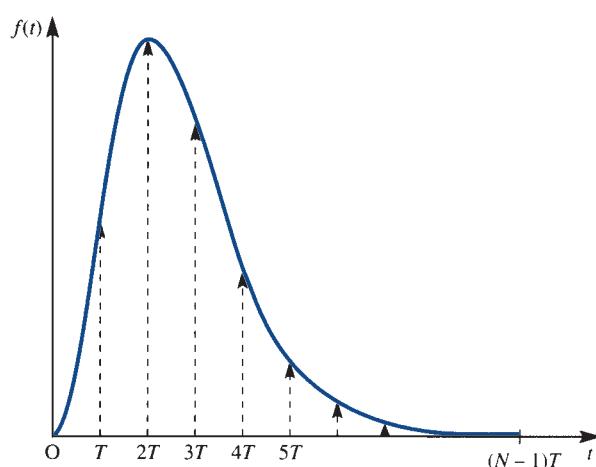
$$T_{\min} = \pi/\omega_m$$

which is known as the **Nyquist interval** and we have in fact deduced a form of the **Nyquist–Shannon sampling theorem**. If $T < T_{\min}$ then the ‘copies’ of $F(j\omega)$ are isolated from each other, and we can focus on just one copy, either for the purpose of signal reconstruction, or for the purposes of the estimation of $F(j\omega)$ itself. Here we are concerned only with the latter problem. Basically, we have established a condition under which the spectrum of the samples of the band-limited signal $f(t)$, that is the spectrum of $f_s(t)$, can be used to estimate $F(j\omega)$.

Suppose we have drawn N samples from a continuous signal $f(t)$ at intervals T , in accordance with the Nyquist criterion, as in Figure 8.27. We then consider

$$f_s(t) = \sum_{k=0}^{N-1} f(kT) \delta(t - kT)$$

Figure 8.27
Sampling of a
continuous-time signal.



or equivalently, the sequence

$$\{f_k\}_{k=0}^{N-1}, \quad \text{where } f_k = f(kT)$$

Note that

$$f_s(t) = 0 \quad (t > (N-1)T)$$

so that

$$f_k = 0 \quad (k > N-1)$$

The Fourier transform of $f_s(t)$ is

$$\begin{aligned} F_s(j\omega) &= \int_{-\infty}^{\infty} f_s(t) e^{-j\omega t} dt = \int_{-\infty}^{\infty} \sum_{k=0}^{N-1} f(kT) \delta(t - kT) e^{-j\omega t} dt \\ &= \sum_{k=0}^{N-1} \int_{-\infty}^{\infty} f(kT) \delta(t - kT) e^{-j\omega t} dt \\ &= \sum_{k=0}^{N-1} f(kT) e^{-j\omega kT} = \sum_{k=0}^{N-1} f_k e^{-j\omega kT} \end{aligned} \quad (8.87)$$

The transform in (8.87) is a function of the continuous variable ω so, as in (8.77), we must now sample the continuous spectrum $F_s(j\omega)$ to permit computer evaluation.

We chose N samples to represent $f(t)$ in the time domain, and for this reason we also choose N samples in the frequency domain to represent $F(j\omega)$. Thus we sample (8.87) at intervals $\Delta\omega$ to generate the sequence

$$\{F_s(jn\Delta\omega)\}_{n=0}^{N-1} \quad (8.88a)$$

where

$$F_s(jn\Delta\omega) = \sum_{k=0}^{N-1} f_k e^{-jkn\Delta\omega T} \quad (8.88b)$$

We must now choose the frequency-domain sampling interval $\Delta\omega$. To see how to do this, recall that the sampled spectrum $F_s(j\omega)$ consisted of repeats of $F(j\omega)$, spaced at intervals $2\pi/T$ apart. Thus to sample just one copy in its entirety, we should choose

$$N\Delta\omega = 2\pi/T$$

or

$$\Delta\omega = 2\pi/NT \quad (8.89)$$

Note that the resulting sequence, defined outside $0 \leq n \leq N-1$, is periodic, as we should expect. However, note also that, following our discussion in Section 8.6, the process of recovering a time signal from samples of its spectrum will result in a periodic waveform, whatever the nature of the original time signal. We should not be surprised by this, since it is exactly in accordance with our introductory discussion in Section 8.1. In view of the scaling factor $1/T$ in (8.86), our estimate of the Fourier transform $F(j\omega)$ of $f(t)$ over the interval

$$0 \leq t \leq (N-1)T$$

will, from (8.88), be the sequence of samples

$$\{TF_s(jn\Delta\omega)\}_{n=0}^{N-1}$$

where

$$TF_s(jn\Delta\omega) = T \sum_{k=0}^{N-1} f_k e^{-jkn\Delta\omega T}$$

which, from the definition of the discrete Fourier transform in (8.77), gives

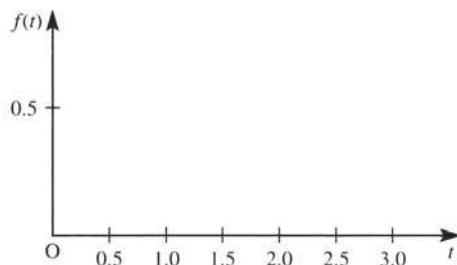
$$TF_s(jn\Delta\omega) = T \times \text{DFT } \{f_k\}$$

where DFT $\{f_k\}$ is the discrete Fourier transform of the sequence $\{f_k\}$. We illustrate the use of this estimate in Example 8.19.

Example 8.19

The delayed triangular pulse $f(t)$ is as illustrated in Figure 8.28. Estimate its Fourier transform using 10 samples and compare with the exact values.

Figure 8.28
The delayed triangular pulse.



Solution Using $N = 10$ samples at intervals $T = 0.2$ s, we generate the sequence

$$\{f_k\}_{k=0}^9 = \{f(0), f(0.2), f(0.4), f(0.6), f(0.8), f(1.0), f(1.2), f(1.4), f(1.6), f(1.8)\}$$

Clearly, from Figure 8.28, we can express the continuous function $f(t)$ as

$$f(t) = \begin{cases} t & (0 \leq t \leq 0.5) \\ 1 - t & (0.5 < t < 1) \\ 0 & (t \geq 1) \end{cases}$$

and so

$$\{f_k\}_{k=0}^9 = \{0, 0.2, 0.4, 0.4, 0.2, 0, 0, 0, 0, 0\}$$

Using (8.77), the discrete Fourier transform $\{F_n\}_{n=0}^9$ of the sequence $\{f_k\}_{k=0}^9$ is generated by

$$F_n = \sum_{k=0}^9 f_k e^{-jkn\Delta\omega T}, \quad \text{where} \quad \Delta\omega = \frac{2\pi}{NT} = \frac{2\pi}{10 \times 0.2} = \pi$$

That is,

$$F_n = \sum_{k=0}^9 f_k e^{-jk\pi(0.2\pi)}$$

or, since $f_0 = f_5 = f_6 = f_7 = f_8 = f_9 = 0$,

$$F_n = \sum_{k=1}^4 f_k e^{-jk\pi(0.2\pi)}$$

The estimate of the Fourier transform, also based on $N = 10$ samples, is then the sequence

$$\{TF_n\}_{n=0}^9 = \{0.2F_n\}_{n=0}^9$$

We thus have 10 values representing the Fourier transform at

$$\omega = n\Delta\omega \quad (n = 0, 1, 2, \dots, 9)$$

or since $\Delta\omega = 2\pi/NT$

$$\omega = 0, \pi, 2\pi, \dots, 9\pi$$

At $\omega = \pi$, corresponding to $n = 1$, our estimate is

$$\begin{aligned} 0.2F_1 &= 0.2 \sum_{k=1}^4 f_k e^{-jk\pi(0.2\pi)} \\ &= 0.2[0.2e^{-j(0.2\pi)} + 0.4(e^{-j(0.4\pi)} + e^{-j(0.6\pi)}) + 0.2e^{-j(0.8\pi)}] \\ &= -0.1992j \end{aligned}$$

At $\omega = 2\pi$, corresponding to $n = 2$, our estimate is

$$\begin{aligned} 0.2F_2 &= 0.2 \sum_{k=1}^4 f_k e^{-jk\pi(0.4\pi)} \\ &= 0.2[0.2e^{-j(0.4\pi)} + 0.4(e^{-j(0.8\pi)} + e^{-j(1.2\pi)}) + 0.2e^{-j(1.6\pi)}] \\ &= -0.1047 \end{aligned}$$

Continuing in this manner, we compute the sequence

$$\{0.2F_0, 0.2F_1, \dots, 0.2F_n\}$$

as

$$\{0.2400, -0.1992j, -0.1047, 0.0180j, -0.0153, 0, -0.0153, -0.0180j, -0.1047, 0.1992j\}$$

This then represents the estimate of the Fourier transform of the continuous function $f(t)$. The exact value of the Fourier transform of $f(t)$ is easily computed by direct use of the definition (8.15) as

$$F(j\omega) = \mathcal{F}\{f(t)\} = \frac{1}{4} e^{-j\omega/2} \operatorname{sinc}^2 \frac{1}{4} \omega$$

which we can use to examine the validity of our result. The comparison is shown in Figure 8.29 and illustrated graphically in Figure 8.30.

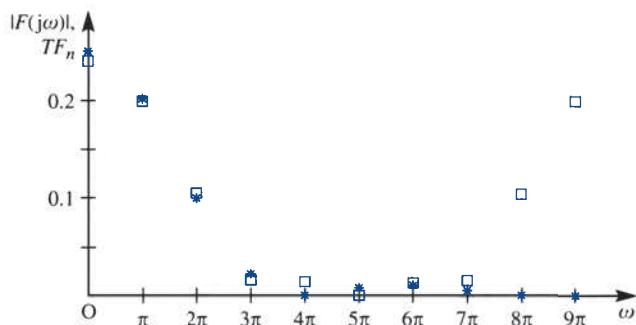
Figure 8.29

Comparison of exact results and DFT estimate for the amplitude spectrum of the signal of Example 8.19.

ω	Exact $F(j\omega)$	DFT estimate	$ F(j\omega) $	$ DFT \text{ estimate} $	% error
0	0.2500	0.2400	0.2500	0.2400	4%
π	-0.2026j	-0.1992j	0.2026	0.1992	1.7%
2π	-0.1013	-0.1047	0.1013	0.1047	3.2%
3π	0.0225j	0.0180j	0.0225	0.0180	20%
4π	0	-0.0153	0	0.0153	—
5π	-0.0081j	0	0.0081	0	—
6π	-0.0113	-0.0153	0.0113	0.0153	—
7π	0.0041	-0.0180j	0.0041	0.0180	—
8π	0	-0.1047	0	0.1047	—
9π	-0.0025j	0.1992j	0.0025	0.1992	—

Figure 8.30

Exact result $|F(j\omega)|$ (*) and DFT estimate TF_n of the Fourier transform in Example 8.19.



From the Nyquist–Shannon sampling theorem, with $T = 0.2$ s, we deduce that our results will be completely accurate if the original signal $f(t)$ is band-limited with a zero spectrum for $|\omega| > |\omega_m| = 5\pi$. Our signal is not strictly band-limited in this way, and we thus expect to observe some error in our results, particularly near $\omega = 5\pi$, because of the effects of aliasing. The estimate obtained is satisfactory at $\omega = 0, \pi, 2\pi$, but begins to lose accuracy at $\omega = 3\pi$. Results obtained above $\omega = 5\pi$ are seen to be images of those obtained for values below $\omega = 5\pi$, and this is to be expected owing to the periodicity of the DFT. In our calculation the DFT sequence will be periodic, with period $N = 10$; thus, for example,

$$|TF_7| = |TF_{7-10}| = |TF_{-3}| = T|F_{-3}|$$

As we have seen many times, for a real signal the amplitude spectrum is symmetric about $\omega = 0$. Thus $|F_{-3}| = |F_3|$, $|F_{-5}| = |F_5|$, and so on, and the effects of the symmetry are apparent in Figure 8.29. It is perhaps worth observing that if we had calculated (say) $\{TF_{-4}, TF_{-3}, \dots, TF_0, TF_1, \dots, TF_5\}$, we should have obtained a ‘conventional’ plot, with the right-hand portion, beyond $\omega = 5\pi$, translated to the left of the origin. However, using the plot of the amplitude spectrum in the chosen form does highlight the source of error due to aliasing.

In this section we have discussed a method by which Fourier transforms can be estimated numerically, at least in theory. It is apparent, though, that the amount of labour involved is significant, and as we observed in Section 8.6.3 an algorithm based on this approach is in general prohibitive in view of the amount of computing time required. The next section gives a brief introduction to a method of overcoming this problem.

8.6.5 The fast Fourier transform

The calculation of a discrete Fourier transform based on N sample values requires, as we have seen, N^2 complex multiplications and $N(N - 1)$ summations. For real signals, symmetry can be exploited, but for large N , $\frac{1}{2}N^2$ does not represent a significant improvement over N^2 for the purposes of computation. In fact, a totally new approach to the problem was required before the discrete Fourier transform could become a practical engineering tool. In 1965 Cooley and Tukey introduced the **fast Fourier transform** to compute the DFT and its inverse, and to compute the FFT in order to reduce the computational complexity (J. W. Cooley and J. W. Tukey, An algorithm for the machine computation of complex Fourier series, *Mathematics of Computation* **19** (1965) 297–301). We shall briefly introduce their approach in this section: for a full discussion see E. E. Brigham, *The Fast Fourier Transform* (Englewood Cliffs, NJ, Prentice Hall, 1974), whose treatment is similar to that adopted here.

Note that FFTs are routinely performed in electrical engineering since processing a signal in the frequency space is sometimes more advantageous than in its natural setting. There is also a large commonality between statistics, FFTs and Laplace transforms in industries involving signal processing that often now results in machine learning approaches. Such approaches are used to make market forecasts and predictions wherever there is a time series, e.g. nonlinear ‘Wave’ patterns over some time interval. See, for example, F. Camstra and A. Vinciarelli, *Machine Learning for Audio Usage and Video Analysis* (second edition, London, Springer, 2015).

We shall restrict ourselves to the situation where $N = 2^\gamma$ for some integer γ , and, rather than examine the general case, we shall focus on a particular value of γ . In proceeding in this way, the idea should be clear and the extension to other values of γ appear credible. We can summarize the approach as being in three stages:

- (1) matrix formulation;
- (2) matrix factorization; and, finally,
- (3) rearranging.

Stage 1: We first consider a matrix formulation of the DFT. From (8.77), the Fourier transform sequence $\{G_k\}_{k=0}^{N-1}$ of the sequence $\{g_n\}_{n=0}^{N-1}$ is generated by

$$G_k = \sum_{n=0}^{N-1} g_n e^{-j2\pi nk/N} \quad (k = 0, 1, \dots, N-1) \quad (8.90)$$

We shall consider the particular case when $\gamma = 2$ (that is, $N = 2^2 = 4$), and define

$$W = e^{-j2\pi/N} = e^{-j\pi/2}$$

so that (8.90) becomes

$$G_k = \sum_{n=0}^{N-1} g_n W^{nk} = \sum_{n=0}^3 g_n W^{nk} \quad (k = 0, 1, 2, 3)$$

Writing out the terms of the transformed sequence, we have

$$G_0 = g_0 W^0 + g_1 W^0 + g_2 W^0 + g_3 W^0$$

$$G_1 = g_0 W^0 + g_1 W^1 + g_2 W^2 + g_3 W^3$$

$$G_2 = g_0 W^0 + g_1 W^2 + g_2 W^4 + g_3 W^6$$

$$G_3 = g_0 W^0 + g_1 W^3 + g_2 W^6 + g_3 W^9$$

which may be expressed in the vector-matrix form

$$\begin{bmatrix} G_0 \\ G_1 \\ G_2 \\ G_3 \end{bmatrix} = \begin{bmatrix} W^0 & W^0 & W^0 & W^0 \\ W^0 & W^1 & W^2 & W^3 \\ W^0 & W^2 & W^4 & W^6 \\ W^0 & W^3 & W^6 & W^9 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} \quad (8.91)$$

or, more generally, as

$$G_k = \mathbf{W}^{nk} \mathbf{g}_n$$

where the vectors \mathbf{G}_k and \mathbf{g}_n and the square matrix \mathbf{W}^{nk} are defined as in (8.91). The next step relates to the special properties of the entries in the matrix \mathbf{W}^{nk} . Note that $W^{nk} = W^{nk+pN}$, where p is an integer, and so

$$W^4 = W^0 = 1$$

$$W^6 = W^2$$

$$W^9 = W^1$$

Thus (8.91) becomes

$$\begin{bmatrix} G_0 \\ G_1 \\ G_2 \\ G_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & W^1 & W^2 & W^3 \\ 1 & W^2 & W^0 & W^2 \\ 1 & W^3 & W^2 & W^1 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} \quad (8.92)$$

Equation (8.92) is the end of the first stage of the development. In fact, we have so far only made use of the properties of the N th roots of unity. Stage two involves the factorization of a matrix, the details of which will be explained later.

Stage 2: We begin by noting that

$$\begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & W^2 & W^0 & W^2 \\ 1 & W^1 & W^2 & W^3 \\ 1 & W^3 & W^2 & W^1 \end{bmatrix} \quad (8.93)$$

where we have used $W^5 = W^1$ and $W^0 = 1$ (in the top row). The matrix on the right-hand side of (8.93) is the coefficient matrix of (8.92), *but with rows 2 and 3 interchanged*. Thus we can write (8.92) as

$$\begin{bmatrix} G_0 \\ G_2 \\ G_1 \\ G_3 \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} \quad (8.94)$$

We now define a vector \mathbf{g}' as

$$\mathbf{g}' = \begin{bmatrix} g'_0 \\ g'_1 \\ g'_2 \\ g'_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix} \quad (8.95)$$

It then follows from (8.95) that

$$\begin{aligned} g'_0 &= g_0 + W^0 g_2 \\ g'_1 &= g_1 + W^0 g_3 \end{aligned}$$

so that g'_0 and g'_1 are each calculated by one complex multiplication and one addition. Of course, in this special case, since $W^0 = 1$, the multiplication is unnecessary, but we are attempting to infer the general situation. For this reason, W^0 has not been replaced by 1.

Also, it follows from (8.95) that

$$\begin{aligned} g'_2 &= g_0 + W^2 g_2 \\ g'_3 &= g_1 + W^2 g_3 \end{aligned}$$

and, since $W^2 = -W^0$, the computation of the pair g'_2 and g'_3 can make use of the computations of $W^0 g_2$ and $W^0 g_3$, with one further addition in each case. Thus the vector \mathbf{g}' is determined by a total of four complex additions and two complex multiplications.

Stage 3: To complete the calculation of the transform, we return to (8.94), and rewrite it in the form

$$\begin{bmatrix} G_0 \\ G_2 \\ G_1 \\ G_3 \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} g'_0 \\ g'_1 \\ g'_2 \\ g'_3 \end{bmatrix} \quad (8.96)$$

It then follows from (8.96) that

$$\begin{aligned} G_0 &= g'_0 + W^0 g'_1 \\ G_2 &= g'_0 + W^2 g'_1 \end{aligned}$$

and we see that G_0 is determined by one complex multiplication and one complex addition. Furthermore, because $W^2 = -W^0$, G_2 follows after one further complex addition.

Similarly, it follows from (8.96) that

$$\begin{aligned} G_1 &= g'_2 + W^1 g'_3 \\ G_3 &= g'_2 + W^3 g'_3 \end{aligned}$$

and, since $W^3 = -W^1$, a total of one further complex multiplication and two further additions are required to produce the re-ordered transform vector

$$[G_0 \ G_2 \ G_1 \ G_3]^T$$

Thus the total number of operations required to generate the (re-ordered) transform is four complex multiplications and eight complex additions. Direct calculation would have required $N^2 = 16$ complex multiplications and $N(N - 1) = 12$ complex additions. Even with a small value of N , these savings are significant, and, interpreting computing time

requirements as being proportional to the number of complex multiplications involved, it is easy to see why the FFT algorithm has become an essential tool for computational Fourier analysis. When $N = 2^r$, the FFT algorithm is effectively a procedure for producing $\gamma N \times N$ matrices of the form (8.93). Extending our ideas, it is possible to see that generally the FFT algorithm, when $N = 2^r$, will require $\frac{1}{2}N\gamma$ (four, when $N = 2^2 = 4$) complex multiplications and $N\gamma$ (eight, when $N = 4$) complex additions. Since

$$\gamma = \log_2 N$$

the demands of the FFT algorithm in terms of computing time, estimated on the basis of the number of complex multiplications, are often given as about $N \log_2 N$, as opposed to N^2 for the direct evaluation of the transform. This completes the second stage of our task, and we are only left with the problem of rearrangement of our transform vector into ‘natural’ order.

The means by which this is achieved is most elegant. Instead of indexing G_0, G_1, G_2, G_3 in decimal form, an alternative binary notation is used, and $[G_0 \quad G_1 \quad G_2 \quad G_3]^T$ becomes

$$[G_{00} \quad G_{01} \quad G_{10} \quad G_{11}]^T$$

The process of ‘bit reversal’ means rewriting a binary number with its bits or digits in reverse order. Applying this process to $[G_{00} \quad G_{01} \quad G_{10} \quad G_{11}]^T$ yields

$$[G_{00} \quad G_{10} \quad G_{01} \quad G_{11}]^T = [G_0 \quad G_2 \quad G_1 \quad G_3]^T$$

with decimal labelling. This latter form is exactly the one obtained at the end of the FFT calculation, and we see that the natural order can easily be recovered by rearranging the output on the basis of bit reversal of the binary indexed version.

We have now completed our introduction to the fast Fourier transform. We shall now consider an example to illustrate the ideas discussed here. We shall then conclude by considering in greater detail the matrix factorization process used in the second stage.

Example 8.20

Use the method of the FFT algorithm to compute the Fourier transform of the sequence

$$\{g_n\}_{n=0}^3 = \{1, 2, 1, 0\}$$

Solution In this case $N = 4 = 2^2$, and we begin by computing the vector $\mathbf{g}'_n = [g'_0 \quad g'_1 \quad g'_2 \quad g'_3]^T$, which, from (8.95), is given by

$$\mathbf{g}'_n = \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ g_3 \end{bmatrix}$$

For $N = 4$

$$W^n = (e^{-j2\pi/4})^n = e^{-jn\pi/2}$$

and so

$$\mathbf{g}'_n = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 2 \end{bmatrix}$$

Next, we compute the ‘bit-reversed’ order transform vector \mathbf{G}' , say, which from (8.96) is given by

$$\mathbf{G}' = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} g'_0 \\ g'_1 \\ g'_2 \\ g'_3 \end{bmatrix}$$

or, in this particular case,

$$\mathbf{G}' = \begin{bmatrix} G_{00} \\ G_{10} \\ G_{01} \\ G_{11} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -j \\ 0 & 0 & 1 & j \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \\ -2j \\ 2j \end{bmatrix} \quad (8.97)$$

Finally, we recover the transform vector $\mathbf{G} = [G_0 \quad G_1 \quad G_2 \quad G_3]^T$ as

$$\mathbf{G} = \begin{bmatrix} 4 \\ -2j \\ 0 \\ 2j \end{bmatrix}$$

and we have thus established the Fourier transform of the sequence $\{1, 2, 1, 0\}$ as the sequence

$$\{4, -2j, 0, 2j\}$$

It is interesting to compare the labour involved in this calculation with that in Example 8.18.

To conclude this section, we fill in the details for the matrix factorization operation, which is at the core of the process of calculating the fast Fourier transform. In a book of this nature it is not appropriate to reproduce a proof of the validity of the algorithm for any N of the form $N = 2^\gamma$. Rather, we shall illustrate how the factorization we introduced in (8.93) was obtained. The factored form of the matrix will not be generated in any calculation: what actually happens is that the various summations are performed using their structural properties. From (8.90), with $W = e^{-j2\pi/N}$, we wish to calculate the sums

$$G_k = \sum_{n=0}^{N-1} g_n W^{nk} \quad k = 0, 1, \dots, N-1 \quad (8.98)$$

In the case $N = 4$, $\gamma = 2$ we see that k and n take only the values 0, 1, 2 and 3, so we can represent both k and n using two-digit binary numbers; in general γ -digit binary numbers will be required.

We write $k = k_1 k_0$ and $n = n_1 n_0$, where k_0, k_1, n_0 and n_1 are digits that may take the values 0 or 1 only. For example, $k = 3$ becomes $k = 11$ and $n = 2$ becomes $n = 10$. The decimal form can always be recovered easily as $k = 2k_1 + k_0$ and $n = 2n_1 + n_0$. This is simply a binary repetition for n and k .

Using binary notation, we can write (8.98) as

$$G_{k_1 k_0} = \sum_{n_0=0}^1 \sum_{n_1=0}^1 g_{n_1 n_0} W^{(2n_1+n_0)(2k_1+k_0)} \quad (8.99)$$

The single summation of (8.98) is now replaced, when $\gamma = 2$, by two summations. Again we see that for the more general case with $N = 2^\gamma$ a total of γ summations replaces the single sum of (8.98).

The matrix factorization operation with which we are concerned is now achieved by considering the term in (8.99).

$$\begin{aligned} W^{(2n_1+n_0)(2k_1+k_0)} &= W^{(2k_1+k_0)2n_1} W^{(2k_1+k_0)n_0} \\ &= W^{4n_1 k_1} W^{2n_1 k_0} W^{(2k_1+k_0)n_0} \end{aligned} \quad (8.100)$$

Since $W = e^{-j2\pi/N}$, and $N = 4$ in this case, the leading term in (8.100) becomes

$$\begin{aligned} W^{4n_1 k_1} &= (e^{-j2\pi/4})^{4n_1 k_1} = (e^{-j2\pi})^{n_1 k_1} \\ &= 1^{n_1 k_1} = 1 \end{aligned}$$

Again we observe that in the more general case such a factor will always emerge.

Thus (8.100) can be written as

$$W^{(2n_1+n_0)(2k_1+k_0)} = W^{2n_1 k_0} W^{(2k_1+k_0)n_0}$$

so that (8.99) becomes

$$G_{k_1 k_0} = \sum_{n_0=0}^1 \left[\sum_{n_1=0}^1 g_{n_1 n_0} W^{2n_1 k_0} \right] W^{(2k_1+k_0)n_0} \quad (8.101)$$

which is the required matrix factorization. This can be seen by writing

$$g'_{k_0 n_0} = \sum_{n_1=0}^1 g_{n_1 n_0} W^{2n_1 k_0} \quad (8.102)$$

so that the sum in the square brackets in (8.101) defines the four relations

$$\begin{cases} g'_{00} = g_{00} W^{2.0.0} + g_{10} W^{2.1.0} = g_{00} + g_{10} W^0 \\ g'_{01} = g_{01} W^{2.0.0} + g_{11} W^{2.1.0} = g_{01} + g_{11} W^0 \\ g'_{10} = g_{00} W^{2.0.1} + g_{10} W^{2.1.1} = g_{00} + g_{10} W^2 \\ g'_{11} = g_{01} W^{2.0.1} + g_{11} W^{2.1.1} = g_{01} + g_{11} W^2 \end{cases} \quad (8.103)$$

which, in matrix form, becomes

$$\begin{bmatrix} g'_{00} \\ g'_{01} \\ g'_{10} \\ g'_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{bmatrix} \begin{bmatrix} g_{00} \\ g_{01} \\ g_{10} \\ g_{11} \end{bmatrix} \quad (8.104)$$

and we see that we have re-established the system of equations (8.95), this time with binary indexing. Note that in (8.103) and (8.104) we distinguished between terms in W^0 depending on how the zero is generated. When the zero is generated through the value of the summation index (that is, when $n_1 = 0$ and thus a zero will always be generated whatever the value of γ) we replace W^0 by 1. When the index is zero because of the value of k_0 , we maintain W^0 as an aid to generalization.

The final stage of the factorization appears when we write the outer summation of (8.101) as

$$G'_{k_0 k_1} = \sum_{n_0=0}^1 g'_{k_0 n_0} W^{(2k_1+k_0)n_0} \quad (8.105)$$

which, on writing out in full, gives

$$\begin{aligned} G'_{00} &= g'_{00} W^{0.0} + g'_{01} W^{0.1} = g'_{00} + g'_{01} W^0 \\ G'_{01} &= g'_{00} W^{2.0} + g'_{01} W^{2.1} = g'_{00} + g'_{01} W^2 \\ G'_{10} &= g'_{10} W^{1.0} + g'_{11} W^{1.1} = g'_{10} + g'_{11} W^1 \\ G'_{11} &= g'_{10} W^{3.0} + g'_{11} W^{3.1} = g'_{10} + g'_{11} W^3 \end{aligned}$$

or, in matrix form,

$$\begin{bmatrix} G'_{00} \\ G'_{01} \\ G'_{10} \\ G'_{11} \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} g'_{00} \\ g'_{01} \\ g'_{10} \\ g'_{11} \end{bmatrix} \quad (8.106)$$

The matrix in (8.106) is exactly that of (8.98), and we have completed the factorization process as we intended. Finally, to obtain the transform in a natural order, we must carry out the bit-reversal operation. From (8.101) and (8.104), we achieve this by simply writing

$$G_{k_1 k_0} = G'_{k_0 k_1} \quad (8.107)$$

We can therefore summarize the Cooley–Tukey algorithm for the fast Fourier transform for the case $N = 4$ by the three relations (8.102), (8.105) and (8.107), that is

$$g'_{k_0 n_0} = \sum_{n_1=0}^1 g_{n_1 n_0} W^{2n_1 k_0}$$

$$G'_{k_0 k_1} = \sum_{n_0=0}^1 g'_{k_0 n_0} W^{(2k_1+k_0)n_0}$$

$$G_{k_1 k_0} = G'_{k_0 k_1}$$

The evaluation of these three relationships is equivalent to the matrix factorization process together with the bit-reversal procedure discussed above.

The fast Fourier transform is essentially a computer-orientated algorithm and highly efficient codes are available in MATLAB and other software libraries, usually requiring a simple subroutine call for their implementation. The interested reader who would prefer to produce ‘home-made’ code may find listings in the textbook by Brigham quoted at the beginning of this section, as well as elsewhere.



8.6.6 Exercises

- 28 Calculate directly the discrete Fourier transform of the sequence

$$\{1, 0, 1, 0\}$$

using the methods of Section 8.6.3 (see Example 8.18).

- 29 Use the fast Fourier transform method to calculate the transform of the sequence of Exercise 28 (follow Example 8.20).

- 30 Use the FFT algorithm in MATLAB (or an alternative) to improve the experiment with the estimation of the spectrum of the signal of Example 8.19.

- 31 Derive an FFT algorithm for $N = 2^3 = 8$ points.
Work from (8.98), writing

$$k = 4k_2 + 2k_1 + k_0, \quad k_i = 0 \text{ or } 1 \quad \text{for all } i$$

$$n = 4n_2 + 2n_1 + n_0, \quad n_i = 0 \text{ or } 1 \quad \text{for all } i$$

to show that

$$g'_{k_0 n_1 n_0} = \sum_{n_2=0}^1 g_{n_2 n_1 n_0} W^{4k_0 n_2}$$

$$g''_{k_0 k_1 n_0} = \sum_{n_1=0}^1 g'_{k_0 n_1 n_0} W^{(2k_1+k_0)2n_1}$$

$$G'_{k_0 k_1 k_2} = \sum_{n_0=0}^1 g''_{k_0 k_1 n_0} W^{(4k_2+2k_1+k_0)n_0}$$

$$G_{k_2 k_1 k_0} = G'_{k_0 k_1 k_2}$$

8.7 Engineering application: the design of analogue filters

In this section we explore the ideas of mathematical design or synthesis. We shall express in mathematical form the desired performance of a system, and, utilizing the ideas we have developed, produce a system design.

This chapter has been concerned with the frequency-domain representation of signals and systems, and the system we shall design will operate on input signals to produce output signals with specific frequency-domain properties. In Figure 8.31 we illustrate the amplitude response of an ideal low-pass filter. This filter passes perfectly signals, or components of signals, at frequencies less than the cut-off frequency ω_c . Above ω_c , attenuation is perfect, meaning that signals above this frequency are not passed by this filter.

The amplitude response of this ideal device is given by

$$|H'(j\omega)| = \begin{cases} 1 & (|\omega| \leq \omega_c) \\ 0 & (|\omega| > \omega_c) \end{cases}$$

Such an ideal response cannot be attained by a real analogue device, and our design problem is to approximate this response to an acceptable degree using a system that can be constructed. A class of functions whose graphs resemble that of Figure 8.31 is the set

$$|H_B(j\omega)| = \frac{1}{\sqrt{[1 + (\omega/\omega_c)^{2n}]}}$$

and we see from Figure 8.32, which corresponds to $\omega_c = 1$, that, as n increases, the graph approaches the ideal response. This particular approximation is known as the **Butterworth approximation**, and is only one of a number of possibilities.

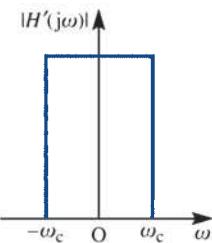
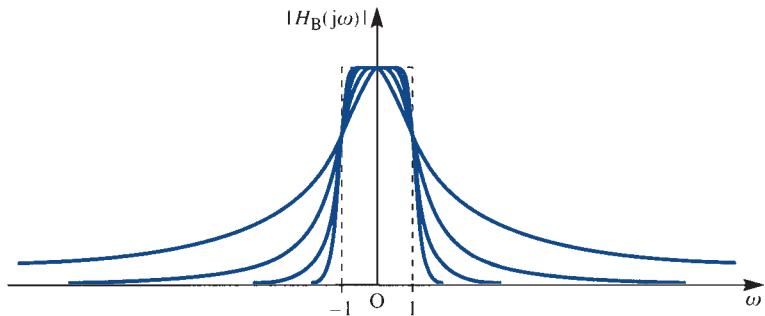


Figure 8.31 Amplitude response of an ideal low-pass filter.

Figure 8.32

Amplitude responses of the Butterworth filters.



To explore this approach further, we must ask the question whether such a response could be obtained as the frequency response of a realizable, stable linear system. We assume that it can, although if our investigation leads to the opposite conclusion then we shall have to abandon this approach and seek another. If $H_B(j\omega)$ is the frequency response of such a system then it will have been obtained by replacing s with $j\omega$ in the system Laplace transfer function. This is at least possible since, by assumption, we are dealing with a stable system. Now

$$|H_B(j\omega)|^2 = \frac{1}{1 + (j\omega/j\omega_c)^{2n}}$$

where $|H_B(j\omega)|^2 = H_B(j\omega)H_B^*(j\omega)$. If $H_B(s)$ is to have real coefficients, and thus be realizable, then we must have $H_B^*(j\omega) = H(-j\omega)$. Thus

$$H_B(j\omega)H_B(-j\omega) = \frac{1}{1 + (\omega/\omega_c)^{2n}} = \frac{1}{1 + (j\omega/j\omega_c)^{2n}}$$

and we see that the response could be obtained by setting $s = j\omega$ in

$$H_B(s)H_B(-s) = \frac{1}{1 + (s/j\omega_c)^{2n}}$$

Our task is now to attempt to separate $H_B(s)$ from $H_B(-s)$ in such a way that $H_B(s)$ represents the transfer function of a stable system. To do this, we solve the equation

$$1 + (s/j\omega_c)^{2n} = 0$$

to give the poles of $H_B(s)H_B(-s)$ as

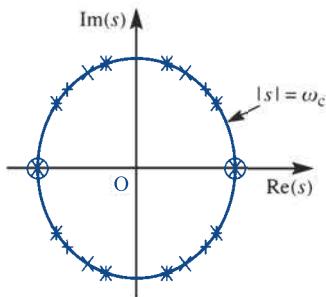
$$s = \omega_c e^{j[(2k+1)\pi/2n+\pi/2]} \quad (k = 0, 1, 2, 3, \dots) \quad (8.108)$$

Figure 8.33 shows the pole locations for the cases $n = 1, 2, 3$ and 5 . The important observations that we can make from this figure are that in each case there are $2n$ poles equally spaced around the circle of radius ω_c in the Argand diagram, and that there are no poles on the imaginary axis. If $s = s_1$ is a pole of $H_B(s)H_B(-s)$ then so is $s = -s_1$, and we can thus select as poles for the transfer function $H_B(s)$ those lying in the left half-plane. The remaining poles are then those of $H_B(-s)$. By this procedure, we have generated a stable transfer function $H_B(s)$ for our filter design.

The transfer function that we have generated from the frequency-domain specification of system behaviour must now be related to a real system, and this is the next step

Figure 8.33

Pole locations for the Butterworth filters:
 (○) $n = 1$; (+) $n = 2$;
 (×) $n = 3$;
 (*) $n = 8$.



in the design process. The form of the transfer function for the filter of order n can be shown to be

$$H_B(s) = \frac{\omega_c^n}{(s - s_1)(s - s_2) \cdots (s - s_n)}$$

where s_1, s_2, \dots, s_n are the stable poles generated by (8.108). The reader is invited to show that the second-order Butterworth filter has transfer function

$$H_B(s) = \frac{\omega_c^2}{s^2 + \sqrt{2}\omega_c s + \omega_c^2}$$

Writing $Y(s) = H_B(s)U(s)$, with $H_B(s)$ as above, we obtain

$$Y(s) = \frac{\omega_c^2}{s^2 + \sqrt{2}\omega_c s + \omega_c^2} U(s)$$

or

$$(s^2 + \sqrt{2}\omega_c s + \omega_c^2)Y(s) = \omega_c^2 U(s) \quad (8.109)$$

If we assume that all initial conditions are zero then (8.109) represents the Laplace transform of the differential equation

$$\frac{d^2y(t)}{dt^2} + \sqrt{2}\omega_c \frac{dy(t)}{dt} + \omega_c^2 y(t) = \omega_c^2 u(t) \quad (8.110)$$

This step completes the mathematical aspect of the design exercise. It is possible to show that a system whose behaviour is modelled by this differential equation can be constructed using elementary circuit components, and the specification of such a circuit would complete the design. For a fuller treatment of the subject the interested reader could consult M. J. Chapman, D. P. Goodall and N. C. Steele, *Signal Processing in Electronic Communications* (Chichester, Horwood Publishing, 1997).

To appreciate the operation of this filter, the use of the Signal Processing Toolbox in MATLAB is recommended. After setting the cut-off frequency ω_c at 4 for example, the output of the system $y(t)$ corresponding to an input signal $u(t) = \sin t + \sin 10t$ will demonstrate the almost-perfect transmission of the low-frequency ($\omega = 1$) term, with nearly total attenuation of the high-frequency ($\omega = 10$) signal. As an extension to this exercise, the differential equation to represent the third- and fourth-order filters should be obtained, and the responses compared. Using a simulation package and an FFT coding, it is possible to investigate the operation of such devices from the viewpoint of the frequency domain by examining the spectrum of samples drawn from both input and output signals.



8.8 Engineering application: direct design of digital filters and windows

This application section provides a brief introduction to some methods of digital filter design. In particular we introduce a transform based on the Fourier transform itself, rather than going via the exponential form of Fourier series and the underlying periodicity implications. The material contained in this section first appeared in *Signal Processing in Electronic Communication* by M. J. Chapman, D. P. Goodall and N. C. Steele, originally published in the Horwood Series in Engineering Science in 1997 and is reproduced by courtesy of the current publishers Woodhead Publishing Limited.

8.8.1 Digital filters

Suppose $f(t)$ is a signal with Fourier transform $F(j\omega)$ so that

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega t} d\omega \quad (8.111)$$

If we now sample $f(t)$ at times $t = kT$, $k \in \mathbb{Z}$, we obtain the sequence $\{f_k\} = \{f(kT)\}$ and (8.111) gives

$$f_k = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega) e^{j\omega kT} d\omega \quad (8.112)$$

Splitting this infinite interval of integration into intervals of length $2\pi/T$, we obtain

$$\begin{aligned} f_k &= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \int_{n\frac{2\pi}{T}}^{(n+1)\frac{2\pi}{T}} F(j\omega) e^{j\omega kT} d\omega \\ &= \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} \int_0^{\frac{2\pi}{T}} F\left(j\left(\omega + n\frac{2\pi}{T}\right)\right) e^{j(\omega+n\frac{2\pi}{T})kT} d\omega \\ &= \frac{1}{2\pi} \int_0^{\frac{2\pi}{T}} \left(\sum_{n=-\infty}^{\infty} F\left(j\left(\omega + n\frac{2\pi}{T}\right)\right) \right) e^{j\omega kT} d\omega \end{aligned}$$

since $e^{j(\omega+n\frac{2\pi}{T})kT} = e^{j\omega kT} e^{jn2\pi} = e^{j\omega kT}$. As usual, we do not attempt to give conditions under which the above interchange between an integral and an infinite sum is valid. In any case, the above is only intended as a formal procedure leading to a definition for the discrete-time Fourier transform.

If we now let θ be the normalized frequency $\theta = \omega T$ and set

$$F(e^{j\theta}) = \frac{1}{T} \sum_{n=-\infty}^{\infty} F\left(j\left(\frac{\theta + n2\pi}{T}\right)\right) \quad (8.113)$$

where we note that the right-hand side has period 2π in θ , we obtain

$$f_k = \frac{1}{2\pi} \int_0^{2\pi} F(e^{j\theta}) e^{jk\theta} d\theta \quad (8.114)$$

The periodic function, $F(e^{j\theta})$, is referred to as the **discrete-time Fourier transform (DTFT)** of the sequence $\{f_k\}$. Equation (8.113) is unsuitable for calculation of the DTFT and so we instead use (8.114) to define the inverse DTFT and invert this in order to define the direct transform. We claim that the DTFT is, in fact, given by

$$F(e^{j\theta}) = \sum_{n=-\infty}^{\infty} f_n e^{-jn\theta} = \sum_{n=-\infty}^{\infty} f_n (e^{j\theta})^{-n} \quad (8.115)$$

Note that this is the same as the transform defined in (6.1), which is known as the bilateral **z transform** of $\{f\}$ reflecting the fact that it is defined for both positive and negative values of the time index k , evaluated at $z = e^{j\theta}$. This fact also explains the use of the notation, $F(e^{j\theta})$. To show that (8.115) is valid, we substitute into the right-hand side of (8.114) to give

$$\begin{aligned} \frac{1}{2\pi} \int_0^{2\pi} F(e^{j\theta}) e^{jk\theta} d\theta &= \frac{1}{2\pi} \int_0^{2\pi} \sum_{n=-\infty}^{\infty} f_n e^{-jn\theta} e^{jk\theta} d\theta \\ &= \sum_{n=-\infty}^{\infty} f_n \frac{1}{2\pi} \int_0^{2\pi} e^{j(k-n)\theta} d\theta \end{aligned}$$

assuming the interchange of summation and integration is permissible. However, it is easy to see that

$$\frac{1}{2\pi} \int_0^{2\pi} e^{j(k-n)\theta} d\theta = \begin{cases} 0, & \text{for } k \neq n \\ 1, & \text{for } k = n \end{cases} = \delta_{k-n}$$

and so the right-hand side of (8.114) reduces to

$$\sum_{n=-\infty}^{\infty} f_n \delta_{k-n} = f_k,$$

as desired. To summarize, we have the two equations

$$\mathbf{DTFT} \quad F(e^{j\theta}) = \sum_{k=-\infty}^{\infty} f_k e^{-jk\theta} \quad (8.116a)$$

$$\mathbf{IDTFT} \quad f_k = \frac{1}{2\pi} \int_0^{2\pi} F(e^{j\theta}) e^{jk\theta} d\theta \quad (8.116b)$$

Example 8.21

Calculate the discrete-time Fourier transform of the finite sequence

$$\{u\} = \{1, 2, 2, 1\}$$

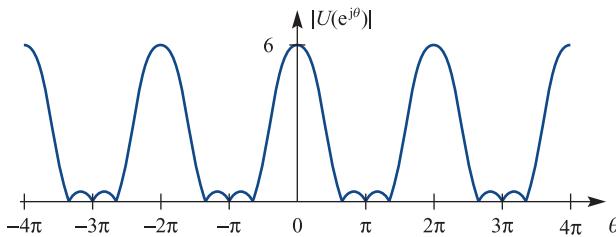
Solution We adopt the convention that the above sequence is ‘padded-out’ with zeros, that is we have $\{u\} = \{u_k\}$ where $u_0 = u_3 = 1$, $u_1 = u_2 = 2$ and $u_k = 0$ otherwise. It follows from (8.116a) that

$$\begin{aligned} U(e^{j\theta}) &= 1 + 2e^{-j\theta} + 2e^{-2j\theta} + e^{-3j\theta} \\ &= e^{-\frac{3}{2}j\theta} \left[\left(e^{\frac{3}{2}j\theta} + e^{\frac{3}{2}j\theta} \right) + 2 \left(e^{\frac{1}{2}j\theta} + e^{-\frac{1}{2}j\theta} \right) \right] \\ &= e^{-\frac{3}{2}j\theta} \left[2 \cos\left(\frac{3\theta}{2}\right) + 4 \cos\left(\frac{\theta}{2}\right) \right] \end{aligned}$$

A sketch of $|U(e^{j\theta})| = \left| 2 \cos\left(\frac{3\theta}{2}\right) + 4 \cos\left(\frac{\theta}{2}\right) \right|$ is given in Figure 8.34. $|U(e^{j\theta})|$ is called the **amplitude spectrum** of the sequence $\{u\}$. Figure 8.34 clearly shows the periodicity of $|U(e^{j\theta})|$, which by now is not surprising. In a similar fashion, we refer to $\arg U(e^{j\theta})$ as the **phase spectrum** of $\{u\}$.

Figure 8.34

Amplitude spectrum $|U(e^{j\theta})|$ for Example 8.21.



We are now in a position to develop a direct approach to the design of digital filters based on a Fourier series approach. Suppose that $D(z)$ is the transfer function of a stable discrete-time system, then, we can write as usual,

$$Y(z) = D(z)U(z)$$

If the input sequence is $\{u_k\} = \{\delta_k\} = \{1, 0, 0, 0, \dots\}$, the unit impulse sequence with z transform $U(z) = 1$, then the transform of the output sequence, namely the impulse response sequence, is

$$Y_\delta(z) = D(z) = \sum_{n=0}^{\infty} d_n z^{-n}$$

Since the system is stable, by assumption, there is a frequency response which is obtained by taking the DTFT of the impulse response sequence. This is achieved by replacing z by $e^{j\omega}$ in $D(z)$ to obtain

$$D(e^{j\omega T}) = D(e^{j\theta}) = \sum_{n=0}^{\infty} d_n e^{-jn\theta} \quad (8.117)$$

where $\theta = \omega T$.

Now (8.117) can be interpreted as the Fourier expansion of $D(e^{j\theta})$, using as basis functions the orthogonal set $\{e^{-jn\theta}\}$. It is then easy to show that the Fourier coefficients relative to this base are given by

$$d_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} D(e^{j\theta}) e^{jn\theta} d\theta$$

We now set $D(e^{j\theta})$ to the desired ideal frequency response function and calculate the resulting Fourier coefficients, $\{h_d(n)\}$ say. It should be noted that, at this stage, we can no longer restrict to $n \geq 0$, i.e. $h_d(n)$, as defined above, is not causal and hence does not correspond with the impulse response of any *realizable* system. If a filter is to be realized using a finite number of delay elements, some form of truncation must take place. It is helpful to think of this truncation being performed by the application of a **window**, defined by a window weighting function $w(n)$. The simplest window is the **rectangular window**, with weighting function $w(n)$ defined by

$$w(n) = \begin{cases} 1, & -n_1 \leq n \leq n_2 \\ 0, & \text{otherwise} \end{cases}$$

Using this window, we actually form

$$\sum_{n=-\infty}^{\infty} w(n)h_d(n)e^{-jn\theta} = \sum_{n=-n_1}^{n_2} h_d(n)e^{-jn\theta} = \tilde{D}(e^{j\theta})$$

where, if n_1 and n_2 are sufficiently large, $\tilde{D}(e^{j\theta})$ will be an adequate approximation to $D(e^{j\theta})$, the desired frequency response. It is important to note that the **filter length**, that is the number of delay elements or terms in the difference equation, depends on the choice of n_1 and n_2 . This means that some accuracy will always have to be sacrificed in order to produce an acceptable design.

We explore this technique by designing a low-pass filter in Example 8.22.

Example 8.22

Use the Fourier series, or direct design method, to produce a low-pass digital filter with cut-off frequency $f_c = 1$ kHz, when the sampling frequency is $f_s = 5$ kHz.

Solution

We wish to make use of the non-dimensional frequency variable θ and, since $T = 1/f_s = 1/5000$, we have

$$\theta = \omega T = 2\pi f T = \frac{2\pi f}{5000}$$

The cut-off frequency is then $\theta_c = 2\pi f_c / 5000 = 2\pi/5$ and the ideal frequency response $D(e^{j\theta})$ is now defined by

$$D(e^{j\theta}) = \begin{cases} 1 & |\theta| \leq 2\pi/5 \\ 0 & |\theta| > 2\pi/5 \end{cases}$$

We now calculate the coefficients $h_d(n)$ as

$$\begin{aligned} h_d(n) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} D(e^{j\theta}) e^{jn\theta} d\theta \\ &= \frac{1}{2\pi} \int_{-2\pi/5}^{2\pi/5} e^{jn\theta} d\theta \\ &= \frac{1}{n\pi} \sin\left(\frac{2n\pi}{5}\right) \quad \text{for } n \neq 0 \\ &= \frac{2}{5} \operatorname{sinc}\left(\frac{2n\pi}{5}\right) \quad (\text{also valid for } n=0) \end{aligned}$$

At this stage, we have to choose the length of the filter. By now, we know that a ‘long’ filter is likely to produce superior results in terms of frequency domain performance. However, experience again tells us that there will be penalties in some form or other. Let us choose a filter of length 9, with the coefficients selected for simplicity as symmetric about $n = 0$. As already discussed, this choice leads to a non-causal system, but we deal with this problem when it arises. This scheme is equivalent to specifying the use of a rectangular window defined by

$$w(n) = \begin{cases} 1 & -4 \leq n \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

We now calculate the coefficients $h_d(-4), h_d(-3), \dots, h_d(0), \dots, h_d(4)$, which are tabulated in Figure 8.35.

Figure 8.35
Coefficients $h_d(k)$,
for $k = -4, -3, \dots, 4$.

$h_d(\pm 4)$	$h_d(\pm 3)$	$h_d(\pm 2)$	$h_d(\pm 1)$	$h_d(0)$
-0.07568	-0.06237	0.09355	0.30273	0.40000

The transfer function of the digital filter is then \tilde{D} , where

$$\begin{aligned}\tilde{D}(z) &= \sum_{n=-4}^4 h_d(n)z^{-n} \\ &= -0.07568z^{-4} - 0.06237z^{-3} + 0.09355z^{-2} + 0.30273z^{-1} + 0.40000 \\ &\quad + 0.30273z + 0.09355z^2 - 0.06237z^3 - 0.07568z^4\end{aligned}$$

Although this system is indeed non-causal, since its impulse response sequence contains terms in positive powers of z , we can calculate the frequency response as

$$\begin{aligned}\tilde{D}(e^{j\theta}) &= -0.15137 \cos(4\theta) - 0.12473 \cos(3\theta) + 0.18710 \cos(2\theta) \\ &\quad + 0.60546 \cos(\theta) + 0.40000\end{aligned}$$

Figure 8.36 illustrates the corresponding amplitude response.

Figure 8.36 Amplitude response of the non-causal filter of Example 8.22.

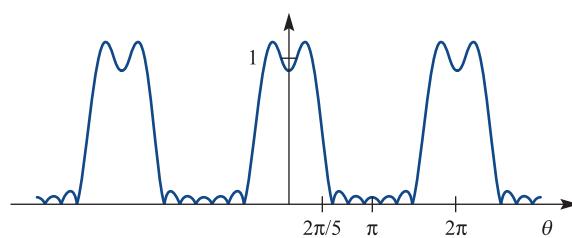


Figure 8.36, of Example 8.22, shows us that the amplitude response of our filter is a reasonable approximation to the design specification. We do, however, notice that there are some oscillations in both pass- and stop-bands. These are due to the abrupt cut-off of the rectangular window function and the effect is known as **Gibbs’ phenomenon**. Thus, the window function generates additional spectral components, which are referred to as **spectral leakage**. A way of improving the performance in this respect is discussed in Section 8.8.2. The immediate problem is the realization of this non-causal design. To see how we can circumvent the difficulty, we proceed as follows.

The transfer function we have derived is of the general form

$$\tilde{D}(z) = \sum_{k=-N}^N h_d(k)z^{-k}$$

$$= z^N [h_d(-N) + h_d(-N+1)z^{-1} + \dots + h_d(0)z^{-N} + \dots + h_d(N)z^{-2N}]$$

Suppose that we implement the system with transfer function

$$\tilde{D}(z) = z^{-N}\tilde{D}(z)$$

which is a causal system. First we notice that, on setting $z = e^{j\omega T}$, the amplitude response $|\tilde{D}(e^{j\omega T})|$ is given by

$$|\tilde{D}(e^{j\omega T})| = |e^{-j\omega NT}| |\tilde{D}(e^{j\omega T})| = |\tilde{D}(e^{j\omega T})|$$

that is, it is identical with that of the desired design. Furthermore,

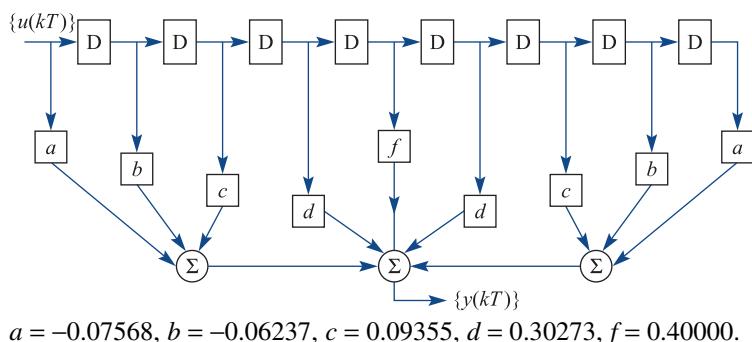
$$\arg\{\tilde{D}(e^{j\omega T})\} = \arg\{\tilde{D}(e^{j\omega T})\} - N\omega T$$

indicating a pure delay of amount NT in the response of the second system. This means that, assuming we are prepared to accept this delay, our design objective can be met by the system with transfer function $Q(z)$ given by

$$\tilde{D}(z) = [-0.07568 - 0.06237z^{-1} + 0.09355z^{-2} + 0.30273z^{-3} + 0.40000z^{-4} \\ + 0.30273z^{-5} + 0.09355z^{-6} - 0.06237z^{-7} - 0.07568z^{-8}]$$

It is evident from Figure 8.37 that the filter designed in Example 8.22 differs from the previous designs. The nature of this difference is the absence of feedback paths in the block diagram realization of Figure 8.37. One effect of this is that the impulse response sequence is finite, a fact which we already know, since the design method involved truncating the impulse response sequence. Filters of this type are known as **finite impulse response (FIR)** designs and may always be implemented using structures not involving feedback loops. Another name used for such structures is **non-recursive**, but it is not correct to assume that the only possible realization of an FIR filter is by use of a non-recursive structure; for details see M. T. Jong, *Methods of Discrete Signals and Systems Analysis* (New York, McGraw-Hill, 1982).

Figure 8.37 A realization of the final system of Example 8.22.

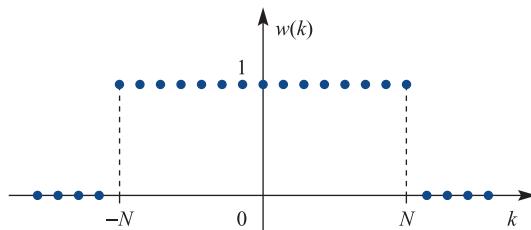


8.8.2 Windows

In this section, we consider the problem identified in Example 8.22 in connection with the sharp cut-off of the rectangular window function.

Figure 8.38

Rectangular window sequence.



The rectangular window sequence, illustrated in Figure 8.38, is defined by

$$w(k) = \begin{cases} 1 & |k| \leq N \\ 0 & \text{otherwise} \end{cases}$$

which can be expressed in the form

$$w(k) = \zeta(k + N) - \zeta(k - (N + 1))$$

where $\zeta(k) = \{h(k)\}$, defined in Example 6.22.

Since

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

the DTFT of the sequence $\{w(k)\}$ is

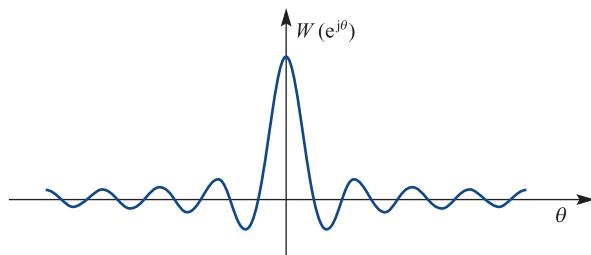
$$\begin{aligned} W(e^{j\theta}) &= \frac{\sin(\frac{1}{2}(2N+1)\theta)}{\sin(\frac{1}{2}\theta)} \quad \text{for } \theta \neq 0 \\ &= \frac{(2N+1)\operatorname{sinc}(\frac{1}{2}(2N+1)\theta)}{\operatorname{sinc}(\frac{1}{2}\theta)} \end{aligned}$$

It is easy to see that $W(e^{j0}) = W(1) = \sum_{n=-N}^N w(n) = 2N + 1$ and so the above formula, using the

since, function, is valid for all θ , including $\theta = 0$. The graph of this function is illustrated in Figure 8.39. The first positive (negative) zero in its spectrum is the positive (negative) value of θ closest to zero such that $W(e^{j\theta}) = 0$. The **main lobe** of the window function is that part of the graph of $W(e^{j\theta})$ that lies between the first positive and first negative zero in $W(e^{j\theta})$. The **main lobe width** is the distance between the first positive and negative zeros in $W(e^{j\theta})$. As the length of the window increases, the main lobe narrows and its peak value rises and, in some sense, $W(e^{j\theta})$ approaches an impulse, which is desirable. However, the main disadvantage is that the amplitudes of the side lobes also increase.

Figure 8.39

DTFT of the rectangular window sequence.



The use of any window leads to distortion of the spectrum of the original signal caused by the size of the side lobes in the window spectrum and the width of the window's

main spectral lobe, producing oscillations in the filter response. The window function can be selected so that the amplitudes of the side lobes are relatively small, with the result that the size of the oscillations is reduced; however, in general, the main lobe width does not decrease. Thus, in choosing a window, it is important to know the trade-off between having narrow main lobe and low side lobes in the window spectrum.

A considerable amount of research has been carried out, aimed at determining suitable alternative window functions which smooth out straight truncation and thus reduce the Gibbs' phenomena effects observed in the amplitude response of Figure 8.36. To minimize the effect of spectral leakage, windows which approach zero smoothly at either end of the sampled signal are used. We do not discuss the derivation of the various window functions, rather we tabulate, in Figure 8.40, some of the more popular examples in a form suitable for symmetric filters of length $2N + 1$. For a more detailed discussion on windows and their properties, see, for example: E. C. Ifeachor and B. W. Jervis, *Digital Signal Processing: A Practical Approach* (Addison-Wesley, Wokingham, UK, 1993); A. V. Oppenheim and R. W. Schafer, *Discrete-time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1989); S. J. Stearns and D. R. Hush, *Digital Signal Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1990).

Window name	$w(k)$
Bartlett	$w(k) = \begin{cases} (k + N)/N & -N \leq k < 0 \\ (N - k)/N & 0 \leq k \leq N \end{cases}$
von Hann or Hanning	$w(k) = 0.5 + 0.5 \cos(\pi k/(N + 1))$
Hamming	$w(k) = 0.54 + 0.46 \cos(\pi k/N)$
Blackman	$w(k) = 0.42 + 0.5 \cos(\pi k/N) + 0.08 \cos(2\pi k/N)$

In each case, $w(k) = 0$ for k outside the range $[-N, N]$.

Figure 8.40 Some popular window functions.

Note: Slight variations on the above definitions may be found in various texts. These tend to involve switching between ‘division by N ’, ‘division by $N + \frac{1}{2}$ ’ and ‘division by $N + 1$ ’. For example, the von Hann or Hanning window is variously defined by $w(k) = 0.5(1 + \cos(\pi k/N))$ or $w(k) = 0.5(1 + \cos(2\pi k/(2N + 1)))$ or $w(k) = 0.5(1 + \cos(\pi k/(N + 1)))$ for $|k| \leq N$ with $w(k) = 0$ for $|k| > N$. The Bartlett window, or one of its variations, is sometimes referred to as a **triangular window**. It should also be noted that both the Bartlett window and the Blackman window, as defined in Figure 8.40, satisfy $w(-N) = w(N) = 0$ and hence give rise to difference equations of order $2N - 2$ rather than $2N$.

Formulations for other configurations can easily be deduced, or may be found in, for example, L. B. Jackson, *Digital Filters and Signal Processing* (Kluwer Academic Publishers, Boston, MA, 1986); R. E. Ziener, W. H. Tranter and D. R. Fannin, *Signals and Systems* (Macmillan, New York, 1983). The section closes with an example of the application to the design of Example 8.22.

Example 8.23

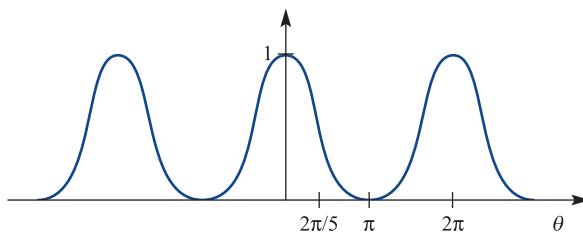
Plot the amplitude response for the filter design of Example 8.22, using (a) the Hamming window and (b) the Blackman window.

Solution (a) The transfer function coefficients are now given by $h_d(k) w_H(k)$, where $w_H(k)$ are the Hamming window coefficients, calculated with $N = 4$ and $-4 \leq k \leq 4$. The Hamming window coefficients are tabulated in Figure 8.41.

Figure 8.41 Hamming window coefficients for $-4 \leq k \leq 4$.

$N = 4$	± 4	± 3	± 2	± 1	0
	0.08000	0.21473	0.54000	0.86527	1.00000

Figure 8.42 Amplitude response of the filter of Example 8.22, with Hamming window.



The transfer function then becomes

$$\hat{D}_H(z) = [-0.00605 - 0.01339z^{-1} + 0.05052z^{-2} + 0.26194z^{-3} + 0.40000z^{-4} + 0.26194z^{-5} + 0.05052z^{-6} - 0.01339z^{-7} - 0.00605z^{-8}]$$

The frequency response is then obtained by writing $z = e^{j\theta}$, as

$$\hat{D}_H(e^{j\theta}) = e^{-j4\theta} (-0.01211 \cos(4\theta) - 0.02678 \cos(3\theta) + 0.10103 \cos(2\theta) + 0.52389 \cos(\theta) + 0.40000)$$

Figure 8.42 illustrates the magnitude of this response and the reduction of oscillations in both the pass- and stop-band is striking. The penalty is the lack of sharpness near the cut-off frequency, although the stop-band characteristics close to $\theta = \pi$ are quite good.

(b) Proceeding as in case (a), we calculate the Blackman window coefficients as shown in Figure 8.43. The Blackman windowed transfer function is thus

$$\hat{D}_B(z) = -0.00414 + 0.03181z^{-1} + 0.23418z^{-2} + 0.40000z^{-3} + 0.23418z^{-4} + 0.03181z^{-5} - 0.00414z^{-6}$$

Figure 8.43 Blackman window coefficients for $-4 \leq k \leq 4$.

$N = 4$	± 4	± 3	± 2	± 1	0
	0.00000	0.06645	0.34000	0.77355	1.00000

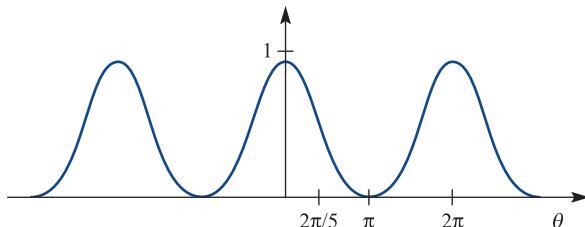
and the frequency response is found as

$$\hat{D}(e^{j\theta}) = e^{-j3\theta} (-0.00829 \cos(3\theta) + 0.06361 \cos(2\theta) + 0.46836 \cos(\theta) + 0.40000)$$

The amplitude response is shown in Figure 8.44 and this design again suffers from a relatively poor performance in terms of sharpness of cut-off. The ripples observed in the pass- and stop-bands with the rectangular window have been

Figure 8.44

Amplitude response
of the filter of
Example 8.22, with
Blackman window.



removed as before. However, the ‘flat’ characteristic of the Hamming design close to $\theta = \pi$ is not evident when using the Blackman window for this particular filter .

8.8.3 Exercises

- 32** Use the direct design method with a rectangular window of length 11 to produce a causal low-pass filter with non-dimensional cut-off frequency

$$\theta_c = \frac{\pi}{2}$$

Plot the frequency response.

- 33** Repeat Exercise 32 but use a Hamming window.

8.9 Review exercises (1–25)

- 1** Calculate the Fourier sine transform of the causal function $f(t)$ defined by

$$f(t) = \begin{cases} t & (0 \leq t \leq 1) \\ 1 & (1 < t \leq 2) \\ 0 & (t > 2) \end{cases}$$

- 2** Show that if $\mathcal{F}\{f(t)\} = F(j\omega)$ then $\mathcal{F}\{f(-t)\} = F(-j\omega)$. Show also that



$$\mathcal{F}\{f(-t - a)\} = e^{ja\omega} F(-j\omega)$$

where a is real and positive.

Find $\mathcal{F}\{f(t)\}$ when

$$f(t) = \begin{cases} -\frac{1}{2}\pi & (t < -2) \\ \frac{1}{4}\pi t & (-2 \leq t \leq 2) \\ \frac{1}{2}\pi & (t > 2) \end{cases}$$

- 3** Use the result



$$\mathcal{F}[H(t + \frac{1}{2}T) - H(t - \frac{1}{2}T)] = T \operatorname{sinc} \frac{1}{2}\omega T$$

and the frequency convolution result to verify that the Fourier transform of the windowed cosine function



$$f(t) = \cos \omega_0 t [H(t + \frac{1}{2}T) - H(t - \frac{1}{2}T)]$$

is

$$\frac{1}{2}T [\operatorname{sinc} \frac{1}{2}(\omega - \omega_0)T + \operatorname{sinc} \frac{1}{2}(\omega + \omega_0)T]$$



Show that



$$\delta(t - t_1) * \delta(t - t_2) = \delta(t - (t_1 + t_2))$$

and hence show that

$$\begin{aligned} \mathcal{F}\{\cos \omega_0 t H(t)\} &= \frac{1}{2}\pi [\delta(\omega + \omega_0) + \delta(\omega - \omega_0)] \\ &\quad + \frac{j\omega}{\omega_0^2 - \omega^2} \end{aligned}$$

- 5 Establish the demodulation property,

$$\begin{aligned}\mathcal{F}\{f(t)\cos\omega_0t\cos\omega_0t\} \\ = \frac{1}{2}F(j\omega) + \frac{1}{4}[F(j\omega+2j\omega_0)+F(j\omega-2j\omega_0)]\end{aligned}$$

- 6 Use the result $\mathcal{F}\{H(t+T)-H(t-T)\}=2T\operatorname{sinc}\omega T$ and the symmetry property to show that

$$\mathcal{F}\{\operatorname{sinc}t\}=\pi[H(\omega+1)-H(\omega-1)]$$

Check your result by use of the inversion integral.

- 7 For a wide class of frequently occurring Laplace transforms it is possible to deduce an inversion integral based on the Fourier inversion integral. If $X(s)=\mathcal{L}\{x(t)\}$ is such a transform, we have

$$x(t)=\frac{1}{j2\pi}\int_{\gamma-j\infty}^{\gamma+j\infty}X(s)e^{st}ds$$

where $\operatorname{Re}(s)=\gamma$, with γ real, defines a line in the s plane to the right of all the poles of $X(s)$. Usually the integral can be evaluated using the residue theorem, and we then have

$$x(t)=\sum \text{residues of } X(s)e^{st} \text{ at all poles of } X(s)$$

- (a) Write down the poles for the transform

$$X(s)=\frac{1}{(s-a)(s-b)}$$

where a and b are real. Calculate the residues of $X(s)e^{st}$ at these poles and invert the transform.

- (b) Calculate

$$(i) \quad \mathcal{L}^{-1}\left\{\frac{1}{(s-2)^2}\right\} \quad (ii) \quad \mathcal{L}^{-1}\left\{\frac{1}{s^2(s+1)}\right\}$$

- (c) Show that

$$\mathcal{L}^{-1}\left\{\frac{2s}{(s^2+1)^2}\right\}=t\sin t$$

- 8 A linear system has impulse response $h(t)$, so that the output corresponding to an input $u(t)$ is

$$y(t)=\int_{-\infty}^{\infty}h(t-\tau)u(\tau)d\tau$$

When $u(t)=\cos\omega_0t$, $y(t)=-\sin\omega_0t$ ($\omega_0 \geq 0$).

Find the output when $u(t)$ is given by

- (a) $\cos\omega_0(t+\frac{1}{4}\pi)$ (b) $\sin\omega_0t$
 (c) $e^{j\omega_0t}$ (d) $e^{-j\omega_0t}$

This system is known as a **Hilbert transformer**.

9

In Section 8.5.1 we established that



$$\mathcal{F}^{-1}\left\{\frac{1}{j\omega}\right\}=\frac{1}{2}\operatorname{sgn}(t)$$

where $\operatorname{sgn}(t)$ is the signum function. Deduce that

$$\mathcal{F}\{\operatorname{sgn}(t)\}=\frac{2}{j\omega}$$

and use the symmetry result to demonstrate that

$$\mathcal{F}\left\{-\frac{1}{\pi t}\right\}=j\operatorname{sgn}(\omega)$$

10

The **Hilbert transform** of a signal $f(t)$ is defined by

$$F_{\text{Hi}}(x)=\mathcal{H}\{f(t)\}=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{f(\tau)}{\tau-x}d\tau$$

Show that the operation of taking the Hilbert transform is equivalent to the convolution

$$-\frac{1}{\pi t}*f(t)$$

and hence deduce that the Hilbert-transformed signal has an amplitude spectrum $F_{\text{Hi}}(j\omega)$ identical with $f(t)$. Show also that the phase of the transformed signal is changed by $\pm\frac{1}{2}\pi$, depending on the sign of ω

11

Show that

$$\begin{aligned}&\frac{t}{(t^2+a^2)(t-x)} \\ &= \frac{1}{x^2+a^2}\left(\frac{a^2}{t^2+a^2}+\frac{x}{t-x}-\frac{xt}{t^2+a^2}\right)\end{aligned}$$

Hence show that the Hilbert transform of

$$f(t)=\frac{t}{t^2+a^2} \quad (a>0)$$

is

$$\frac{a}{x^2+a^2}$$

- 12** If $F_{\text{Hi}}(x) = \mathcal{H}\{f(t)\}$ is the Hilbert transform of $f(t)$, establish the following properties:

- $\mathcal{H}\{f(a+t)\} = F_{\text{Hi}}(x+a)$
- $\mathcal{H}\{f(at)\} = F_{\text{Hi}}(ax) \quad (a > 0)$
- $\mathcal{H}\{f(-at)\} = -F_{\text{Hi}}(-ax) \quad (a > 0)$
- $\mathcal{H}\left\{\frac{df}{dt}\right\} = \frac{d}{dx} F_{\text{Hi}}(x)$
- $\mathcal{H}\{tf(t)\} = xF_{\text{Hi}}(x) + \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) dt$

- 13** Use Exercises 9 and 10 to deduce the inversion formula

$$f(t) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{F_{\text{Hi}}(x)}{x-t} dx$$

- 14** Define the **analytic signal** associated with the real signal $f(t)$ as

$$f_a(t) = f(t) - jF_{\text{Hi}}(t)$$

where $F_{\text{Hi}}(t)$ is the Hilbert transform of $f(t)$. Use the method of Review exercise 13 to show that

$$\mathcal{F}\{f_a(t)\} = F_a(j\omega) = \begin{cases} 2F(j\omega) & (\omega \neq 0) \\ 0 & (\omega = 0) \end{cases}$$

- 15** Use the result $\mathcal{F}\{H(t)\} = 1/j\omega + \pi\delta(\omega)$ and the symmetry property to show that

$$\mathcal{F}^{-1}\{H(\omega)\} = \frac{1}{2}\delta(t) + \frac{j}{2\pi t}$$

(Hint: $H(-\omega) = 1 - H(\omega)$.)

Hence show that if $\phi(t)$ is defined by $\mathcal{F}\{\phi(t)\} = 2H(\omega)F(j\omega)$ then $\phi(t) = f(t) - jF_{\text{Hi}}(t)$, the analytic signal associated with $f(t)$, where $F(j\omega) = \mathcal{F}\{f(t)\}$ and $F_{\text{Hi}}(t) = \mathcal{H}\{f(t)\}$.

If $f(t) = \cos \omega_0 t$ ($\omega_0 > 0$), find $\mathcal{F}\{f(t)\}$ and hence $\hat{f}(t)$. Deduce that

$$\mathcal{H}\{\cos \omega_0 t\} = -\sin \omega_0 t$$

By considering the signal $g(t) = \sin \omega_0 t$ ($\omega_0 > 0$), show that

$$\mathcal{H}\{\sin \omega_0 t\} = \cos \omega_0 t$$

- 16** A causal system has impulse response $\bar{h}(t)$, where $\bar{h}(t) = 0$ ($t < 0$). Define the even part $\bar{h}_e(t)$ of $\bar{h}(t)$ as

$$\bar{h}_e(t) = \frac{1}{2} [\bar{h}(t) + \bar{h}(-t)]$$

and the odd part $\bar{h}_o(t)$ as

$$\bar{h}_o(t) = \frac{1}{2} [\bar{h}(t) - \bar{h}(-t)]$$

Since $\bar{h}(t) = 0$ ($t < 0$) deduce that

$$\bar{h}_o(t) = \text{sgn}(t)\bar{h}_e(t)$$

and that

$$\bar{h}(t) = \bar{h}_e(t) + \text{sgn}(t)\bar{h}_e(t) \quad \text{for all } t$$

Verify this result for $\lambda(t) = \sin t H(t)$. Take the Fourier transform of this result to establish that

$$\bar{H}(j\omega) = \bar{H}_e(j\omega) + j\mathcal{H}\{\bar{H}_e(j\omega)\}$$

Let $\bar{h}(t) = e^{-at}H(t)$ be such a causal impulse response. By taking the Fourier transform, deduce the Hilbert transform pair

$$\mathcal{H}\left\{\frac{a}{a^2 + t^2}\right\} = -\frac{x}{a^2 + x^2}$$

Use the result

$$\mathcal{H}\{tf(t)\} = x\mathcal{H}\{f(t)\} + \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) dt$$

to show that

$$\mathcal{H}\left\{\frac{t}{a^2 + t^2}\right\} = \frac{a}{x^2 + a^2}$$

17 The **Hartley transform** is defined as

$$F_H(s) = \mathsf{H}\{f(t)\} = \int_{-\infty}^{\infty} f(t) \text{cas} 2\pi st dt$$

where $\text{cas } t = \cos t + \sin t$. Find the Hartley transform of the functions

- $f(t) = e^{-at}H(t) \quad (a > 0)$
- $f(t) = \begin{cases} 0 & (|t| > T) \\ 1 & (|t| \leq T) \end{cases}$

An alternative form of the Fourier transform pair is given by

$$F(jp) = \int_{-\infty}^{\infty} f(t) e^{-j2\pi pt} dt$$

$$g(t) = \int_{-\infty}^{\infty} G(jp) e^{j2\pi pt} dt$$

where the frequency p is now measured in hertz. Define the even part of the Hartley transform as

$$E(s) = \frac{1}{2} [F_H(s) + F_H(-s)]$$

and the odd part as

$$O(s) = \frac{1}{2} [F_H(s) - F_H(-s)]$$

Show that the Fourier transform of $f(t)$ is given by

$$F(jp) = E(p) - jO(p)$$

and confirm your result for $f(t) = e^{-2t}H(t)$.

- 19 Prove the **time-shift result** for the Hartley transform in the form

$$\begin{aligned}\mathcal{F}\{f(t - T)\} &= \sin(2\pi Ts) F_H(-s) \\ &\quad + \cos(2\pi Ts) F_H(s)\end{aligned}$$

- 20 Using the alternative form of the Fourier transform given in Review exercise 18, it can be shown that the Fourier transform of the Heaviside step function is

$$\mathcal{F}\{H(t)\} = \frac{1}{jp\pi} + \frac{1}{2}\delta(p)$$

Show that the Hartley transform of $H(t)$ is then

$$\frac{1}{2}\delta(s) + \frac{1}{s\pi}$$

and deduce that the Hartley transform of $H(t - \frac{1}{2})$ is

$$\frac{1}{2}\delta(s) + \frac{\cos \pi s - \sin \pi s}{s\pi}$$

- 21 Show that $\mathsf{H}\{\delta(t)\} = 1$ and deduce that $\mathsf{H}\{1\} = \delta(s)$. Show also that $\mathsf{H}\{\delta(t - t_0)\} = \text{cas } 2\pi s_0 t_0$ and that

$$\begin{aligned}\mathsf{H}\{\text{cas } 2\pi s_0 t\} &= \mathsf{H}\{\cos 2\pi s_0 t\} + \mathsf{H}\{\sin 2\pi s_0 t\} \\ &= \delta(s - s_0)\end{aligned}$$

- 22 Prove the Hartley transform **modulation theorem** in the form

$$\mathsf{H}\{f(t) \cos 2\pi s_0 t\} = \frac{1}{2} F_H(s - s_0) + \frac{1}{2} F_H(s + s_0)$$

Hence show that

$$\mathsf{H}\{\cos 2\pi s_0 t\} = \frac{1}{2} [\delta(s - s_0) + \delta(s + s_0)]$$

$$\mathsf{H}\{\sin 2\pi s_0 t\} = \frac{1}{2} [\delta(s - s_0) - \delta(s + s_0)]$$

- 23 Show that

$$\mathcal{F}\{\tan^{-1} t\} = \frac{\pi e^{-|s|}}{js}$$

(Hint: Consider $\int_{-\infty}^t (1+t^2)^{-1} dt$.)

- 24 Show that

$$x(t) = \frac{1}{2} (1 + \cos \omega_0 t)[H(t + \frac{1}{2}T) - H(t - \frac{1}{2}T)]$$

has Fourier transform

$$T[\text{sinc } \omega + \frac{1}{2} \text{sinc}(\omega - \omega_0) + \frac{1}{2} \text{sinc}(\omega + \omega_0)]$$

- 25 The **discrete Hartley transform** of the sequence $\{f(r)\}_{r=0}^{N-1}$ is defined by

$$\begin{aligned}\mathsf{H}(v) &= \frac{1}{N} \sum_{r=0}^{N-1} f(r) \text{cas}\left(\frac{2\pi vr}{N}\right) \\ (v &= 0, 1, \dots, N-1)\end{aligned}$$

where the function ‘cas’ is defined as in Exercise 17.

The inverse transform is

$$f(r) = \sum_{v=0}^{N-1} \mathsf{H}(v) \text{cas}\left(\frac{2\pi vr}{N}\right) \quad (r = 0, \dots, N-1)$$

Show that in the case $N = 4$,

$$\mathsf{H} = \mathbf{T}f$$

$$\mathsf{H} = [\mathsf{H}(0) \quad \mathsf{H}(1) \quad \mathsf{H}(2) \quad \mathsf{H}(3)]^T$$

$$f = [f(0) \quad f(1) \quad f(2) \quad f(3)]^T$$

$$\mathbf{T} = \frac{1}{4} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

Hence compute the discrete Hartley transform of the sequence $\{1, 2, 3, 4\}$. Show that $\mathbf{T}^2 = \frac{1}{4} \mathbf{I}$ and hence that $\mathbf{T}^{-1} = 4\mathbf{T}$, and verify that applying the \mathbf{T}^{-1} operator regains the original sequence.



9

Partial Differential Equations

Chapter 9 Contents

9.1	Introduction	616
9.2	General discussion	617
9.3	Solution of the wave equation	634
9.4	Solution of the heat-conduction/diffusion equation	660
9.5	Solution of the Laplace equation	677
9.6	Finite elements	694
9.7	Integral solutions	707
9.8	General considerations	716
9.9	Engineering application: wave propagation under a moving load	723
9.10	Engineering application: blood-flow model	726
9.11	Review exercises (1–21)	730

9.1 Introduction

Many physical processes fundamental to science and engineering are governed by **partial differential equations**, that is equations involving partial derivatives. The most familiar of these processes are heat conduction and wave propagation. To describe such phenomena, we make assumptions about gradients (for instance, the Fourier law that heat flow is proportional to temperature gradient) and we write down balance equations; partial differential equations are thus produced in a natural way. Unless the situation is very simple, there will be many independent variables, for example a time variable t and a space variable x , and the differential equations *must* involve partial derivatives. Most engineering applications involve the use of two or three spatial dimensions and this means using partial derivatives. Hence solving engineering problems means solving partial differential equations as well as the use of some vector calculus (Chapter 3), particularly the use of the gradient operator.

The application of partial differential equations is much wider than the simple situations already mentioned. Maxwell's equations (see Example 3.16) comprise a set of partial differential equations that form the basis of electromagnetic theory, and are fundamental to electrical engineers and physicists. The equations of fluid flow are partial differential equations and are widely used in aeronautical engineering, acoustics, the study of groundwater flows in civil engineering, the development of most fluid handling devices used in mechanical engineering and in investigating flame and combustion processes in chemical engineering. Quantum mechanics is yet another theory governed by a partial differential equation, the Schrödinger equation, which forms the basis of much of physics, chemistry and electronic engineering. Stress analysis is important in large areas of civil and mechanical engineering, and again requires a complicated set of partial differential equations. This is by no means an exhaustive list, but it does illustrate the importance of partial differential equations and their solution.

One of the major difficulties with partial differential equations is that it is extremely difficult to illustrate their solutions geometrically, in contrast to single-variable problems, where a simple curve can be used. For instance, the temperature in a room, particularly if it is time-varying, is not at all easy to draw or visualize, but such information is of crucial importance to a heating engineer. Modern graphics packages have improved the situation considerably in two and three dimensions and the displays can often give a good qualitative understanding. A second basic problem with partial differential equations is that it is intrinsically more difficult to solve them or even to decide whether a solution exists. The driving force of most physical systems that can be modelled by partial differential equations is determined by either what happens on the boundary of the region under consideration or how the system is started at zero time. Boundaries, therefore, play a very significant role, and we shall see that a problem can have a solution for one set of boundary conditions but not for another. Finding a solution to a partial differential equation is often quite straightforward but finding *the* solution that fits the boundary conditions is very difficult.

The solution of partial differential equations has been greatly eased by the use of computers, which have allowed the rapid numerical solution of problems that would otherwise have been intractable. Such methods have generally been integrated into this chapter, since they are now one of the standard techniques available. However, the finite-element method is considered separately, since it is more complicated, and requires a lot of careful thought and work (the section dealing with it can be omitted on a first reading). The finite-element method originated in stress analysis in civil engineering work, but has now spread into most areas where complicated boundaries are encountered.

There are three basic types of equation that appear in most areas of science and engineering, and it is essential to understand their solutions before any progress can be made on more complicated sets of equations, nonlinear equations or equations with variable coefficients.

9.2

General discussion

The three basic types of equation are referred to as the **wave equation**, the **heat-conduction or diffusion equation**, and the **Laplace equation**. In this section we briefly discuss the formulation of these three basic forms, and then consider each in more detail in later sections. The various sections will concentrate on finding and understanding solutions of the three types of equations in simple regions. The treatment of advanced methods, more complicated equations and other regions will be left to more comprehensive books on partial differential equations (see, for example, R. Haberman, *Applied Partial Differential Equations*, Upper Saddle River, NJ, Prentice Hall, 2003).

9.2.1 Wave equation

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \nabla^2 u \quad (9.1)$$

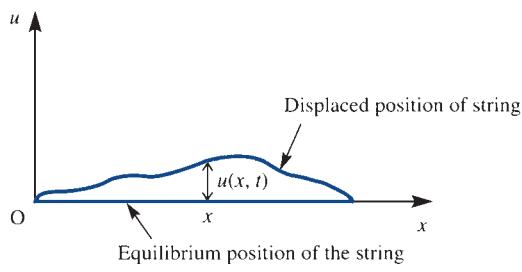
where the notation ∇^2 for Laplacian was introduced in Chapter 3, equation (3.20)

Many phenomena that involve propagation of a signal require the wave equation (9.1) to be solved in the appropriate number of space dimensions. Perhaps the simplest, in one space dimension, is the vibration of a taut string stretched to a uniform tension T between two fixed points as illustrated in Figure 9.1(a), where u is the displacement, x is measured along the equilibrium position of the string and t is time. Applying Newton's law of motion to an element Δs of the string (Figure 9.1b), for motion in the u direction, we have

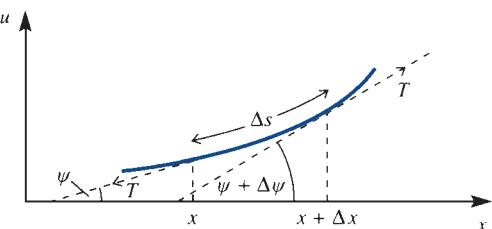
$$\text{net force in } u \text{ direction} = \text{mass element} \times \text{acceleration in } u \text{ direction}$$

that is,

$$T \sin(\psi + \Delta\psi) - T \sin \psi = \rho \Delta s \frac{\partial^2 u}{\partial t^2} \quad (9.2)$$



(a)



(b)

Figure 9.1 Displacement of an element of a taut string.

where ρ is the mass per unit length of the string. Neglecting terms quadratic in small quantities and using the Taylor series expansions

$$\cos \Delta\psi = 1 + O(\Delta\psi^2) \approx 1, \sin \Delta\psi = \Delta\psi + O(\Delta\psi^3) \approx \Delta\psi$$

and the expression

$$\Delta s = \sqrt{\left[1 + \left(\frac{\partial u}{\partial x}\right)^2\right]} \Delta x \approx \Delta x$$

for the arclength, (9.2) becomes

$$T \sin \psi + T \cos \psi \Delta\psi - T \sin \psi = \rho \Delta x \frac{\partial^2 u}{\partial t^2}$$

or

$$T \cos \psi \frac{\Delta\psi}{\Delta x} = \rho \Delta x \frac{\partial^2 u}{\partial t^2}$$

which in the limit as $\Delta x \rightarrow 0$ becomes

$$T \cos \psi \frac{\partial \psi}{\partial x} = \rho \frac{\partial^2 u}{\partial t^2} \quad (9.3)$$

Again assuming that ψ itself is small for small oscillations of the string, we have $\cos \psi \approx 1$, and the gradient of the string

$$\frac{\partial u}{\partial x} = \tan \psi \approx \psi$$

and hence from (9.3) we obtain

$$T \frac{\partial^2 u}{\partial x^2} = \rho \frac{\partial^2 u}{\partial t^2}$$

Thus the displacement of the string satisfies the one-dimensional wave equation

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \quad (9.4)$$

and the propagation of the disturbance in the string is given by a solution of this equation where $c^2 = T/\rho$.

By considering the theory of small displacements of a compressible fluid, sound waves can likewise be shown to propagate according to (9.1). The one-dimensional form (9.4) will model the propagation of sound in an organ pipe, while the spherically symmetric version of (9.1) will give a solution for waves emanating from an explosion. Because it is known that most wave phenomena satisfy the wave equation, it is reasonable, from a physical standpoint, that the propagation of electromagnetic waves will also satisfy (9.1). A careful analysis of Maxwell's equations in free space is required to show this result (see Example 3.16). We could give further examples of physical

phenomena that have (9.1) as a basic equation, but we have described enough here to establish its importance and the need to look at methods of solution. An aspect of the wave equation that is not often discussed is its bad behaviour. Any discontinuities in a variable or its derivatives will, according to the wave equation, propagate with time. An obvious physical manifestation of this is a shock wave. When an aircraft breaks the sound barrier, a shock is produced and the sonic boom can be heard many miles away. How the shock is produced is a complicated nonlinear effect, but once it has been produced it propagates according to the wave equation.

Example 9.1

Show that

$$u = u_0 \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi c t}{L}\right)$$

satisfies the one-dimensional wave equation and the conditions

- (a) a given initial displacement $u(x, 0) = u_0 \sin(\pi x/L)$, and
- (b) zero initial velocity, $\partial u(x, 0)/\partial t = 0$.

Solution

Clearly the condition (a) is satisfied by inspection. If we now partially differentiate u with respect to t ,

$$\frac{\partial u}{\partial t} = -\frac{u_0 \pi c}{L} \sin\left(\frac{\pi x}{L}\right) \sin\left(\frac{\pi c t}{L}\right)$$

so that at $t = 0$ we have $\partial u/\partial t = 0$ and (b) is satisfied.

It remains to show that (9.4) is also satisfied. Using the standard subscript notation for partial derivatives,

$$u_{xx} = \frac{\partial^2 u}{\partial x^2} = -\frac{u_0 \pi^2}{L^2} \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi c t}{L}\right)$$

$$u_{tt} = \frac{\partial^2 u}{\partial t^2} = -\frac{u_0 \pi^2 c^2}{L^2} \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi c t}{L}\right)$$

so that the equation is indeed satisfied.

This solution corresponds physically to the fundamental mode of vibration of a taut string plucked at its centre.

Example 9.2

Verify that the function

$$u = a \exp\left[-\left(\frac{x}{h} - \frac{ct}{h}\right)^2\right]$$

satisfies the wave equation (9.4). Sketch the graphs of the solution u against x at $t = 0$, $t = 2h/c$ and $t = 4h/c$.

Solution Evaluate the partial derivatives as

$$u_x = \frac{-2a(x-ct)}{h^2} \exp\left[-\left(\frac{x}{h} - \frac{ct}{h}\right)^2\right]$$

$$u_t = \frac{2ac(x-ct)}{h^2} \exp\left[-\left(\frac{x}{h} - \frac{ct}{h}\right)^2\right]$$

and

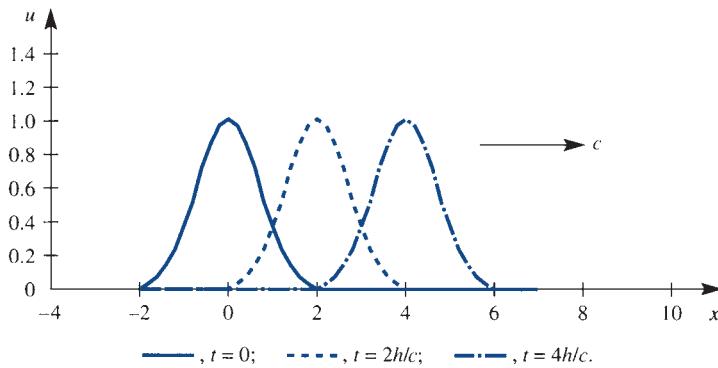
$$u_{xx} = \frac{-2a}{h^2} \exp\left[-\left(\frac{x}{h} - \frac{ct}{h}\right)^2\right] + \frac{4a(x-ct)^2}{h^4} \exp\left[-\left(\frac{x}{h} - \frac{ct}{h}\right)^2\right]$$

$$u_{tt} = \frac{-2ac^2}{h^2} \exp\left[-\left(\frac{x}{h} - \frac{ct}{h}\right)^2\right] + \frac{4a(x-ct)^2 c^2}{h^4} \exp\left[-\left(\frac{x}{h} - \frac{ct}{h}\right)^2\right]$$

Clearly (9.4) is satisfied by these second derivatives.

The curves of u against x are plotted in Figure 9.2, and show a wave initially centred at the origin moving with a constant speed c to the right. That is, the solution represents a wave travelling to the right.

Figure 9.2
Propagating wave
in Example 9.2.



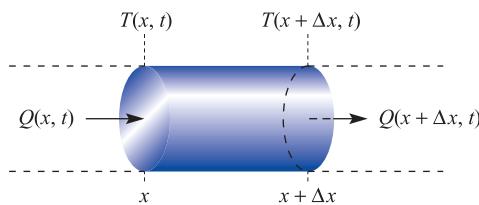
9.2.2 Heat-conduction or diffusion equation

$$\boxed{\frac{1}{\kappa} \frac{\partial u}{\partial t} = \nabla^2 u} \quad (9.5)$$

This equation arises most commonly when heat is transferred from a hot area to a cold one by conduction, when the temperature satisfies (9.5), where $\nabla^2 = \nabla \cdot \nabla$ is the Laplacian of the scalar point function $u(x, t)$, (see Chapter 3, equation (3.20)).

In Section 3.6 a full derivation of the equation (9.5) is made. Here we shall investigate the one-dimensional version in the context of the heat flow along a thin bar. The bar is assumed to have a uniform cross-sectional area and an insulated outer surface through which no heat is lost. It is also assumed that, at any cross-section $x = \text{constant}$, the temperature $T(x, t)$ is uniform. Consider an element of the bar from x to $x + \Delta x$, where x is measured along the length of the bar, as illustrated in Figure 9.3. An amount of heat $Q(x, t)$ per unit time per unit area enters the left-hand face and an

Figure 9.3 Heat flow in an element.



amount $Q(x + \Delta x, t)$ leaves the right-hand face of the element. The net increase per unit cross-sectional area in unit time is

$$Q(x, t) - Q(x + \Delta x, t)$$

If c is the specific heat of the bar and ρ is its density then the amount of heat in the element is $c\rho T \Delta x$. The net increase in heat in the element in unit time is

$$c\rho \frac{\partial T}{\partial t} \Delta x$$

and is equated to the net amount entering. Thus

$$c\rho \frac{\partial T}{\partial t} \Delta x = Q(x, t) - Q(x + \Delta x, t)$$

which in the limit as $\Delta x \rightarrow 0$ gives

$$c\rho \frac{\partial T}{\partial t} = -\frac{\partial Q}{\partial x} \quad (9.6)$$

The **Fourier law** for the conduction of heat states that the heat transferred across unit area is proportional to the temperature gradient. Thus

$$Q = -k \frac{\partial T}{\partial x}$$

where k is the thermal conductivity and the minus sign takes into account the fact that heat flows from hot to cold. Substitution for Q in (9.6) gives the **one-dimensional heat equation**

$$\frac{\partial T}{\partial t} = \kappa \frac{\partial^2 T}{\partial x^2} \quad (9.7)$$

where $\kappa = k/c\rho$ is called the **thermal diffusivity**.

An entirely similar derivation for the diffusion equation can be made. The only difference is that the Fourier law is replaced by **Fick's law** that the diffusional flow of a material is proportional to the concentration gradient.

The equations describing more complicated phenomena, such as the time-dependent electromagnetic equations or the equations of fluid mechanics, have the same basic structure as (9.5), but with additional terms or with coupling to other equations of the same type. We certainly need to know how to solve (9.5) before even contemplating solving these more complex versions.

An essential feature of the heat-conduction equation is that, given a long enough time and assuming that there are no time-varying inputs, the temperature will eventually settle down to a steady state. Thus the final solution is independent of time, and hence

will satisfy $\partial u / \partial t = 0$ or $\nabla^2 u = 0$. The transient behaviour tells how this solution is approached from its given starting value. Physically it is reasonable that any initial temperature, however complicated, will move to a smooth final solution, and we should not expect the severe difficulties with discontinuities that occur with the wave equation. Exactly how initial discontinuities are treated in a numerical solution, however, can affect the accuracy in the early development of the solution.

Example 9.3

Show that

$$T = T_\infty + (T_m - T_\infty) e^{-U(x-Ut)/\kappa} \quad (x \geq Ut)$$

satisfies the one-dimensional heat-conduction equation (9.7), together with the boundary conditions $T \rightarrow T_\infty$ as $x \rightarrow \infty$ and $T = T_m$ at $x = Ut$.

Solution

The second term vanishes as $x \rightarrow \infty$, for any fixed t , and hence $T \rightarrow T_\infty$. When $x = Ut$, the exponential term is unity, so the T_∞ s cancel and $T = T_m$. Hence the two boundary conditions are satisfied. Checking both sides of the heat-conduction equation (9.7),

$$\begin{aligned} \frac{1}{\kappa} \frac{\partial T}{\partial t} &= \frac{1}{\kappa} (T_m - T_\infty) \frac{U^2}{\kappa} e^{-U(x-Ut)/\kappa} \\ \frac{\partial^2 T}{\partial x^2} &= (T_m - T_\infty) \frac{U^2}{\kappa^2} e^{-U(x-Ut)/\kappa} \end{aligned}$$

which are obviously equal, so that the equation is satisfied.

The example models a block of material being melted at a temperature T_m , with the melting boundary having constant speed U , and with a steady temperature T_∞ at great distances. An application of this model would be a heat shield on a re-entry capsule ablated by frictional heating.

Example 9.4

Show that the function

$$T = \frac{1}{\sqrt{t}} \exp\left(-\frac{x^2}{4\kappa t}\right)$$

satisfies the one-dimensional heat-conduction equation (9.7). Plot T against x for various times t , and comment.

Solution

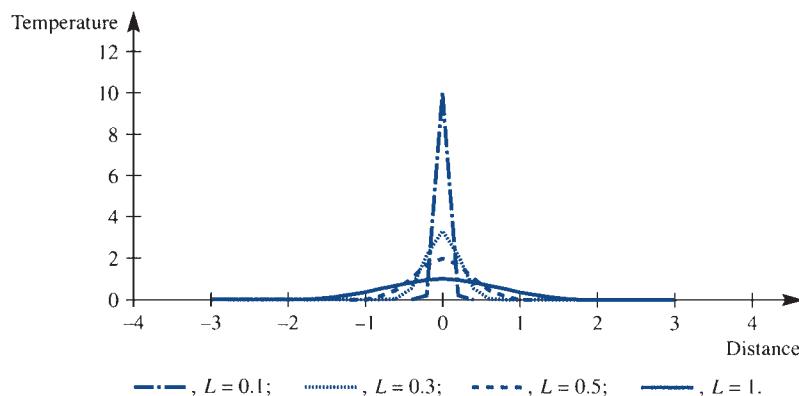
We first calculate the partial derivatives

$$\begin{aligned} \frac{\partial T}{\partial t} &= -\frac{1}{2} \frac{1}{t^{3/2}} \exp\left(-\frac{x^2}{4\kappa t}\right) + \frac{1}{\sqrt{t}} \frac{-x^2}{4\kappa} \frac{-1}{t^2} \exp\left(-\frac{x^2}{4\kappa t}\right) \\ \frac{\partial T}{\partial x} &= \frac{1}{\sqrt{t}} \frac{-2x}{4\kappa t} \exp\left(-\frac{x^2}{4\kappa t}\right) \end{aligned}$$

and

$$\frac{\partial^2 T}{\partial x^2} = \frac{-1}{2\kappa t^{3/2}} \exp\left(-\frac{x^2}{4\kappa t}\right) + \frac{-x}{2\kappa t^{3/2}} \frac{-2x}{4\kappa t} \exp\left(-\frac{x^2}{4\kappa t}\right)$$

Figure 9.4 Solution of the heat-conduction equation starting from an initial spike in Example 9.4.



It is easily checked that (9.7) is satisfied except at the time $t = 0$, where T is not properly defined. The graph of reduced temperature $T/\sqrt{4\kappa t}$ against distance x at various times $t = L^2/4\kappa$ can be seen in Figure 9.4. Physically, the problem corresponds to a very hot weld being applied instantaneously to the bar. The initial temperature ‘spike’ at $x = 0$ is seen to spread out as time progresses, and, as expected from the physical interpretation, T tends to zero for all x as the time becomes large. Alternatively the problem describes the diffusion of a large pulse of contaminant into a thin tube of fluid.

9.2.3 Laplace equation

$$\nabla^2 u = 0 \quad (9.8)$$

The simplest physical interpretation of this equation has already been mentioned, namely as the steady-state heat equation. So, for example, the two-dimensional Laplace equation

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$$

could represent the steady-state distribution of temperature over a thin rectangular plate in the (x, y) plane.

Heat transfer is well understood intuitively, and good guesses at steady-state solutions can usually be made. Perhaps less commonly understood is the case of the electrostatic potential in a uniform dielectric, which also satisfies the Laplace equation. Working out the electrical behaviour of a capacitor that is charged in a certain manner simply implies solving (9.8) subject to appropriate boundary conditions. Possibly the least obvious, but extremely important, application of the Laplace equation is in inviscid, irrotational fluid mechanics. To a large extent, subsonic aerodynamics is based on (9.8) as an approximate model. The lift and drag on an aerofoil in a fluid stream can be evaluated accurately from suitable solutions of this equation. It is only close to the aerofoil that viscous and rotational effects become important. Example 9.6 shows an application of the Laplace equation to inviscid fluid flow. Further details about fluid dynamics can be found in specialist texts, for example D. Acheson, *Elementary Fluid Dynamics* (Oxford, Oxford University Press, 2002).

The Laplace equation is a ‘smoother’ in the sense that it irons out peaks and troughs. Physically, the steady-state heat-conduction context tells us that if a particular point has a higher temperature than neighbouring points then heat will flow from hot to cold until the ‘hot spot’ is eliminated. Thus there are no interior points at which the solution u of (9.8) is smaller or larger than all of its neighbours. This result can be confirmed mathematically, and establishes that smooth solutions are obtained, see Section 9.7.1.

Example 9.5

Show that

$$u = x^4 - 2x^3y - 6x^2y^2 + 2xy^3 + y^4$$

satisfies the Laplace equation.

Solution Differentiating

$$u_x = 4x^3 - 6x^2y - 12xy^2 + 2y^3, \quad u_y = -2x^3 - 12x^2y + 6xy^2 + 4y^3$$

$$u_{xx} = 12x^2 - 12xy - 12y^2, \quad u_{yy} = -12x^2 + 12xy + 12y^2$$

so clearly

$$u_{xx} + u_{yy} = 0$$

and the two-dimensional Laplace equation is satisfied.

Example 9.6

Show that the function

$$\psi = Uy \left(1 - \frac{a^2}{x^2 + y^2} \right)$$

satisfies the Laplace equation, and sketch the curves $\psi = \text{constant}$.

Solution First calculate the partial derivatives:

$$\psi_x = \frac{2xyUa^2}{(x^2 + y^2)^2}$$

$$\psi_y = U - \frac{Ua^2}{x^2 + y^2} + \frac{2y^2Ua^2}{(x^2 + y^2)^2}$$

$$\psi_{xx} = \frac{2yUa^2}{(x^2 + y^2)^2} - \frac{8x^2yUa^2}{(x^2 + y^2)^3}$$

$$\psi_{yy} = \frac{2yUa^2}{(x^2 + y^2)^2} + \frac{4yUa^2}{(x^2 + y^2)^2} - \frac{8y^3Ua^2}{(x^2 + y^2)^3}$$

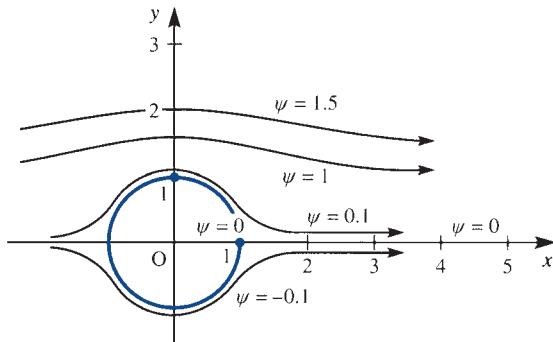
Substituting into (9.8) gives

$$\nabla^2\psi = \frac{8yUa^2}{(x^2 + y^2)^2} - \frac{8y(x^2 + y^2)Ua^2}{(x^2 + y^2)^3} = 0$$

and hence the Laplace equation is satisfied.

Figure 9.5

Streamlines for flow past a cylinder of radius 1, from the Laplace equation in Example 9.6.



Secondly, to sketch the contours, we note that $\psi = 0$ on $y = 0$ and on the circle $x^2 + y^2 = a^2$. On keeping $y = y_0$ and letting $x \rightarrow \pm\infty$, the second term vanishes, so the curves tend to $\psi = Uy_0$. Figure 9.5 shows the solution, which corresponds physically to the flow of an inviscid, irrotational fluid past a cylinder placed in a uniform stream.



Computer packages can verify the differentiations and the plotting in any of the examples in this section. For instance, the MAPLE instructions

```
psi:=U*y*(1-a^2/(x^2+y^2));
diff(psi,x,x); diff(psi,y,y); simplify(%+%%);
```

verify the Laplace equation in Example 9.6. The plotting in Figure 9.5 can be achieved from the instructions

```
with(plots):
g:=y*(1-1/(x^2+y^2));
aa:=[g=0.001,g=0.1,g=1,g=1.5];
implicitplot({seq(aa[i],i=1..4)},x=-2..4,y=0..2,
scaling=constrained);
```

9.2.4 Other and related equations

We discussed in Section 9.1 applications in science and engineering. Many such applications are governed by equations that are closely related to the three basic equations discussed above. For example, consider the equations of slow, steady, viscous flow in two dimensions, which take the form

$$\left. \begin{aligned} \frac{\partial p}{\partial x} &= \frac{1}{\mathcal{R}} \nabla^2 u, & \frac{\partial p}{\partial y} &= \frac{1}{\mathcal{R}} \nabla^2 v \\ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0, \end{aligned} \right\} \quad (9.9)$$

where u , v and p are the non-dimensional velocities and pressure, and \mathcal{R} is the **Reynolds number** (see the book by Acheson mentioned above). The system has a familiar look about it, and indeed a little simple manipulation gives $\nabla^2 p = 0$, so that the pressure satisfies the Laplace equation. If p can be calculated then $\partial p / \partial x$ and $\partial p / \partial y$ are known, so we have equations of the form

$$\nabla^2 u = f(x, y) \quad (9.10)$$

This equation is called the **Poisson equation**, and is clearly closely related to the Laplace equation. It can be interpreted physically as steady heat conduction with heat sources in the region. A careful study of the solution of the Laplace equation is required before either (9.10) or (9.9) can be attacked. In Sections 9.5 and 9.7 some discussions of the Poisson equation take place.

If there is good knowledge about the time behaviour of the wave or diffusion equation then we can often obtain important information from them without solving the full equations. For instance, if we put a periodic time dependence $u = e^{j\omega t}v(x, y, z)$ into (9.1), or if we put an exponentially decaying solution $u = e^{-\beta t}v(x, y, z)$ into (9.5), then the variable v , in both cases, satisfies an equation of the form

$$\nabla^2 v + \lambda v = 0 \quad (9.11)$$

Equation (9.11) is called the **Helmholtz equation**, and plays an important role in the solution of eigenvalue problems. It is perhaps of relevance that the best studied eigenvalue equation, the **Schrödinger equation**, is almost the same, namely

$$\frac{\hbar^2}{8\pi^2 m} \nabla^2 u - V(x, y, z)u + Eu = 0$$

It is a bit more complicated than (9.11), but it forms the basis of quantum mechanics, on which whole industries are built.

So far, all of the equations that we have considered are *linear*, since they have not included any quadratic (or higher) terms in u or its derivatives. As soon as we move from linear to *nonlinear* problems, a whole new crop of theoretical and computational difficulties arises. Very few such equations can be solved analytically, and devising computational schemes is not easy. Even worse, mathematicians cannot always tell whether or not a solution even exists. An act of faith is usually made by scientists and engineers that their problem is modelled correctly and therefore there must be a mathematical solution reflecting the physics. Often the faith is well founded, but modelling is an imperfect art and there are many things that can go wrong. It may be thought that nonlinear problems do not occur in practice, but this is certainly not the case. For some phenomena, like the behaviour of thermionic valves or avalanche semiconductors or pulsed lasers, it is the nonlinearity that produces the desired effects. Other situations arise where the nonlinearity of the system may or may not be important. For instance, the full two-dimensional fluid equations are the Navier–Stokes equations

$$\left. \begin{aligned} \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} &= -\frac{\partial p}{\partial x} + \frac{1}{\mathcal{R}} \nabla^2 u \\ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} &= -\frac{\partial p}{\partial y} + \frac{1}{\mathcal{R}} \nabla^2 v \\ \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0 \end{aligned} \right\} \quad (9.12)$$

Using the vector notation of Chapter 3, equations (9.12) can be written

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} = -\nabla p + \frac{1}{\mathcal{R}} \nabla^2 \mathbf{u}$$

where u , v , p and \mathcal{R} are defined as for (9.9). These equations are *nonlinear* because of the presence of quadratic convection terms such as $u \partial u / \partial x$. It can be seen that (9.12) reduce to (9.9) for slow flow when quadratic terms are neglected. While (9.9) would be applicable to the flow of molten glass, we would need the full equations (9.12) to look

at flow close to an aerofoil. Indeed, as \Re becomes large, the flow becomes turbulent, that is unstable, and the applicability of these equations comes into question.

9.2.5 Arbitrary functions and first-order equations

In each of the examples in this section, a solution has been given; it has been checked that the solution satisfies the appropriate partial differential equation. In no case has the boundary condition been part of the specification of the problem, although in several cases boundary conditions were checked. In the next sections the boundary conditions are given as part of the set-up of the example. This is the natural way that a physical problem is specified and it proves to be a much tougher proposition.

The most significant difference between ordinary and partial differential equations is the treatment of the ‘arbitrary constants’. Consider the examples:

ODE

Solve the ordinary differential equation

$$\frac{dy(t)}{dt} = 3t^2$$

Integrating gives

$$y(t) = t^3 + K$$

where K is an **arbitrary constant**, since differentiating $y(t)$ with respect to t produces $3t^2$ whatever the value of the constant K .

PDE

Solve the partial differential equation

$$\frac{\partial z(x, t)}{\partial t} = 3t^2$$

Integrating gives

$$z(x, t) = t^3 + f(x)$$

where $f(x)$ is an **arbitrary function**. Differentiating with respect to t produces $3t^2$ for any function $f(x)$ because x is kept constant in the partial differentiation.

Extending this idea it can be seen that each partial integration introduces an arbitrary function into the solution. Sufficient conditions must be given to determine these arbitrary functions. It is not always easy to decide exactly what conditions are required, but in subsequent sections an idea will be given for the three classic equations, the wave equation, the heat-conduction equation and the Laplace equation. An extended discussion can be found in Section 9.8.

Consider for the moment a first-order equation. Such equations are of less interest in applications to engineering and science, but there is a comprehensive theory for their solution which will illustrate the use of arbitrary functions.

Example 9.7

Find the general solution, $u(x, t)$, of the partial differential equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0$$

and find the particular solution when $u(x, 0) = x^2$.

Solution

Change the variables $z = x - t$ and $T = t$ and use the chain rule to evaluate the terms in the equation

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial z} \frac{\partial z}{\partial t} + \frac{\partial u}{\partial T} \frac{\partial T}{\partial t} = -\frac{\partial u}{\partial z} + \frac{\partial u}{\partial T}$$

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial z} \frac{\partial z}{\partial x} + \frac{\partial u}{\partial T} \frac{\partial T}{\partial x} = \frac{\partial u}{\partial z}$$

Putting these differentials into the equation

$$0 = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = \frac{\partial u}{\partial T}$$

Thus $u(z, T)$ can be deduced as

$$u(z, T) = f(z), \text{ where } f \text{ is an arbitrary function}$$

Reverting to the original variables

$$u(x, t) = f(x - t)$$

and a general solution of the partial differential equation has been obtained.

For the particular solution with initial conditions written in parametric form, $x = s$, $t = 0$, $u = s^2$, it is easily deduced that $s^2 = f(s)$ and hence

$$u(x, t) = (x - t)^2$$

The solution of quasi-linear first-order equations with two variables, x and y , is comparatively straightforward

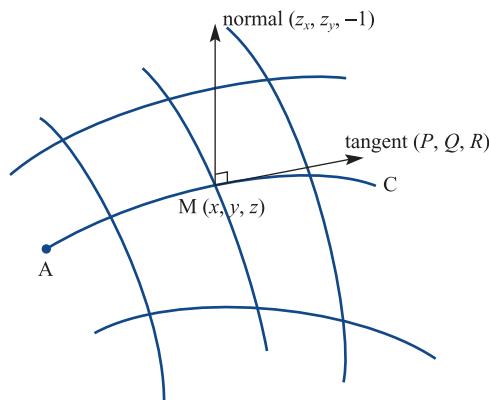
$$P(x, y, z) \frac{\partial z}{\partial x} + Q(x, y, z) \frac{\partial z}{\partial y} = R(x, y, z) \quad (9.13)$$

Provided P , Q and R are ‘well behaved’ a method of solution can be deduced, although the resulting integrals cannot always be obtained explicitly. Extension to many variable problems is similar, but the geometrical interpretation is more difficult.

In Section 3.2.1 it was seen that the function $z = f(x, y)$, illustrated in Figure 9.6, has a normal $(\partial z / \partial x, \partial z / \partial y, -1)$ at a typical point M , having coordinates (x, y, z) . Equation (9.13) says that the normal to the surface is perpendicular to the vector (P, Q, R) at the point M and thus (P, Q, R) must lie in the tangent plane. Now examine the curve C in the surface starting at the point A and moving along C in a direction that is always parallel to (P, Q, R) at the current point. The direction therefore remains perpendicular to the normal $(\partial z / \partial x, \partial z / \partial y, -1)$ at all points and must move in a tangential direction to the surface; such curves are called **characteristic curves**. The point must remain in the surface and this tangential direction (dx, dy, dz) must therefore be parallel to (P, Q, R) so that

$$\frac{dx}{P} = \frac{dy}{Q} = \frac{dz}{R} \quad (9.14)$$

Figure 9.6 Surface $z = f(x, y)$ showing the tangent, normal and characteristic curve C .



Starting from (9.13) we have shown that $z(x, y)$ can be obtained from the two ordinary differential equations (9.14). If we start from (9.14) we know that the normal direction $(\partial z / \partial x, \partial z / \partial y, -1)$ is perpendicular to the tangent vector (dx, dy, dz) and hence perpendicular to (P, Q, R) so

$$P(x, y, z) \frac{\partial z}{\partial x} + Q(x, y, z) \frac{\partial z}{\partial y} - R(x, y, z) = 0$$

and (9.13) is satisfied.

From a particular starting point, $x = a, y = b, z = c$, the solution of (9.13) is obtained as the characteristic curve obtained from the solution of the ordinary differential equations (9.14). Usually there is a starting curve; then essentially the process calculates the characteristic curve from each point of the starting curve and the solution surface is generated.

To illustrate the method return to Example 9.7 when the equations (9.14) become

$$\frac{dt}{1} = \frac{dx}{1} = \frac{du}{0}$$

using the variables given. The two ordinary differential equations are

$$\frac{dx}{dt} = 1 \quad \text{with solution } x - t = A$$

$$\frac{du}{dt} = 0 \quad \text{with solution } u = B$$

The constants A and B are arbitrary and are determined from a given initial data point. Usually the initial data is given on a curve $t = 0, x = f(s), y = g(s)$, so for each s there are arbitrary constants A and B ; in this case, the constants depend on s , that is $A(s)$ and $B(s)$. In the current example (9.7) the initial data is $t = 0, x = s, u = s^2$ giving

$$s = A \quad \text{so} \quad x - t = s$$

$$s^2 = B \quad \text{so} \quad u = s^2$$

Eliminating s gives $u = (x - t)^2$ as deduced earlier. A further example shows how the method is applied.

Example 9.8

Solve the equation

$$x \frac{\partial z}{\partial x} + y \frac{\partial z}{\partial y} = xy$$

for $z(x, y)$ given that $z = f(s)$ when $x = s$ and $y = 1 - s$.

Solution

The two ordinary equations obtained from (9.14) are

$$\frac{dx}{x} = \frac{dy}{y} \quad \text{and} \quad \frac{dx}{x} = \frac{dz}{xy} \tag{9.15a,b}$$

Solving (9.15a) gives $\ln x = \ln y + C$ which reduces to $x = Ay$. Putting this result into (9.15b) gives

$$xdx = Adz$$

which on solving gives

$$\frac{1}{2}x^2 = Az + B$$

To obtain the arbitrary constants A and B we insert the initial conditions

$$A = \frac{x}{y} = \frac{s}{1-s} \quad (9.16a)$$

$$B = \frac{1}{2}x^2 - \frac{x}{y}z = \frac{1}{2}s^2 - \frac{s}{1-s}f(s) \quad (9.16b)$$

Thus A and B have been obtained in terms of s . From (9.16a) we get

$$\frac{x}{y} = \frac{s}{1-s} \quad \text{and hence} \quad s = \frac{x}{x+y}$$

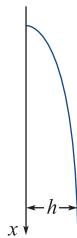
So that (9.16b) becomes

$$\frac{1}{2}x^2 - \frac{x}{y}z = \frac{1}{2}\left(\frac{x}{x+y}\right)^2 - \frac{x}{y}f\left(\frac{x}{x+y}\right)$$

Rearranging, z is then calculated as

$$z = \frac{1}{2}xy - \frac{1}{2}\frac{xy}{(x+y)^2} + f\left(\frac{x}{x+y}\right)$$

The solution can be checked by substitution into the original differential equation.



 Because there is a comprehensive theory of the solution of some classes of first-order partial differential equations computer packages can be used to solve these equations with comparative ease. The MAPLE instructions

```
with(PDEtools):
PDE:=x*diff(z(x,y),x) + y*diff(z(x,y),y) - x*y;
pdsolve(PDE);
```

produce the general solution in Example 9.8.

Figure 9.7 Draining of liquid down the side of the vessel of Example 9.9.

Example 9.9

A practical example of first-order equations involves the draining of liquid from a vessel, a procedure common to many industrial processes. The thickness of the liquid layer is required as time progresses.

A thin layer of liquid drains down the side of a vessel, as illustrated in Figure 9.7. From the theory of thin layers, the equation for the fluid motion is given by

$$\frac{\partial h}{\partial t} + ah^2 \frac{\partial h}{\partial x} = 0$$

where $h(x, t)$ is the thickness of the layer and a is a constant that depends on the viscosity, density and the gravity constant. Find the solution for $h(x, t)$ given the initial condition $h(x, 0) = \alpha\sqrt{x}$ where α is a constant.

Solution The ordinary differential equations (9.14) are

$$\frac{dt}{1} = \frac{dx}{ah^2} = \frac{dh}{0}$$

Clearly $h = C$, a constant, solves one of the equations and the other is

$$\frac{dx}{dt} = ah^2 = aC^2 \quad \text{with solution } x = atC^2 + K$$

where C and K are arbitrary constants. Thus the solution of the equation is

$$\begin{cases} C = h \\ K = x - ath^2 \end{cases}$$



Using MAPLE to try for a solution, the instructions

```
with(PDEtools):
drain:=diff(h(x,t),t)+a*h(x,t)^2*diff(h(x,t),x);
pdsolve(drain);
```

give h as a solution of the equation

$$f(h) = x - ath^2$$

Note that the package has combined the arbitrary constants C and K into one arbitrary function f , determined by the initial data.

Clearly the package can go no further without the specification of the initial condition. Putting in the conditions $x = s$, $t = 0$, $h = \alpha\sqrt{s}$ gives f as

$$f(\alpha\sqrt{s}) = s \quad \text{or} \quad f(p) = \left(\frac{p}{\alpha}\right)^2$$

The function h can now be calculated as

$$h(x, t) = \alpha \sqrt{\frac{x}{1 + a\alpha^2 t}}$$

and shows that, for large t , the layer thins at a rate proportional to $t^{-\frac{1}{2}}$. The solution can be checked by direct substitution into the original equations. A plot of h against x at successive times or a three-dimensional plot of $h(x, t)$ using `PDEplot` in MAPLE can be used to illustrate the solution.



The MATLAB instructions that produce the surface shape, h/α , at successive times $a\alpha^2 t = 0, 2, 4, 8$ are

```
y=0:0.1:4;
X=[sqrt(y)', sqrt(y/3)', sqrt(y/5)', sqrt(y/9)'];
plot(X,y)
```

Further applications of first-order equations occur in the study of the time evolution of the probability distribution of the position of a particle, for instance in Brownian motion. The equation is

$$\frac{\partial f(x, t)}{\partial t} + \frac{\partial}{\partial x}[D_r(x, t)f(x, t)] = \frac{\partial^2}{\partial x^2}[D_f(x, t)f(x, t)]$$

where D_r and D_f are drift and diffusion coefficients. If diffusion can be neglected then the equation is just a first-order partial differential equation.

In gas dynamics and also in traffic flow problems a similar equation, the **Burgers' equation**, can be shown to apply

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = v \frac{\partial^2 u}{\partial x^2}$$

In the two situations u is the gas velocity or the traffic density and v is a viscosity coefficient. For the inviscid case, again the equation reduces to a first-order partial differential equation. The derivation of these equations is lengthy and beyond the scope of this text but can be found in specialist books.

The solution of inviscid Burgers' equation is obtained from

$$\frac{dt}{1} = \frac{dx}{u} = \frac{du}{0}$$

The two equations give one obvious solution $u = A$ and the second is

$$\frac{dx}{dt} = u = A \quad \text{and hence } x = At + B$$

Taking initial conditions $t = 0$, $u = V(s)$ and $x = s$ we obtain

$$A = u = V(s)$$

$$B = x - ut = s$$

Eliminating s gives the solution for $u(x, t)$ in implicit form

$$V(x - ut) = u$$

9.2.6 Exercises

- 1 Find the possible values of a and b in the expression

$$u = \cos at \sin bx$$

such that it satisfies the wave equation

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

- 2 Taking

$$u = f(x + \alpha t)$$

where f is any function, find the values of α that will ensure that u satisfies the wave equation

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

- 3 Verify that the function

$$u(x, y) = x^4 - 6x^2y^2 + y^4$$

satisfies the Laplace equation.

- 4 The function $z(r, t)$ depends only on the radial distance in spherical polar coordinates and on the time. The wave equation in this coordinate system is

$$\frac{\partial^2 z}{\partial r^2} + \frac{2}{r} \frac{\partial z}{\partial r} = \frac{1}{c^2} \frac{\partial^2 z}{\partial t^2}$$

Show that $z(r, t) = r^{-1} \cos(r - ct)$ satisfies the equation ($r \neq 0$).

- 5 Find all the possible solutions of the heat-conduction equation

$$\frac{1}{\kappa} \frac{\partial u}{\partial r} = \frac{\partial^2 u}{\partial x^2}$$

of the form

$$u(x, t) = e^{\alpha t} V(x)$$

- 6 Find the values of the constant n for which

$$V = r^n(3 \cos^2 \theta - 1)$$

satisfies the Laplace equation (in spherical polar coordinates and independent of ϕ)

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

for all values of the variables r and θ .

- 7 Show that $u = e^{-kt} \cos mx \cos nt$ is a solution of the equation

$$c^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2} + 2k \frac{\partial u}{\partial t}$$

provided that the constants k, m, n and c are related by the equation $n^2 + k^2 = c^2 m^2$.

- 8 If $V = x^3 + axy^2$, where a is a constant, show that

$$x \frac{\partial V}{\partial x} + y \frac{\partial V}{\partial y} = 3V$$

Find the value of a if V is to satisfy the equation

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0$$

Taking this value of a , show that if $u = r^3 V$, where $r^2 = x^2 + y^2$, then

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 27rV$$

- 9 The **telegraph equation** has the form

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{1}{c^2} \left(\frac{\partial^2 \phi}{\partial t^2} + k \frac{\partial \phi}{\partial t} \right)$$

where c^2 is the speed of light and k is usually small. Given that $\Phi(x, t)$ is a solution of the wave equation

$$\frac{\partial^2 \Phi}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 \Phi}{\partial t^2}$$

show that $\phi(x, t) = \Phi(x, t) e^{-kt/2}$ is a solution of the telegraph equation, if terms of order k^2 can be neglected.

- 10 The transmission-line equations represent the flow of current along a long, leaky wire such as a transatlantic cable. The equations take the form

$$-\frac{\partial I}{\partial x} = gv + c \frac{\partial v}{\partial t}$$

$$-\frac{\partial v}{\partial x} = rI + L \frac{\partial I}{\partial t}$$

where g, c, r and L are constants and I and v are the current and voltage respectively.

- (a) Show that when $r = g = 0$, the equations reduce to the wave equation.
- (b) Show that when $L = 0$, the equations reduce to a heat-conduction equation with a forcing term. Write $W = v e^{gtx/c}$ to reduce to the normal form of the equation.

- (c) Put $a = \frac{1}{2}(r/L + g/c)$ and then $w = v e^{at}$. Show that when $rc = gL$, w satisfies the wave equation.

Show that if f is a function of x only then

$$u = f(x) \sin(ay + b)$$

where a and b are constants, is a solution of the partial differential equation

$$\frac{\partial^2 u}{\partial y^2} = \frac{\partial^2 u}{\partial x^2} - 2a \frac{\partial u}{\partial x}$$

provided that $f(x)$ satisfies the ordinary differential equation

$$\frac{d^2 f}{dx^2} - 2a \frac{df}{dx} + a^2 f = 0$$

Hence show that

$$u = (A + Bx) e^{ax} \sin(ay + b)$$

where A and B are arbitrary constants, is a solution of the partial differential equation.

- 12 Show that $f(x, y) = x^2 y^2 + g(x/y)$ satisfies the partial differential equation

$$x \frac{\partial f}{\partial x} + y \frac{\partial f}{\partial y} = 4x^2 y^2$$

for any arbitrary function g . It is given that $f = t^2$ on the line with parametric equation $x = 1 - t$, $y = t$; find the function g .

- 13 Show that the partial differential equation

$$\frac{\partial^2 u}{\partial x \partial y} + \frac{\partial u}{\partial x} = 0$$

has the general solution

$$u(x, y) = e^{-y}[f(x) + g(y)]$$

where f and g are arbitrary functions.

- 14 Find the general solution for $u(x, y)$ in the equation (check using MAPLE)

$$x^2 \frac{\partial u}{\partial x} + y^2 \frac{\partial u}{\partial y} = (x + y)u$$

Show that the solution that satisfies the conditions $u = s^2$, $x = s$, $y = 1$ takes the form

$$u = \frac{x^2 y^2}{xy - x + y}$$

9.3

Solution of the wave equation

In this section we consider methods of solving the wave equation introduced in Section 9.2.1.

9.3.1 D'Alembert solution and characteristics

A classical solution of the one-dimensional wave equation

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \quad (9.4)$$

is obtained by changing the axes to reduce the equation to a particularly simple form. Let

$$r = x + ct, \quad s = x - ct$$

Then, using the chain rule procedure for transformation of coordinates (see Section 3.1.1),

$$u_{xx} = u_{rr} + 2u_{rs} + u_{ss}$$

$$u_{tt} = c^2(u_{rr} - 2u_{rs} + u_{ss})$$

so that the wave equation (9.4) becomes

$$4c^2 u_{rs} = 0$$

This equation can now be integrated once with respect to s to give

$$u_r = \frac{\partial u}{\partial r} = \theta(r)$$

where θ is an arbitrary function of r . Now, integrating with respect to r , we obtain

$$u = f(r) + g(s)$$

which, on substituting for r and s , gives the solution of the wave equation (9.4) as

$$u = f(x + ct) + g(x - ct) \quad (9.17)$$

where f and g are *arbitrary functions* and f is just the integral of the arbitrary function θ .

The solution (9.17) is one of the few cases where the general solution of a partial differential equation can be found. However, finding the precise form of the arbitrary functions f and g that satisfy given initial data is not always easy. The initial conditions must give just enough information to evaluate f and g , which are functions of the *single* variables $r = x + ct$ and $s = x - ct$ respectively.

In Example 9.2 we have already seen a simple example of a wave of this type. We first deduced that a function of $x - ct$ satisfied the wave equation, and then showed in Figure 9.2 that it represented a wave travelling in the x direction with velocity c .

The next example is similar.

Example 9.10

Check that $u = 1/[1 + (x + ct)^2]$ satisfies the wave equation (9.4) and show that it represents a travelling wave in the $-x$ direction.

Solution

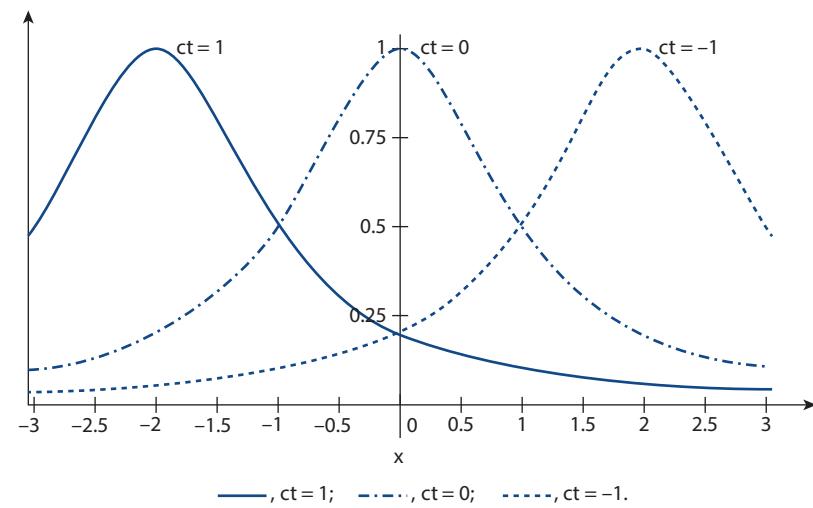
Differentiating partially with respect to x and t

$$u_x = \frac{-2(x + ct)}{[1 + (x + ct)^2]^2}, \quad u_{xx} = \frac{2[-1 + 3(x + ct)^2]}{[1 + (x + ct)^2]^3}$$

$$u_t = \frac{-2c(x + ct)}{[1 + (x + ct)^2]^2}, \quad u_{tt} = \frac{2c^2[-1 + 3(x + ct)^2]}{[1 + (x + ct)^2]^3}$$

and the wave equation is satisfied. Plots of the function u against x for various values of ct are shown in Figure 9.8. The same curve can be seen to be just translated to the left, i.e. a travelling wave.

Figure 9.8 Solution to Example 9.10 showing u against x for various values of ct .



In Example 9.11 we attempt the more difficult task of fitting initial conditions to the solution.

Example 9.11

Solve the wave equation (9.4) subject to the conditions

- (a) zero initial velocity, $\partial u(x, 0)/\partial t = 0$ for all x , and
- (b) an initial displacement given by

$$u(x, 0) = F(x) = \begin{cases} 1-x & (0 \leq x \leq 1) \\ 1+x & (-1 \leq x \leq 0) \\ 0 & \text{otherwise} \end{cases}$$

Solution

This example corresponds physically to an infinite string initially at rest, and displaced as in Figure 9.9, which is then released.

From (9.17) we have a solution of the wave equation as

$$u = f(x + ct) + g(x - ct)$$

We now fit the given boundary data. Condition (a) gives

$$0 = cf'(x) - cg'(x) \quad \text{for all } x$$

so that

$$f(x) - g(x) = K = \text{an arbitrary constant}$$

and thus

$$u = f(x + ct) + f(x - ct) - K$$

Similarly, condition (b) gives

$$F(x) = 2f(x) - K$$

so that

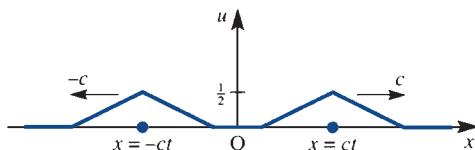
$$u = \frac{1}{2}F(x + ct) + \frac{1}{2}F(x - ct) \quad (9.18)$$

We now have the solution to the equation in terms of the function F defined in condition (b). (Note that the same is true for any function F .)

The solution is plotted in Figure 9.10 as u against x for given times. It may be observed from this example that we have two **travelling waves**, one propagating to the right and one to the left. The initial shape is propagated exactly, except for a factor of two, and the shape discontinuities are not smoothed out, as noted in Section 9.2.1.

Figure 9.9 Initial displacement in Example 9.11.

Figure 9.10 Solution to Example 9.11 showing two waves propagating in the $+x$ and $-x$ directions with velocity c .



The analysis in Example 9.11 can be extended to solve the wave equation subject to the general conditions

- (a) an initial velocity, $\partial u(x, 0)/\partial t = G(x)$, and
- (b) an initial displacement, $u(x, 0) = F(x)$ for all x .

Condition (a) gives, from (9.17),

$$G(x) = c[f'(x) - g'(x)]$$

so that

$$c[f(x) - g(x)] = \int_0^x G(x) dx + Kc$$

Condition (b) gives

$$f(x) + g(x) = F(x)$$

and we can solve for $f(x)$ and $g(x)$ as

$$f(x) = \frac{1}{2}F(x) + \frac{1}{2c} \int_0^x G(x) dx + \frac{1}{2}K$$

$$g(x) = \frac{1}{2}F(x) - \frac{1}{2c} \int_0^x G(x) dx - \frac{1}{2}K$$

The solution thus becomes

$$u = \frac{1}{2}[F(x+ct) + F(x-ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} G(z) dz \quad (9.19)$$

which is commonly called the **d'Alembert solution**. As in Examples 9.2, 9.10 and 9.11, it gives rise to waves propagating in the $+x$ and $-x$ directions.

As mentioned in Section 9.1, a difficulty is to illustrate the solution of a partial differential equation in a simple way. Figure 9.10 is a ‘snapshot’ at a particular time t , and if we wish to look at the solution over all (x, t) then we have to draw u as a function of the two variables x and t . We can draw the solution to Example 9.11 in a three-dimensional diagram as in Figure 9.11, but for any higher-dimensional problem such a diagram is clearly impossible. The ‘snapshot’ in Figure 9.10 corresponds to a plane slice parallel to the (u, x) plane.

The d'Alembert solution, which reduces to an integral along the boundary, does not have any simple extension other than for the x axis. In Section 9.7.2 the Green's function is introduced and it can be interpreted as an extension since it involves integrals round the boundary of a general region. However, the calculation of the Green's function is a tough proposition for any but the simplest regions. Following from the idea of the d'Alembert solution, **characteristics** (which will be studied in the next few paragraphs) can be used to extend the range of boundaries that can be dealt with.

The idea of using an (x, t) plane is a very useful one for the wave equation, since the solution

$$u = f(x+ct) + g(x-ct)$$

gives a representation by characteristics. If we plot the lines $x+ct = \text{constant}$ and $x-ct = \text{constant}$ as in Figure 9.12 then we see that the line AP has equation $x-ct = x_0$ and the line BP has equation $x+ct = x_1$. Thus

on the whole of AP $g(x-ct) = g(x_0)$

on the whole of BP $f(x+ct) = f(x_1)$

Figure 9.11 Solution to Example 9.11 in (x, t, u) space.

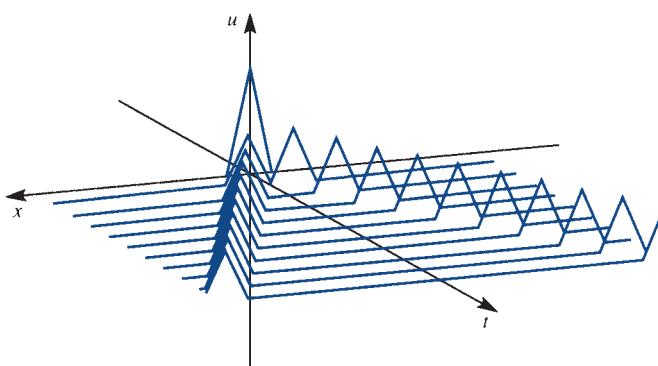
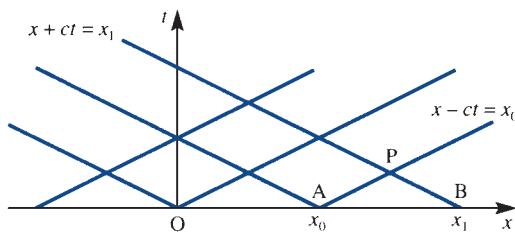


Figure 9.12

Characteristics
 $x + ct = \text{constant}$ and
 $x - ct = \text{constant}$.



Thus g takes a constant value on AP and f takes a constant value on BP. If we can calculate f and g on the initial line $t = 0$ then we know the value of u at P, namely

$$u(P) = f(x_1) + g(x_0) \quad (9.20)$$

Since P is an arbitrary point, the solution at any point would be known. The essential problem is to calculate $f(x)$ and $g(x)$ on the line $t = 0$.

Typical conditions on $t = 0$ are

- (a) $u(x, 0) = F(x)$, and
- (b) $\partial u(x, 0)/\partial t = G(x)$,

which specify the initial position and velocity of the system. Now

$$c \frac{\partial u}{\partial x} + \frac{\partial u}{\partial t} = cf'(x + ct) + cg'(x - ct) + cf'(x + ct) - cg'(x - ct) = 2cf'(x + ct)$$

and similarly

$$c \frac{\partial u}{\partial x} - \frac{\partial u}{\partial t} = 2cg'(x - ct)$$

On $t = 0$ we know that $\partial u/\partial x = F'(x)$ and $\partial u/\partial t = G(x)$, so we can deduce that

$$cF'(x) + G(x) = 2cf'(x)$$

$$cF'(x) - G(x) = 2cg'(x)$$

Since F and G are given, we can compute

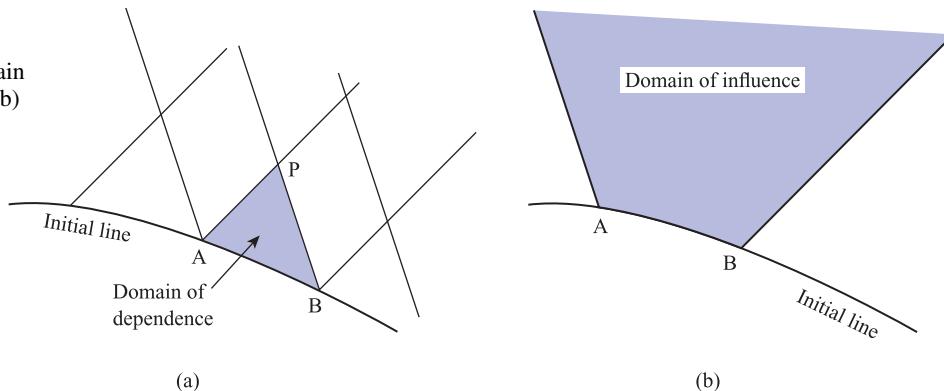
$$f'(x) = \frac{1}{2}[F'(x) + G(x)/c]$$

$$g'(x) = \frac{1}{2}[F'(x) - G(x)/c]$$

and hence $f(x)$ and $g(x)$ can be computed by straightforward integration.

This method is essentially the same as the d'Alembert method, but it concentrates on calculating $f(x)$ and $g(x)$ on the initial line and then constructing the solution at P by the characteristics AP and BP. The method gives great insight into the behaviour of the solution of such equations, but it is not an easy technique to use in practice. Perhaps the best that can be obtained from characteristics is an idea of how the solution depends on the initial data. In Figure 9.13 the characteristics emanating from the initial line are drawn. To evaluate the solution at P, we must have information on the section of the initial line AB, and the rest of the initial line is irrelevant to the solution at P. This is called the **domain of dependence**. The section of the initial line AB has a **domain of influence** determined by the characteristics through the points A and B. The data on AB cannot influence the solution outside the shaded region in Figure 9.13(b).

Figure 9.13
Characteristics showing (a) the domain of dependence and (b) the domain of influence.



Example 9.12

Use characteristics to compute the solution of the one-dimensional wave equation (9.4), with speed $c = 1$, given (a) the initial conditions that $u = V(x)$ and $\partial u / \partial t = 0$, for $x > 0$ and $t = 0$, and (b) the boundary condition that $u = 0$ at $x = 0$ and $t > 0$. Describe the solutions in the particular cases

$$(i) \quad V(x) = 1 \quad \text{and} \quad (ii) \quad V(x) = \frac{x}{x^3 + 1}.$$

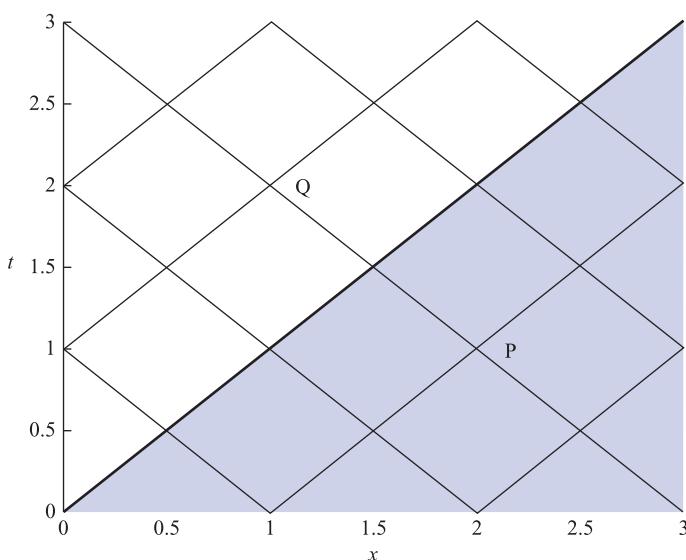
Solution

The characteristics are plotted in Figure 9.14. It can be seen that for $x > t$, at a typical point P two characteristics emanating from the initial line, $t = 0$, meet and the solution can be computed at P from data on the initial line. However, for $x < t$, at a typical point Q the characteristics emanating from the boundary, $x = 0$, are required. These observations will be borne out in the mathematical computations.

For the region $x > t$ the characteristic analysis described in the previous analysis can be followed. It was shown that f and g in the solution

$$u = f(x - t) + g(x + t) \quad (9.21)$$

Figure 9.14
Characteristics for Example 9.12.



can be calculated, taking $G(x) = 0$, as

$$f'(x) = \frac{1}{2} V'(x) \quad \text{and} \quad g'(x) = \frac{1}{2} V'(x)$$

Integrating, and putting the arbitrary constant to be zero, gives

$$f(x) = \frac{1}{2} V(x) \quad \text{and} \quad g(x) = \frac{1}{2} V(x) \quad \text{for } x > 0$$

and hence the solution

$$u(x, t) = \frac{1}{2} [V(x+t) + V(x-t)] \quad \text{for } x > t \quad (9.22)$$

For the region $x < t$, (9.21) requires $f(z)$ at negative values of z . The function is not yet known for negative values and must be determined by the other condition (b) on $x = 0$. The condition $u = 0$ at $x = 0$ and $t > 0$ implies in (9.21)

$$0 = f(-t) + g(t)$$

and hence for a general variable z

$$f(-z) = -g(z) = -\frac{1}{2} V(z) \quad \text{for } z > 0$$

Using this result, (9.21) gives the required solution

$$u(x, t) = \frac{1}{2} [V(x+t) - V(t-x)] \quad \text{for } x > t \quad (9.23)$$

The complete solution for all $x > 0$ and $t > 0$ is now known from (9.22) and (9.23).

Case (i)

In this case $V(x) = 1$ so (9.22) gives $u = 1$, in the shaded region of Figure 9.14, and (9.23) gives $u = 0$, in the unshaded region of Figure 9.14. Thus

$$u(x, t) = \begin{cases} 1 & \text{for } x > t \\ 0 & \text{for } x < t \end{cases}$$

Note that the discontinuity in the boundary data at $x = 0$, $t = 0$ is propagated along the characteristic $x = t$.

Case (ii)

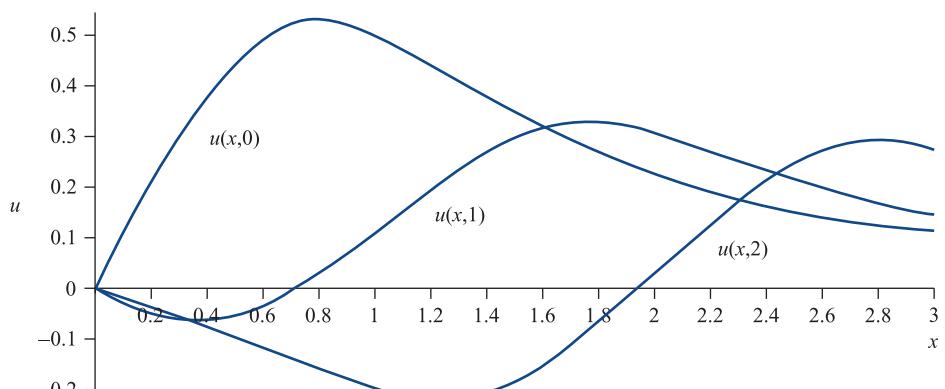
Putting the function $V(x) = x/(x^3 + 1)$ into (9.22) and (9.23) gives the solution

$$u(x, t) = \begin{cases} \frac{1}{2} \frac{x-t}{(x-t)^3 + 1} + \frac{1}{2} \frac{x+t}{(x+t)^3 + 1} & \text{for } x > t \\ \frac{1}{2} \frac{x-t}{(t-x)^3 + 1} + \frac{1}{2} \frac{x+t}{(x+t)^3 + 1} & \text{for } x < t \end{cases}$$

The boundary data are now smooth so the function $u(x, t)$ remains smooth as illustrated for three cases in Figure 9.15.

The basic physical problem described in this example is a very long string held at one end and initially at rest. The string is then displaced at $t = 0$ in the shape of the function $V(x)$ and released.

Figure 9.15
Smooth solutions for Example 9.12, Case (ii); string displacements at various times.



In more complicated problems, the evaluation of the arbitrary functions f and g in equation (9.17) and the use of characteristics is no longer straightforward. We do not have a d'Alembert type of solution; a great deal of thought and care is needed.

The idea of characteristics can be applied to more general classes of second-order partial differential equations. Example 9.13 illustrates a case of a constant-coefficient equation.

Example 9.13

Find the characteristics of the equation

$$0 = u_{xx} + 2u_{xt} + 2\alpha u_{tt}$$

Study the case when $\alpha = \frac{3}{8}$ and the solution satisfies the boundary conditions

(a) $\partial u(x, 0)/\partial t = 0$ for $x \geq 0$

(b) $u(x, 0) = F(x) = \begin{cases} 1 & (0 < x < 1) \\ 0 & (x \geq 1) \end{cases}$

(c) $u(0, t) = 0$ for all t

(d) $\partial u(0, t)/\partial x = 0$ for all t

Solution

Since the coefficients of the equation are constants, we know that the characteristics are straight lines, so we look for solutions of the form

$$u = u(x + at)$$

Putting $z = x + at$ and writing $u' = du/dz$ and so on, we obtain

$$0 = u_{xx} + 2u_{xt} + 2\alpha u_{tt} = (1 + 2a + 2a^2\alpha)u''$$

Hence for a solution we require

$$1 + 2a + 2a^2\alpha = 0$$

or

$$a = \frac{-1 \pm \sqrt{(1 - 2\alpha)}}{2\alpha}$$

If $\alpha > \frac{1}{2}$ then the two values of a (a_1 and a_2 say) are complex, and the characteristics $x + a_1 t = \text{constant}$ and $x + a_2 t = \text{constant}$ do not make sense in the real plane.

If $\alpha = \frac{1}{2}$ then both roots give $a = 1$, and we only have a single characteristic $x + t = \text{constant}$, which is not useful for further computation.

For the case $\alpha < \frac{1}{2}$, we find two real values for a and two sets of characteristics.

It is precisely for this reason that characteristics serve no useful purpose for the heat-conduction or Laplace equations. A further discussion can be found in Section 9.8 after the formal classification of equations has been completed.

Take the case $\alpha = \frac{3}{8}$; then we obtain $a_1 = -2$ and $a_2 = -\frac{2}{3}$, so the solution has the form

$$u = f(x - 2t) + g(x - \frac{2}{3}t)$$

where f and g are arbitrary functions and the characteristics are the straight lines $x - 2t = \text{constant}$ and $x - \frac{2}{3}t = \text{constant}$.

The boundary conditions given in the problem are a little more complicated than in the d'Alembert solution. Conditions (a) and (b) give

$$\left. \begin{array}{l} 0 = \frac{\partial u(x, 0)}{\partial t} = -2f'(x) - \frac{2}{3}g'(x) \\ F(x) = u(x, 0) = f(x) + g(x) \end{array} \right\} \quad (x \geq 0)$$

Taking $f(0) = g(0) = 0$, we can integrate the first of these expressions and then solve for $f(x)$ and $g(x)$ on the line $t = 0$ as

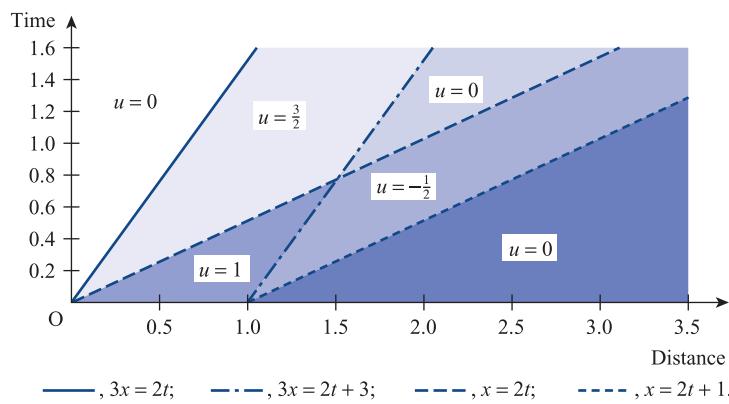
$$f(x) = -\frac{1}{2}F(x), \quad g(x) = \frac{3}{2}F(x) \quad (x \geq 0)$$

Conditions (c) and (d) say that $u(0, t) = 0$, and $\partial u(0, t)/\partial x = 0$. Thus on the line $x = 0$ we deduce

$$f(z) = g(z) = 0 \quad (z < 0)$$

We can now construct the solution by characteristics. Figure 9.16 illustrates this solution. Because $f(x)$ and $g(x)$ are constant along the respective characteristics, we deduce $u(A) = 0$, $u(B) = -\frac{1}{2}$, $u(C) = 1$, $u(D) = 0$, $u(E) = \frac{3}{2}$, $u(F) = 0$ at typical points in the six regions that divide up the first quadrant of the (x, t) plane.

Figure 9.16
Characteristic solution of Example 9.13.
The solution u takes the constant values shown in the six regions of the first quadrant.



For non-constant-coefficient equations the characteristics are not usually straight lines, which causes computational difficulties. In particular, there are some fundamental problems when characteristics of the same family intersect. The solution loses its uniqueness, and ‘shocks’ can be generated. The classical wave equation (9.4) will propagate these shocks, but it requires ‘curved characteristics’ to generate them.

9.3.2 Separation of variables

A method of considerable importance is the **method of separation of variables**. The basis of the method is to attempt to look for solutions $u(x, y)$ of a partial differential equation as a product of functions of single variables

$$u(x, y) = X(x)Y(y)$$

The advantage of this approach is that it is sometimes possible to find X and Y as solutions of *ordinary* differential equations. These are very much easier to solve than partial differential equations, and it may be possible to build up solutions of the full equation in terms of the solutions for X and Y . A simple example illustrates the general strategy. Suppose that we wish to solve

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 0$$

Then we should write $u = X(x)Y(y)$ and substitute into the above equation to obtain

$$Y \frac{dX}{dx} + X \frac{dY}{dy} = 0, \quad \text{or} \quad \frac{1}{X} \frac{dX}{dx} = -\frac{1}{Y} \frac{dY}{dy}$$

Note that the partial differentials become ordinary differentials, since the functions are just functions of a single variable. Now

$$\text{LHS} = \frac{1}{X} \frac{dX}{dx} = \text{a function of } x \text{ only}$$

$$\text{RHS} = -\frac{1}{Y} \frac{dY}{dy} = \text{a function of } y \text{ only}$$

Since LHS = RHS for all x and y , the only way that this can be achieved is for each side to be a *constant*. We thus have two ordinary differential equations

$$\frac{1}{X} \frac{dX}{dx} = \lambda, \quad -\frac{1}{Y} \frac{dY}{dy} = \lambda$$

These equations can be solved easily as

$$X = B e^{\lambda x}, \quad Y = C e^{-\lambda y}$$

and thus the solution of the original partial differential equation is

$$u(x, y) = X(x)Y(y) = A e^{\lambda(x-y)}$$

where $A = BC$. The constants A and λ are arbitrary. The crucial question is whether the boundary conditions imposed by the problem can be satisfied by a sum of solutions of this type.

The method of separation of variables can be a very powerful technique, and we shall see it used on all three of the basic partial differential equations. It should be noted, however, that not all equations have separable solutions, see Example 9.2, and even

when they have it is not always possible to satisfy the boundary conditions with such solutions.

In the case of the heat-conduction equation and the wave equation, the form of one of the functions in the separated solution is dictated by the physics of the problem. We shall see that the separation technique becomes a little simpler when such physical arguments are used. However, for the Laplace equation there is no help from the physics, so the method just described needs to be applied.

In most wave equation problems we are looking for either a travelling-wave solution as in Section 9.3.1 or for periodic solutions, as a result of plucking a violin string for instance. It therefore seems natural to look for specific solutions that have periodicity built into them. These will not be general solutions, but they will be seen to be useful for a whole class of problems. The essential mathematical simplicity of the method comes from only having to solve ordinary differential equations.

The above argument suggests that we seek solutions of the wave equation

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} \quad (9.4)$$

of the form either

$$u = \sin(c\lambda t)v(x) \quad (9.24a)$$

or

$$u = \cos(c\lambda t)v(x) \quad (9.24b)$$

both of which when substituted into (9.4) give the ordinary differential equation

$$\frac{d^2 v}{dx^2} = -\lambda^2 v$$

This is a simple harmonic equation with solutions $v = \sin \lambda x$ or $v = \cos \lambda x$. We can thus build up a general solution of (9.4) from linear multiples of the four basic solutions

$$u_1 = \cos \lambda c t \sin \lambda x \quad (9.25a)$$

$$u_2 = \cos \lambda c t \cos \lambda x \quad (9.25b)$$

$$u_3 = \sin \lambda c t \sin \lambda x \quad (9.25c)$$

$$u_4 = \sin \lambda c t \cos \lambda x \quad (9.25d)$$

and try to satisfy the boundary conditions using appropriate linear combinations of solutions of this type. We saw an example of such a solution in Example 9.1.

Example 9.14

Solve the wave equation (9.4) for the vibration of a string stretched between the points $x = 0$ and $x = l$ and subject to the boundary conditions

- (a) $u(0, t) = 0$ ($t \geq 0$) (fixed at the end $x = 0$);
- (b) $u(l, t) = 0$ ($t \geq 0$) (fixed at the end $x = l$);
- (c) $\partial u(x, 0)/\partial t = 0$ ($0 \leq x \leq l$) (with zero initial velocity);
- (d) $u(x, 0) = F(x)$ (given initial displacement).

Consider the two cases

$$(i) \quad F(x) = \sin(\pi x/l) + \frac{1}{4} \sin(3\pi x/l)$$

$$(ii) \quad F(x) = \begin{cases} x & (0 \leq x \leq \frac{1}{2}l) \\ l-x & (\frac{1}{2}l \leq x \leq l) \end{cases}$$

Solution Clearly, we are solving the problem of a stretched string, held at its ends $x = 0$ and $x = l$ and released from rest.

By inspection, we see that the solutions (9.25b, d) cannot satisfy condition (a). We see that condition (b) is satisfied by the solutions (9.25a, c), provided that

$$\sin \lambda l = 0, \quad \text{or} \quad \lambda l = n\pi \quad (n = 1, 2, 3, \dots)$$

It may be noted that only specific values of λ in (9.25) give permissible solutions. Thus the string can only vibrate with given frequencies, $nc/2l$. The solution (9.25) appropriate to this problem takes the form either

$$u = \cos\left(\frac{nc\pi t}{l}\right) \sin\left(\frac{n\pi x}{l}\right) \quad (9.26a)$$

or

$$u = \sin\left(\frac{nc\pi t}{l}\right) \sin\left(\frac{n\pi x}{l}\right) \quad (9.26b)$$

($n = 1, 2, 3, \dots$). To satisfy condition (c) for all x , we must choose the solution (9.26a) and omit (9.26b). Clearly, it is not possible to satisfy the initial condition (d) with (9.26a). However, because the wave equation is linear, any *sum* of such solutions is also a solution. Thus we build up a solution

$$u = \sum_{n=1}^{\infty} b_n \cos\left(\frac{nc\pi t}{l}\right) \sin\left(\frac{n\pi x}{l}\right) \quad (9.27)$$

Case (i)

The initial condition (d) for $u(x, 0)$ gives

$$\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{l}\right) = \sin\left(\frac{\pi x}{l}\right) + \frac{1}{4} \sin\left(\frac{3\pi x}{l}\right)$$

and the values of b_n can be evaluated by inspection as

$$b_1 = 1, b_2 = 0, b_3 = \frac{1}{4}, b_4 = b_5 = \dots = 0$$

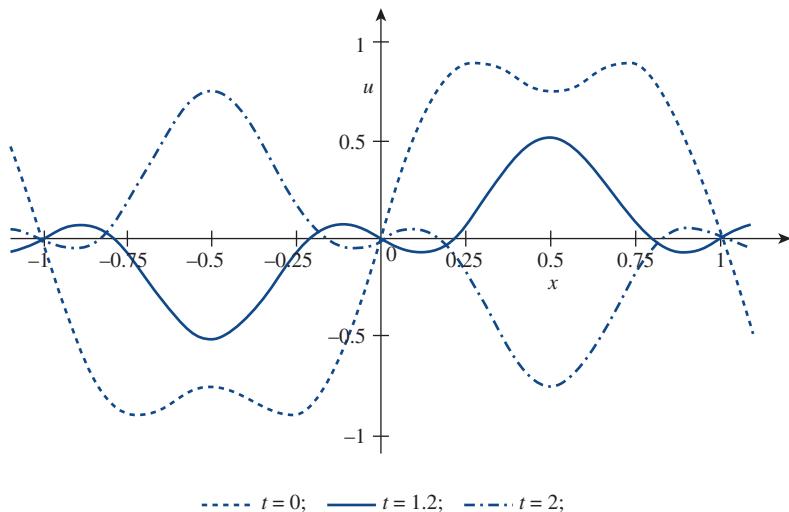
The full solution is therefore

$$u = \cos\left(\frac{\pi ct}{l}\right) \sin\left(\frac{\pi x}{l}\right) + \frac{1}{4} \cos\left(\frac{3\pi ct}{l}\right) \sin\left(\frac{3\pi x}{l}\right)$$

The solution is illustrated in Figure 9.17.

Figure 9.17

Sketch of the solution to Example 9.14 (i) with $c = \frac{1}{3}$ and $l = 1$.



Case (ii)

The condition (d) for $u(x, 0)$ simply gives

$$\sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi x}{l}\right) = f(x) = \begin{cases} x & (0 \leq x \leq \frac{1}{2}l) \\ l-x & (\frac{1}{2}l \leq x \leq l) \end{cases}$$

and thus to determine b_n we must find the Fourier sine series expansion of the function $f(x)$ over the finite interval $0 \leq x \leq l$. We have from (7.17) that

$$\begin{aligned} b_n &= \frac{2}{l} \int_0^l f(x) \sin\left(\frac{n\pi x}{l}\right) dx \\ &= \frac{2}{l} \int_0^{l/2} x \sin\left(\frac{n\pi x}{l}\right) dx + \frac{2}{l} \int_{l/2}^l (l-x) \sin\left(\frac{n\pi x}{l}\right) dx \\ &= \frac{4l}{\pi^2 n^2} \sin\left(\frac{1}{2}n\pi\right) \quad (n = 1, 2, 3, \dots) \end{aligned}$$

The complete solution of the wave equation in this case is therefore

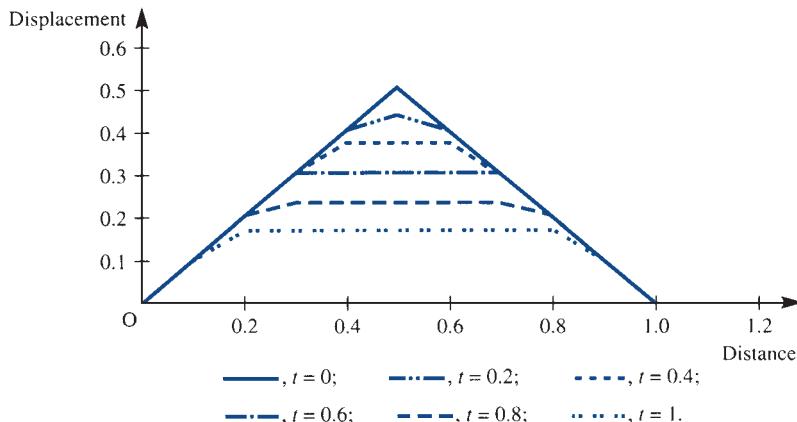
$$u(x, t) = \frac{4l}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \sin\left(\frac{1}{2}n\pi\right) \cos\left(\frac{nc\pi t}{l}\right) \sin\left(\frac{n\pi x}{l}\right) \quad (9.28)$$

or

$$\begin{aligned} u(x, t) &= \frac{4l}{\pi^2} \left[\cos\left(\frac{c\pi t}{l}\right) \sin\left(\frac{\pi x}{l}\right) - \frac{1}{9} \cos\left(\frac{3c\pi t}{l}\right) \sin\left(\frac{3\pi x}{l}\right) \right. \\ &\quad \left. + \frac{1}{25} \cos\left(\frac{5c\pi t}{l}\right) \sin\left(\frac{5\pi x}{l}\right) + \dots \right] \end{aligned}$$

The complete solution to Example 9.14 Case (ii) gives some very useful information. We see that all the even ‘harmonics’ have disappeared from the solution and the amplitudes of the harmonics decrease like $1/n^2$. A beautiful theory of musical instruments can be built up from such solutions. We see that for different instruments different

Figure 9.18 Solution of Example 9.14 (ii) with $c = \frac{1}{3}$ and $l = 1$.



harmonics are important and have different amplitudes. It is this that gives an instrument its characteristic sound. Convergence of Fourier series is not, in general, easy to prove, but suffice to say that since $f(x)$ is a bounded periodic function with a finite number of optima and discontinuities, it has pointwise convergence.

A sensible question that we can ask is whether we can use the sum of the series in (9.28) to plot u , see E. C. Zachmanoglou and O. W. Thoe *Introduction to Partial Differential Equations with Applications* (New York, Dover, 2003). At discontinuities in particular, Fourier series can be very slow to converge, so that, although (9.28) is a complete solution, does it provide us with any useful information? In the present case there is no particular problem, but the general comment should be noted. The solution (9.28) is plotted in Figure 9.18 with $l = 1$ and $c = \frac{1}{3}$. We note that even with 10 terms $u(0.5, 0) = 0.4899$ instead of the correct value 0.5, so there is a 2% error in the calculated value. Perhaps the most pertinent comment that we can make is that a good number of terms in the series are required to obtain a solution, and exact solutions may not be as useful as we might expect.

Separated solutions depend on judicious use of the known solutions (9.25) of the wave equation to fit the boundary conditions. Although it is not always possible to solve any particular problem using separated solutions, the idea is sufficiently straightforward that it is always worth a try. The extension to other equations and coordinate systems is possible. The use of other orthogonal functions was introduced in Section 7.5 and some of these will be discussed in Sections 9.4 and 9.5.

Example 9.15

Solve the wave equation (9.4) for vibrations in an organ pipe subject to the boundary conditions

- $u(0, t) = 0$ ($t \geq 0$) (the end $x = 0$ is closed);
- $\partial u(l, t)/\partial x = 0$ ($t \geq 0$) (the end $x = l$ is open);
- $u(x, 0) = 0$ ($0 \leq x \leq l$) (the pipe is initially undisturbed);
- $\partial u(x, 0)/\partial t = v = \text{constant}$ ($0 \leq x \leq l$) (the pipe is given an initial uniform blow).

Solution

From the solution (9.25), we deduce from condition (a) that solutions (9.25b, d) must be omitted, and similarly from condition (c) that solution (9.25a) is not useful. We are left with the solution (9.25c) to satisfy the boundary condition (b). This can only be satisfied if

$$\cos \lambda l = 0, \quad \text{or} \quad \lambda l = (n + \frac{1}{2})\pi \quad (n = 0, 1, 2, \dots)$$

Thus we obtain solutions of the form

$$u = b_n \sin\left[\frac{(n + \frac{1}{2})\pi ct}{l}\right] \sin\left[\frac{(n + \frac{1}{2})\pi x}{l}\right] \quad (n = 0, 1, 2, \dots)$$

giving a general solution

$$u = \sum_{n=0}^{\infty} b_n \sin\left[\frac{(n + \frac{1}{2})\pi ct}{l}\right] \sin\left[\frac{(n + \frac{1}{2})\pi x}{l}\right]$$

The condition (d) gives

$$v = \sum_{n=0}^{\infty} b_n \frac{(n + \frac{1}{2})\pi c}{l} \sin\left[\frac{(n + \frac{1}{2})\pi x}{l}\right]$$

which, on using (7.17) to obtain the coefficients of the Fourier sine series expansion of the constant v over the finite interval $0 \leq x \leq l$, gives

$$b_n = \frac{2v}{(n + \frac{1}{2})\pi} \frac{l}{(n + \frac{1}{2})\pi c} = \frac{8lv}{\pi^2 c} \frac{1}{(2n + 1)^2}$$

Our complete solution of the wave equation is therefore

$$u = \frac{8lv}{\pi^2 c} \sum_{n=0}^{\infty} \frac{1}{(2n + 1)^2} \sin\left[(n + \frac{1}{2})\pi \frac{ct}{l}\right] \sin\left[(n + \frac{1}{2})\pi \frac{x}{l}\right]$$

or,

$$\begin{aligned} u = & \frac{8lv}{\pi^2 c} \left[\sin\left(\frac{\pi ct}{2l}\right) \sin\left(\frac{\pi x}{2l}\right) + \frac{1}{9} \sin\left(\frac{3\pi ct}{2l}\right) \sin\left(\frac{3\pi x}{2l}\right) \right. \\ & \left. + \frac{1}{25} \sin\left(\frac{5\pi ct}{2l}\right) \sin\left(\frac{5\pi x}{2l}\right) + \dots \right] \end{aligned}$$

It would be instructive to compute this solution and compare it with Figure 9.18, which corresponds to the solution of Example 9.14.

9.3.3 Laplace transform solution

For linear problems that are time-varying from 0 to ∞ , as in the case of the wave equation, Laplace transforms provide a formal method of solution. The only difficulty is whether the final inversion can be performed.

First we obtain the Laplace transforms of the partial derivatives

$$\frac{\partial u}{\partial x}, \quad \frac{\partial u}{\partial t}, \quad \frac{\partial^2 u}{\partial x^2}, \quad \frac{\partial^2 u}{\partial t^2}$$

of the function $u(x, t)$, $t \geq 0$. Using the same procedure as that used to obtain the Laplace transform of standard derivatives in Section 11.3.1 in *Modern Engineering Mathematics* (MEM), we have the following:

$$(a) \quad \mathcal{L}\left\{\frac{\partial u}{\partial x}\right\} = \int_0^\infty e^{-st} \frac{\partial u}{\partial x} dt = \frac{d}{dx} \int_0^\infty e^{-st} u(x, t) dt$$

using Leibniz's rule (see MEM) for differentiation under an integral sign. Noting that

$$\mathcal{L}\{u(x, t)\} = U(x, s) = \int_0^\infty e^{-st} u(x, t) dt$$

we have

$$\mathcal{L}\left\{\frac{\partial u}{\partial x}\right\} = \frac{d}{dx} U(x, s) \quad (9.29)$$

(b) Writing $y(x, t) = \partial u / \partial x$, repeated application of the result (9.29) gives

$$\mathcal{L}\left\{\frac{\partial y}{\partial x}\right\} = \frac{d}{dx} \mathcal{L}\{y(x, t)\} = \frac{d}{dx} \left(\frac{d}{dx} U(x, s) \right)$$

so that

$$\mathcal{L}\left\{\frac{\partial^2 u}{\partial x^2}\right\} = \frac{d^2 U(x, s)}{dx^2} \quad (9.30)$$

$$(c) \quad \begin{aligned} \mathcal{L}\left\{\frac{\partial u}{\partial t}\right\} &= \int_0^\infty e^{-st} \frac{\partial u}{\partial t} dt \\ &= [e^{-st} u(x, t)]_0^\infty + s \int_0^\infty e^{-st} u(x, t) dt = [0 - u(x, 0)] + sU(x, s) \end{aligned}$$

so that

$$\mathcal{L}\left\{\frac{\partial u}{\partial t}\right\} = sU(x, s) - u(x, 0) \quad (9.31)$$

where we have assumed that $u(x, t)$ is of exponential order.

(d) Writing $v(x, t) = \partial u / \partial t$, repeated application of (9.31) gives

$$\begin{aligned} \mathcal{L}\left\{\frac{\partial v}{\partial t}\right\} &= sV(x, s) - v(x, 0) \\ &= s[sU(x, s) - u(x, 0)] - v(x, 0) \end{aligned}$$

so that

$$\mathcal{L}\left\{\frac{\partial^2 u}{\partial t^2}\right\} = s^2 U(x, s) - su(x, 0) - u_t(x, 0) \quad (9.32)$$

where $u_t(x, 0)$ denotes the value of $\partial u / \partial t$ at $t = 0$.

Let us now return to consider the wave equation (9.4)

$$c^2 \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}$$

subject to the boundary conditions $u(x, 0) = f(x)$ and $\partial u(x, 0)/\partial t = g(x)$. Taking Laplace transforms on both sides of (9.4) and using the results (9.30) and (9.32) gives

$$c^2 \frac{d^2 U(x, s)}{dx^2} = s^2 U(x, s) - g(x) - sf(x) \quad (9.33)$$

The problem has thus been reduced to an ordinary differential equation in $U(x, s)$ of a straightforward type. It can be solved for given conditions at the ends of the x range, and the solution can then be inverted to give $u(x, t)$.



MAPLE or MATLAB can be used to assist with the transforms and the inverse transforms but considerable experience is needed to convert results to a simple form.

Example 9.16

Solve the wave equation (9.4) for a semi-infinite string by Laplace transforms, given that

- (a) $u(x, 0) = 0$ ($x \geq 0$) (string initially undisturbed);
- (b) $\partial u(x, 0)/\partial t = x e^{-x/a}$ ($x \geq 0$) (string given an initial velocity);
- (c) $u(0, t) = 0$ ($t \geq 0$) (string held at $x = 0$);
- (d) $u(x, t) \rightarrow 0$ as $x \rightarrow \infty$ for $t \geq 0$ (string held at infinity).

Solution

Using conditions (a) and (b) and substituting for $f(x)$ and $g(x)$ in the result (9.33), the transformed equation in this case is

$$c^2 \frac{d^2}{dx^2} U(x, s) = s^2 U(x, s) - x e^{-x/a}$$

By seeking a particular integral of the form

$$U = \alpha x e^{-x/a} + \beta e^{-x/a}$$

we obtain a solution of the differential equation as

$$U(x, s) = A e^{sx/c} + B e^{-sx/c} - \frac{e^{-x/a}}{c^2/a^2 - s^2} \left[x + \frac{2c^2/a}{c^2/a^2 - s^2} \right]$$

where A and B are arbitrary constants.

Transforming the given boundary conditions (c) and (d), we have $U(0, s) = 0$ and $U(x, s) \rightarrow 0$ as $x \rightarrow \infty$, which can be used to determine A and B . From the second condition $A = 0$, and the first condition then gives

$$B = \frac{2c^2/a}{(c^2/a^2 - s^2)^2}$$

so that the solution becomes

$$U(x, s) = \frac{2c^2/a}{(c^2/a^2 - s^2)^2} e^{-sx/c} - \frac{e^{-x/a}}{(c^2/a^2 - s^2)} \left[x + \frac{2c^2/a}{(c^2/a^2 - s^2)} \right]$$

Fortunately in this case these transforms can be inverted from tables of Laplace transforms.

Using the second shift theorem (5.7) together with the Laplace transform pairs

$$\mathcal{L}\{\sinh \omega t\} = \frac{\omega}{s^2 - \omega^2}, \quad \mathcal{L}\{\cosh \omega t\} = \frac{s}{s^2 - \omega^2}$$

$$\mathcal{L}\left\{\frac{\omega t \cosh \omega t - \sinh \omega t}{2\omega^3}\right\} = \frac{1}{(s^2 - \omega^2)^2}$$

we obtain the solution as

$$u = \frac{a}{c} \left[(ct - x) \cosh\left(\frac{ct - x}{a}\right) H(ct - x) - ct e^{-x/a} \cosh\left(\frac{ct}{a}\right) \right] \\ + \frac{a}{c} \left[e^{-x/a} (x + a) \sinh\left(\frac{ct}{a}\right) - a \sinh\left(\frac{ct - x}{a}\right) H(ct - x) \right]$$

where $H(t)$ is the Heaviside step function defined in Section 5.2.1.

9.3.4 Exercises

- 15 Solve the wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}$$

subject to the initial conditions

(a) $u(x, 0) = \sin x$ (all x)

(b) $\frac{\partial u}{\partial t}(x, 0) = 0$ (all x)

Use both the d'Alembert solution and the separation of variables method and show that they both give the same result.

- 16 Find the separated solution of the wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}$$

that satisfies the initial conditions

$$u(x, 0) = 0, \quad \frac{\partial u}{\partial t}(x, 0) = \sin x(1 + \cos x)$$

- 17 Show that



$$2 \frac{x - ct(x^2 - c^2 t^2)}{1 - 4cxt - (x^2 - c^2 t^2)^2} = \frac{x - ct}{1 + (x - ct)^2} + \frac{x + ct}{1 - (x + ct)^2}$$

Hence deduce that the function satisfies the wave equation. Check that this differential equation is satisfied using MAPLE.

- 18

The spherically symmetric version of the wave equation (9.4) takes the form

$$\frac{1}{c^2} \frac{\partial^2 u}{\partial r^2} = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r}$$

Show, by putting $v = ru$, that it has a solution

$$ru = f(ct - r) + g(ct + r)$$

Interpret the terms as spherical waves.

- 19

Using the trigonometric identity

$$\sin A \cos B = \frac{1}{2} \sin(A - B) + \frac{1}{2} \sin(A + B)$$

rewrite the solution (9.28) to Example 9.14 as a progressive wave.

- 20

Solve the wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}$$

subject to the initial conditions

(a) $u(x, 0) = 0$ (all x)

(b) $\frac{\partial u}{\partial t}(x, 0) = x e^{-x^2}$ (all x)

- 21 Find the solutions to the wave equation (9.4) subject to the boundary conditions

(a) $\partial u(x, 0)/\partial t = 0$ for all x

(b) $u = \begin{cases} 1-x & (0 \leq x \leq 1) \\ 1+x & (-1 \leq x \leq 0) \\ 0 & (|x| \geq 1) \end{cases}$ at $t = 0$

using d'Alembert's method. Compare with Example 9.11.

- 22 Compute the characteristics of the equation

$$3u_{xx} + 6u_{xy} + u_{yy} = 0$$

- 23 Show that the partial differential equation

$$6\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial x \partial t} - \frac{\partial^2 u}{\partial t^2} = 0$$

has solutions of the form $u(x, t) = f(x + \lambda t)$ provided either $\lambda = 2$ or $\lambda = -3$. Given

$$u(x, 0) = x^2 - 1 \quad \text{and} \quad \frac{\partial u}{\partial t}(x, 0) = 2x \quad \text{for all } x$$

find the solution for u .

- 24 The function $u(r, t)$ satisfies the partial differential equation

$$\frac{\partial^2 u}{\partial r^2} + \frac{2}{r} \frac{\partial u}{\partial r} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}$$

where c is a positive constant. Show that this equation has a solution of the form

$$u = \frac{g(r)}{r} \cos \omega t$$

where ω is a constant and g satisfies

$$\frac{d^2 g}{dr^2} + \frac{\omega^2 g}{c^2} = 0$$

Show that, if u satisfies the conditions

$$u(a, t) = \beta \cos \omega t$$

$$u(b, t) = 0$$

then the solution is

$$u(r, t) = \frac{\beta a \cos \omega t \sin[\omega(b-r)/c]}{r \sin[\omega(b-a)/c]}$$

- 25 Use characteristics to compute the solution of the wave equation (9.4), with speed $c = 1$, given the initial conditions that for all x and $t = 0$

(a) $u = 0$ (b) $\partial u/\partial t = \exp(-|x|)$

Use a time step of 0.5 to compute (on a spreadsheet or other package) the first four steps over the range $-3 < x < 3$.

- 26 Using the separated solution approach of Section 9.3.2, obtain a series solution of the wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}$$

subject to the boundary conditions

- (a) $u(0, t) = 0$ ($t > 0$)
 (b) $\partial u(x, 0)/\partial t = 0$ ($0 < x < \pi$)
 (c) $u(\pi, t) = 0$ ($t > 0$)
 (d) $u(x, 0) = \pi x - x^2$ ($0 < x < \pi$)

- 27 The end at $x = 0$ of an infinitely long string, initially at rest along the x axis, undergoes a periodic displacement $a \sin \omega t$, for $t > 0$, transverse to the x axis. The displacement $u(x, t)$ of any point on the string at any time is given by the solution of the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \quad (x > 0, t > 0)$$

subject to the boundary conditions

- (a) $u(x, 0) = 0$ ($x > 0$)
 (b) $\partial u(x, 0)/\partial t = 0$ ($x > 0$)
 (c) $u(0, t) = a \sin \omega t$ ($t > 0$)
 (d) $|u(x, t)| < L$, L constant

where the last condition specifies that the displacement is bounded.

Using the Laplace transform method, show that the displacement is given by

$$u(x, t) = a \sin \left[\omega \left(t - \frac{x}{c} \right) \right] H \left(t - \frac{x}{c} \right)$$

where $H(t)$ is the Heaviside step function.

Plot a graph of $u(x, t)$, and discuss.

9.3.5 Numerical solution

For all but the simplest problems, we have to find a numerical solution. In Section 9.3.1 we saw that characteristics give a possible numerical way of working by extending the solution away from the initial line. While this method is possible, it is difficult to program except for the simplest problems, where other methods would be preferred anyway. In particular, when characteristics are curved it becomes difficult to keep track of the solution front. However, calculus methods suffer because they cannot cope with discontinuities, so that, should these occur, the methods described in this section will tend to ‘smear out’ the shocks. Characteristics provide one of the few methods that will trap the shocks when we use the fact that the latter are propagated along the characteristics.

The numerical solution of ordinary differential equations was studied in some detail in Chapter 2. The basis of the methods was to construct approximations to differentials in terms of values of the required function at discrete points. The commonest approximation was the ‘central difference approximation’

$$\frac{df(a)}{dx} \approx \frac{f(a+h) - f(a-h)}{2h}$$

and for the second derivative

$$\frac{d^2f(a)}{dx^2} \approx \frac{f(a+h) - 2f(a) + f(a-h)}{h^2}$$

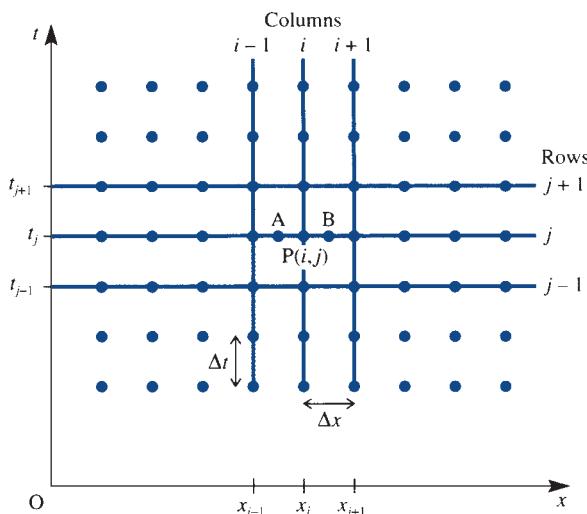
The justification of these approximations and the computation of the errors involved depend on the Taylor expansions of the functions. In partial differentiation the approximations are the same except that there is a partial derivative in both x and t for the function $u(x, t)$.

Figure 9.19 illustrates a mesh of points, or **nodes**, with spacing Δx in the x direction and Δt in the t direction. Each node is specified by a pair of integers (i, j) , so that the coordinates of the nodal points take the form

$$x_i = a + i\Delta x, \quad t_j = b + j\Delta t$$

and a and b specify the origin chosen. The mesh points or nodes lie on the intersection of the **rows** ($j = \text{constant}$) and **columns** ($i = \text{constant}$).

Figure 9.19 Mesh points for a numerical solution of the wave equation.



The approximations are applied to a typical point P, with discretized coordinates (i, j) , and with increments $\Delta x = x_{i+1} - x_i$ and $\Delta t = t_{j+1} - t_j$, which are taken to be uniform through the mesh. We know that at the points A and B we can approximate

$$\left(\frac{\partial u}{\partial x}\right)_A \simeq \frac{u(i, j) - u(i-1, j)}{\Delta x}$$

$$\left(\frac{\partial u}{\partial x}\right)_B \simeq \frac{u(i+1, j) - u(i, j)}{\Delta x}$$

so that the second derivative at P has the numerical form

$$\frac{\partial^2 u}{\partial x^2} = \frac{(\partial u / \partial x)_B - (\partial u / \partial x)_A}{\Delta x} = \frac{u(i+1, j) - 2u(i, j) + u(i-1, j)}{\Delta x^2}$$

Similarly,

$$\frac{\partial^2 u}{\partial t^2} = \frac{u(i, j+1) - 2u(i, j) + u(i, j-1)}{\Delta t^2}$$

Thus the wave equation $\partial^2 u / \partial t^2 = c^2 \partial^2 u / \partial x^2$ becomes

$$\begin{aligned} & \frac{u(i, j+1) - 2u(i, j) + u(i, j-1)}{\Delta t^2} \\ &= c^2 \frac{u(i+1, j) - 2u(i, j) + u(i-1, j)}{\Delta x^2} \end{aligned}$$

which can be rearranged as

$$\begin{aligned} u(i, j+1) &= 2u(i, j) - u(i, j-1) \\ &+ \lambda^2 [u(i+1, j) - 2u(i, j) + u(i-1, j)] \end{aligned} \quad (9.34)$$

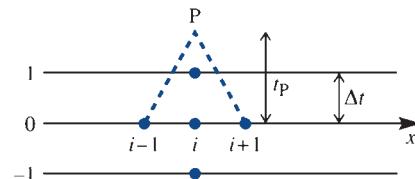
where

$$\lambda = c \Delta t / \Delta x$$

Equation (9.34) is a **finite-difference representation** of the wave equation, and provided that u is known on rows $j-1$ and j then $u(i, j+1)$ can be computed on row $j+1$ from (9.34) and thus the solution continued. On the zeroth row the boundary conditions $u(x, 0) = f(x)$ and $\partial u(x, 0) / \partial t = g(x)$ are known, so that $f_i = u(i, 0)$ and $g_i = \partial u(i, 0) / \partial t$ are also known at each node on this row, and these are used to start the process off. From Figure 9.20, we see that

$$g_i = \frac{\partial u}{\partial t} = \frac{u(i, 1) - u(i, -1)}{2\Delta t} \quad (9.35)$$

Figure 9.20 The first rows of mesh points in a numerical solution of the wave equation.



Now (9.34) with $j = 0$ becomes

$$u(i, 1) = 2u(i, 0) - u(i, -1) + \lambda^2 [u(i+1, 0) - 2u(i, 0) + u(i-1, 0)]$$

Since $u(i, 0) = f_i$ and $u(i, -1) = u(i, 1) - 2\Delta t g_i$, (9.34) now takes the form

$$u(i, 1) = (1 - \lambda^2)f_i + \frac{1}{2}\lambda^2(f_{i+1} + f_{i-1}) + \Delta t g_i \quad (9.36)$$

Thus the basic strategy is to compute row zero from $u(i, 0) = f_i$, evaluate row one from (9.36), and then march forward for general row j by (9.34).

Example 9.17

Solve the wave equation $\partial^2 u / \partial t^2 = c^2 \partial^2 u / \partial x^2$ numerically with the conditions

- (a) $u(x, 0) = \sin(\pi x)$ ($0 \leq x \leq 1$) (initial displacement);
- (b) $\partial u(x, 0) / \partial t = 0$ ($0 \leq x \leq 1$) (initially at rest);
- (c) $u(0, t) = u(1, t) = 0$ ($t \geq 0$) (the two ends held fixed).

Use the values $c = 1$, $\Delta x = 0.25$, $\Delta t = 0.1$.

Solution Note that $\lambda^2 = 0.16$. The values at $t = 0$ are given by condition (a)

x	0	0.25	0.5	0.75	1
u	0	0.7071	1	0.7071	0

The values at $t = 0.1$ (or $j = 1$) are computed from (9.36) with $f_i = \sin(\pi x)$

$$u(i, 1) = 0.84f_i + 0.08(f_{i+1} + f_{i-1})$$

and give

x	0	0.25	0.5	0.75	1
u	0	0.674	0.9531	0.674	0

The first two rows are now complete, so formula (9.34) can be used for each of the subsequent times, for $t = 0.2$ (or $j = 2$)

$$u(i, 2) = 2u(i, 1) - u(i, 0) + 0.16[u(i+1, 1) - 2u(i, 1) + u(i-1, 1)]$$

which gives

x	0	0.25	0.5	0.75	1
u	0	0.5777	0.8169	0.5777	0

and for $t = 0.3$ (or $j = 3$)

x	0	0.25	0.5	0.75	1
u	0	0.4272	0.6042	0.4272	0

and so on.

This problem has an exact solution so the results can be compared with $u(x, t) = \sin(\pi x) \cos(\pi t)$.



Numerical calculations can be performed very efficiently with MATLAB: the ‘colon’ notation allows complex manipulations of sub-matrices to be done and makes the package ideally suited to this type of computation. The instructions, for n mesh points and general parameter $L = \lambda^2$,

```

n=5;L=0.16;                                % values in the example
x=[0:1/(n-1):1]; z=sin(x*pi);    %sets up initial line
zz=[0,(1-L)*z([2:n-1])+L*(z([1:n-2])+z([3:n]))/2,0]
                                    %sets up second line
zzz=[0,2*zz([2:n-1])-z([2:n-1])+L*(zz([1:n-2])-
2*zz([2:n-1])+zz([3:n])),0]           %sets up the third line
z=zz;zz=zzz;                                % prepares for subsequent
                                                % lines

```

produce the solution to this problem. Repeating the last two lines continues the solution for t incremented by Δt .

Example 9.18

Solve the wave equation $\partial^2 u / \partial t^2 = c^2 \partial^2 u / \partial x^2$ for a semi-infinite string, given the initial conditions

- $u(x, 0) = x \exp[-5(x - 1)^2]$ ($x \geq 0$) (string given an initial displacement);
- $\partial u(x, 0) / \partial t = 0$ ($x \geq 0$) (string at rest initially);
- $u(0, t) = 0$ ($t \geq 0$) (string held at the point $x = 0$).

Solution

Since $g_i = 0$ in (9.36), only the one parameter λ needs to be specified. Figure 9.21 shows the solution of u over eight time steps with $\lambda = 0.5$. It can be seen that the solution splits into two waves, one moving in the $+x$ direction and the other in the $-x$ direction. At a given time $t = 0.8/c$, the u values are presented in the table shown in Figure 9.22 for various values of λ . We see that for $\lambda < 1$ the solution is reasonably consistent, and we have errors of a few per cent.

Figure 9.21 Solution of Example 9.18 with $\Delta x = 0.2$, $\lambda = 0.5$ for successive values of ct .

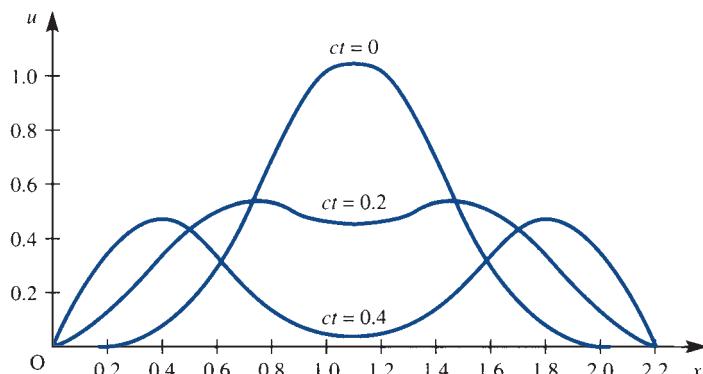


Figure 9.22

Table of values of u for a numerical solution of Example 9.18 with $\Delta x = 0.2$ and $ct = 0.8$.

x	0	0.2	0.4	0.6	0.8	1.0
$u(\lambda = 0.25)$	0	0.3451	0.4674	0.3368	0.1353	0.0236
$u(\lambda = 0.5)$	0	0.3487	0.4665	0.3318	0.1340	0.0272
$u(\lambda = 1)$	0	0.3652	0.4582	0.3105	0.1322	0.0408
$u(\lambda = 2)$	0	0.1078	0.3571	0.6334	0.5742	0.2749

However, for $\lambda = 2$ the solution looks very suspect. A further two time steps gives, at $ct = 1.6$, the solution

x	0	0.2	0.4	0.6	0.8	1
$u(\lambda = 2)$	0	-3.12	21.75	-10.25	-34.70	32.72

Clearly the solution has gone wild!

Looking back to Figure 9.20, we can attempt an explanation for the apparent divergence of the solution in Example 9.18. The characteristics through the points $(x_{i-1}, 0)$ and $(x_{i+1}, 0)$ are

$$x_{i-1} = x - ct$$

$$x_{i+1} = x + ct$$

which can be solved to give, at the point P,

$$x_P = \frac{1}{2}(x_{i+1} + x_{i-1})$$

$$ct_P = \frac{1}{2}(x_{i+1} - x_{i-1}) = \Delta x$$

Recalling the work done on characteristics, we should require the new point to be *inside* the domain of dependence defined by the interval (x_{i-1}, x_{i+1}) . Hence we require

$$t_P \geq \Delta t$$

so

$$\frac{c\Delta t}{\Delta x} \leq 1 \quad (9.37)$$

Indeed, a careful analysis, found in many specialist numerical analysis books, shows that this is precisely the condition for convergence of the method; it is commonly called the Courant, Friedrichs and Levy (CFL) condition.

The stringent condition on the time step Δt has always been considered to be a limitation on so-called **explicit methods** of the type described here, but such methods have the great merit of being very simple to program. As computers get faster, the very short time step is becoming less of a problem, and vector or array processors allow nodes to be dealt with simultaneously, thus making such methods even more competitive.

There are, however, clear advantages in the stability of calculations if an **implicit method** is used. In Figure 9.19 the approximation to u_{xx} may be formed by the average of the approximations from rows $j + 1$ and $j - 1$. Thus

$$\begin{aligned}
 & [u(i, j+1) - 2u(i, j) + u(i, j-1)]/c^2\Delta t^2 \\
 & = \frac{1}{2} [u(i+1, j+1) - 2u(i, j+1) + u(i-1, j+1) \\
 & \quad + u(i+1, j-1) - 2u(i, j-1) + u(i-1, j-1)]/\Delta x^2
 \end{aligned}$$

Assuming that u is known on rows j and $j-1$, we can rearrange the equation into the convenient form

$$\begin{aligned}
 & -\lambda^2 u(i+1, j+1) + 2(1+\lambda^2)u(i, j+1) - \lambda^2 u(i-1, j+1) \\
 & = 4u(i, j) + \lambda^2 u(i+1, j-1) - 2(1+\lambda^2)u(i, j-1) + \lambda^2 u(i-1, j-1)
 \end{aligned} \tag{9.38}$$

The right-hand side of (9.38) is known, since it depends only on rows j and $j-1$. The unknowns on row $j+1$ appear on the left-hand side. The equations must now be solved simultaneously using the Thomas algorithm for a tridiagonal matrix (described in Section 5.5.2 of MEM). This algorithm is very rapid and requires little storage. It can be shown that the method will proceed satisfactorily for any λ , so that the time step is unrestricted. The evaluation of rows 0 and 1 is the same as for the explicit method, so this can reduce the accuracy, and clearly the algorithm needs a finite x region to allow the matrix inversion.

Example 9.19

Solve the wave equation $\partial^2 u / \partial t^2 = c^2 \partial^2 u / \partial x^2$ by an implicit method given

- (a) $u(0, t) = 0$ ($t \geq 0$) (fixed at $x = 0$);
- (b) $u(1, t) = 0$ ($t \geq 0$) (fixed at $x = 1$);
- (c) $\partial u(x, 0) / \partial t = 0$ ($0 \leq x \leq 1$) (zero initial velocity);
- (d) $u(x, 0) = \begin{cases} 1 & (x = \frac{1}{4}) \\ 0 & \text{otherwise} \end{cases}$ (displaced at the one point $x = \frac{1}{4}$).

Compare the solutions at a fixed time for various λ .

Solution

Here we have a wave equation solved for a string stretched between two points and displaced at a single point.

The numerical solution shows the expected behaviour of a wave splitting into two waves, one moving in the $-x$ direction and the other in the $+x$ direction. The waves are reflected from the ends, and eventually give a complicated wave shape.

The computations were performed with $\Delta x = 0.125$ and various λ or Δt , with $\lambda = c\Delta t/\Delta x$. The values of u are given at the same time, $T = \Delta x/c$, for various λ :

x	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875	1
$u(\lambda = 0.2)$	0	0.3394	0.2432	0.3412	0.0352	0.0019	0.0001	0	0
$u(\lambda = 0.1)$	0	0.3479	0.2297	0.3493	0.0344	0.0014	0	0	0
$u(\lambda = 0.05)$	0	0.3506	0.2254	0.3519	0.0341	0.0013	0	0	0
$u(\lambda = 0.025)$	0	0.3514	0.2243	0.3526	0.0340	0.0014	0	0	0

Although the method converges for all λ , the accuracy still requires a small λ (or time step), but the value $\lambda = 0.05$ certainly gives an accuracy of less than 1%. It may be noted that at the chosen value of T the wave has split but has not progressed far enough to be reflected from the end $x = 1$.



Equation (9.38) can be written in matrix form

$$\mathbf{A}\mathbf{U}_{j+1} = 4\mathbf{U}_j - \mathbf{A}\mathbf{U}_{j-1} \quad \text{or} \quad \mathbf{U}_{j+1} = 4\mathbf{A}^{-1}\mathbf{U}_j - \mathbf{U}_{j-1}$$

where the U vectors represent the whole row of u values. This makes the problem ideal for MATLAB. The instructions, for n mesh points and general $L = \lambda^2$,

```

n=9;L=0.01;                                %values for the example
a=[-L 2*(1+L) -L]; A=eye(n); for i=2:n-1, A(i,i-1:i+1)=a; end
%sets up matrix A
b=[L/2 1-L L/2]; C=eye(n); for i=2:n-1, C(i,i-1:i+1)=b; end
u=[0 0 1 0 0 0 0 0 0]'; v=C*u
%sets up lines one and two
B=inv(A);w=4*B*v-u
%evaluates line three
u=v;v=w;w=4*B*v-u
%continues the solution

```

compute the solution. Repeating the last line continues the solution by an increment Δt .

The methods described in this section all extend to higher dimensions, and some to nonlinear problems. The work involved is correspondingly greater of course.

9.3.6 Exercises

- 28 Use an explicit method to solve the wave equation $\partial^2 u / \partial t^2 = c^2 \partial^2 u / \partial x^2$ for the boundary conditions

(a) $u(0, t) = 0 \quad (t \geq 0)$

(b) $u(1, t) = 0 \quad (t \geq 0)$

(c) $u(x, 0) = 0 \quad (0 \leq x \leq 1)$

(d) $\frac{\partial u(x, 0)}{\partial t} = \begin{cases} x & (0 \leq x \leq \frac{1}{2}) \\ 1-x & (\frac{1}{2} \leq x \leq 1) \end{cases}$

Use $\Delta x = \Delta t = \frac{1}{4}$ and study the behaviour for a variety of values of λ for the first three time steps. Compare your result with the implicit version in (9.38).

- 29 An oscillator is started at the end of a tube, and oscillations propagate according to the wave equation. The displacement $u(x, t)$ satisfies

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 u}{\partial t^2}$$

in $0 < x < l$, for $t > 0$, with the boundary conditions

(a) $u(0, t) = a \sin \omega t, \quad u(l, t) = 0 \quad (t > 0)$

(b) $u(x, 0) = \frac{\partial u(x, 0)}{\partial t} = 0 \quad (0 \leq x \leq l)$

where c, a and ω are real positive constants.

Show that the solution of the partial differential equation is

$$u(x, t) = \frac{a \sin \omega t \sin [\omega(l-x)/c]}{\sin(\omega l/c)}$$

$$+ \sum_{n=1}^{\infty} \frac{2l a c \omega}{\omega^2 l^2 - n^2 \pi^2 c^2} \sin\left(\frac{n\pi x}{l}\right) \sin\left(\frac{n\pi c t}{l}\right)$$

provided that $\omega l/c$ is not an integer.

Compare this solution with one computed using the explicit numerical method. Use $a = 1$, $l = 1$, $c = 1$, $\omega = \frac{1}{2}\pi$, $\Delta x = 0.2$ and $\Delta t = 0.02$ to evaluate $u(x, 0.06)$.

30

Solve the equation



$$c^2 \frac{\partial^2 u}{\partial x^2} + 2 = \frac{\partial^2 u}{\partial t^2}$$

numerically, subject to the conditions

$$u = x(1-x), \quad \frac{\partial u}{\partial t} = 0 \quad (0 < x < 1) \quad \text{at } t = 0$$

$$u = 0 \quad (x = 0, 1) \quad \text{for } t > 0$$

Use

- (a) an explicit method with $\Delta x = \Delta t = 0.2$ and $\lambda = 0.5$;
- (b) an implicit method with $\Delta x = \Delta t = 0.2$ and $\lambda = 0.5$.

Compare your solution with that in Exercise 31.

31

Solve the equation



$$c^2 \frac{\partial^2 u}{\partial x^2} + 2 = \frac{\partial^2 u}{\partial t^2}$$

numerically, subject to the conditions

$$u = x(1-x), \quad \frac{\partial u}{\partial t} = 0 \quad \text{for all } x \text{ at } t = 0$$

Use

- (a) an explicit method with $\Delta x = \Delta t = 0.2$ and $\lambda = 0.5$;
- (b) an implicit method with $\Delta x = \Delta t = 0.2$ and $\lambda = 0.5$.

9.4

Solution of the heat-conduction/diffusion equation

In this section we consider methods for solving the heat-conduction/diffusion equation introduced in Section 9.2.2.

9.4.1 Separation of variables

It was with the aim of solving heat-conduction problems that Fourier (*c.* 1800) first used the idea of separation of variables and Fourier series. As indicated in Section 12.1 in MEM, many mathematicians at the time argued about the validity of his approach, while he continued to solve many practical problems.

In Section 9.2.2 we noted that the heat-conduction equation

$$\frac{1}{\kappa} \frac{\partial u}{\partial t} = \nabla^2 u \tag{9.5}$$

has a steady-state solution U , provided there are no time-varying inputs, satisfying

$$\nabla^2 U = 0$$

and appropriate boundary conditions. One useful way to write the general solution is

$$u = U + v$$

where v also satisfies (9.5) and the boundary conditions for $u - U$. Certainly the heat-conduction interpretation supports this idea, and we base our strategy on first finding U and then determining the transient v that takes the solution from its initial to its final state. We note that $v \rightarrow 0$ as $t \rightarrow \infty$, so that $u \rightarrow U$, and an obvious method is to try an exponential decay to zero. Thus, in the one-dimensional form of the heat-conduction equation

$$\frac{1}{\kappa} \frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} \tag{9.39}$$

we seek a separated solution of the special type discussed in Section 9.3.2, where the physics indicates a solution

$$v = e^{-\alpha t} w(x)$$

which on substitution gives

$$-\frac{\alpha}{\kappa} w = \frac{d^2 w}{dx^2}$$

Letting $\alpha/\kappa = \lambda^2$, we can solve this simple-harmonic equation to give

$$w = A \sin \lambda x + B \cos \lambda x$$

and hence

$$v = e^{-\alpha t} (A \sin \lambda x + B \cos \lambda x) \quad (9.40)$$

Taking the hint from Section 9.3.2, we expect in general to take sums of terms like (9.40) to satisfy all the boundary conditions. Thus we build up a solution

$$v = \sum_{n=1}^{\infty} e^{-\alpha_n t} (A_n \sin \lambda_n x + B_n \cos \lambda_n x) \quad (9.41)$$

Example 9.20

Solve the heat-conduction equation $\partial T / \partial t = \kappa \partial^2 T / \partial x^2$ subject to the boundary conditions

- (a) $T = 0$ at $x = 0$ and for all $t > 0$ (held at zero temperature);
- (b) $\partial T / \partial x = 0$ at $x = l$ and for all $t > 0$ (no heat loss from this end);
- (c) $T = T_0 \sin(3\pi x/2l)$ at $t = 0$ and for $0 \leq x \leq l$ (given initial temperature profile).

Solution

We first note that as $t \rightarrow \infty$ the solution will be $T = 0$, so the steady-state solution is zero, and so from (9.40) we consider a solution of the form

$$T = e^{-\alpha t} (A \sin \lambda x + B \cos \lambda x) \quad (9.42)$$

In order to satisfy the boundary condition (a), it is clear that it is not possible to include the cosine term, so $B = 0$. To satisfy the condition (b) then requires

$$\frac{\partial T}{\partial x} = A e^{-\alpha t} \lambda \cos \lambda x = 0 \quad (x = l)$$

so that

$$\cos \lambda l = 0, \quad \text{or} \quad \lambda l = (n + \frac{1}{2})\pi \quad (n = 0, 1, 2, 3, \dots)$$

leading to the solution

$$T = A e^{-\alpha t} \sin \left[(n + \frac{1}{2})\pi \frac{x}{l} \right]$$

We now compare the T from condition (c) with the solution just obtained at time $t = 0$, giving

$$A e^{-0} \sin \left[(n + \frac{1}{2})\pi \frac{x}{l} \right] = T_0 \sin \left(\frac{3\pi x}{2l} \right)$$

The unknown parameters can now be identified as $n = 1$ and $A = T_0$, and hence the final solution is

$$T = T_0 \exp \left(-\frac{9\kappa\pi^2 t}{4l^2} \right) \sin \left(\frac{3\pi x}{2l} \right)$$

Example 9.21

Solve the heat-conduction equation $\partial u / \partial t = \kappa \partial^2 u / \partial x^2$ subject to the boundary conditions

- (a) $u(0, t) = 0$ ($t \geq 0$) (zero temperature at the end $x = 0$);
- (b) $u(l, t) = 0$ ($t \geq 0$) (zero temperature at the end $x = l$);
- (c) $u(x, 0) = u_0(\frac{1}{2} - x/l)$ ($0 < x < l$) (a given initial temperature profile).

Solution

We are solving the problem of heat conduction in a bar that is held at zero temperature at its ends and with a given initial temperature profile.

It is clear that the final solution as $t \rightarrow \infty$ is $u = 0$, so that $U = 0$ is the steady-state solution, and hence from (9.40) we seek a solution of the form

$$u = e^{-\alpha t}(A \sin \lambda x + B \cos \lambda x)$$

subject to the given boundary conditions. The first of these conditions (a) gives $B = 0$, while the second condition (b) gives

$$\sin \lambda l = 0, \quad \text{or} \quad \lambda l = n\pi \quad (n = 1, 2, \dots)$$

Recalling that $\lambda^2 = \alpha/\kappa$, we find solutions of the form

$$u = A e^{-\kappa n^2 \pi^2 t / l^2} \sin\left(\frac{n\pi x}{l}\right) \quad (n = 1, 2, \dots)$$

Clearly we cannot satisfy (c) from a single solution, so, as indicated in (9.41), we revert to the sum

$$u = \sum_{n=1}^{\infty} A_n e^{-\kappa n^2 \pi^2 t / l^2} \sin\left(\frac{n\pi x}{l}\right) \tag{9.43}$$

which is also a solution.

Using the boundary condition (c), we then have

$$u_0\left(\frac{1}{2} - \frac{x}{l}\right) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi x}{l}\right)$$

and hence, by (7.17),

$$\frac{1}{2}l A_n = u_0 \int_0^l \left(\frac{1}{2} - \frac{x}{l}\right) \sin\left(\frac{n\pi x}{l}\right) dx = \begin{cases} 0 & (\text{odd } n) \\ u_0 l / n\pi & (\text{even } n) \end{cases}$$

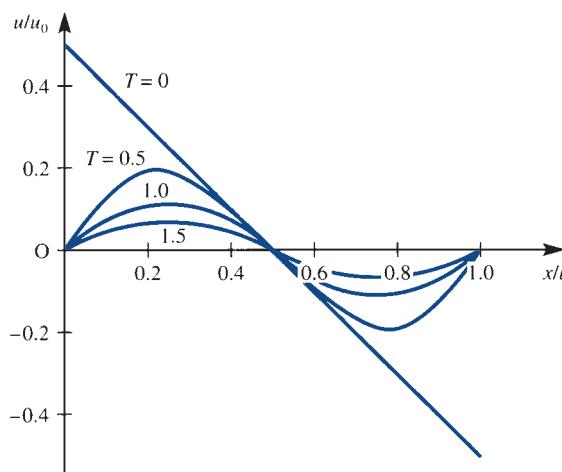
Note again that we have used a periodic extension of the given function to obtain a Fourier sine series valid over the interval $0 \leq x \leq l$. Outside the interval $0 \leq x \leq l$ we have no physical interest in the solution. Substituting back into (9.43) gives as a final solution

$$u = \frac{u_0}{\pi} \sum_{m=1}^{\infty} \frac{1}{m} e^{-4\kappa m^2 \pi^2 t / l^2} \sin\left(\frac{2m\pi x}{l}\right) \tag{9.44}$$

or, in an expanded form,

$$u = \frac{u_0}{\pi} \left[e^{-4\kappa \pi^2 t / l^2} \sin\left(\frac{2\pi x}{l}\right) + \frac{1}{2} e^{-16\kappa \pi^2 t / l^2} \sin\left(\frac{4\pi x}{l}\right) + \frac{1}{3} e^{-36\kappa \pi^2 t / l^2} \sin\left(\frac{6\pi x}{l}\right) + \dots \right]$$

Figure 9.23 Solution of Example 9.21 with $T = t(4\kappa\pi^2/l^2)$.



In Figure 9.23, u/u_0 is plotted against x/l at successive times $T = t(4\kappa\pi^2/l^2) = 0, 0.5, 1, 1.5$.

Taking successive terms in the series for the values $T = t(4\kappa\pi^2/l^2) = 0.5$ and $x/l = 0.2$, we get

	1 term	2 terms	3 terms	4 terms
u/u_0	0.1836	0.1963	0.1956	0.1953

and we see that three terms of this series would probably be sufficient to give three-figure accuracy. In all such problems some numerical experimentation is required to determine how many terms are required. For small t and x we should expect to need a large number of terms, since the temperature at the end switches from $\frac{1}{2}u_0$ to 0 at time $t = 0$. It is well known that discontinuities cause convergence difficulties for Fourier series.

It may be noted in Example 9.21 that the initial discontinuity is smoothed out, as we expected from the physical ideas that we outlined in Section 9.2.2.

Example 9.22

Solve the heat-conduction equation $\partial u / \partial t = \kappa \partial^2 u / \partial x^2$ in a bar subject to the boundary conditions

- (a) $u(0, t) = 0$ ($t \geq 0$) (the end $x = 0$ is held at zero temperature);
- (b) $u(1, t) = 1$ ($t \geq 0$) (the end $x = 1$ is at temperature 1);
- (c) $u(x, 0) = x(2 - x)$ ($0 \leq x \leq 1$) (the initial temperature profile is given).

Solution

First it is clear that the final steady-state solution is $U = x$, since this satisfies (a) and (b) and also $\nabla^2 U = 0$. Secondly putting $u = U + v$, the new variable v satisfies (9.39), but now the boundary conditions on v are

(a') $v(0, t) = 0$

(b') $v(1, t) = 0$

(c') $v(x, 0) = x(2 - x) - x = x - x^2$

The appropriate solutions in (9.40) can now be selected; the condition (a') gives $B = 0$, while the condition (b'), $v(1, t) = 0$, gives

$$\sin \lambda = 0, \quad \text{or} \quad \lambda = n\pi \quad (n = 1, 2, \dots)$$

From (9.41) we then have

$$v = \sum_{n=1}^{\infty} a_n e^{-kn^2\pi^2 t} \sin n\pi x$$

and condition (c') gives

$$x - x^2 = \sum_{n=1}^{\infty} a_n \sin n\pi x$$

Determining the Fourier coefficient using (7.17),

$$\frac{1}{2} a_n = \int_0^1 (x - x^2) \sin n\pi x \, dx = \begin{cases} 4/n^3\pi^3 & (\text{odd } n) \\ 0 & (\text{even } n) \end{cases}$$

Thus the complete solution is

$$u = x + \frac{8}{\pi^3} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^3} e^{-\kappa(2n-1)^2\pi^2 t} \sin(2n-1)\pi x$$

or, in expanded form,

$$u = x + \frac{8}{\pi^3} [e^{-\kappa\pi^2 t} \sin \pi x + \frac{1}{27} e^{-9\kappa\pi^2 t} \sin 3\pi x + \frac{1}{125} e^{-25\kappa\pi^2 t} \sin 5\pi x + \dots]$$

9.4.2 Laplace transform method

As we saw for the wave equation in Section 9.3.3, Laplace transforms provide an alternative method of solution for the heat-conduction or diffusion equation. The method has the merit of dealing with the boundary conditions easily, but it suffers from the usual difficulty of performing the final inversion. The following example serves to illustrate these points.

Example 9.23

Using the Laplace transform method, solve the diffusion equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2} \quad (\text{all } x, t > 0)$$

given that $u(x, t)$ remains bounded and satisfies the boundary conditions

- (a) $u(x, 0) = 0$ (all x)
 (b) $u(a, t) = T\delta(t)$

Solution The problem models an infinite pipe, coincident with the x axis, that is initially filled with clean fluid. It is subjected to an instantaneous burst of contaminant injected at the point $x = a > 0$; the concentration of the contaminant, as it diffuses, is required.

Using (9.30) and (9.31) and taking Laplace transforms gives

$$sU(x, s) - u(x, 0) = \kappa \frac{d^2 U(x, s)}{dx^2}$$

which, on using condition (a), leads to the ordinary differential equation

$$\frac{d^2 U}{dx^2} - \frac{s}{\kappa} U = 0$$

This is readily solved to give

$$U(x, s) = Ae^{\sqrt{(s/\kappa)}x} + Be^{-\sqrt{(s/\kappa)}x}$$

Since concentration remains bounded

$$U(x, s) = Be^{-\sqrt{(s/\kappa)}x} \quad \text{for } x > a \text{ and}$$

$$U(x, s) = Ae^{\sqrt{(s/\kappa)}x} \quad \text{for } x < a$$

Condition (b) then gives

$$U(a, s) = \mathcal{L}\{T\delta(t)\} = T = Be^{-\sqrt{(s/\kappa)}a}$$

so that

$$B = Te^{\sqrt{(s/\kappa)}a} \quad \text{and similarly} \quad A = Te^{-\sqrt{(s/\kappa)}a}$$

giving

$$U(x, s) = Te^{-(x-a)\sqrt{(s/\kappa)}} \quad \text{for } x > a \text{ and} \quad U(x, s) = Te^{-(a-x)\sqrt{(s/\kappa)}} \quad \text{for } x < a$$

To find the solution $u(x, t)$, we must invert the Laplace transform. However, in this case, the methods discussed in Section 11.2.7 in MEM do not suffice, and it is necessary to resort to the use of the complex inversion integral, which is dealt with in specialist texts on Laplace transforms (see also Chapter 8, Review exercise 7). Alternatively, we can turn to the extensive tables that exist of Laplace transform pairs, to find that

$$\mathcal{L}^{-1}\{e^{-b\sqrt{s}}\} = \frac{b}{2\sqrt{\pi}} t^{-3/2} e^{-b^2/4t}, \quad b > 0$$

We can then carry out the required inversion to give the solution

$$u(x, t) = \frac{T|x-a|}{2\sqrt{\pi\kappa}} t^{-3/2} e^{-(x-a)^2/4\kappa t} \quad (t > 0)$$



It is possible to solve this example by using the extensive tables of Laplace transforms in MAPLE. It does, however, require some knowledge of the manipulative skills contained in the package to progress easily.

It should be noted that the solution of Example 9.23 is not variable separable and could not be obtained from the methods of Section 9.4.1.

To date all the problems studied have been restricted to heat flow in a rod. The ideas can be extended to spherical or cylindrical regions. Here one example will be considered that involves radially symmetric regions; see also Exercises 4 and 33. First the heat-conduction equation is written for a problem that only depends on the radial distance r from the origin and the time t . Now

$$r^2 = x^2 + y^2 + z^2 \quad \text{so} \quad 2r \frac{\partial r}{\partial x} = 2x$$

To obtain the Laplacian

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}$$

when $f = f(r, t)$ evaluate

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial r} \frac{\partial r}{\partial x} = \frac{x}{r} \frac{\partial f}{\partial r}$$

and then the second derivative

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial}{\partial x} \left(\frac{x}{r} \frac{\partial f}{\partial r} \right) = \frac{1}{r} \frac{\partial f}{\partial r} - \frac{1}{r^2} \frac{x}{r} \left(x \frac{\partial^2 f}{\partial r^2} \right) + \frac{x}{r} \frac{x}{r} \frac{\partial^2 f}{\partial r^2}$$

The y and z derivatives follow in a similar manner so the Laplacian becomes

$$\begin{aligned} \nabla^2 f &= \frac{3}{r} \frac{\partial f}{\partial r} - \frac{1}{r^3} (x^2 + y^2 + z^2) \frac{\partial f}{\partial r} + \frac{x^2 + y^2 + z^2}{r^2} \frac{\partial^2 f}{\partial r^2} \\ &= \frac{2}{r} \frac{\partial f}{\partial r} + \frac{\partial^2 f}{\partial r^2} = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial f}{\partial r} \right) \end{aligned}$$

The radially symmetric heat conduction equation for the temperature $T(r, t)$ is therefore

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial T}{\partial r} \right) = \frac{1}{\kappa} \frac{\partial T}{\partial t} \tag{9.45}$$

This radially symmetric form can be made to look very similar to the one-dimensional equation by writing

$$T = \frac{\theta(r, t)}{r}$$

With this substitution

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(-\theta + r \frac{\partial \theta}{\partial r} \right) = \frac{1}{r \kappa} \frac{\partial \theta}{\partial t}$$

and differentiating once more gives

$$\frac{1}{r} \left(-\frac{\partial \theta}{\partial r} + \frac{\partial \theta}{\partial r} + r \frac{\partial^2 \theta}{\partial r^2} \right) = \frac{1}{\kappa} \frac{\partial \theta}{\partial t}$$

Thus the equation for $\theta(r, t)$ is just the usual one-dimensional heat-conduction equation

$$\frac{\partial^2 \theta}{\partial r^2} = \frac{1}{\kappa} \frac{\partial \theta}{\partial t} \quad (9.46)$$

Methods developed earlier can now be used on this equation, but note that care must be taken at the origin where $r = 0$ since

$$T = \frac{\theta(r, t)}{r}$$

An example will illustrate some important points.

Example 9.24

Solve the radially symmetric heat-conduction equation (9.45)

- (a) in the region $0 < r < a$ subject to the boundary conditions $T = T_0$ on $r = a$ and T is finite in the region. The initial condition is $T(r, 0) = 0$.
- (b) in the region $r > a$ subject to the boundary conditions $T = T_0$ on $r = a$ and T tends to zero as r tends to infinity. The initial condition is $T(r, 0) = 0$.

Solution (a) To solve (9.45) put

$$T = T_0 + \frac{\theta(r, t)}{r}$$

then $\theta(r, t)$ satisfies (9.46) with the modified boundary conditions $\theta(a, t) = 0$ and $\theta(r, t)/r$ is finite at the origin. Clearly the only separated solution in (9.40) that can satisfy these conditions is

$$\theta = e^{-\alpha t} \sin \lambda r \quad \text{where} \quad \alpha/\kappa = \lambda^2$$

since $\sin \lambda r/r \rightarrow \lambda$ as $r \rightarrow 0$. The condition at $r = a$ gives $\sin(\lambda a) = 0$ so $\lambda a = n\pi$ where n is a positive integer. Summing all solutions of this type gives

$$T = T_0 + \frac{1}{r} \sum_{n=1}^{\infty} A_n e^{-\kappa n^2 \pi^2 t/a^2} \sin(n\pi r/a)$$

The coefficients A_n are given by the initial condition which reduces to the Fourier series problem

$$-rT_0 = \sum_{n=1}^{\infty} A_n \sin(n\pi r/a)$$

giving

$$A_n = \frac{2aT_0(-1)^n}{\pi n}$$

and finally

$$T = T_0 - \frac{2aT_0}{\pi} \frac{1}{r} \left[e^{-\kappa\pi^2 t/a^2} \sin(\pi r/a) - \frac{1}{2} e^{-\kappa^2\pi^2 t/a^2} \sin(2\pi r/a) + \dots \right]$$

- (b) As in part (a) we look to previous methods of solution. First note that it is only necessary to solve (9.46) with the modified boundary conditions $\theta(a, t) = aT_0$ and $\theta(r, t)$ is finite for infinite r . The initial condition is $\theta(r, 0) = 0$. The problem is now very similar to Example 9.23 and Laplace transforms appears to be a sensible approach. The set up is precisely as in Example 9.23 (except for a change in the names of the parameters) to the point where

$$s\bar{\theta}(r, s) = \kappa \frac{d^2\theta(r, s)}{dr^2}$$

where it should be noted that the zero initial condition has been used. The solution is

$$\bar{\theta}(r, s) = A e^{\sqrt{(s/\kappa)}r} + B e^{-\sqrt{(s/\kappa)}r}$$

which is now subject to the condition that $\bar{\theta}(r, s)$ is finite as r gets large implying that $A = 0$. To calculate B , the transformed condition at $r = a$ gives $\bar{\theta}(a, s) = aT_0/s$ and hence

$$B = \frac{aT_0}{s} e^{\sqrt{(s/\kappa)}a}$$

Thus the solution for the transformed equation is

$$\bar{\theta}(r, s) = \frac{aT_0}{s} e^{-\sqrt{(s/\kappa)}(r-a)}$$

The expression requires either access to extensive tables of Laplace transforms or to the tables in MATLAB or MAPLE. They give



$$\bar{\theta}(r, t) = aT_0 \operatorname{erfc}\left(\frac{1}{2} \frac{r-a}{\sqrt{\kappa t}}\right)$$

so that

$$T(r, t) = \frac{aT_0}{r} \operatorname{erfc}\left(\frac{1}{2} \frac{r-a}{\sqrt{\kappa t}}\right)$$

The error function

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-v^2} dv$$

and $\operatorname{erfc}(z) = 1 - \operatorname{erf}(z)$ occur commonly in the solution of the heat-conduction equation (see also Section 7.4), and in statistics. It is well documented and appears as standard in MATLAB and MAPLE. The erfc function is illustrated in Figure 9.24 and the solution T/T_0 is plotted against r/a in Figure 9.25 for various times.

Figure 9.24 The function $\text{erfc}(z)$ of Example 9.24.

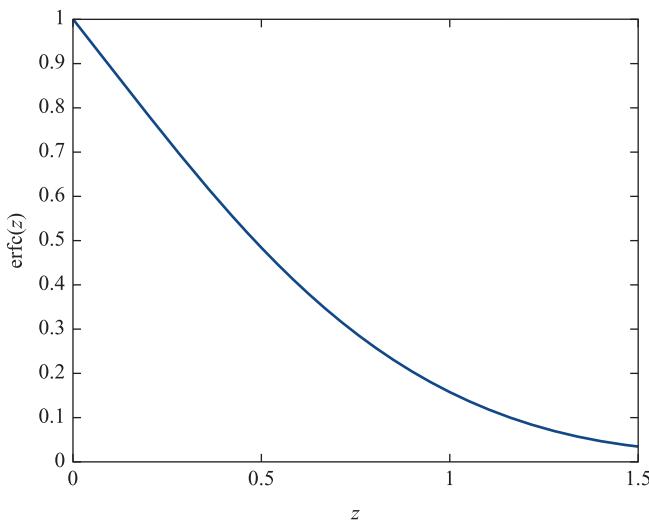
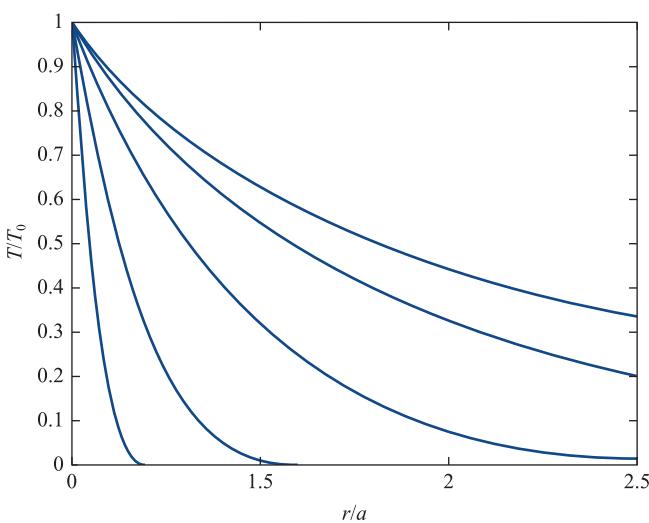


Figure 9.25 Plot of the solution to Example 9.24 for times $4\kappa t/a^2 = 0.01, 0.1, 1, 10$ and 100 .



9.4.3 Exercises

32 Find the solution to the equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}$$

satisfying the conditions

- (a) $\partial u / \partial x = 0$ at $x = 0$ for all t
- (b) $u = 0$ at $x = 1$ for all t
- (c) $u = a \cos(\pi x) \cos(\frac{1}{2}\pi x)$ for $0 \leq x \leq 1$
when $t = 0$

33 The spherically symmetric form of the heat-conduction equation is

$$u_{rr} + \frac{2}{r}u_r = \frac{1}{\kappa}u_t$$

By putting $ru = v$, show that v satisfies the standard one-dimensional heat-conduction equation. What can we expect of a solution as $r \rightarrow \infty$?

Solve the equation in the annulus $a \leq r \leq b$ given that $u(a, t) = T_0$, $u(b, t) = 0$ for all $t > 0$ and

the initial condition $u(r, 0) = 0$ for $a \leq r \leq b$. Show that the solution has the form

$$T = \frac{aT_0}{r} \left[\frac{b-r}{b-a} - \sum_{N=1}^{\infty} A_N e^{-\kappa \lambda^2 t} \sin \left(\frac{r-a}{b-a} N\pi \right) \right]$$

where $\lambda(b-a) = N\pi$. Evaluate the Fourier coefficients A_N .

- 34 Show that $u(x, t) = t^\alpha F(\eta)$, where $\eta = x^2/t$ is a solution of the partial differential equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

if F satisfies

$$4\eta \frac{d^2F}{d\eta^2} + (2+\eta) \frac{dF}{d\eta} - \alpha F = 0$$

Find non-zero values of α and κ for which $F = e^{\kappa\eta}$ is a solution.

- 35 Show that $u(x, t) = f(x) \sin(x - \beta t)$ is a solution of the heat-conduction equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

provided that f and the constant β are chosen suitably. For a semi-infinite slab of uniform material occupying the region $x \geq 0$ construct the solution that satisfies (a) u is bounded as $x \rightarrow \infty$ and (b) $u(0, t) = u_0 \sin 2t$ (that is, a periodic temperature is imposed at $x = 0$).

- 36 Show that the equation

$$\theta_t = \kappa \theta_{xx} - h(\theta - \theta_0)$$

can be reduced to the standard heat-conduction equation by writing $u = e^{ht}(\theta - \theta_0)$. How do you interpret the term $h(\theta - \theta_0)$?

- 37 Use separation of variables to obtain a solution to the heat-conduction equation $\partial u / \partial t = \kappa \partial^2 u / \partial x^2$, given

(a) $\partial u(0, t) / \partial x = 0$ ($t \geq 0$)

(b) $u(l, t) = 0$ ($t \geq 0$)

(c) $u(x, 0) = u_0(\frac{1}{2} - x/l)$ ($0 \leq x \leq l$)

Compare the solution with that obtained in Example 9.21.

- 38 The voltage v at a time t at a distance x along an electric cable of length L with capacitance and resistance only, satisfies

$$\frac{\partial^2 v}{\partial x^2} = \frac{1}{\kappa} \frac{\partial v}{\partial t}$$

Verify that a form of the solution appropriate to the conditions that $v = v_0$ when $x = 0$, and $v = 0$ when $x = L$, for all values of t , is given by

$$v = v_0 \left(1 - \frac{x}{L} \right) + \sum_{n=1}^{\infty} c_n \exp \left(-\frac{\kappa n^2 \pi^2 t}{L^2} \right) \sin \left(n\pi \frac{x}{L} \right)$$

where v_0 and the c_n are constants.

Show that if, in addition, $v = 0$ when $t = 0$ for $0 < x < L$,

$$c_n = -\frac{2v_0}{n\pi}$$

39

A uniform bar of length l has its ends maintained at a temperature of 0°C . Initially, the temperature at any point between the ends of the bar is 10°C , and, after a time t , the temperature $u(x, t)$ at a distance x from one end of the bar satisfies the one-dimensional heat-conduction equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{\kappa} \frac{\partial u}{\partial t} \quad (x > 0)$$

Write down boundary conditions for the bar and show that the solution corresponding to these conditions is

$$u(x, t) = \frac{20}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} (1 - \cos n\pi) \exp \left(-\frac{\kappa n^2 \pi^2 t}{l^2} \right) \times \sin \left(\frac{\pi n x}{l} \right)$$

40

The function $\phi(x, t)$ satisfies the equation

$$\frac{\partial \phi}{\partial t} = a \frac{\partial^2 \phi}{\partial x^2} + b \quad (-h < x < h, t > 0)$$

with the boundary conditions

(a) $\phi(-h, t) = \phi(h, t) = 0$ ($t > 0$)

(b) $\phi(x, 0) = 0$ ($-h < x < h$)

where a , b and h are positive real constants.

Show that the Laplace transform of the solution $\phi(x, t)$ is

$$\frac{b}{s^2} \left\{ 1 - \frac{\cosh[(s/a)^{1/2}x]}{\cosh[(s/a)^{1/2}h]} \right\}$$

9.4.4 Numerical solution

As for the wave equation, except for the most straightforward problems, we must resort to numerical solutions of the heat-conduction equation. Even when analytical solutions are known, they are not always easy to evaluate because of convergence difficulties near to singularities. They are, of course, crucial in testing the accuracy and efficiency of numerical methods.

We can write the heat-conduction equation

$$\frac{1}{\kappa} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \quad (9.47)$$

in the usual *finite-difference* form, using the notation of Figure 9.26.

We assume that we know the solution up to time step j and we wish to calculate the solution at time step $j + 1$. In Section 9.3.5 we showed how to approximate the second derivative as

$$\frac{\partial^2 u}{\partial x^2} = \frac{u(i+1, j) - 2u(i, j) + u(i-1, j)}{\Delta x^2}$$

To obtain the time derivative, we use the approximation between rows j and $j + 1$:

$$\frac{\partial u}{\partial t} = \frac{u(i, j+1) - u(i, j)}{\Delta t}$$

Putting these into (9.47) gives

$$\frac{u(i, j+1) - u(i, j)}{\kappa \Delta t} = \frac{u(i-1, j) - 2u(i, j) + u(i+1, j)}{\Delta x^2}$$

or, on rearranging,

$$u(i, j+1) = \lambda u(i-1, j) + (1 - 2\lambda)u(i, j) + \lambda u(i+1, j) \quad (9.48)$$

where $\lambda = \kappa \Delta t / \Delta x^2$. Equation (9.48) gives a finite-difference representation of (9.47), and provided that all the values are known on row j , we can then compute u on row $j + 1$ from the simple **explicit formula** (9.48).

First a simple example on a coarse grid.

Example 9.25

Use an explicit numerical method to solve the heat-conduction equation (9.47) subject to the boundary conditions

- (a) $u(0, t) = u(0, 1) = 0$ ($t \geq 0$) (both ends held at zero temperature);
- (b) $u(x, 0) = \sin(\pi x)$ ($0 \leq x \leq 1$) (a given initial temperature distribution).

Use the parameters $\Delta t = 0.1$, $\Delta x = 0.25$, $\kappa = 0.1$.

Solution

This problem has the exact solution $u = e^{-\pi^2 \kappa t} \sin(\pi x)$, so the accuracy of the numerical solution can be checked easily.

At $t = 0$ (or $j = 0$) the initial values come from the boundary condition (b)

x	0	0.25	0.5	0.75	1
u	0	0.7071	1	0.7071	0

Before proceeding further we first note that $\lambda = 0.16$. At $t = 0.1$ (or $j = 1$) (9.48) becomes

$$u(i, 1) = 0.16[u(i - 1, 0) + u(i + 1, 0)] + 0.68u(i, 0)$$

which, on calculation for the values $x = 0.25, 0.5, 0.75$ or $i = 1, 2, 3$, gives

x	0	0.25	0.5	0.75	1
u	0	0.6408	0.9063	0.6048	0

Note that at the ends, $x = 0$ and 1 , the boundary condition (a) $u = 0$ is imposed.

At $t = 0.2$ (or $j = 2$)

$$u(i, 2) = 0.16[u(i - 1, 1) + u(i + 1, 1)] + 0.68u(i, 1)$$

and performing the calculations

x	0	0.25	0.5	0.75	1
u	0	0.5808	0.8213	0.5808	0

Similarly for $t = 0.3$ (or $j = 3$)

x	0	0.25	0.5	0.75	1
u	0	0.5263	0.7444	0.5263	0
u exact	0	0.5259	0.7437	0.5259	0

and so on.

In the last table the exact values have been included for comparison.

Example 9.26

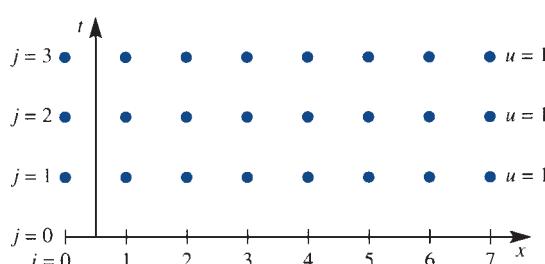
Solve the heat-conduction equation (9.47) subject to the boundary conditions

- (a) $\partial u(0, t)/\partial x = 0$ ($t \geq 0$) (no heat flow through the end $x = 0$);
- (b) $u(1, t) = 1$ ($t \geq 0$) (unit temperature held at $x = 1$);
- (c) $u(x, 0) = x^2$ ($0 \leq x \leq 1$) (a given initial temperature distribution).

Solution

To fit the condition (a) most easily, we allow the first mesh space to straddle the t axis as illustrated in Figure 9.27, where six intervals are used in the x direction. The mesh implies that $\Delta x = 1/6.5 = 0.1538$; condition (a) gives $u(0, j) = u(1, j)$ while condition (b) gives $u(7, j) = 1$.

Figure 9.27 Mesh for Example 9.26:
 $u(7, j) = 1$ and
 $u(0, j) = u(1, j)$ for all
 j ; $\Delta x = 1/6.5 = 0.1538$.





Some simple MATLAB code produces the solution very quickly:

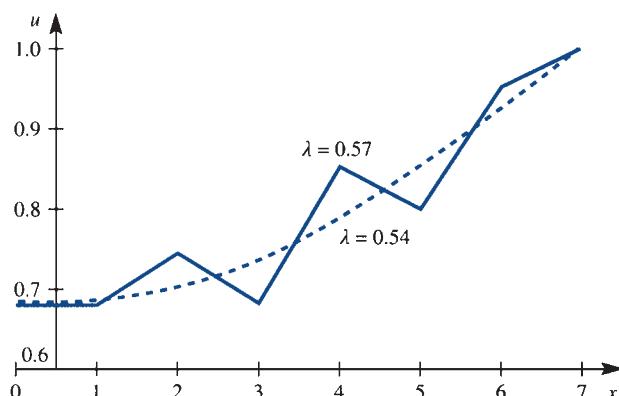
```
n=8;L=0.2; %values of the Example
x=[-0.5/(n-1.5):1/(n-1.5):1],u=x.^2 %initial data
v=[0,L*(u([1:n-2])+u([3:n]))+(1-2*L)*u([2:n-1]),1];
v(1)=v(2); %computes the first row
u=v; v=[0,L*(u([1:n-2])+u([3:n]))+(1-2*L)*u([2:n-1]),1];
v(1)=v(2);
%repeat this line of code to obtain successive rows
```

The following table gives values for u at the three times $t = 0, 0.2\Delta x^2/\kappa$ and $20\Delta x^2/\kappa$, calculated with $\lambda = 0.2$. These results are then compared at $t = 20\Delta x^2/\kappa$, computed with λ taken to be 0.5. It may be noted that there are errors in the third significant figure between the two cases.

i	0	1	2	3	4	5	6	7
$\lambda = 0.2$	$j = 0$	0.0059	0.0059	0.0533	0.1479	0.2899	0.4793	0.7160
	$j = 1$	0.0154	0.0154	0.0627	0.1574	0.2994	0.4888	0.7255
	$j = 100$	0.6817	0.6817	0.7002	0.7362	0.7874	0.8510	0.9233
$\lambda = 0.5$	$j = 40$	0.6850	0.6850	0.7033	0.7389	0.7896	0.8526	0.9241

For $\lambda = 0.54$ we can obtain a solution that compares with the solution given, but for $\lambda = 0.6$ the solution diverges wildly. Figure 9.28 shows a plot of the solution near the critical value of λ at a fixed time $t = 20\Delta x^2/\kappa$. As in the table, the solutions are accurate to about 1% for small λ , but when λ gets much above 0.5, oscillations creep in and the solution is meaningless. In Figure 9.29 a further graph illustrates the development of the solution in the two cases $\lambda = 0.2$ and $\lambda = 0.55$. One solution progresses smoothly as time advances, while the other produces oscillations that will eventually lead to divergence.

Figure 9.28
Numerical solution
of Example 9.26 at
time $t = 20 \Delta x^2/\kappa$, for
two values of λ .



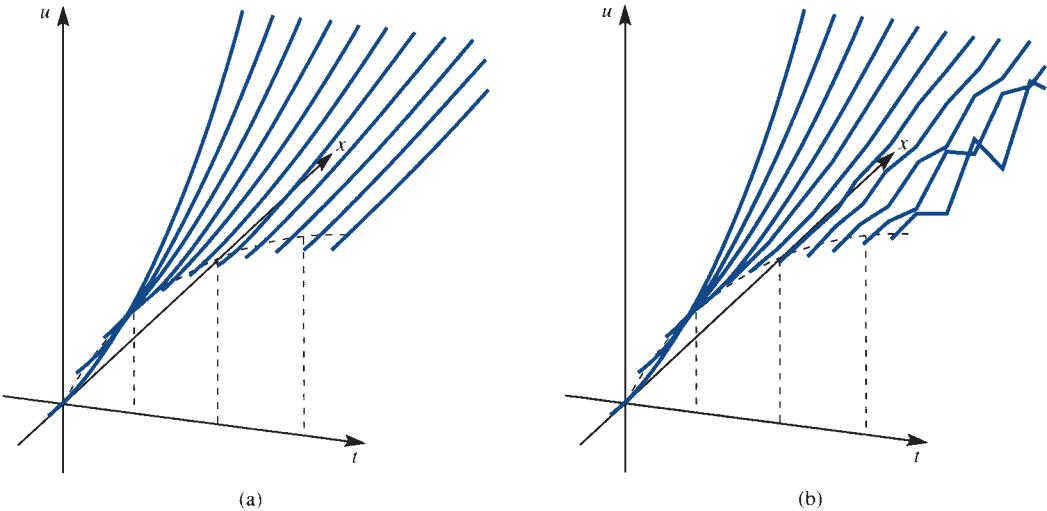


Figure 9.29 Plots of u against x and t from the solution of Example 9.26 with (a) $\lambda = 0.2$; (b) $\lambda = 0.55$.

Comparing Example 9.26 with the numerical solution of the wave equation undertaken in Example 9.18, we observe similar behaviour for the explicit scheme, namely that the method will only converge for small enough time steps or λ . From (9.48) it may be noted that the middle term changes sign at $\lambda = 0.5$, and above this value we might anticipate difficulties. Indeed, some straightforward numerical analysis shows that convergence is certain for $\lambda < 0.5$. It is sufficient here to note that λ must not be too large.

To avoid the limitation on λ , we can again look at an **implicit** formulation of the numerical equations. Returning to Figure 9.26, the idea is to approximate the x derivative by an average of row j and row $j + 1$:

$$\begin{aligned} u(i, j + 1) - u(i, j) &= \lambda\{(1 - \alpha)[u(i - 1, j) - 2u(i, j) + u(i + 1, j)] \\ &\quad + \alpha[u(i - 1, j + 1) - 2u(i, j + 1) + u(i + 1, j + 1)]\} \end{aligned}$$

where $0 \leq \alpha \leq 1$ is an averaging parameter. The case $\alpha = 0$ corresponds to the explicit formulation (9.48), while $\alpha = \frac{1}{2}$ is the best known implicit formulation, and constitutes the **Crank–Nicolson method**. With $\alpha = \frac{1}{2}$, we have

$$\begin{aligned} &- \lambda u(i - 1, j + 1) + 2(1 + \lambda)u(i, j + 1) - \lambda u(i + 1, j + 1) \\ &= \lambda u(i - 1, j) + 2(1 - \lambda)u(i, j) + \lambda u(i + 1, j) \end{aligned} \tag{9.49}$$

We know the solution on row j , so the right-hand side of (9.49) is known, and the unknowns on row $(j + 1)$ have to be solved for simultaneously.

Fortunately the system is tridiagonal, so the very rapid Thomas algorithm can be used. The method performs extremely well: it converges for all λ , and is the best known approach to heat-conduction equations.

Example 9.27

Repeat Example 9.25 but with an implicit scheme.

Solution

At $t = 0$ (or $j = 0$) the initial values come from the boundary condition and are identical for both the implicit and explicit formulations.

At $t = 0.1$ (or $j = 1$) the three equations of the type (9.49) corresponding to $x = 0.25, 0.5, 0.75$ or $i = 1, 2, 3$ are

$$-0.16u(0, 1) + 2.32u(1, 1) - 0.16u(2, 1) = 0.16[u(0, 0) + u(2, 0)] + 1.68u(1, 0)$$

$$-0.16u(1, 1) + 2.32u(2, 1) - 0.16u(3, 1) = 0.16[u(1, 0) + u(3, 0)] + 1.68u(2, 0)$$

$$-0.16u(2, 1) + 2.32u(3, 1) - 0.16u(4, 1) = 0.16[u(2, 0) + u(4, 0)] + 1.68u(3, 0)$$

After noting that the end boundary conditions give

$$u(0, 0) = u(0, 1) = u(4, 0) = u(4, 1) = 0$$

and the right-hand sides evaluated from the initial values, the equations can be written in matrix form as

$$\begin{bmatrix} 2.32 & -0.16 & 0 \\ -0.16 & 2.32 & -0.16 \\ 0 & -0.16 & 2.32 \end{bmatrix} \begin{bmatrix} u(1, 1) \\ u(2, 1) \\ u(3, 1) \end{bmatrix} = \begin{bmatrix} 1.348 \\ 1.906 \\ 1.348 \end{bmatrix}$$

The tridiagonal system can be solved to give

x	0	0.25	0.5	0.75	1
u	0	0.6438	0.9105	0.6438	0

For the next time steps the matrix equation is identical, with the j -suffix advanced by 1 at each time step and the right-hand sides re-evaluated from the most recently computed values of u . Subsequent values are

x	0	0.25	0.5	0.75	1
u at $t = 0.2$	0	0.5862	0.829	0.5862	0
u at $t = 0.3$	0	0.5337	0.7547	0.5337	0

and should be compared with the explicit solution in Example 9.25.

Example 9.28

Solve the heat-conduction equation (9.47) using the implicit formulation (9.49) for the boundary conditions

- $u(0, t) = 0 \quad (t \geq 0)$ (the end $x = 0$ is kept at zero temperature);
- $u(1, t) = 1 \quad (t \geq 0)$ (the end $x = 1$ is kept at unit temperature);
- $u(x, 0) = 0 \quad (0 \leq x \leq 1)$ (initially the bar has zero temperature).

Solution

Here a bar is initially at zero temperature. At one end the temperature is raised to the value 1 and kept at that value.

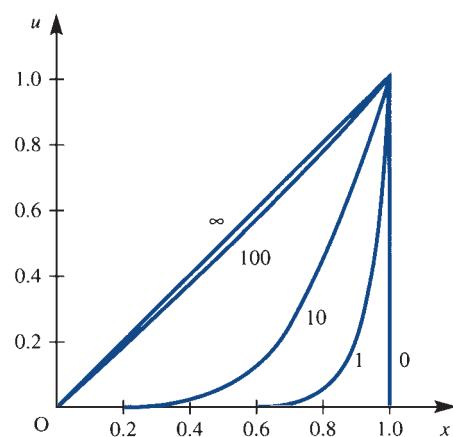


The matrix inversions are very tedious to perform, so again a package such as MATLAB solves the equations very quickly; note the use of the ‘colon’ notation.

```
L=0.3; M=2*(1+L); N=2*(1-L); n=11; %values for the Example
u=zeros(1,n)'; u(n)=1; %initial data
p=[-L M -L]; A=eye(n); for i=2:n-1, A(i,i-1:i+1)=p; end
q=[L N L]; B=eye(n); for i=2:n-1, B(i,i-1:i+1)=q; end
%sets up the matrices in equation (9.49)
DD=inv(A)*B; v=DD*u %solves for first row
u=v; v=DD*u; %repeat this line of code for subsequent rows
```

The results of the calculation are presented in Figure 9.30. At time step 0 the temperature distribution is discontinuous. The successive time steps 1, 10, 100 are shown, and the final distribution $u = x$ is labelled ∞ .

Figure 9.30 Solution of Example 9.28 with $\Delta x = 0.1$, $\lambda = 0.3$ using the Crank–Nicolson scheme.



9.4.5 Exercises

- 41 Derive the usual explicit finite-difference representation of the equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

Using this scheme with $\Delta t = 0.02$ and $\Delta x = 0.2$, determine an approximate solution of the equation at $t = 0.06$, given that

$$u = x^2 \quad \text{when } t = 0 \quad (0 \leq x \leq 1)$$

$$u = 0 \quad \text{when } x = 0 \quad (t > 0)$$

$$u = 1 \quad \text{when } x = 1 \quad (t > 0)$$

- 42 Use both explicit and implicit numerical formulations to obtain solutions of the heat-conduction equation subject to the boundary conditions

- (a) $u(0, t) = 0 \quad (t \geq 0)$ (b) $u(1, t) = e^{-t} \quad (t \geq 0)$
 (c) $u(x, 0) = 0 \quad (0 \leq x < 1)$

Compare the two results for $t = 1$.

- 43 Given that u satisfies the equation



$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

and is subject to the boundary conditions

$$\frac{\partial u}{\partial x} = 1 \quad (x = 0, t > 0)$$

$$u = 0 \quad (x = 1, t > 0)$$

$$u = x(1 - x) \quad (t = 0, 0 \leq x \leq 1)$$

derive a set of algebraic equations from the implicit formulation in Section 9.4.4. Use the implicit method by adapting the MATLAB segment in Example 9.28. Find the solution at $t = 0.02$ and 0.04 using the values $\Delta x = 0.2$ and $\Delta t = 0.02$.

9.5 Solution of the Laplace equation

In this section we consider methods of solving the Laplace equation introduced in Section 9.2.3.

9.5.1 Separation of variables

It is much less obvious how to construct separated solutions for the Laplace equation, since there is less physical feel for the behaviour except that the solution will be smooth. We shall therefore work more formally, as in Section 9.3.2, and seek a solution of the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{9.50}$$

in the form

$$u = X(x)Y(y)$$

which gives on substitution

$$Y \frac{d^2 X}{dx^2} + X \frac{d^2 Y}{dy^2} = 0$$

or

$$\frac{1}{X} \frac{d^2X}{dx^2} = -\frac{1}{Y} \frac{d^2Y}{dy^2} = \lambda \quad (9.51)$$

Now $(d^2X/dx^2)/X$ is a function of x only and $-(d^2Y/dy^2)/Y$ is a function of y only. Since they must be equal for all x and y , both sides of (9.51) must be a constant, λ . We therefore obtain two equations of simple-harmonic type

$$\frac{d^2X}{dx^2} = \lambda X, \quad \frac{d^2Y}{dy^2} = -\lambda Y$$

The type of solution depends on the sign of λ , and we have a variety of possible solutions:

$$\lambda = -\mu^2 < 0: \quad u = (A \sin \mu x + B \cos \mu x)(C e^{\mu y} + D e^{-\mu y}) \quad (9.52a)$$

$$\lambda = \mu^2 > 0: \quad u = (A e^{\mu x} + B e^{-\mu x})(C \sin \mu y + D \cos \mu y) \quad (9.52b)$$

$$\lambda = 0: \quad u = (Ax + B)(Cy + D) \quad (9.52c)$$

where A, B, C and D are arbitrary constants. Using the definitions of the hyperbolic functions, it is sometimes more convenient to express the solution (9.52a) as

$$u = (A \sin \mu x + B \cos \mu x)(C \cosh \mu y + D \sinh \mu y) \quad (9.52d)$$

and (9.52b) as

$$u = (A \sinh \mu x + B \cosh \mu x)(C \sin \mu y + D \cos \mu y) \quad (9.52e)$$

The actual form of the solution depends on the problem in hand, as illustrated in the following examples.

Example 9.29

Use the separated solutions (9.52) of the Laplace equation to find the solution to (9.50) satisfying the boundary conditions

$$u(x, 0) = 0 \quad (0 < x < 2)$$

$$u(x, 1) = 0 \quad (0 < x < 2)$$

$$u(0, y) = 0 \quad (0 < y < 1)$$

$$u(2, y) = a \sin 2\pi y \quad (0 < y < 1)$$

Solution To satisfy the first two conditions, we need to choose the separated solutions that include the $\sin \mu y$ terms. Thus we take solution (9.52b)

$$u = (A e^{\mu x} + B e^{-\mu x})(C \sin \mu y + D \cos \mu y)$$

The first boundary condition gives

$$(A e^{\mu x} + B e^{-\mu x})D = 0 \quad (0 < x < 2)$$

so that $D = 0$. Thus

$$u = (A' e^{\mu x} + B' e^{-\mu x}) \sin \mu y$$

where $A' = AC$ and $B' = BC$. The second boundary condition then gives

$$(A' e^{\mu x} + B' e^{-\mu x}) \sin \mu y = 0 \quad (0 < x < 2)$$

so that $\sin \mu = 0$, or $\mu = n\pi$ with n as integer. Thus

$$u = (A' e^{n\pi x} + B' e^{-n\pi x}) \sin n\pi y$$

From the third boundary condition,

$$(A' + B') \sin n\pi y = 0 \quad (0 < y < 1)$$

so that $B' = -A'$, giving

$$\begin{aligned} u &= A'(e^{n\pi x} - e^{-n\pi x}) \sin n\pi y \\ &= 2A' \sinh n\pi x \sin n\pi y \end{aligned}$$

The final boundary condition then gives

$$2A' \sinh 2n\pi \sin n\pi y = a \sin 2\pi y \quad (0 < y < 1)$$

We must therefore choose $n = 2$, and $a = 2A' \sinh 2n\pi = 2A' \sinh 4\pi$, or $2A' = a/\sinh 4\pi$. The solution is therefore

$$u = a \sin 2\pi y \frac{\sinh 2\pi x}{\sinh 4\pi}$$

Example 9.30

Solve the Laplace equation (9.50) for steady heat conduction in the semi-infinite region $0 \leq y \leq 1$, $x \geq 0$ and subject to the boundary conditions

- | | | |
|-----------------------------|---------------------------|--|
| (a) $u(x, 0) = 0$ | $(x \geq 0)$ | (temperature kept at zero on two sides and at infinity); |
| (b) $u(x, 1) = 0$ | $(x \geq 0)$ | |
| (c) $u(x, y) \rightarrow 0$ | as $x \rightarrow \infty$ | |
- (d) $u(0, y) = 1$ ($0 \leq y \leq 1$) (unit temperature on the fourth side).

Solution Clearly from condition (c) we need a solution that is exponential in x , so we take (9.52b):

$$u = (A e^{\mu x} + B e^{-\mu x})(C \sin \mu y + D \cos \mu y)$$

and since the solution must tend to zero as $x \rightarrow \infty$, we have $A = 0$, giving

$$u = e^{-\mu x}(C' \sin \mu y + D' \cos \mu y)$$

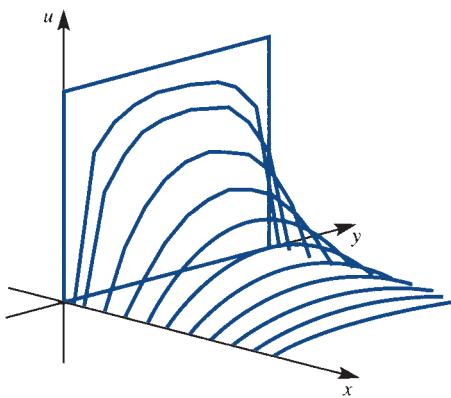
where $C' = BC$ and $D' = BD$. Condition (a) then gives $D' = 0$, and (b) gives $\sin \mu = 0$, or $\mu = n\pi$ ($n = 1, 2, \dots$), so the solution becomes $u = C' e^{-n\pi x} \sin n\pi y$ ($n = 1, 2, \dots$). Because of the linearity of the Laplace equation, we sum over n to obtain the more general solution

$$u = \sum_{n=1}^{\infty} C'_n e^{-n\pi x} \sin n\pi y$$

Condition (d) then gives, as before, a classic Fourier series problem

$$1 = \sum_{n=1}^{\infty} C'_n \sin n\pi y \quad (0 \leq y \leq 1)$$

Figure 9.31 Solution of the Laplace equation in Example 9.30.



so that, using (7.17),

$$C'_n = 2 \int_0^1 \sin n\pi y \, dy = \begin{cases} 4/n\pi & (\text{odd } n) \\ 0 & (\text{even } n) \end{cases}$$

The complete solution is therefore

$$u = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} e^{-(2n-1)\pi x} \sin(2n-1)\pi y$$

or, in expanded form,

$$u = \frac{4}{\pi} \left(e^{-\pi x} \sin \pi y + \frac{1}{3} e^{-3\pi x} \sin 3\pi y + \frac{1}{5} e^{-5\pi x} \sin 5\pi y + \dots \right)$$

In Figure 9.31 the solution $u(x, t)$ is plotted in the (x, y) plane. Because of the discontinuity at $x = 0$, for $x = 0.05$ thirty terms of the series were required to compute u to four-figure accuracy, while for $x = 1$, one or two terms were quite sufficient.

It is clear from Example 9.30 that the solutions (9.52) can only be used for rectangular regions. For various cylindrical and spherically symmetric regions separated solutions can be constructed, but they need more complicated Bessel and Legendre functions. The solutions have the same structure as for rectangular regions and follow the general theory of orthogonal functions discussed in Section 7.5.2. For instance the study of Legendre polynomials is required for problems similar to Example 9.24 when angular dependence is included. There is great merit in calculating exact solutions where we can, since they give significant insight. However, with modern computing techniques it is certainly not necessarily quicker than a straight numerical solution.

Example 9.31

Solve the Laplace equation (9.50) in the region $0 \leq x \leq 1$, $0 \leq y \leq 2$ with the conditions

- (a) $u(x, 0) = x$ ($0 \leq x \leq 1$)
- (b) $u(x, 2) = 0$ ($0 \leq x \leq 1$)
- (c) $u(0, y) = 0$ ($0 \leq y \leq 2$)
- (d) $\partial u(1, y)/\partial x = 0$ ($0 \leq y \leq 2$)

Solution

The steady heat-conduction interpretation of this problem, looking at Figure 9.32, gives a zero temperature on ABC, an insulated boundary on CD and a linear temperature on AD.

Of the solutions (9.52), we require zeros on AB and zero derivative on CD, so we might expect to use trigonometric solutions in the x direction and exponential (or equivalently sinh and cosh) solutions in the y direction. We therefore take a solution of the form (9.52d):

$$u = (A \sin \mu x + B \cos \mu x)(C \cosh \mu y + D \sinh \mu y)$$

From condition (c), we must take $B = 0$, giving

$$u = (C' \cosh \mu y + D' \sinh \mu y) \sin \mu x$$

where $C' = AC$ and $D' = AD$. Condition (d) then gives $\cos \mu = 0$ or

$$\mu = (n + \frac{1}{2})\pi \quad (n = 0, 1, 2, \dots)$$

so the solution becomes

$$u = [C' \cosh(n + \frac{1}{2})\pi y + D' \sinh(n + \frac{1}{2})\pi y] \sin(n + \frac{1}{2})\pi x \quad (n = 0, 1, 2, \dots) \quad (9.53)$$

To satisfy condition (b), it is best to rewrite (9.53) in the equivalent form

$$u = \sin[(n + \frac{1}{2})\pi x] \{E \cosh[(n + \frac{1}{2})\pi(2 - y)] \\ + F \sinh[(n + \frac{1}{2})\pi(2 - y)]\} \quad (n = 0, 1, 2, \dots)$$

We see that (b) now implies $E = 0$, so that our basic solution, summed over all n , is

$$u = \sum_{n=0}^{\infty} F_n \sin[(n + \frac{1}{2})\pi x] \sinh[(n + \frac{1}{2})\pi(2 - y)]$$

The final condition (a) then gives the standard Fourier series problem

$$x = \sum_{n=0}^{\infty} F_n \sinh[(2n + 1)\pi] \sin[(n + \frac{1}{2})\pi x]$$

so that, using (7.17),

$$\frac{1}{2} F_n \sinh(2n + 1)\pi = \int_0^1 x \sin[(n + \frac{1}{2})\pi x] dx = \frac{\sin(n + \frac{1}{2})\pi}{\pi^2(n + \frac{1}{2})^2}$$

The solution in expanded form is therefore

$$u = \frac{8}{\pi^2} \left[\sin \frac{1}{2}\pi x \frac{\sinh \frac{1}{2}\pi(2 - y)}{\sinh \pi} - \frac{\sin \frac{3}{2}\pi x \sinh \frac{3}{2}\pi(2 - y)}{9 \sinh 3\pi} \right. \\ \left. + \sin \frac{5}{2}\pi x \frac{\sinh \frac{5}{2}\pi(2 - y)}{25 \sinh 5\pi} + \dots \right]$$

Curiously, Laplace transform solutions are not natural for the Laplace equation, since there is no obvious semi-infinite parameter. Even in cases like Example 9.30, where we have a semi-infinite region, the Laplace transform in x requires information that is not available, see Section 9.8.2.

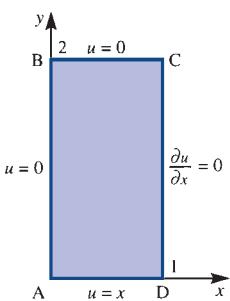


Figure 9.32
Region and boundary conditions for Example 9.31.

Another technique for solution of the Laplace equation involves complex variables and was discussed in Section 4.3.2. It is a method that was very widely used in aerodynamics and in electrostatic problems. Since the advent of modern computers, with highly efficient Laplace solvers, the method has fallen somewhat into disuse. It was the cornerstone of all early flight calculations and, for those interested either in the historical context or in the beautiful mathematical theory, the study of complex-variable solutions is essential. The real and imaginary parts of a differentiable function $f(z)$ of the complex variable $z = x + jy$ automatically satisfy the Laplace equation. Example 9.6 showed a solution that was interpreted as the flow past a cylinder; the function was obtained from the imaginary part of

$$f(z) = U \left(z + \frac{a^2}{z} \right)$$

It is then possible to use the Kutta–Joukowski transformation (see D. Acheson, *Elementary Fluid Dynamics*, Oxford, Oxford University Press, 2002) to transform the circle to an aerofoil shape and the lift and drag on the aerofoil can be computed. Example 9.32 illustrates a much simpler situation.

Example 9.32

If $f(z) = \phi(x, y) + j\psi(x, y)$ is a complex function of the complex variable $z = x + jy$, verify that ϕ and ψ satisfy the Laplace equation for the case $f(z) = z^2$. Sketch the contours of $\phi = \text{constant}$ and $\psi = \text{constant}$.

Solution

Now

$$f(z) = z^2 = (x^2 - y^2) + j2xy$$

and thus

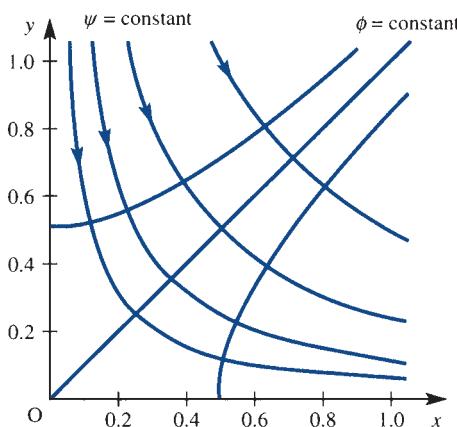
$$\phi = x^2 - y^2, \quad \psi = 2xy$$

It is trivial to differentiate these functions, and both clearly satisfy the Laplace equation:

$$\nabla^2 \phi = 0, \quad \nabla^2 \psi = 0$$

The contours of ϕ and ψ are plotted in Figure 9.33. They are both hyperbolas, which intersect at right angles. The usual interpretation of these solutions is as irrotational inviscid fluid flow into a corner.

Figure 9.33
Complex-variable solution to the Laplace equation in Example 9.32, showing the streamlines of flow into a corner.



If we now try to solve the Poisson equation (9.10), which is an extension of the Laplace equation with heat sources/sinks $f(x, y)$ on the right-hand side

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$

the problem becomes harder as illustrated in Example 9.33.

Example 9.33

Solve the Poisson equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -\sin \frac{\pi x}{a} \cos \frac{\pi y}{b}$$

in the rectangle $0 \leq x \leq a$, $0 \leq y \leq b$ given the boundary conditions

$$u = 0 \text{ on } x = 0, \quad u = f(y) \text{ on } x = a$$

$$\frac{\partial u}{\partial y} = 0 \quad \text{on } y = 0 \text{ and } y = b$$

Solution

Physically the problem can be interpreted as a heated plate with the temperature specified on the two boundaries $x = 0$ and $x = a$ and with insulated boundaries $y = 0$ and $y = b$.

The general strategy is to find a ‘particular integral’ to eliminate the term on the right-hand side, compute the new boundary conditions and then solve the residual Laplace equation. In the present case choose

$$U = K \sin \frac{\pi x}{a} \cos \frac{\pi y}{b}$$

Substitute into the Poisson equation to give

$$\nabla^2 U = -K \left(\frac{\pi^2}{a^2} + \frac{\pi^2}{b^2} \right) \sin \frac{\pi x}{a} \cos \frac{\pi y}{b}$$

and hence

$$K^{-1} = \pi^2 \left(\frac{1}{a^2} + \frac{1}{b^2} \right)$$

Now put

$$u = U + v$$

so that v satisfies the Laplace equation

$$\nabla^2 v = 0$$

and the boundary conditions remain the same

$$v = 0 \text{ on } x = 0, \quad v = f(y) \text{ on } x = a$$

$$\frac{\partial v}{\partial y} = 0 \quad \text{on } y = 0 \text{ and } y = b$$

We are now back to a standard Laplace equation problem that can be solved by separation of variables. From the solutions (9.52) choose

$$v = \frac{1}{2}A_0x$$

$$v = A_n \sinh \frac{n\pi x}{a} \cos \frac{n\pi y}{b} \quad n = 1, 2, 3, \dots$$

These solutions satisfy three of the boundary conditions, just leaving

$$v = f(y) \quad \text{on } x = a$$

to be satisfied. The usual infinite sum of terms is constructed

$$v = \frac{1}{2}A_0x + \sum_{n=1}^{\infty} A_n \sinh \frac{n\pi x}{a} \cos \frac{n\pi y}{b}$$

so that on $x = a$ the remaining boundary condition gives the usual Fourier cosine series problem

$$f(y) = \frac{1}{2}A_0a + \sum_{n=1}^{\infty} (A_n \sinh n\pi) \cos \frac{n\pi y}{b}$$

In Chapter 7 (7.14) and (7.15) give

$$a_0 = A_0a = \frac{2}{b} \int_0^b f(y) dy$$

$$\text{and } a_n = A_n \sinh n\pi = \frac{2}{b} \int_0^b f(y) \cos \left(\frac{n\pi y}{b} \right) dy \quad n = 1, 2, 3, \dots$$

The final solution is given by

$$u = \frac{a^2 b^2 \sin \left(\frac{\pi x}{a} \right) \cos \left(\frac{\pi y}{b} \right)}{a^2 + b^2} + \frac{a_0}{a} x + \sum_{n=1}^{\infty} \frac{a_n}{\sinh n\pi} \sinh \frac{n\pi x}{a} \cos \frac{n\pi y}{b}$$

The method of solution described in Example 9.33 can be quite difficult and clumsy. Finding the ‘particular integral’, U , is not always easy even for simple right-hand sides. If U can be found, the new boundary conditions on v can become very awkward, often further substitutions need to be made to bring the problem to tractable form. In Example 9.33 the right-hand side was carefully chosen to avoid this extra difficulty. An alternative method using Green’s functions will be considered in Section 9.7.2. The solution turns out to be very neat with the right-hand side and the boundary conditions appearing naturally in various integrals. However, although neat, the computation of the Green’s function is just as difficult as the method described in Example 9.33.

9.5.2 Exercises

- 44 Use the separated solutions (9.52) to solve the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

in the region $0 \leq x \leq 1, y \geq 0$ given the boundary conditions

- (a) $u = 0$ on $x = 0$ and $x = 1$ ($y \geq 0$)
- (b) $u \rightarrow 0$ as $y \rightarrow \infty$ ($0 \leq x \leq 1$)
- (c) $u = \sin^5(\pi x)$ on $y = 0$ ($0 \leq x \leq 1$)

(Note: The identity $\sin^5 \theta = \frac{1}{16} (\sin 5\theta - 5 \sin 3\theta + 10 \sin \theta)$.)

- 45 Show that the function $u(x, y) = e^{-\pi y/2} [y \cos(\pi y/2) - x \sin(\pi y/2)]$ satisfies the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

and the boundary conditions

- (a) $u = 0$ on $y = 0$ ($x \geq 0$)
- (b) $u = -xe^{-\pi y/2}$ on $y = 1$ ($x \geq 0$)
- (c) $u = y \cos(\pi y/2)$ on $x = 0$ ($0 \leq y \leq 1$)
- (d) $u \rightarrow 0$ as $x \rightarrow \infty$

- 46 Show that the function $\phi(x, y) = x^2 y$ satisfies the Poisson equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 2y$$

By putting $\phi = u + x^2 y$, show that u satisfies the Laplace equation. Find the solution for ϕ in the unit square which satisfies the boundary conditions

$$\left. \begin{aligned} \phi(x, 0) &= 0 \\ \phi(x, 1) &= x^2 + \sin \pi x \end{aligned} \right\} \text{for } 0 \leq x \leq 1$$

$$\left. \begin{aligned} \phi(0, y) &= 0 \\ \phi(1, y) &= y \end{aligned} \right\} \text{for } 0 \leq y \leq 1$$

- 47 Show that

$$u(r, \theta) = Br^n \sin n\theta$$

satisfies the Laplace equation in polar coordinates,

$$u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0$$

Determine u that is both finite for $r \leq a$ and periodic in θ , given that

$$u(a, \theta) = \sin^3 \theta = \frac{3}{4} \sin \theta - \frac{1}{4} \sin 3\theta$$

- 48 Verify that

$$u = \frac{-2y}{x^2 + y^2 + 2x + 1}, \quad v = \frac{x^2 + y^2 - 1}{x^2 + y^2 + 2x + 1}$$

both satisfy the Laplace equation, and sketch the curves $u = \text{constant}$ and $v = \text{constant}$. Show that

$$u + jv = \frac{j(z-1)}{z+1}$$

where $z = x + jy$.

- 49

(Hadamard example) Show that the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

with $u(0, y) = 0$, $u_x(0, y) = (1/n) \sin ny$ ($n > 0$) has the solution

$$u(x, y) = \frac{1}{n^2} \sinh nx \sin ny$$

Compare this solution, for large n , with the solution to the ‘neighbouring’ problem, when $u(0, y) = 0$, $u_x(0, y) = 0$, and the solution $u(x, y) = 0$.

- 50

Solve the Laplace equation $\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$ in the region $0 < x < 1, 0 < y < 1$ subject to the boundary conditions $u(0, y) = 0$, $u(x, 0) = 0$, $u(1, y) = 1$, $u(x, 1) = 1$ by separation methods.

- 51

A long bar of square cross-section $0 \leq x \leq a$, $0 \leq y \leq a$ has the faces $x = 0$, $x = a$ and $y = 0$ maintained at zero temperature, and the face $y = a$ at a control temperature u_0 . Under steady-state conditions the temperature $u(x, y)$ at a point in a cross-section satisfies the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

Write down the boundary conditions for $u(x, y)$, and hence show that $u(x, y)$ is given by

$$\begin{aligned} u(x, y) &= \frac{4u_0}{\pi} \sum_{n=0}^{\infty} \frac{\operatorname{cosech}(2n+1)\pi}{2n+1} \\ &\quad \times \sinh\left[(2n+1)\frac{\pi y}{a}\right] \sin\left[(2n+1)\frac{\pi x}{a}\right] \end{aligned}$$

- 52 Heat is flowing steadily in a metal plate whose shape is an infinite rectangle occupying the region $-a < x < a$, $y > 0$ of the (x, y) plane. The temperature at the point (x, y) is denoted by $u(x, y)$. The sides $x = \pm a$ are insulated, the temperature approaches zero as $y \rightarrow \infty$, while the side $y = 0$ is maintained at a fixed temperature $-T$ for $-a < x < 0$ and T for $0 < x < a$. It is known that $u(x, y)$ satisfies the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

and the boundary conditions

- (a) $u \rightarrow 0$ as $y \rightarrow \infty$ for all x in $-a < x < a$
 (b) $\partial u / \partial x = 0$ when $x = \pm a$

(c) $u(x, 0) = \begin{cases} -T & (-a < x < 0) \\ T & (0 < x < a) \end{cases}$

Using the method of separation, obtain the solution $u(x, y)$ in the form

$$u(x, y) = \frac{4T}{\pi} \sum_{n=0}^{\infty} \frac{1}{2n+1} \exp\left[-(n+\frac{1}{2})\frac{\pi y}{a}\right] \times \sin\left[(n+\frac{1}{2})\frac{\pi x}{a}\right]$$

- 53 A thin semicircular plate of radius a has its bounding diameter kept at zero temperature and its curved boundary at a constant temperature T_0 . The steady-state temperature $T(r, \theta)$ at a point having polar coordinates (r, θ) , referred to the centre of the circle as origin, is given by the Laplace equation

$$\frac{\partial^2 T}{\partial r^2} + \frac{1}{r} \frac{\partial T}{\partial r} + \frac{1}{r^2} \frac{\partial^2 T}{\partial \theta^2} = 0$$

Assuming a separated solution of the form

$$T = R(r)\Theta(\theta)$$

show that

$$T(r, \theta) = \frac{4T_0}{\pi} \sum_{n=0}^{\infty} \frac{(r/a)^{2n+1}}{2n+1} \sin(2n+1)\theta$$

The Laplace equation in spherical polar coordinates (r, θ, ϕ) takes the form

$$\frac{\partial}{\partial r} \left(r^2 \frac{\partial V}{\partial r} \right) + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial V}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2 V}{\partial \phi^2} = 0$$

If V is only a function of r and θ , and V takes the form

$$V = R(r)y(x), \quad \text{where } x = \cos \theta$$

show that

$$\frac{d}{dr} \left(r^2 \frac{dR}{dr} \right) = k(k+1)R$$

$$(1-x^2) \frac{d^2y}{dx^2} - 2x \frac{dy}{dx} + k(k+1)y = 0$$

where k is a constant.

The function V satisfies the Laplace equation in the region $a \leq r \leq b$. On $r = a$, $V = 0$ and on $r = b$, $V = \alpha \sin^2 \theta$, where α is a constant. Given that solutions for y are

$$y = \begin{cases} 1 & (k=0) \\ x & (k=1) \\ \frac{1}{2}(3x^2 - 1) & (k=2) \end{cases}$$

find V throughout the region.

9.5.3 Numerical solution

Of the three classical partial differential equations, the Laplace equation proves to be the most difficult to solve. The other two have a natural time variable in them, and it is possible, with a little care, to march forward either by a simple explicit method or by an implicit procedure. In the case of the Laplace equation, information is given around the whole of the boundary of the solution region, so the field variables at *all* mesh points must be solved simultaneously. This in turn leads to a solution by matrix inversion.

The usual numerical approximation for the partial derivatives, discussed in Section 9.3.5, are employed, so that the equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{9.54}$$

at a typical point, illustrated in Figure 9.34, becomes

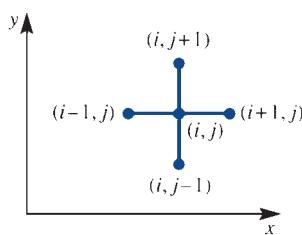
$$\frac{u(i+1, j) - 2u(i, j) + u(i-1, j)}{\Delta x^2} + \frac{u(i, j+1) - 2u(i, j) + u(i, j-1)}{\Delta y^2} = 0$$

For the case $\Delta x = \Delta y$ rearranging gives

$$4u(i, j) = u(i+1, j) + u(i-1, j) + u(i, j+1) + u(i, j-1) \quad (9.55)$$

In the typical five-point module (9.55) the increments Δx and Δy are taken to be the same and it is noted that the middle value $u(i, j)$ is the average of its four neighbours. This corresponds to the absence of ‘hot spots’. We now examine how (9.55) can be implemented.

Figure 9.34 Five-point computational module for the Laplace equation.



Example 9.34

Solve the Laplace equation (9.54) in the square region $0 \leq x \leq 1$, $0 \leq y \leq 1$ with the boundary conditions

- (a) $u = 0$ on $x = 0$
- (b) $u = 1$ on $x = 1$
- (c) $u = 0$ on $y = 0$
- (d) $u = 0$ on $y = 1$

Solution

For a first solution we take the simplest mesh, illustrated in Figure 9.35(a), which contains only four interior points labelled u_1 , u_2 , u_3 and u_4 . The four equations obtained from (9.55) are

$$4u_1 = 0 + 0 + u_2 + u_4$$

$$4u_2 = 0 + 1 + u_3 + u_1$$

$$4u_3 = 1 + 0 + u_4 + u_2$$

$$4u_4 = 0 + 0 + u_1 + u_3$$

which in turn can be written in matrix form as

$$\begin{bmatrix} 4 & -1 & 0 & -1 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ -1 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

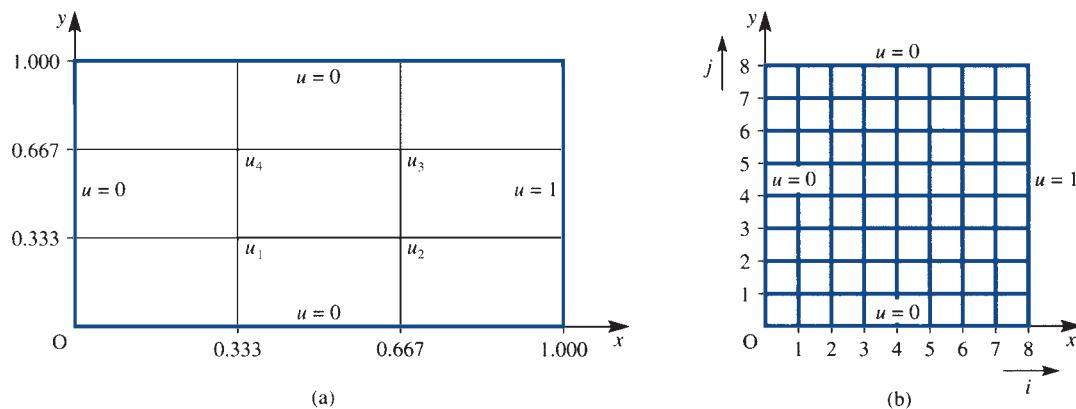


Figure 9.35 Meshes for the solution of the Laplace equation in Example 9.34: (a) a simple mesh containing 4 interior points; (b) a larger mesh with 49 interior points.

This has the solution $u_1 = 0.125$, $u_2 = 0.375$, $u_3 = 0.375$, $u_4 = 0.125$. A larger mesh obtained by dividing the sides up into eight equal parts is indicated in Figure 9.35(b). The equations now take the form

$$\begin{aligned}
 4u(1, 1) &= u(2, 1) + 0 && + u(1, 2) + 0 \\
 4u(2, 1) &= u(3, 1) + u(1, 1) + u(2, 2) + 0 \\
 &\vdots && \vdots \\
 4u(7, 1) &= 1 && + u(6, 1) + u(7, 2) + 0 \\
 4u(1, 2) &= u(2, 2) + 0 && + u(1, 3) + u(1, 1) \\
 4u(2, 2) &= u(3, 2) + u(1, 2) + u(2, 3) + u(2, 1) \\
 &\vdots && \vdots \\
 4u(7, 2) &= 1 && + u(6, 2) + u(7, 3) + u(7, 1) \\
 &\vdots && \vdots
 \end{aligned}$$

We thus generate 49 linear equations in 49 unknowns, which can be solved by any convenient matrix inverter. The matrices take the block form

$$\mathbf{A} = \begin{bmatrix} 4 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 4 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 4 \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

so that the equations become

The matrix equation (9.56) can be solved by an elimination technique or an iterative method like **successive over-relaxation** (SOR). As indicated in Section 5.5.4 of MEM, SOR is the simplest to program, and elimination techniques are best performed by a package from a computer library. For the current problem we present the solution in Figure 9.36, where the cases $\Delta x = \frac{1}{8}$ and $\Delta x = \frac{1}{4}$ are both shown. It may be seen from this example that the accuracy of the solution is quite tolerable when the cases $\Delta x = \frac{1}{4}$ and $\Delta x = \frac{1}{8}$ are compared.

Figure 9.36

The solution of Example 9.34. The solution is symmetric about the line $j = 4$; the solution with $\Delta x = 0.125$ is given in the upper half and the solution with $\Delta x = 0.25$ is shown in parentheses in the lower half.

<i>j</i> values	0	1	2	3	4	5	6	7	8	<i>i</i> values
8	0	0	0	0	0	0	0	0	1	
7	0	0.017	0.038	0.064	0.103	0.164	0.269	0.483	1	
6	0	0.032	0.069	0.117	0.184	0.282	0.431	0.661	1	
5	0	0.042	0.089	0.150	0.233	0.350	0.512	0.731	1	
4	0	0.045	0.096	0.162	0.250	0.371	0.536	0.749	1	
		(0.098)		(0.250)		(0.527)				
3										
2	0		(0.071)		(0.188)		(0.429)		1	
1										
0	0		0		0		0		1	

Note the averaging behaviour of the Laplace equation and observe that the discontinuity in the corner does not spread into the solution. The corner nodes are never used in the numerical calculation, so the discontinuity is avoided.



The solution of the Laplace equation is often required so packages like MATLAB have the machinery to set up the solution for simple regions. For the 9×9 problem, with 49 unknowns, the code is listed; note that the MATLAB numbering of the nodes is different from the text.

```
G=numgrid('S',9) % sets up the numbering for a 9 x 9 square
A=delsq(G); % stored as a sparse matrix
rhs=zeros(49,1);rhs(43:49,1)=ones(7,1); % computes rhs
A\rhs % gives the quoted solution
```

Because of its simplicity, SOR is an attractive method for solving Laplace-type problems. Equations (9.55) are rewritten with an iteration superscript as

$$\begin{aligned} u^{n+1}(i,j) = & u^n(i,j) + \frac{1}{4}w[u^n(i+1,j) + u^n(i-1,j) + u^n(i,j+1) \\ & + u^n(i,j-1) - 4u^n(i,j)] \end{aligned} \quad (9.57)$$

and w is a relaxation factor, discussed in Chapter 5 of MEM.

Knowing all the $u(i,j)$ at iteration n , we can use (9.57) to evaluate $u(i,j)$ at iteration $n+1$. Normally the $u(i,j)$ are over-written in the computer as they are computed, so that some of the ns in the right-hand side of (9.57) become $(n+1)s$. The order of evaluation of the is and js in (9.57) is critical, but the most obvious methods by rows or columns prove to be satisfactory.

A great deal is known about the optimum relaxation factor w . It is closely related to the value of the maximum eigenvalue of the matrix associated with the problem. For square regions with unit side and with u given on the boundary and equal mesh spacing it can be shown that $w = 2/(1 + \sin \Delta x)$ is the best value. For other problems this is usually used as a starting guess, but numerical experimentation is required to determine an optimum or near-optimum value.

We have only considered u to be given on the boundary, and it is essential to know how to deal with derivative boundary conditions, since these are very common. Let us consider a typical example:

$$\frac{\partial u}{\partial x} = g(y) \quad \text{on } x=0$$

We then insert a fictitious line of nodes, as shown in Figure 9.37. Approximately, the boundary condition gives

$$u(1,j) - u(-1,j) = g(y_j)2\Delta x$$

so that

$$u(-1,j) = u(1,j) - 2\Delta x g(y_j) \quad (9.58)$$

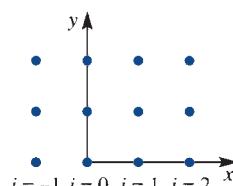


Figure 9.37

Fictitious nodes, $i = -1$, introduced outside the boundary.

Equations (9.55) or (9.57) are now solved for $i = 0$ as well as $i > 0$, but at the end of a sweep $u(-1,j)$ will be updated via (9.58).

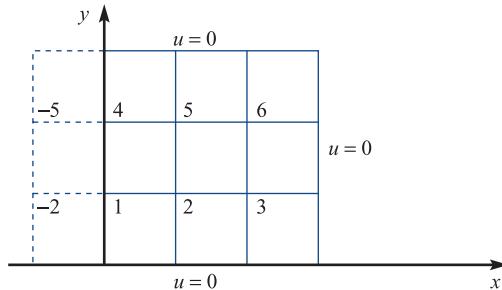
Example 9.35

Solve the Laplace equation (9.54) for steady-state heat conduction in the unit square, given that

- (a) $\partial u / \partial x = \frac{1}{2} - y$, $x = 0$ (steady heat supply on this boundary);
 (b) $u = 0$ on $x = 1$, $y = 0$, $y = 1$ (zero temperature on the other three sides).

Use $\Delta x = \Delta y = \frac{1}{3}$.

Figure 9.38 The mesh for Example 9.35.

**Solution**

Labelling the six unknown values u_1, u_2, \dots, u_6 as shown in Figure 9.38, equation (9.55) gives

$$\begin{aligned} 4u_1 &= 0 + u_2 + u_4 + u_{-2} \\ 4u_2 &= 0 + u_3 + u_5 + u_1 \\ 4u_3 &= 0 + 0 + u_6 + u_2 \\ 4u_4 &= u_1 + u_5 + 0 + u_{-5} \\ 4u_5 &= u_2 + u_6 + 0 + u_4 \\ 4u_6 &= u_3 + 0 + 0 + u_5 \end{aligned}$$

The values u_{-2} and u_{-5} are evaluated from boundary condition (a)

$$u_5 - u_{-5} = 2h\left(\frac{1}{2} - \frac{2}{3}\right) = -\frac{1}{9}, \quad u_2 - u_{-2} = 2h\left(\frac{1}{2} - \frac{1}{3}\right) = \frac{1}{9}$$

so the equations become

$$\begin{aligned} 4u_1 &= 2u_2 + u_4 - \frac{1}{9} \\ 4u_2 &= u_1 + u_3 + u_5 \\ 4u_3 &= u_2 + u_6 \\ 4u_4 &= u_1 + 2u_5 + \frac{1}{9} \\ 4u_5 &= u_2 + u_4 + u_6 \\ 4u_6 &= u_3 + u_5 \end{aligned}$$

Thus there are six linear equations in six unknowns, which can be solved by any convenient method. For instance, SOR as suggested in (9.57) gives the set of equations with iteration counter n and relaxation factor w

$$\begin{aligned}
 u_1^{n+1} &= u_1^n + \frac{w}{4}(2u_2^n + u_4^n - \frac{1}{9} - 4u_1^n) \\
 u_2^{n+1} &= u_2^n + \frac{w}{4}(u_1^{n+1} + u_3^n + u_5^n - 4u_2^n) \\
 u_3^{n+1} &= u_3^n + \frac{w}{4}(u_2^{n+1} + u_6^n - 4u_3^n) \\
 u_4^{n+1} &= u_4^n + \frac{w}{4}(u_1^{n+1} + 2u_5^n + \frac{1}{9} - 4u_4^n) \\
 u_5^{n+1} &= u_5^n + \frac{w}{4}(u_2^{n+1} + u_4^{n+1} + u_6^n - 4u_5^n) \\
 u_6^{n+1} &= u_6^n + \frac{w}{4}(u_3^{n+1} + u_5^{n+1} - 4u_6^n)
 \end{aligned}$$

The equations are diagonally dominant so the iterations converge quickly; six significant figures are obtained in 11 iterations with $w = 1$ and at near optimum $w = 1.2$ in 8 iterations.

	u_1	u_2	u_3	u_4	u_5	u_6
$h = \frac{1}{3}$	-0.024 24	-0.005 05	-0.001 01	0.024 24	0.005 05	0.001 01
$h = \frac{1}{6}$	-0.031 21	-0.005 17	-0.000 77	0.031 21	0.005 17	0.000 77
$h = \frac{1}{12}$	-0.033 57	-0.005 22	-0.000 68	0.033 57	0.005 22	0.000 68

The expected symmetry is observed from the solution, physically heat is supplied to the bottom half of the left-hand boundary and an equal amount is extracted from the top half. For comparison of the accuracy, the calculations with $h = \frac{1}{6}$ and $\frac{1}{12}$ have been included in the table.

Example 9.36

Solve the Poisson equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -1$$

for steady heat conduction with a heating term, in the unit square given that

- (a) $\partial u / \partial x = 0$ on $x = 0$ (insulated along this boundary);
 - (b) $u = y^2$ on $x = 1$
 - (c) $u = 0$ on $y = 0$
 - (d) $u = x$ on $y = 1$
- (temperature given on three sides).

Solution

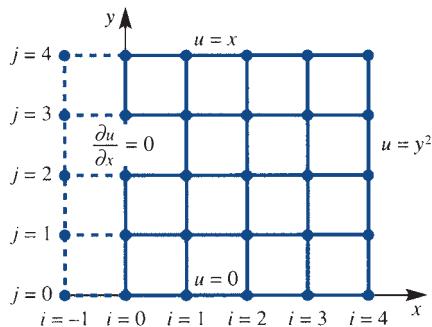
The region is illustrated in Figure 9.39, with mesh spacing $\Delta x = \Delta y = \frac{1}{4}$. Equation (9.58) just gives $u(-1, j) = u(1, j)$ for each j . Equation (9.55) is modified to take into account the right-hand side of the Poisson equation. The value Δx^2 is added to the right-hand side of (9.55) for each of the interior points. We can write the equations as

$$4u(0, 1) = u(-1, 1) + u(1, 1) + 0 + u(0, 2) = 2u(1, 1) + u(0, 2)$$

$$4u(1, 1) = u(0, 1) + u(2, 1) + 0 + u(1, 2) + \frac{1}{16}$$

$$4u(2, 1) = u(1, 1) + u(3, 1) + 0 + u(2, 2) + \frac{1}{16}$$

Figure 9.39 The mesh used in Example 9.36.



and so on, and hence obtain 12 equations in 12 unknowns. These can be solved by any convenient method to give a solution as shown in Figure 9.40.

Figure 9.40 Data from the solution of Example 9.36 using a step length 0.25 in each direction.

$j = 4$	0.0000	0.2500	0.5000	0.7500	1.0000
$j = 3$	0.2050	0.3021	0.4278	0.5329	0.5625
$j = 2$	0.2160	0.2630	0.3137	0.3290	0.2500
$j = 1$	0.1329	0.1578	0.1727	0.1567	0.0625
$j = 0$	0.0000	0.0000	0.0000	0.0000	0.0000
	$i = 0$	$i = 1$	$i = 2$	$i = 3$	$i = 4$

9.5.4 Exercises

- 55 Use the five-point difference approximation in (9.55) to solve



$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (0 \leq x \leq 1, 0 \leq y \leq 1)$$

where $u(x, 0) = u(0, y) = 0$, $u(x, 1) = x$, $u(1, y) = y(2 - y)$. Find the approximations for $u(\frac{1}{2}, \frac{1}{2})$ for grid sizes $\Delta x = \Delta y = \frac{1}{2}$ and $\Delta x = \Delta y = \frac{1}{4}$.

- 56 Use a mesh $\Delta x = \Delta y = \frac{1}{2}$ and $\Delta x = \Delta y = \frac{1}{4}$ to solve



$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad (0 < x < 1, 0 < y < 1)$$

satisfying $u(0, y) = 1$, $\partial u(1, y)/\partial x = 0$, $u(x, 0) = 0$, $u(x, 1) = 1$.

- 57 A numerical solution is to be determined for the loading of a uniform plate, where the displacement w satisfies the equation



$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + 20 = 0$$

and a square mesh of side h is used. Show that, at an interior point 0 with neighbours 1, 2, 3 and 4, the approximation to the equation is

$$4w_0 = w_1 + w_2 + w_3 + w_4 + 20h^2$$

The plate is in the shape of a trapezium whose vertices can be represented by the points $(0, 0)$, $(5, 0)$, $(2, 3)$ and $(0, 3)$. The plate is held on its edges so that on the boundary $w = 0$. Compute the solution for w at the five interior points if h is taken as 1.

- 58

The function $\phi(x, y)$ satisfies the equation



$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = x$$

and the boundary conditions (see Figure 9.41)

$$\phi = 3 - y^2 \quad \text{on OA } (x = 0, 0 \leq y \leq 1)$$

$$\frac{\partial \phi}{\partial y} = -\phi \quad \text{on AB } (y = 1, 0 < x < 1)$$

$$\phi = 1 \quad \text{on BC } (x = 1, 0 \leq y \leq 1)$$

$$\phi = 3 - x \quad \text{on CO } (y = 0, 0 < x < 1)$$

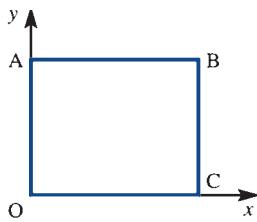


Figure 9.41 Region for Exercise 58.

Solve the equation numerically, using a mesh of
 (a) $h = \frac{1}{2}$ in each direction, (b) $h = \frac{1}{4}$ in each direction.

- 59 The function $\phi(x, y)$ satisfies the Laplace equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0$$

inside the region shown in Figure 9.42. The function ϕ takes the value $\phi = 9x^2$ at all points on the boundary. Making full use of symmetry, formulate

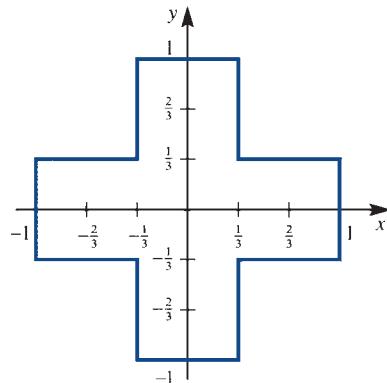


Figure 9.42 Region for Exercise 59.

a set of finite-difference equations to solve for the nodal values of ϕ on a square grid of side $h = \frac{1}{3}$. Solve for ϕ at the nodal points.

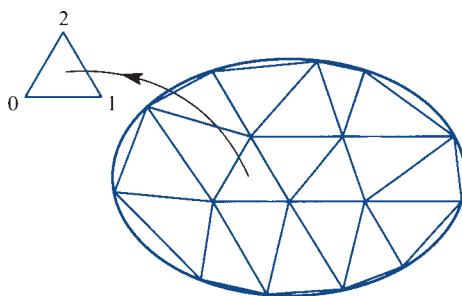
9.6 Finite elements

In Section 9.5.3 we sought numerical solutions of the Laplace equation, but we noted that only simple geometries could be handled using finite differences. The region described in Exercise 57 is about as difficult as can be treated easily. To adapt methods to awkward regions is not easy, so alternative strategies have been sought. Great advances took place in the 1960s, when civil engineers pioneered the method of finite elements. To solve plate bending problems, they solved the appropriate equations for small patches and then ‘stitched’ the latter together to form an overall solution. The job is not an easy one, and requires a large amount of arithmetic. It was only when large, fast computers became available that the method was viable. This method is now very widely used, and forms the basis of most calculations in stress analysis and for many fluid flows. It is very adaptable and physically satisfying, but is very difficult to program. This is in contrast to finite differences, which are reasonably easy. In general the advice to anyone employing this technique is to use a finite-element ‘package’, available in most computer libraries, and not to write one’s own program. For instance, in the Partial Differential Equation Toolbox in MATLAB, finite elements is the standard method of solution even for solutions in a rectangular region. As with many toolboxes of this type they need a lot of work to master all the details. It is important, however, to understand the basis of the method. We shall illustrate this method for a simple situation, but refer to specialist books for details and extensions: for example, see J. Whiteley *Finite Element Methods: A Practical Guide* (Berlin, Springer, 2017).

We consider solutions of the Poisson equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \rho(x, y) \quad (9.59)$$

Figure 9.43
Triangular finite-element mesh, with the local numbering of a typical element.



in a region R with u given on the boundary of R . The region R is divided up into a triangular mesh as in Figure 9.43. We aim to calculate the value of u at the nodal points of the mesh, but with the function suitably interpolated in each triangle. The simplest situation is obtained if, in a typical triangle, u is approximated as a linear function

$$u = ax + by + c \quad (9.60)$$

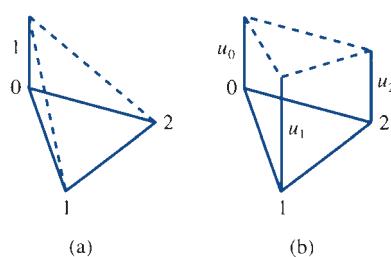
taking the values u_0 , u_1 and u_2 at the corners. This function can be written explicitly in terms of the functions

$$\begin{aligned} L_0 &= \begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} / \begin{vmatrix} x_0 & y_0 & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} \\ L_1 &= \begin{vmatrix} x & y & 1 \\ x_2 & y_2 & 1 \\ x_0 & y_0 & 1 \end{vmatrix} / \begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_0 & y_0 & 1 \end{vmatrix} \quad (9.61) \\ L_2 &= \begin{vmatrix} x & y & 1 \\ x_0 & y_0 & 1 \\ x_1 & y_1 & 1 \end{vmatrix} / \begin{vmatrix} x_2 & y_2 & 1 \\ x_0 & y_0 & 1 \\ x_1 & y_1 & 1 \end{vmatrix} \end{aligned}$$

each of which is taken to be zero outside the triangle with vertices (x_0, y_0) , (x_1, y_1) and (x_2, y_2) . The denominators are just $2A$, where A is the area of the triangle. The function L_0 , illustrated in Figure 9.44(a), takes the values 1 at (x_0, y_0) , 0 at (x_1, y_1) and (x_2, y_2) ; it is linear in the triangle and is taken to be zero elsewhere. The functions L_1 and L_2 behave similarly. The field variable u in the element, denoted by u^e , can now be written as

$$u^e(x, y) = u_0 L_0 + u_1 L_1 + u_2 L_2 \quad (9.62)$$

Figure 9.44
(a) The function L_0 ;
(b) u approximated as a linear function in the element.



The situation is illustrated in Figure 9.44(b). Note that $u^e(x, y)$ is a linear function in x and y , $u^e(x_0, y_0) = u_0$, $u^e(x_1, y_1) = u_1$ and $u^e(x_2, y_2) = u_2$, and hence gives an explicit form for the function in (9.60) that has the correct values at the nodes.

Example 9.37

Find the linear function that has the values u_0 at $(0, 0)$, u_1 at $(1, 1)$ and u_2 at $(\frac{1}{2}, 1)$.

Solution

From (9.61), the functions L_0 , L_1 and L_2 are given by

$$L_0 = \begin{vmatrix} x & y & 1 \\ 1 & 1 & 1 \\ \frac{1}{2} & 1 & 1 \end{vmatrix} \begin{vmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ \frac{1}{2} & 1 & 1 \end{vmatrix} = (\frac{1}{2} - \frac{1}{2}y)/\frac{1}{2} = 1 - y$$

$$L_1 = \begin{vmatrix} x & y & 1 \\ \frac{1}{2} & 1 & 1 \\ 0 & 0 & 1 \end{vmatrix} \begin{vmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{vmatrix} = 2x - y$$

$$L_2 = \begin{vmatrix} x & y & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{vmatrix} \begin{vmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{2} \end{vmatrix} = 2y - 2x$$

Thus, from (9.62), the required linear function is

$$u = (1 - y)u_0 + (2x - y)u_1 + 2(y - x)u_2$$

or

$$u = x(2u_1 - 2u_2) + y(-u_0 - u_1 + 2u_2) + u_0$$

We build up the solution of (9.59) as the sum over all the elements of the functions constructed to be linear in an element and zero outside the element. Thus

$$u = \sum u^e$$

To be of use, this function must satisfy (9.59) in some approximate sense. The function cannot be differentiated across the element boundaries, since it has discontinuous behaviour. We therefore have to satisfy the equation in an integrated or '**weak**' form.

We use the well-known result that if V is continuous and

$$\iint_R V\phi \, dx \, dy = 0 \tag{9.63}$$

for a complete set of functions ϕ (that is, a set of functions that will approximate any continuous function as accurately as desired) then $V \equiv 0$ in R . Using the residual of (9.59) in (9.63) gives

$$\begin{aligned}
0 &= \iint_R \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} - \rho \right) \phi dx dy \\
&= \iint_R \left[\frac{\partial}{\partial x} \left(\phi \frac{\partial u}{\partial x} \right) + \frac{\partial}{\partial y} \left(\phi \frac{\partial u}{\partial y} \right) - \frac{\partial u}{\partial x} \frac{\partial \phi}{\partial x} - \frac{\partial u}{\partial y} \frac{\partial \phi}{\partial y} - \rho \phi \right] dx dy \\
&= - \iint_R (u_x \phi_x + u_y \phi_y + \rho \phi) dx dy - \int_C (\phi u_y dx - \phi u_x dy)
\end{aligned}$$

where the final integral is obtained over the boundary C of R using Green's theorem

$$\iint_R \left(\frac{\partial N}{\partial x} - \frac{\partial M}{\partial y} \right) dx dy = \int_C (M dx + N dy)$$

described in Section 3.4.5.

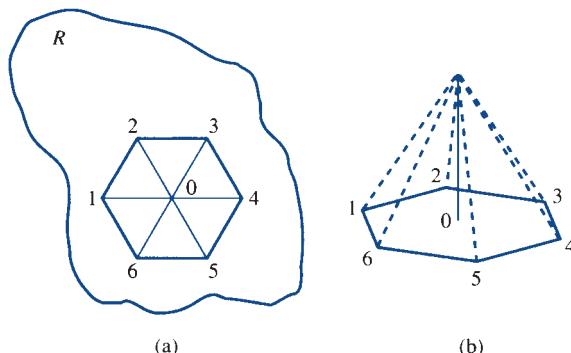
Choosing ϕ to be zero on the boundary C , we have an integrated form for (9.59) as

$$0 = \iint_R (u_x \phi_x + u_y \phi_y + \rho \phi) dx dy \quad (9.64)$$

It is this integrated or ‘weak’ form of (9.59) that we shall satisfy. We are therefore satisfying the equation in a global sense over the whole region. In comparison, finite-difference approximations are local to the mesh point. Clearly, we cannot use (9.64) with a complete set of functions ϕ , since there must be infinitely many of them. The best that can be done is to use N test functions ϕ_n ($n = 1, \dots, N$) when there are N interior nodes. There will then be N equations for the N unknowns u_n ($n = 1, \dots, N$) at the node points. As $N \rightarrow \infty$, the functions ϕ_n must form a complete set, and then the weak form of (9.64) will be satisfied identically rather than approximately. The most popular set of functions ϕ_i is that due to Galerkin, who used the pyramid functions illustrated in Figure 9.45. At a typical point 0 with neighbours 1, 2, 3, ..., m we have

$$\phi = \begin{cases} 1 & \text{at node 0} \\ 0 & \text{at nodes } 1, 2, \dots, m \end{cases} \quad \begin{array}{l} \text{and piecewise-linear in each} \\ \text{of the neighbouring triangles} \\ \text{identically zero outside the neighbouring triangles} \end{array} \quad (9.65)$$

Figure 9.45
(a) A typical node and its neighbours in the region R ;
(b) the pyramid function used at the typical node.



If there are N nodes in the mesh then there are N pyramid functions of the type (9.65). We substitute each of these functions in turn into (9.64) to satisfy the weak form of our original Poisson equation. Taking a typical node, we see that ϕ is piecewise-linear in the neighbouring triangles. For a typical such triangle 012, ϕ is just the linear function L_0 defined earlier. We substitute $\phi = L_0$ into the right-hand side of (9.64) and use the fact that

$$u = u_0 L_0 + u_1 L_1 + u_2 L_2$$

to obtain the contribution from this particular triangle as

$$\begin{aligned} I_e = \iint_{\Delta_{012}} & \left[\left(u_0 \frac{\partial L_0}{\partial x} + u_1 \frac{\partial L_1}{\partial x} + u_2 \frac{\partial L_2}{\partial x} \right) \frac{\partial L_0}{\partial x} \right. \\ & \left. + \left(u_0 \frac{\partial L_0}{\partial y} + u_1 \frac{\partial L_1}{\partial y} + u_2 \frac{\partial L_2}{\partial y} \right) \frac{\partial L_0}{\partial y} + \rho_e L_0 \right] dx dy \end{aligned}$$

where ρ_e is taken to be constant in the triangle. Since L_i ($i = 0, 1, 2$) are linear, $\partial L_i / \partial x$ and $\partial L_i / \partial y$ are constants, and hence the integrals can be performed explicitly, giving

$$\begin{aligned} I_e = \{u_0[(y_1 - y_2)^2 + (x_1 - x_2)^2] &+ u_1[(y_2 - y_0)(y_1 - y_2) + (x_2 - x_0)(x_1 - x_2)] \\ &+ u_2[(y_0 - y_1)(y_1 - y_2) + (x_0 - x_1)(x_1 - x_2)]\}/4A + \frac{1}{3}A\rho_e \end{aligned}$$

From (9.64) with ϕ chosen as (9.65), we obtain for the point 0 the sum of such terms over neighbouring elements:

$$\sum_e I_e = 0$$

This is just an equation of the form

$$\sum_{i=0}^m a_i u_i + b = 0$$

where the coefficients a_i and b depend only on the geometry and *not* on the field variables.

A similar computation is performed for each internal node, with ϕ taken to be of the form (9.65). The Poisson equation is linear in the u_i , and since there is one such equation for each internal node, we obtain N equations in the N unknowns u_i ($i = 1, \dots, N$). These form a matrix (called the **stiffness matrix**) equation, which can be solved for the u_i . The general strategy is

- (i) calculate all the coefficients;
- (ii) assemble the stiffness matrix;
- (iii) invert the matrix to obtain the unknowns u_i ($i = 1, 2, 3, \dots, N$);
- (iv) calculate any required data from the solution.

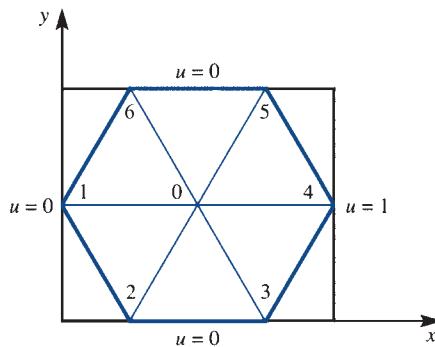


For the Laplace equation with linear approximating functions in triangular elements, a MATLAB implementation will be developed. In this development, note the complexity of the programming even for this very simple situation and also the organization of the input data required by the program. There has been no attempt at efficiency in the programming of the sections of the code.

Example 9.38

Solve the Laplace equation $\partial^2 u / \partial x^2 + \partial^2 u / \partial y^2 = 0$ in the unit square shown in Figure 9.46, subject to the boundary conditions indicated.

Figure 9.46
Mesh for the finite-element solution to Example 9.38.



Solution Here we have a simple rectangular region with $\rho = 0$, and which is the same problem as Example 9.34.

$$\begin{array}{lll}
 \Delta 012 & \text{gives a contribution} & 0.625u_0 - 0.375u_1 - 0.250u_2 \\
 \Delta 023 & " & 0.500u_0 - 0.250u_2 - 0.250u_3 \\
 \Delta 034 & " & 0.625u_0 - 0.250u_3 - 0.375u_4 \\
 \Delta 045 & " & 0.625u_0 - 0.375u_4 - 0.250u_5 \\
 \Delta 056 & " & 0.500u_0 - 0.250u_5 - 0.250u_6 \\
 \Delta 061 & " & 0.625u_0 - 0.250u_6 - 0.375u_1
 \end{array}$$

Adding all these contributions for the point 0, which is the only unspecified point, gives

$$0 = 3.5u_0 - 0.75u_1 - 0.50u_2 - 0.50u_3 - 0.75u_4 - 0.50u_5 - 0.50u_6$$

so that, knowing $u_1 = u_2 = u_3 = u_5 = u_6 = 0$ and $u_4 = 1$, we obtain $u_0 = 0.2143$. Comparing with Example 9.34, we see that our result is not particularly accurate, which is not surprising, since the mesh chosen here is particularly crude.

It is clear from Example 9.38 that the contributions from each triangle need considerable computational effort and the finite-element method is unsuitable for hand computations.



The coefficients in Example 9.38 can be computed from the MATLAB M-file stored under the name *coeff.m*. The coordinates of the vertices of the triangle are inserted as $a = [p \ q]$, $b = [r \ s]$ and $c = [u \ v]$. The coefficients are produced in a_0, a_1, a_2 .

```

function [a0,a1,a2]=coeff(a,b,c)
A=[a 1]; B=[b 1]; C=[c 1]; lx=[1 0 0]; ly=[0 1 0];
den=0.5/det([A;B;C]);
L0=[det([lx;B;C]) det([ly;B;C])];
L1=[det([lx;C;A]) det([ly;C;A])];
L2=[det([lx;A;B]) det([ly;A;B])];
a0=L0'*L0*den; a1=L1'*L0*den; a2=L2'*L0*den;

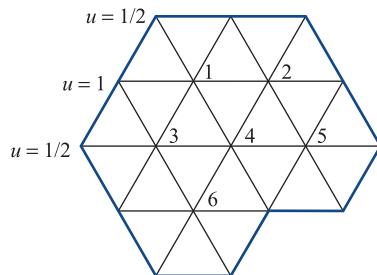
```

Such a function file will be used in a more general program.

Example 9.39

Solve the Laplace equation in the region shown in Figure 9.47 subject to zero boundary conditions except for the three points indicated. All the triangles are equilateral of side a .

Figure 9.47 Mesh for Example 9.39. The unmarked boundary points are given as $u = 0$.



Solution

Note that in the region in Figure 9.47 it would be very difficult to implement a standard finite-difference mesh. We now follow the general strategy.

- (i) Calculate the coefficients. When all the triangles are equilateral, the coefficients are all identical so the amount of computation is greatly reduced. For a typical point $\Delta 012$ gives a contribution $(4u_0 - 2u_1 - 2u_2)/4\sqrt{3}$

$\Delta 023$	"	$(4u_0 - 2u_2 - 2u_3)/4\sqrt{3}$
$\Delta 034$	"	$(4u_0 - 2u_3 - 2u_4)/4\sqrt{3}$
⋮		

and hence adding the six contributions gives the total for the typical point

$$(6u_0 - u_1 - u_2 - u_3 - u_4 - u_5 - u_6)/\sqrt{3}$$

- (ii) Assemble the stiffness matrix. Apply the results in (i) to each of the six active points

$$6u_1 = u_3 + u_4 + u_2 + 0 + \frac{1}{2} + 1$$

$$6u_2 = u_4 + u_5 + 0 + 0 + 0 + u_1$$

$$6u_3 = 0 + u_6 + u_4 + u_1 + \frac{1}{2} + 1$$

$$6u_4 = u_6 + 0 + u_5 + u_2 + u_1 + u_3$$

$$6u_5 = 0 + 0 + 0 + 0 + u_7 + u_4$$

$$6u_6 = 0 + 0 + 0 + u_4 + u_3 + 0$$

and the matrices take the form

- (iii) We now need to solve the matrix equation $\mathbf{A}\mathbf{u} = \mathbf{b}$ for the vector \mathbf{u} . It may be noted that the matrix does not have much ‘structure’, except that it is diagonally dominant, so a direct inversion is usually preferred unless the dimension of the matrix is very large. The equations were solved using MATLAB to give

$$\mathbf{u} = \begin{bmatrix} 0.3481 \\ 0.0900 \\ 0.3471 \\ 0.1514 \\ 0.0402 \\ 0.0831 \end{bmatrix}$$

- (iv) Calculate any required data from the solution in (iii).



To construct the rows of the stiffness matrix, \mathbf{A} , in a MATLAB implementation, the coordinates of the nodes are put into a matrix

$$\mathbf{CO} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$$

with the internal nodes first followed by the boundary nodes (note in the MATLAB program this matrix is declared as global). The neighbours of each of the internal nodes are placed in the matrix *link*, one row for each node. The MATLAB function, stored in the M-file *stiff.m*, computes the contribution to the stiffness matrix from the k th internal point, the output *a* gives the contribution to the k th row in the full stiffness matrix.

```
function a=stiff(mm,k,L)
%mm=no of neighbours, k=current point, L=row of k's
neighbours
global CO
a=zeros(1,mm+1);
for p=1:mm-1
    [l,m,n]=coeff(CO(k,:),CO(L(p),:),CO(L(p+1),:));
    % note that coeff.m is used
    a(1)=a(1)+l; a(p+1)=a(p+1)+m; a(p+2)=a(p+2)+n;
end
[l,m,n]=coeff(CO(k,:),CO(L(mm),:),CO(L(1),:));
a(1)=a(1)+l; a(mm+1)=a(mm+1)+m; a(2)=a(2)+n;
```

The following example illustrates the use of MATLAB in the solution of the Laplace equation.

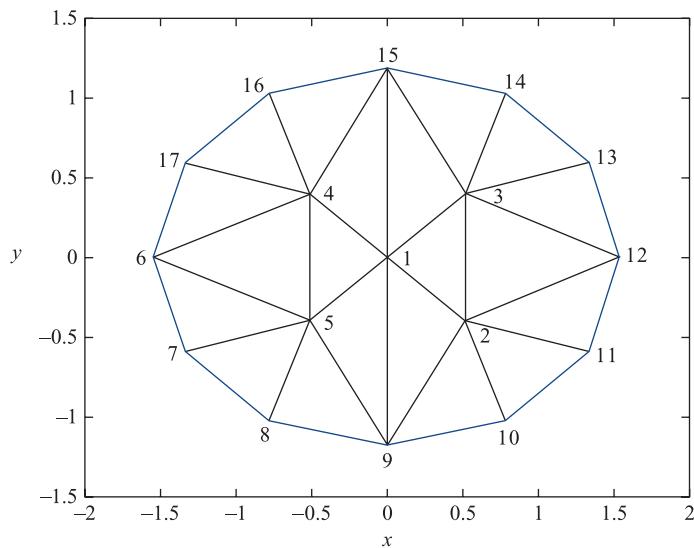
Example 9.40

Solve the Laplace equation $\nabla^2\phi=0$ in the elliptical region

$$\frac{x^2}{\cosh^2 1} + \frac{y^2}{\sinh^2 1} = 1$$

with $\phi=1$ on the upper half of the ellipse and $\phi=0$ on the lower half. The situation is illustrated in Figure 9.48.

Figure 9.48 Mesh for the elliptical plate in Example 9.40.

**Solution**

The problem corresponds physically to an elliptical plate, hot on one side and cold on the other. In Figure 9.48, the triangulated mesh illustrated is the one used in the program. Note how well the mesh fits the boundary even for a small number of boundary nodes.

Data for the problem is placed in a script file stored as *inform.m*.



```
nin=5; nbdry=12; %number of internal and boundary nodes
global CO
v=-pi:pi/6:5*pi/6; X=[cosh(1)*cos(v), sinh(1)*sin(v)];
X1=(X(1,:)+X(2,:)+X(3,:)+X(4,:))/7;
CO=[0 0;-X1(1,1) X1(1,2);-X1(1,1) -X1(1,2);X1(1,1)
-X1(1,2);X1;X];
%coords of points, internal first then bdry
link=[2 3 15 4 5 9; 1 9 10 11 12 3; 1 2 12 13 14 15; 1 15
16 17 6 5; 1 4 6 7 8 9];
%links from interior points to neighbours, in CO order
bdry=[0.5 0 0 0 0 0.5 1 1 1 1 1]; %boundary values, in CO order
A=zeros(nin); rhs=zeros(nin,1);
```

The solution is then computed from the following code, which should be stored in some convenient place so that it can be edited easily. The complete stiffness matrix, \mathbf{A} , is assembled and all the boundary data is transferred to the right-hand side vector called *rhs*.



```

inform %inserts the data from inform.m
for k=1:nin %for each internal node transfers info to A or
rhs
    r=link(k,:); m=nnz(r);
    z=stiff(m,k,r);
    %uses stiff.m, which in turn uses coeff.m, to compute the
    contributions from row k
    A(k,k)=A(k,k)+z(1);
    for i=1:m
        if r(i)<=nin
            A(k,r(i))=A(k,r(i))+z(i+1);
        else
            rhs(k)=rhs(k)-z(i+1)*bdry(r(i)-nin);
        end
    end
end
A, rhs, A\rhs %prints out the stiffness matrix, the rhs
and the final solution

```

The print-out is

A=	4.3857	-1.1266	-1.1266	-1.1266	-1.1266	rhs=	-0.0604
	-1.1266	3.9085	-0.6836	0	0		0.0699
	-1.1266	-0.6836	3.9085	0	0		2.0284
	-1.1266	0	0	3.9085	-0.6836		2.0284
	-1.1266	0	0	-0.6836	3.9085		0.0699

which gives the final solution as

0.5000 0.2868 0.7132 0.7132 0.2868

The solution has all the correct symmetries about the x and y axes. An exact solution to the problem can be obtained by separation of variables in an appropriate coordinate system in terms of Fourier series. For node 3 this method gives the value 0.7076 compared with the FE value of 0.7132; an accuracy of less than 1% has been obtained.

For the solution of the Poisson equation with $\rho \neq 0$ in (9.59) all the segments of the MATLAB programs need to be modified. A similar problem in a rectangular region was studied in Example 9.36 using finite differences.

Example 9.41

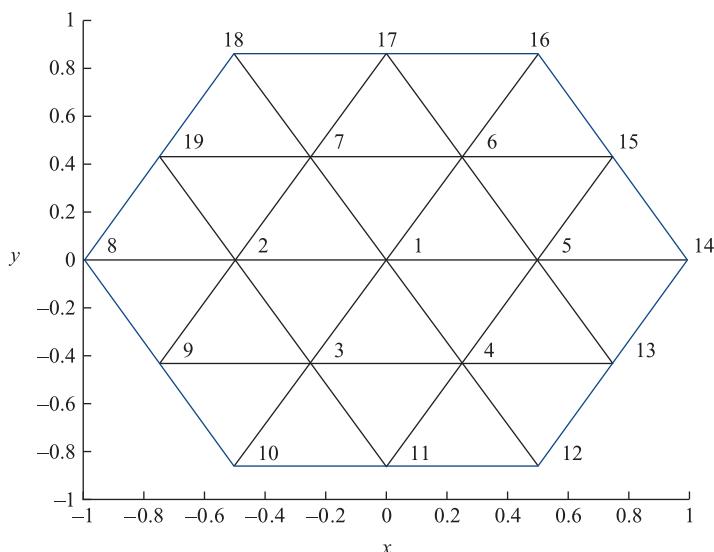
Solve the Poisson equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = -2$$

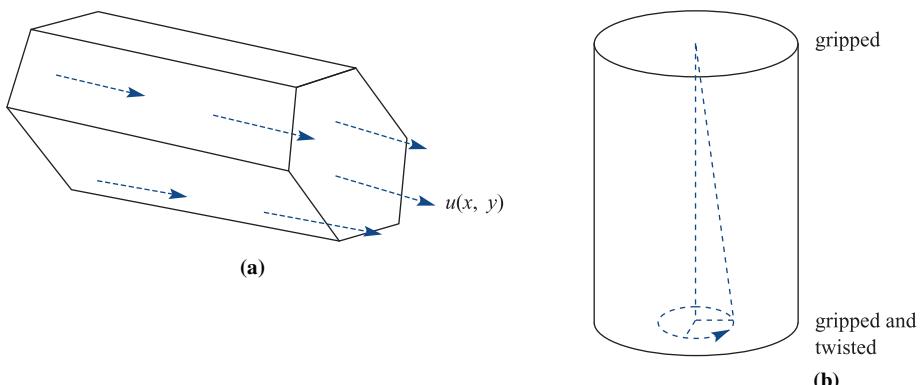
in the hexagonal region illustrated in Figure 9.49 and with $u = 0$ on the boundary of the region.

Figure 9.49

Hexagonal region with mesh used in Example 9.41.

**Figure 9.50**

(a) Flow in hexagonal pipe;
(b) cylinder in torsion.

**Solution**

There are several physical interpretations of this problem. In the context of heat flow, the boundary is kept at a fixed temperature and heat supplied at a uniform rate to the plate. For the unidirectional flow of a viscous fluid in a long hexagonal pipe, u is the velocity and the constant right-hand side is related to the pressure gradient along the tube, see Figure 9.50(a). A honeycomb of tubes in a heat exchanger is a possible application.

When a cylinder is in torsion, by gripping at one end and gripping and twisting at the other, the stresses can be computed from the same Poisson equation; for an illustration see Figure 9.50(b). For a detailed description of these physical problems and the derivations a specialist book should be consulted (S. C. Hunter, *Mechanics of Continuous Media*, Chichester, Ellis Horwood, 1983).

The modifications to the MATLAB implementations can be checked easily against the same problem with an elliptical region since an exact solution is known to be

$$\phi = \frac{a^2 b^2}{a^2 + b^2} \left(1 - \frac{x^2}{a^2} - \frac{y^2}{b^2} \right)$$

for the region

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

The stiffness matrix is the same as in Example 9.40 and only the right-hand sides are modified.



For the hexagonal region with the mesh shown in Figure 9.49 the matrices are computed using the modified MATLAB segments and the print-out is

```
A= 3.4641 - 0.5774 - 0.5774 - 0.5774 - 0.5774 - 0.5774 - 0.5774 - 0.5774  
- 0.5774 3.4641 - 0.5774 0 0 0 - 0.5774  
- 0.5774 - 0.5774 3.4641 - 0.5774 0 0 0  
- 0.5774 0 - 0.5774 3.4641 - 0.5774 0 0  
- 0.5774 0 0 - 0.5774 3.4641 - 0.5774 0  
- 0.5774 0 0 0 - 0.5774 3.4641 - 0.5774  
- 0.5774 - 0.5774 0 0 0 - 0.5774 3.4641  
  
rhsT = 0.4330 0.4330 0.4330 0.4330 0.4330 0.4330 0.4330 0.4330
```

and gives the solution

```
0.4167 0.2917 0.2917 0.2917 0.2917 0.2917 0.2917 0.2917
```

The problem has a great deal of symmetry which is reflected in the solution and, indeed, symmetries could have been built into the program to reduce the computational effort. In the current situation the solution is computed almost instantly, but in most engineering problems every bit of symmetry should be used to its fullest.

In this section no more than the ‘flavour’ of the finite-element method has been given. The intimate connection with computers makes it difficult to do more than show the complications that occur and give an outline of how they are dealt with.

For many problems a linear approximation is not good enough, for example in stress analysis for finite deformations, when high derivatives are required. Also, there is no good reason why a triangle is chosen rather than a quadrilateral. There are several choices that must be made at the start of the calculation:

- (i) division of the region into triangles, quadrilaterals, . . . ;
- (ii) level of approximation, linear, quadratics, cubics, . . . ;
- (iii) choice of test function;
- (iv) method of integration over elements: exact, Gaussian, . . . ;
- (v) method of labelling nodes;
- (vi) method of solution of the resulting matrix equation.

As indicated in (iv), once we have abandoned linear approximations, the integrals cannot be performed exactly, and we need to use an approximate method. Gaussian integration for triangles works very well, and is commonly used. Usually, there is no obvious labelling of nodes and it is necessary for each node to keep a list of which nodes are neighbours, as in the matrix link in the MATLAB segment *inform.m*. In the Partial Differential Equation Toolbox in MATLAB an automatic triangulation of a region can be found; it automatically chooses the nodes and their coordinates, the labelling, as in (v), and the list of neighbours. The complexity can be appreciated for a simple rectangular mesh from the MATLAB commands

```
g='squareg';  
[p,e,t]=poimesh(g,4); % p,e,t represent the points, edges,  
triangles respectively  
pdemesh(p,e,t)
```



Because the labelling is not straightforward, the resulting stiffness matrix rarely has a simple structure, and the most usual method of inversion is by a full frontal attack with Gaussian elimination.

The method has been illustrated for only one equation, the Poisson equation. A similar analysis has to be undertaken for each new equation considered.

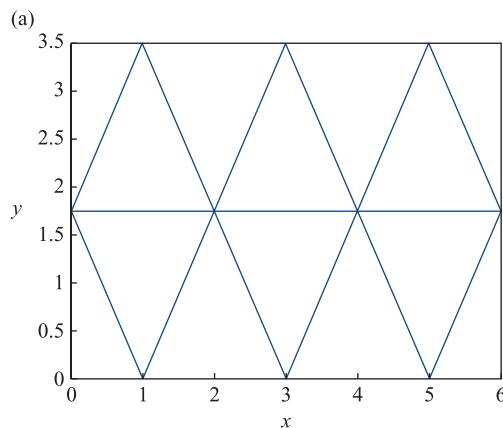
9.6.1 Exercises



All these exercises require substantial programming expertise. Alternatively the Partial Differential Equations Toolbox in MATLAB can be used.

60

Solve the Laplace equation for the rectangular region $0 \leq x \leq 6$ and $0 \leq y \leq 2\sqrt{3}$ using finite elements. On the right-hand boundary $u = 1$ and zero on the remainder of the boundary.



- (a) Use the mesh in Figure 9.51(a) with two interior points.
- (b) Use the mesh in Figure 9.51(b) with five interior points.

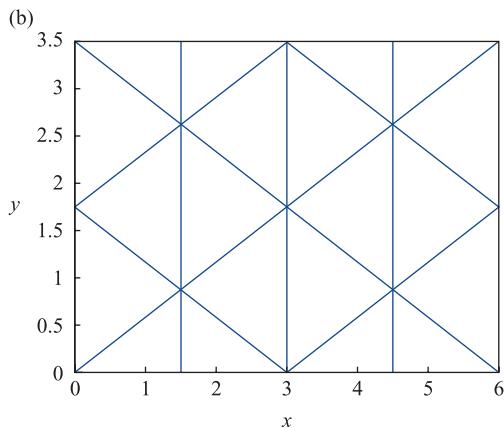


Figure 9.51 Mesh for Exercise 60: (a) with two interior points; (b) with five interior points.

61

Solve the problem in Exercise 57 using the triangular finite-element mesh shown in Figure 9.52.

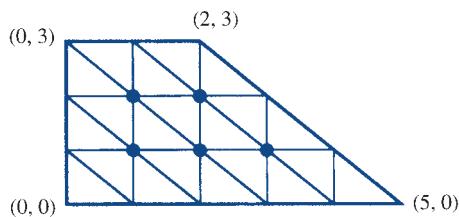


Figure 9.52 Finite-element mesh for Exercise 61.

62

Solve the problem in Exercise 59 using the triangular finite-element mesh shown in Figure 9.53.

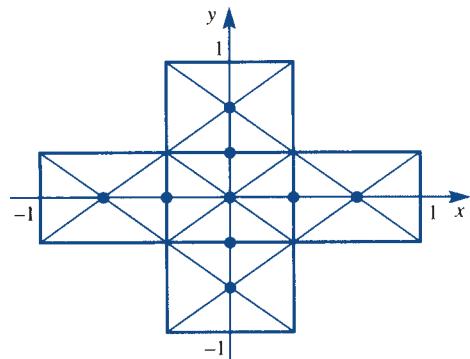


Figure 9.53 Finite-element mesh for Exercise 62.

9.7**Integral solutions**

In previous sections solutions were built up from elementary solutions of various partial differential equations, the most obvious of which was from separated solutions. There are many problems where separated solutions are not available but building up from elementary solutions may still be possible. This section will show methods of solution which can lead to very important ideas that can be exploited practically and importantly to some proofs of existence and uniqueness of solutions. Some numerical methods also use these ideas, for instance the boundary element method is an extension of finite elements with the advantage that the dimension of the calculations is reduced by one. On the whole the mathematics is quite demanding so the interested reader is left to explore the full power of the new methods in specialist books.

9.7.1 Separation of variables

In plane polar coordinates with $x = r \cos \theta$, $y = r \sin \theta$, the Laplace equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

becomes (see Example 3.6 in Section 3.1.1)

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \left(\frac{\partial^2 u}{\partial \theta^2} \right) = 0 \quad (9.66)$$

Writing $u = F(r)G(\theta)$ and substituting gives the separated equations

$$\frac{1}{F} \frac{d}{dr} \left(r \frac{dF}{dr} \right) = -\frac{1}{G} \frac{d^2 G}{d\theta^2} = \mu^2$$

or

$$r \frac{d}{dr} \left(r \frac{dF}{dr} \right) = \mu^2 F \quad \text{and} \quad \frac{d^2 G}{d\theta^2} + \mu^2 G = 0$$

For periodic solutions of the equation in G the constant μ must be an integer n . Thus the solution is

$$G = A \cos n\theta + B \sin n\theta$$

and for the F equation it follows easily that

$$F = Cr^n + \frac{D}{r^n}$$

where A, B, C, D are arbitrary constants. We now choose a specific problem to illustrate the use of these solutions.

Example 9.42

Solve the Laplace equation $\nabla^2 u = 0$ inside the circle $r = a$ with u specified on the boundary as $u(a, \theta) = f(\theta)$, a continuous periodic function with period 2π .

Solution

If the solution is finite inside the circle then F must be of the form $F = Cr^n$. A sum of all the terms then becomes

$$u(r, \theta) = \frac{1}{2}A_0 + \sum_{n=1}^{\infty} r^n (A_n \cos n\theta + B_n \sin n\theta) \quad (9.67)$$

and the solution is now a standard Fourier series problem (see Chapter 12 in MEM) namely

$$u(a, \theta) = f(\theta) = \frac{1}{2}A_0 + \sum_{n=1}^{\infty} a^n (A_n \cos n\theta + B_n \sin n\theta)$$

The coefficients are

$$a^n A_n = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos(nt) dt \quad \text{and} \quad B_n = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin(nt) dt$$

These coefficients are substituted into (9.67) and the summation and integration are interchanged; this is permissible since $f(\theta)$ is a continuous and bounded periodic function

$$u(r, \theta) = \frac{1}{\pi} \int_0^{2\pi} f(t) \left[\frac{1}{2} + \sum_{n=1}^{\infty} \left(\frac{r}{a} \right)^n (\cos n\theta \cos nt + \sin n\theta \sin nt) \right] dt$$

or

$$u(r, \theta) = \frac{1}{\pi} \int_0^{2\pi} f(t) \left[\frac{1}{2} + \sum_{n=1}^{\infty} \left(\frac{r}{a} \right)^n \cos n(\theta - t) \right] dt \quad (9.68)$$

Now consider the series

$$\frac{1}{2} + z + z^2 + z^3 + \dots \quad \text{where } z = R e^{j\phi}$$

which has the sum, for $|z| < 1$,

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

Take the real part and we obtain, after a little algebra

$$\frac{1}{2} + R \cos \phi + R^2 \cos 2\phi + \dots = \frac{1-R^2}{2(1-2R \cos \phi + R^2)}$$

Use this expression in (9.68) to obtain a final result

$$u(r, \theta) = \frac{1}{2\pi} \int_0^{2\pi} f(t) \frac{a^2 - r^2}{a^2 - 2ra \cos(\theta - t) + r^2} dt \quad (9.69)$$

which is called the Poisson integral formula.

Example 9.42 shows that the solution of a complicated partial differential equation can be reduced to an integral. The problem has been reduced essentially from

a two-dimensional problem to a one-dimensional problem on the boundary. Integrals are well understood and there is a vast array of methods that can be used to solve the problem either explicitly or numerically. In general if a method can be developed to convert a partial differential equation to an integral form on the boundary then a great deal of progress has been made to obtaining a solution.

Not least of the advantages of an integral formulation is that bounds on integrals are much easier to obtain than on differentials. These are used almost exclusively to prove uniqueness and existence of solutions. A result that follows easily from (9.69) is obtained by putting $r = 0$

$$u(0, \theta) = \frac{1}{2\pi} \int_0^{2\pi} f(t) dt \quad (9.70)$$

so that the value at the centre of the circle is the average of the values on the bounding circle. Thus the value at the centre of the circle can never be the maximum (or minimum) value in the region. For a general Laplace equation problem, since every interior point can be placed at the centre of a small circle, it can never be the maximum. We can therefore deduce that for the Laplace equation the maximum (or minimum) value cannot be in the interior but must be on the boundary. This result is one of the keystones of the proof of uniqueness of solution.

9.7.2 Use of singular solutions

Again consider the two-dimensional Laplace equation (9.66) in plane polar coordinates. If we look for solutions, $f(r)$, that only depend on r , then the equation becomes

$$\frac{1}{r} \frac{d}{dr} \left(r \frac{df}{dr} \right) = 0$$

which can be integrated as

$$r \frac{df}{dr} = A \quad \text{and then} \quad f = A \ln r + B$$

where A, B are arbitrary constants. The solution has a singularity at the origin which can be exploited to obtain more general solutions and reduce the problem again to an integral round the boundary, as in Section 9.7.1. The method is based on Green's theorem discussed in Section 3.4.5

$$\oint_C (P dx + Q dy) = \iint_S \left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} \right) dx dy \quad (9.71)$$

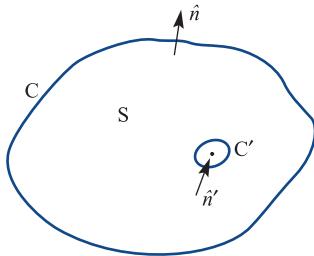
where the curve C encloses the region S . Put

$$P = -u \frac{\partial v}{\partial y} + v \frac{\partial u}{\partial y} \quad \text{and} \quad Q = u \frac{\partial v}{\partial x} - v \frac{\partial u}{\partial x}$$

into (9.71) to give

$$\oint_C u \left(\frac{\partial v}{\partial x} dy - \frac{\partial v}{\partial y} dx \right) - v \left(\frac{\partial u}{\partial x} dy - \frac{\partial u}{\partial y} dx \right) = \iint_S (u \nabla^2 v - v \nabla^2 u) dx dy$$

Figure 9.54 Region S bounded by the curve C, ‘punctured’ by the small circle C’.



and finally

$$\oint_C \left(u \frac{\partial v}{\partial n} - v \frac{\partial u}{\partial n} \right) ds = \iint_S (u \nabla^2 v - v \nabla^2 u) dx dy \quad (9.72)$$

where n is the normal direction and s is the length parameter along C.

We now need to choose u and v in the extended Green’s theorem (9.72) to obtain useful results. The region considered is the interior of S , which is bounded by the curve C , ‘punctured’ by a small circle C' with centre \mathbf{r}_0 and radius ε (which we will eventually tend to zero), $\mathbf{r} = \mathbf{r}_0 + \varepsilon(\cos \phi, \sin \phi)$ (see Figure 9.54). Take

$$u = -\frac{1}{2\pi} \ln |\mathbf{r} - \mathbf{r}_0|$$

and consider the second term in (9.72) on the small circle C'

$$\frac{\partial u}{\partial n'} = -\frac{\partial u}{\partial \varepsilon} = \frac{1}{2\pi} \frac{\partial}{\partial \varepsilon} \ln \varepsilon = \frac{1}{2\pi} \frac{1}{\varepsilon} \quad \text{and} \quad ds = \varepsilon d\phi$$

Thus the term becomes

$$\oint_{C'} v \frac{\partial u}{\partial n'} ds = \frac{1}{2\pi} \int_0^{2\pi} v \frac{1}{\varepsilon} \varepsilon d\phi = \frac{1}{2\pi} \int_0^{2\pi} v d\phi = v(\mathbf{r}_0)$$

The last result just comes from (9.70) and we see that this term just picks out the value of v at the point \mathbf{r}_0 . It remains to construct appropriate u and v to exploit this idea. It is left to specialist texts to consider the choices for general problems; here we will concentrate on the Dirichlet problem (Section 9.8.2) for the Poisson equation, where

$$\nabla^2 \psi = -\rho(x, y) \quad \text{in the region } S \text{ and } \psi \text{ given on } C$$

and let

$$\psi' = G(x, y; x_0, y_0) = -\frac{1}{2\pi} \ln |\mathbf{r} - \mathbf{r}_0| + H(x, y; x_0, y_0) \quad (9.73)$$

so that

$$\nabla^2 \psi' = 0 \quad \text{in the region } S \text{ and } \psi' = 0 \text{ on } C$$

and H has no singularities in the region. In (9.72) put $u = \psi'$ and $v = \psi$. Because ψ' satisfies the Laplace equation and ψ the Poisson equation in the region S we can write (9.72) as

$$\oint_C \left(\psi' \frac{\partial \psi}{\partial n} - \psi \frac{\partial \psi'}{\partial n} \right) ds - \oint_{C'} \left[\psi' \frac{\partial \psi}{\partial r} - \psi \frac{\partial}{\partial r} \left(-\frac{1}{2\pi} \ln |r - r_0| + H(r, r_0) \right) \right] \varepsilon d\phi \\ = - \iint_S G\rho \, dx \, dy$$

Taking the left-hand side terms one by one: the first term is zero since $\psi' = 0$ from the conditions given; the second gives the required integral round the boundary; the third term is of order $\varepsilon \ln \varepsilon$ so tends to zero as $\varepsilon \rightarrow 0$; the fourth term was treated above and gives $\psi(r_0)$; and the fifth term is of order ε and tends to zero. Collecting up the terms gives

$$\psi(x_0, y_0) = - \oint_C \psi(x, y) \frac{\partial}{\partial n} G(x, y; x_0, y_0) \, ds + \iint_S G(x, y; x_0, y_0) \rho(x, y) \, dx \, dy \quad (9.74)$$

We can now find ψ at any point in the region from the value of ψ on the boundary, the right-hand side of the Poisson equation $\rho(x, y)$, together with the function $G(x, y; x_0, y_0)$, called **Green's function**. At the moment it is assumed that Green's function exists and can be calculated. For simple geometries it can often be found and advanced books show how and when this can be done. The whole theory of Green's functions can be applied to many different equations and boundary conditions but this is the province of advanced books on partial differential equations (see R. Haberman, *Partial Differential Equations with Fourier Series and Boundary Value Problems*, fifth edition, Pearson, 2013). An example will illustrate the method.

Example 9.43

Solve the Laplace equation

$$\nabla^2 f = 0 \quad \text{in the region } y > 0$$

given that $f(x, 0) = F(x)$, a known function, on the x axis and that f is zero at infinity.

Solution Green's function (9.73) can be constructed by reflection as

$$G[(x, y; x_0, y_0)] = -\frac{1}{2\pi} \ln |(x - x_0, y - y_0)| + \frac{1}{2\pi} \ln |(x - x_0, y + y_0)| \\ = -\frac{1}{4\pi} \ln \left\{ \frac{(x - x_0)^2 + (y - y_0)^2}{(x - x_0)^2 + (y + y_0)^2} \right\}$$

Note that the added term has no singularities in the region $y > 0$; the function is zero on $y = 0$ and tends to zero as x and y tend to infinity. Now

$$\frac{\partial G}{\partial n} = -\frac{\partial G}{\partial y} = \frac{1}{4\pi} \left[\frac{2(y - y_0)}{(x - x_0)^2 + (y - y_0)^2} - \frac{2(y + y_0)}{(x - x_0)^2 + (y + y_0)^2} \right]$$

Putting $y = 0$ gives

$$\frac{\partial G}{\partial n} = -\frac{y_0}{\pi} \left[\frac{1}{(x-x_0)^2 + y_0^2} \right]$$

The solution is then obtained from (9.74) as

$$\psi(x_0, y_0) = \int_{-\infty}^{\infty} F(x) \frac{y_0}{\pi} \left[\frac{1}{(x-x_0)^2 + y_0^2} \right] dx$$

Again we see that the solution has been reduced to an integral along the boundary. Finding the exact form of Green's function is known for some classic problems, like the one in Example 9.43, but in general it is a difficult calculation. The problem is closely connected to finding a solution of the partial differential equation with zero boundary values (at least for the Dirichlet problem) and a Dirac delta function imposed at a point in the interior.

9.7.3 Sources and sinks for the heat-conduction equation

In Example 9.4 a solution to the one-dimensional heat-conduction equation was obtained which corresponded to a pulse of heat being supplied at a particular point at zero time. The subsequent dissipation of the heat pulse is illustrated in Figure 9.4. A similar solution can be obtained for the three-dimensional problem. The radial symmetric equation of the heat-conduction equation is

$$\frac{\partial T}{\partial t} = \kappa \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial T}{\partial r} \right) \quad (9.75)$$

where r is the radial distance from the origin and $\kappa = k/(\rho c)$ is the thermal diffusivity. The parameter k is the thermal conductivity, ρ is the density of the medium and c is the specific heat capacity. The solution corresponding to the one-dimensional solution of Example 9.4 is

$$T = \frac{Q/(\rho c)}{8(\pi \kappa t)^{3/2}} \exp\left(-\frac{r^2}{4\kappa t}\right) \quad (9.76)$$

The solution can be verified by direct substitution, as in Example 9.4. It is noted that there is a singularity at zero time which corresponds to a point source releasing an instantaneous amount of heat Q , calculated as follows:

The total amount of heat in the whole of the space at time t is computed from the amount of heat in the shell of radius r and thickness dr and then integrating

$$H = \int_0^{\infty} 4\pi r^2 (\rho c T) dr = \int_0^{\infty} 4\pi r^2 \frac{Q}{8(\pi \kappa t)^{3/2}} \exp\left(-\frac{r^2}{4\kappa t}\right) dr$$

The integration can be obtained by differentiating with respect to α the well known integral

$$\int_0^{\infty} e^{-\alpha z^2} dz = \frac{1}{2} \sqrt{\frac{\pi}{\alpha}} \quad \text{to get} \quad \int_0^{\infty} z^2 e^{-\alpha z^2} dz = \frac{1}{4} \sqrt{\frac{\pi}{\alpha^3}}$$

Thus

$$H = \frac{Q}{2\sqrt{\pi(\kappa t)^{3/2}}} \frac{\sqrt{\pi}}{4} (4\kappa t)^{3/2} = Q$$

The solution (9.76) is usually called an **instantaneous source**, releasing an amount of heat Q at time zero. As expected for fixed t and as r tends to infinity the temperature T tends to its resting temperature of zero. Also for fixed r the temperature tends to zero as the time t tends to infinity and all the heat is conducted away.

For a **continuous source**, heat is released continuously at some given rate. In a time interval ds at time s , assume that heat released is $q(s)ds$ then the temperature is given by (9.76) with the time starting at s . Temperature due to release of heat $q(s)ds$ at time s is given by

$$\frac{q(s)/(\rho c)}{8(\pi\kappa(t-s))^{3/2}} \exp\left(-\frac{r^2}{4\kappa(t-s)}\right) ds \quad t > s$$

Thus over the whole interval then

$$T(r, t) = \frac{1}{8\rho c(\pi\kappa)^{3/2}} \int_0^t \frac{q(s)}{(t-s)^{3/2}} \exp\left(-\frac{r^2}{4\kappa(t-s)}\right) ds \quad (9.77)$$

Again we see that the solution of the heat-conduction equation can be written as an integral with all the advantages listed earlier. For most functions $q(s)$, the integral in (9.77) cannot be performed explicitly, but for $q = q_0$, a constant, it is possible. Making the substitution

$$u = r/[4\kappa(t-s)]^{1/2}$$

reduces the integral to

$$\begin{aligned} T(r, t) &= \frac{q_0}{8\rho c(\pi\kappa)^{3/2}} \int_{r/\sqrt(4\kappa t)}^{\infty} \exp(-u^2) 2(4\kappa)^{1/2} \frac{1}{r} du \\ &= \frac{q_0}{2\rho c\kappa\pi^{3/2} r} \int_{r/\sqrt(4\kappa t)}^{\infty} \exp(-u^2) du = \frac{q_0}{4\rho c\kappa\pi r} \operatorname{erfc}\left(\frac{r}{\sqrt(4\kappa t)}\right) \end{aligned} \quad (9.78)$$

where erfc is the known function defined in Example 9.24 and can be found in all computer packages.

It can be seen that for large times the erfc function tends to one so the steady temperature due to the source decays like $1/r$ and the steady temperature T_s is

$$T_s(r) = \frac{q_0}{4\rho c\kappa\pi r} \frac{1}{r} \quad (9.79)$$

This function must satisfy the Laplace equation; except at the origin, and gives the three-dimensional singular solution that is used to construct Green's function in an exactly similar manner to the two-dimensional version described in Section 9.7.2.

Many situations can be tackled using (9.76)–(9.79) and the solution can be reduced to an integral, which usually requires a numerical quadrature.

Example 9.44

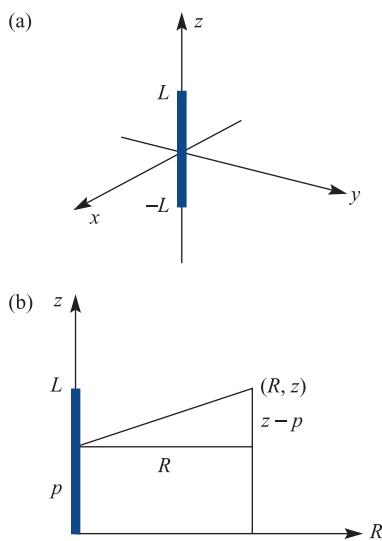
Find the steady temperature due a constant line source of length $2L$, placed in an infinite conducting medium with constant thermal properties.

Solution

We will use the steady point solution in (9.79). The axes are set up in Figure 9.55, with (R, z) being the cylindrical polar coordinates of the field point relative to the origin at the centre of the rod – note that there is no angular coordinate since the solution is clearly symmetrical about the z axis.

Figure 9.55

(a) Line source
and (b) cylindrical
coordinates (R, z) for
Example 9.44.



Take an element of the line at $(p, 0)$ of length dp releasing heat at a rate of $q_L dp$ then from (9.79) the temperature at (R, z) due to this element is

$$\frac{q_L}{4\rho c \kappa \pi} \frac{1}{\sqrt{[R^2 + (z-p)^2]}} dp$$

Hence the temperature due the whole of the line is

$$\begin{aligned} T_L(R, z) &= \frac{q_L}{4\rho c \kappa \pi} \int_{-L}^{L} \frac{1}{\sqrt{[R^2 + (z-p)^2]}} dp \\ &= \frac{q_L}{4\rho c \kappa \pi} \left[\sinh^{-1}\left(\frac{z+L}{R}\right) - \sinh^{-1}\left(\frac{z-L}{R}\right) \right] \end{aligned}$$

Such a calculation can be used to model the temperature due to a heated pipe or cable buried underground or diffusion of contaminant from a section of a steadily leaking pipe. The effect of the burial of a line source at a distance below a surface with a fixed temperature can be calculated by adding a parallel line sink at an equal distance above the surface (see Exercise 66). Note that for large R the square bracket behaves like

$2L/R$ to first order so that at large distances the line source just looks like a point source with strength $2Lq_L$. A wide range of applications of these ideas can be found in H. S. Carslaw and J. C. Jaeger, *Conduction of Heat in Solids* (New York, Oxford University Press, 1959).

9.7.4 Exercises

- 63 Use the Poisson formula (9.69) to solve the Laplace equation in the disk $r \leq a$ with the temperature given as $T = T_0$ for $0 < \theta < \pi$ and $T = 0$ for $\pi < \theta < 2\pi$, where (r, θ) are plane polar coordinates.

- 64 Show that

$$u(x, y) = \frac{1}{4\pi} \ln[(x - a)^2 + (y - b)^2]$$

satisfies the Laplace equation at all points except (a, b) . Check that the function

$$G(x, y; x_0, y_0) =$$

$$-\frac{1}{4\pi} \ln \left\{ \frac{[(x - x_0)^2 + (y - y_0)^2][(x + x_0)^2 + (y + y_0)^2]}{[(x - x_0)^2 + (y + y_0)^2][(x + x_0)^2 + (y - y_0)^2]} \right\}$$

satisfies all the properties of the Green's function for the Dirichlet problem for the Laplace equation in the quarter region $x \geq 0, y \geq 0$. Hence solve the Laplace equation $\nabla^2 T = 0$ in the region $x \geq 0, y \geq 0$ with the boundary conditions $T(x, 0) = f(x)$ for $x \geq 0$ and $T(0, y) = g(y)$ for $y \geq 0$ and T remains bounded at infinity. Show that

$$T(x_0, y_0) =$$

$$\begin{aligned} & \frac{y_0}{\pi} \int_0^\infty \left\{ \frac{1}{(x - x_0)^2 + y_0^2} - \frac{1}{(x + x_0)^2 + y_0^2} \right\} f(x) dx \\ & + \frac{x_0}{\pi} \int_0^\infty \left\{ \frac{1}{(y + y_0)^2 + x_0^2} - \frac{1}{(y - y_0)^2 + x_0^2} \right\} g(y) dy \end{aligned}$$

Evaluate T when $f(x) = 1$ and $g(y) = 0$.

- 65 Green's function of the Dirichlet problem for the Laplace equation in the disk, $r \leq a$, can be written in

terms of polar coordinates of the point (r_0, θ_0) and its inverse point $(a^2/r_0, \theta_0)$. Check that the function

$$G(r, \theta, r_0, \theta_0) =$$

$$\frac{1}{4\pi} \ln \left\{ \frac{r_0^2}{a^2} \frac{r^2 + \frac{a^4}{r_0^2} - \frac{2ra^2}{r_0} \cos(\theta - \theta_0)}{r^2 + r_0^2 - 2rr_0 \cos(\theta - \theta_0)} \right\}$$

satisfies the conditions of Green's function with $G = 0$ on $r_0 = a$. Deduce that the solution of the Laplace equation in the region $r \leq a$ and $u(a, \theta) = f(\theta)$ is given by the Poisson formula (9.69).

Find the steady temperature $T(x, y, z)$ due to a constant line source of length $2L$, placed at $x = a$, $y = 0$, $-L \leq z \leq L$ with the plane $x = 0$ maintained at zero temperature. Use the result in Example 9.44.

- 66

A uniform ring source consists of instantaneous point sources at the points of the circle $z = 0, x^2 + y^2 = a^2$ or $x = a \cos \theta, y = a \sin \theta$. Each element of the ring, $d\theta$, releases an amount of heat $qa d\theta$ at time $t = 0$. Use (9.76) to show that the temperature at any point $(R \cos \phi, R \sin \phi, z)$ is

$$\begin{aligned} & T(R \cos \phi, R \sin \phi, z) \\ & = \frac{q2\pi a}{8\rho c(\pi\kappa t)^{3/2}} \exp\left(-\frac{R^2 + z^2 + a^2}{4\kappa t}\right) I_0\left(\frac{aR}{2\kappa t}\right) \end{aligned}$$

where I_0 is a modified Bessel function, which is a known function available in MAPLE and MATLAB. It can be defined as

$$I_0(\alpha) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\alpha \cos \psi) d\psi$$

9.8

General considerations

There are properties of a general nature that can be deduced without reference to any particular partial differential equation. The formal classification of second-order equations and their intimate connection with the appropriateness of boundary conditions will be considered in this section. The much more difficult problems of the existence and uniqueness are left to specialist texts.

9.8.1 Formal classification

In the preceding sections we have discussed in general terms the three classic partial differential equations. We shall now show that second-order linear equations can be reduced to one of these three types.

Consider the general form of a second-order equation:

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} + Du_x + Eu_y + F = 0 \quad (9.80)$$

where A, B, C, \dots are constants. If we make a change of variable

$$r = ax + y, \quad s = x + by$$

then the chain rule gives

$$\begin{aligned} u_{xx} &= a^2u_{rr} + 2au_{rs} + u_{ss} \\ u_{xy} &= au_{rr} + (1+ab)u_{rs} + bu_{ss} \\ u_{yy} &= u_{rr} + 2bu_{rs} + b^2u_{ss} \end{aligned}$$

Substituting into (9.80) gives

$$\begin{aligned} u_{rr}(Aa^2 + 2Ba + C) + 2u_{rs}(aA + B + abB + bC) + u_{ss}(A + 2Bb + b^2C) \\ + (aD + E)u_r + (D + Eb)u_s + F = 0 \end{aligned}$$

If we choose to eliminate the u_{rs} term then we must put

$$a(A + bB) = -(B + bC) \quad (9.81)$$

and we can eliminate a by substitution to obtain

$$(A + 2bB + b^2C) \left[u_{ss} + \frac{AC - B^2}{(A + Bb)^2} u_{rr} \right] + \dots = 0 \quad (9.82)$$

We can see immediately that the behaviour of (9.82) depends critically on the sign of $AC - B^2$ and this leads to the following classification.

Case 1: $AC - B^2 > 0$, elliptic equations

On putting $(AC - B^2)/(A + Bb)^2 = \lambda^2$, (9.82) becomes

$$\alpha(u_{ss} + \lambda^2 u_{rr}) + \dots = 0$$

and on further putting $q = r/\lambda$,

$$u_{ss} + u_{qq} + \dots = 0$$

The second-order terms are just the same as the Laplace operator. Equations such as (9.80) with $AC - B^2 > 0$ are called **elliptic equations**.

Case 2: $AC - B^2 = 0$, parabolic equations

In this case (9.82) simply becomes

$$u_{ss} + \dots = 0$$

with only one second-order term surviving. The equation is almost identical to the heat-conduction equation. Equations such as (9.80) with $AC - B^2 = 0$ are called **parabolic equations**.

Case 3: $AC - B^2 < 0$, hyperbolic equations

On putting $(AC - B^2)/(A + Bb)^2 = -\mu^2$, (9.82) becomes

$$\alpha(u_{ss} - \mu^2 u_{rr}) + \dots = 0$$

and on further putting $t = r/\mu$,

$$u_{ss} - u_{tt} + \dots = 0$$

which we can identify with the terms of the wave equation. Equations such as (9.80) with $AC - B^2 < 0$ are called **hyperbolic equations**.

Thus we see that simply by changing axes and adjusting length scales, the general equation (9.80) is reduced to one of the three standard types. We therefore have strong reasons for studying the three classical equations very closely. An example illustrates the process.

Example 9.45

Discuss the behaviour of the equation

$$u_{xx} + 2u_{xy} + 2\alpha u_{yy} = 0$$

for various values of the constant α .

Solution In the notation of (9.80), $A = 1$, $B = 1$ and $C = 2\alpha$, so from (9.81)

$$a = -\frac{1 + 2\alpha b}{1 + b}$$

and the change of variables $r = ax + y$, $s = x + by$ gives

$$u_{ss} + \frac{2\alpha - 1}{(1 + b)^2} u_{rr} = 0$$

Thus if $\alpha > \frac{1}{2}$ and $q = r(1 + b)/\sqrt{(2\alpha - 1)}$, we have the elliptic equation

$$u_{ss} + u_{qq} = 0$$

If $\alpha = \frac{1}{2}$, we have the parabolic equation

$$u_{ss} = 0$$

If $\alpha < \frac{1}{2}$ and $t = r(1 + b)/\sqrt{(1 - 2\alpha)}$, we have the hyperbolic equation

$$u_{ss} - u_{tt} = 0$$

In (9.80) the assumption was that A, B, C, \dots were constants. Certainly for many problems this is not the case, and A, B, \dots are functions of x and y , and possibly u also. Therefore the analysis described may not hold globally for variable-coefficient equations. However, we can follow the same analysis at each point of the region under consideration. If the equation at *every* point is of one type, say elliptic, then we call the equation elliptic. There are good physical problems where its type can change. One of the best known examples is for transonic flow, where the equation is of the form

$$\left(1 - \frac{u^2}{c^2}\right)\psi_{xx} - \frac{2uv}{c^2}\psi_{xy} + \left(1 - \frac{v^2}{c^2}\right)\psi_{yy} + f(\psi) = 0$$

where u and v are the velocity components and c is a constant. We calculate

$$AC - B^2 = \left(1 - \frac{u^2}{c^2}\right)\left(1 - \frac{v^2}{c^2}\right) - \left(\frac{uv}{c^2}\right)^2 = 1 - \frac{u^2 + v^2}{c^2} = 1 - \frac{q^2}{c^2}$$

If we put $q/c = M$, the Mach number, then for $M > 1$ the flow is hyperbolic and supersonic, while for $M < 1$ it is elliptic and subsonic. It is easy to appreciate that transonic flows are very difficult to compute, since different boundary conditions and techniques are required on the subsonic and supersonic sides.

9.8.2 Boundary conditions

In the preceding sections we chose natural boundary conditions for the three classical partial differential equations. We can formalize these ideas a bit further and look at appropriate boundary conditions and the consequences of choosing inappropriate conditions. We shall confine ourselves to two-variable situations, but it is possible to extend the theory to problems with more variables.

Suppose that we are trying to obtain the solution $u(x, y)$ to a partial differential equation in a region R with boundary C . The commonest boundary conditions involve u or the normal derivative $\partial u / \partial n$ on C . The normal derivative (which is discussed in Section 3.2.1) at a point P on C is the rate of change of u with respect to the variable n along the line that is normal to C at P . The three conditions that are found to occur most regularly are

Cauchy conditions

u and $\frac{\partial u}{\partial n}$ given on C

Dirichlet conditions

u given on C

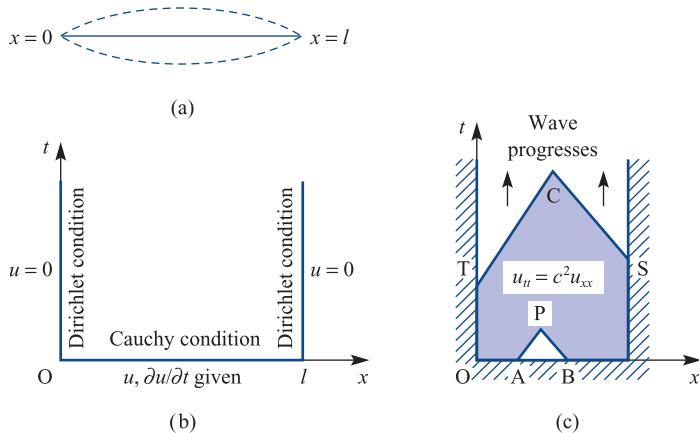
Neumann conditions

$\frac{\partial u}{\partial n}$ given on C

It is common for different conditions to apply to different parts of the boundary C . A boundary is said to be **closed** if conditions are specified on the whole of it, or **open** if conditions are only specified on part of it. The boundary can of course include infinity; conditions at infinity are specified if the boundary is closed or unspecified if it is open.

Figure 9.56

(a) A vibrating string fixed at its ends $x = 0$ and $x = l$;
 (b) the corresponding region and boundary conditions in the (x, t) plane; (c) wave moving forward with time.



The natural conditions for the *wave equation* are Cauchy conditions on an open boundary. The d'Alembert solution (9.15) in the (x, t) plane in Section 9.3.1 corresponds to u and $\partial u / \partial t$ given on the open boundary, $t = 0$. Physically these conditions correspond to a given displacement and velocity at time $t = 0$. However, the vibrations of a finite string, say a violin string, will be given by mixed conditions (Figure 9.56a). On the initial line $0 \leq x \leq l$, $t = 0$ (Figure 9.56b) Cauchy conditions will hold, with both u and $\partial u / \partial t$ given. The ends of the string are held fixed, so we have Dirichlet conditions $u = 0$ on $x = 0$, $t \geq 0$ and $u = 0$ on $x = l$, $t \geq 0$.

Figure 9.56 is typical of the hyperbolic-type equations in the two variables x and t that arise in wave propagation problems. For the second-order equation

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} = 0$$

the characteristics are defined by

$$\frac{dy}{dx} = \frac{B \pm \sqrt{(B^2 - AC)}}{A} \quad (9.83)$$

For a hyperbolic equation, $B^2 - AC > 0$, so there are two characteristics, which for constant A , B and C are straight lines. Each of the characteristics carries one piece of information from the boundary into the solution region. This is illustrated in Figure 9.56(c), where the solution at P is completely determined from the information on AB . The pair of characteristics, TC and SC , then allows us to push the solution further into the region. It is clear from the d'Alembert solution that Cauchy data is required on the line $t = 0$ but a single condition is required on the lines $x = 0, l$.

There is no reason why the boundaries cannot be at infinity – an extremely long string can sensibly be modelled in this way. Care at such infinite boundaries must be taken, since the modelling of what happens there is not always obvious; certainly it requires thought.

We have mentioned the commonest boundary conditions, but it is possible to conceive of others. However, such conditions do not always give a unique solution; a physical example will illustrate this point.

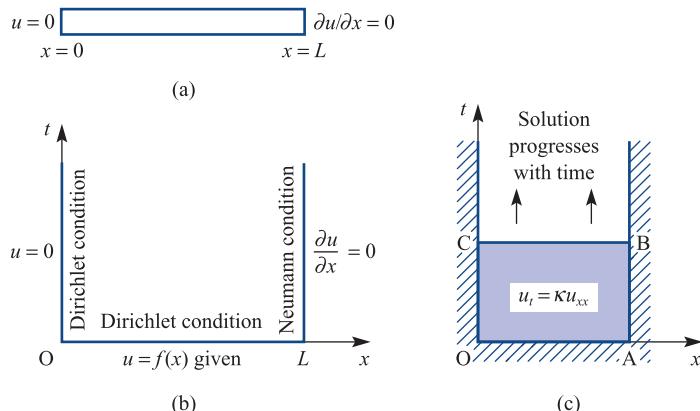
Consider the problem in Example 9.1, which has the solution

$$u = u_0 \sin\left(\frac{\pi x}{L}\right) \cos\left(\frac{\pi ct}{L}\right)$$

Suppose that a photograph of the string is taken at the times $t = L/2c$ and $t = 3L/2c$. Can the solution then be constructed from these two photographs? At the two times the

Figure 9.57

(a) A heated bar with a temperature $u = 0$ at $x = 0$ and insulated, $\partial u / \partial x = 0$, at $x = L$;
 (b) the corresponding region and boundary conditions in the (x, t) plane. (c) Solution can be computed at successive times.



string has the same shape $u = 0$; that is, the string is in its resting position. One possible solution is therefore that the string has not moved. We know that a non-zero solution to Example 9.1 is possible, so we have two solutions to our problem, and we have lost uniqueness. Specifying the displacement at two successive times is not a sensible set of boundary conditions. Although we have stated an extreme case for clarity, the same problem is there for any T , and non-unique solutions can occur if incorrect boundary conditions are imposed.

The *boundary conditions* for the heat-conduction equation (9.7) for the one-dimensional case are given by specifying u or the normal derivative $\partial u / \partial n$ on a curve C in the (x, t) plane, that is *Dirichlet* or *Neumann conditions* respectively. Because there is only one time derivative in (9.7), we need only specify one function at $t = 0$ (say), rather than two as in the wave equation. In the simplest one-dimensional problem, at time $t = 0$ the temperature in a bar is given, $u(x, 0) = f(x)$, and at the ends some temperature condition is satisfied for all time. Typical conditions might be $u(0, t) = 0$, so that the end $x = 0$ is kept at zero temperature, and $\partial u(L, t) / \partial x = 0$, which implies no heat loss from the end $x = L$. The situation is illustrated in Figure 9.57. It is clear that, no matter what the starting temperature $f(x)$ is, the solution must tend to $u = 0$ as the final solution.

In the case of a parabolic equation we have $B^2 - AC = 0$, so the characteristics in (9.83) coalesce. Imagine that there are two characteristics very close together. The information on the boundaries will propagate a long way, since the two lines will meet ‘close to infinity’. We should therefore expect that information on the initial line would propagate forward in time, and because there is only one characteristic that one piece of information on the boundary curve would be sufficient. Figure 9.57(c) illustrates the situation, with the solution on CB being determined by a single boundary condition on each of CO, OA and AB.

Again, as with the wave equation, there is no reason why the bar cannot be of infinite length, at least in a mathematical idealization, so that the initial curve C can include infinite parts. The conditions at infinity are usually quite clear and cause little difficulty.

An interesting feature is that it is very difficult to integrate the heat-conduction equation backwards in time. Suppose we are given a temperature distribution at time $t = T$ and seek the initial distribution at $t = 0$ that produces such a distribution of temperature at $t = T$. If there is an exact solution then the problem can be solved, but it is unstable in the sense that small changes at $t = T$ can lead to huge changes at $t = 0$. Consider, for instance, the solution to the heat-conduction equation with $\kappa = 0.5$ in the following two situations:

Given $u = 0$ on $x = 0$ and 1 , and at $t = 5$
 $u(x, 5) = \sin(\pi x) e^{-2.5\pi^2}$
 find $u(x, 0)$.

The solution is just one of the separated solutions in (9.40), namely

$$u(x, t) = \sin(\pi x) e^{-0.5\pi^2 t}$$

At $t = 5$ $u < 2 \times 10^{-11}$
 At $t = 0$ $u = \sin(\pi x)$

Given $u = 0$ on $x = 0$ and 1 , and at $t = 5$
 $u(x, 5) = \sin(2\pi x) e^{-10\pi^2}$
 find $u(x, 0)$.

The solution is just one of the separated solutions in (9.40), namely

$$u(x, t) = \sin(2\pi x) e^{-2\pi^2 t}$$

At $t = 5$ $u < 1.4 \times 10^{-43}$
 At $t = 0$ $u = \sin(2\pi x)$

The two conditions at $t = 5$ differ by a very small amount ($< 10^{-11}$) but, integrating backwards to $t = 0$, the two solutions are significantly different. Although this analysis is physically artificial it indicates why integrating backwards in time is unstable. This phenomenon, in particular, leads to almost insuperable difficulties when a numerical solution is sought, since errors are inherent in any numerical method. Such a situation applies, for instance, when a space capsule is required to have a specified temperature distribution on reaching its final orbit. The designer wants to know an initial temperature distribution that will achieve this end.

The boundary conditions most relevant to the Laplace equation are *Dirichlet* or *Neumann conditions*. These specify respectively u or the normal derivative $\partial u / \partial n$ on a closed physical boundary. One condition around the whole boundary, which may include an infinite part, is sufficient for this equation. Typically, on a rectangular plate as shown in Figure 9.58, the temperature is maintained at 1 on CD, at 0 on AB and AD, and there is no heat loss from CB.

However, it should be noted that for Neumann conditions, $\partial u / \partial n = f(s)$, on the whole boundary C , where s is a measure of length along the boundary, the function $f(s)$ must satisfy an integral condition. Just consider the Laplace equation $\nabla^2 u = 0$ in the region A with this boundary condition. In Section 3.4.5 Green's theorem was written

$$\oint_C P dx + Q dy = \iint_A \frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y} dx dy$$

This can be re-written by putting $P = -(\partial u / \partial y)$, $Q = \partial u / \partial x$ to give

$$\oint_C -\frac{\partial u}{\partial y} dx + \frac{\partial u}{\partial x} dy = \iint_A \nabla^2 u dx dy = 0$$

The right-hand side is put equal to zero since u satisfies the Laplace equation. Thus

$$0 = \oint_C \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right) \cdot (dy, -dx) = \oint_C \frac{\partial u}{\partial n} ds$$

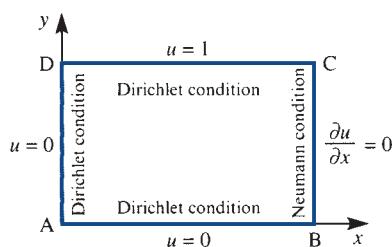
and therefore

$$\oint_C f(s) ds = 0$$

For the steady heat-conduction interpretation, $\partial u / \partial n$ is proportional to the heat entering through an element of the boundary. This result says that for a steady state to be achieved the net amount of heat entering the region must be zero.

Figure 9.58 is typical of an elliptic equation where we have conditions on a closed boundary. In (9.83) we have the condition that $B^2 - AC < 0$, so the characteristics associated with the solution do not make sense in the real plane. There is no ‘time’ in elliptic problems; such problems are concerned with steady-state behaviour and not propagation with time. We are dealing with a fundamentally different situation from the hyperbolic and parabolic cases. The interpretation of the idea of characteristics is unclear physically, and does not prove to be a useful direction to explore, although advanced theoretical treatments do use the concept.

Figure 9.58 Typical boundary conditions for the Laplace equation in a rectangular plate.



It is possible to solve the Laplace equation with other boundary conditions, for instance Cauchy conditions u and $\partial u / \partial x$ on the y axis. However, an example due to Hadamard (see Exercise 49) shows that the solution is unstable in the sense that small changes in the boundary conditions cause large changes in the solution. This type of problem is not well posed, and should not occur in a physical situation; however, mistakes are made and this type of behaviour should be carefully noted.

Figure 9.59 gives in tabular form a summary of the appropriate boundary conditions for these problems.

Figure 9.59
Appropriateness of boundary conditions to the three classical partial differential equations (adapted from P. M. Morse and H. Feshbach, *Methods of Theoretical Physics*, Volume I. New York, McGraw-Hill, 1953).

Data	Boundary	$\nabla^2 u = u_t$ Hyperbolic	$\nabla^2 u = 0$ Elliptic	$\nabla^2 u = u_t$ Parabolic
Dirichlet or Neumann	Open	Insufficient data	Insufficient data	Unique, stable solution for $t > 0$
	Closed	Not unique	Unique, stable (to an arbitrary constant in the Neumann case)	Overspecified
Cauchy	Open	Unique, stable	Solution may exist, but is unstable	Overspecified
	Closed	Overspecified	Overspecified	Overspecified

9.8.3 Exercises

- 68** Determine the type of each of the following partial differential equations, and reduce them to the standard form by change of axes:

$$\begin{aligned} \text{(a)} \quad & u_{xx} + 2u_{xy} + u_{yy} = 0 \\ \text{(b)} \quad & u_{xx} + 2u_{xy} + 5u_{yy} + 3u_x + u = 0 \\ \text{(c)} \quad & 3u_{xx} - 5u_{xy} - 2u_{yy} = 0 \end{aligned}$$

- 69** Find the general solution of the equation Exercise 68(c).

- 70** Use the change of variable $u = x + y$, $v = x - y$ to transform the partial differential equation

$$\frac{\partial^2 f}{\partial x^2} - 2 \frac{\partial^2 f}{\partial x \partial y} + \frac{\partial^2 f}{\partial y^2} = 0 \quad (9.84)$$

to

$$\frac{\partial^2 f}{\partial v^2} = 0$$

Hence compute the general solution of (9.84) as

$$f = (x - y)F(x + y) + G(x + y)$$

where F and G are arbitrary functions.

- 71** Establish the nature of the Tricomi equation

$$yu_{xx} + u_{yy} = 0$$

in the regions (a) $y > 0$, (b) $y = 0$ and (c) $y < 0$. Use (9.83) to determine the characteristics of the equation where they are real.

Verify that the function $f = [Ax^3 + (B/x^2)]y(1 - y^2)$, with A and B constants, satisfies the partial differential equation

$$x^2 \frac{\partial^2 f}{\partial x^2} + (1 - y^2) \frac{\partial^2 f}{\partial y^2} = 0$$

In which regions is the equation elliptic, parabolic and hyperbolic?

Determine the nature of the equation

$$2q \frac{\partial^2 v}{\partial p^2} + 4p \frac{\partial^2 v}{\partial p \partial q} + 2q \frac{\partial^2 v}{\partial q^2} + 2 \frac{\partial v}{\partial q} = 0$$

Show that if $p = \frac{1}{2}(x^2 - y^2)$ and $q = \frac{1}{2}(x^2 + y^2)$, the equation reduces to the Laplace equation in x and y .

Show that the equation

$$x^2 \frac{\partial^2 u}{\partial x^2} - y^2 \frac{\partial^2 u}{\partial y^2} = 0$$

is hyperbolic. Sketch the domain of dependence and range of influence from the characteristics.

9.9 Engineering application: wave propagation under a moving load

A wide range of practical problems can be studied under the general heading of moving loads. Cable cars that carry passengers, buckets that remove spoil to waste tips, and cable cranes are very obvious examples, while electric train pantographs on overhead wires are perhaps less obvious. Extending the problem to beams opens up a whole range of new problems, such as trains going over bridges, gantry cranes and the like. An excellent general discussion and wide range of applications is given by L. Fryba, *Vibration of Solids and Structures under Moving Loads* (Groningen, Noordhoff, 1973), and *Initial Value Problems, Fourier Series, Overhead Wires, Partial Differential Equations of Applied Mathematics, Open University Mathematics Unit M321, 5, 6 and 7* (Milton Keynes, 1974) treat pantographs on overhead wires. See also S. Howsion, *Practical Applied Mathematics* (Cambridge, Cambridge University Press, 2005).

A straightforward linearized theory of a cable tightly stretched between fixed supports provides important information about the behaviour of such systems. Certainly, if such a theory could not be solved then more complicated problems involving large deformations, slack cables or beams would be beyond reach. The basic assumptions are

- the deflections from the horizontal are small compared with the length of the cable;
- deflections due to the weight of the cable itself are neglected;
- the horizontal tension in the cable is so large compared with the perturbations caused by the load that it may be regarded as constant.

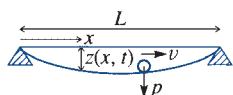


Figure 9.60 Moving load across a taut wire.

Figure 9.60 shows the situation under study and the coordinate system used.

Because the problem is one of small deflections, the basic equation is the wave equation with a forcing term from the moving load:

$$-c^2 \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial t^2} = p(x, t)$$

Since the two ends are fixed, we have

$$z(0, t) = z(l, t) = 0 \quad (t \geq 0) \quad (9.85)$$

A trolley is assumed to start at $x = 0$, with the cable initially at rest; that is,

$$z(x, 0) = \frac{\partial}{\partial t} z(x, 0) = 0 \quad (0 \leq x \leq l) \quad (9.86)$$

It remains to specify the forcing function $p(x, t)$ due to the moving load. We use the simplest assumption of the delta function and step function, as defined in Section 5.2, namely

$$-c^2 \frac{\partial^2 z}{\partial x^2} + \frac{\partial^2 z}{\partial t^2} = P \delta(t - \frac{x}{v}) H(l - x) \quad (9.87)$$

The delta function models the impulse of the trolley at time t at distance x , while the step function switches off the forcing function when the trolley reaches the end $x = l$.

There are several ways of solving this equation, but here the Laplace transform method will be used. Taking the transform of (9.87) using (9.30) and (9.32) together with the initial condition (9.86), we obtain the ordinary differential equation

$$-c^2 Z'' + s^2 Z = P e^{-sx/v} H(l - x)$$

Since we have no interest in the case $x > l$, the final term can be omitted, since it is just 1 if $x < l$ and 0 if $x > l$. It is now straightforward to solve this equation as

$$Z = A e^{sx/c} + B e^{-sx/c} + \frac{P e^{-sx/v}}{s^2(1 - c^2/v^2)}$$

Before evaluating A and B , it is clear that the speed $v = c$ causes problems, since the third term is then infinite, and the solution is not valid for this case. The solution is going to depend on whether the trolley speed is subcritical $v < c$ or supercritical $v > c$.

Equation (9.85) gives $Z(0, s) = Z(l, s) = 0$ for the boundary conditions, so that A and B can now be evaluated from

$$0 = A + B + \frac{P}{s^2(1 - c^2/v^2)}$$

$$0 = A e^{s l/c} + B e^{-s l/c} + \frac{P e^{-s l/v}}{s^2(1 - c^2/v^2)}$$

Some straightforward algebra gives A and B , and hence Z , as

$$Z = \frac{P}{s^2(1 - c^2/v^2)} \left\{ e^{-sx/v} - e^{-sl/v} \frac{\sinh(sx/c)}{\sinh(sl/c)} + \frac{\sinh[s(x-l)/c]}{\sinh(sl/c)} \right\}$$

It is easy to check that the two boundary conditions at $x = 0$ and $x = l$ are satisfied. As with all transform solutions, the main question is whether the inversion can be performed. Fortunately the three terms can be found in tables of transforms, to give

$$\begin{aligned} \frac{z}{P} \left(1 - \frac{c^2}{v^2} \right) &= \left(t - \frac{x}{v} \right) H \left(t - \frac{x}{v} \right) \\ &\quad - H \left(t - \frac{l}{v} \right) \left\{ \frac{x}{l} \left(t - \frac{l}{v} \right) + \frac{2l}{c\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \sin \frac{n\pi x}{l} \sin \left[\left(\frac{n\pi c}{l} \right) \left(t - \frac{l}{v} \right) \right] \right\} \\ &\quad + \left\{ \left(\frac{x-l}{l} \right) t + \frac{2l}{c\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} \sin \left[\frac{n\pi(x-l)}{l} \right] \sin \left[\frac{n\pi ct}{l} \right] \right\} \end{aligned} \quad (9.88)$$

The three terms can be identified immediately. The first is the displacement caused by the trolley moving with speed v and hitting the value x after a time $t = x/v$; the second term only appears for $t > l/v$, and gives the reflected wave from $x = l$; while the third term is the wave caused by the trolley disturbance propagating in the cable with wave speed c .

To look a little more closely at the solution (9.88), we shall consider the case $x = \frac{1}{2}l$. Thus the motion of the midpoint will be considered as a function of time. Plotting such waves is easier in non-dimensional form, so we first rewrite (9.88) in terms of

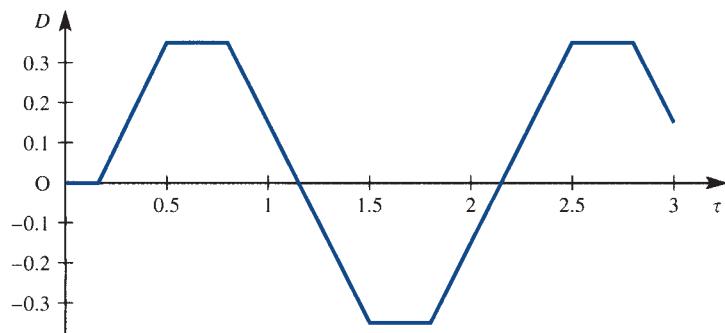
$$D = \frac{cz}{Pl} \left(1 - \frac{c^2}{v^2} \right), \quad \tau = \frac{ct}{l}, \quad \lambda = \frac{c}{v}$$

so that D is the non-dimensional displacement, τ is the non-dimensional time, with $\tau = 1$ corresponding to the time for the wave to propagate the length of the cable, and λ is the ratio of wave speed to trolley speed. The second step is then to take $x = \frac{1}{2}l$ to give

$$\begin{aligned} D &= (\tau - \frac{1}{2}\lambda) H(\tau - \frac{1}{2}\lambda) \\ &\quad - H(\tau - \lambda) \left\{ \frac{1}{2}(\tau - \lambda) - \frac{2}{\pi^2} [\sin \pi(\tau - \lambda) - \frac{1}{9} \sin 3\pi(\tau - \lambda) \right. \\ &\quad \left. + \frac{1}{25} \sin 5\pi(\tau - \lambda) \dots] \right\} - \frac{1}{2}\tau + \frac{2}{\pi^2} (\sin \pi\tau - \frac{1}{9} \sin 3\pi\tau + \frac{1}{25} \sin 5\pi\tau \dots) \end{aligned}$$

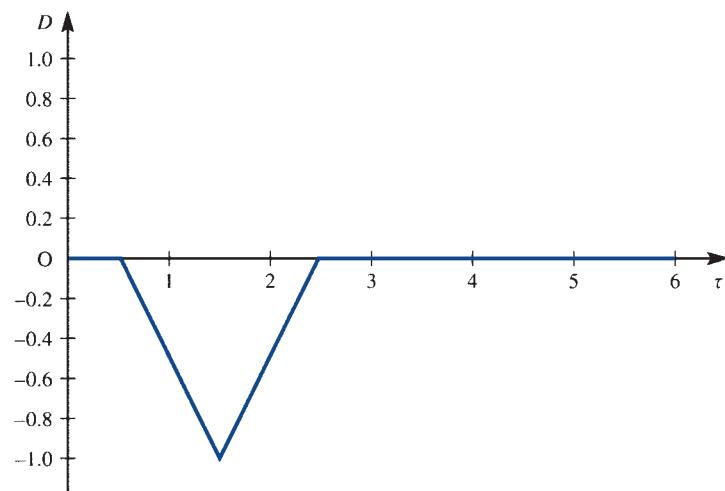
In Figure 9.61 the supercritical case, $\lambda = 0.3$, is displayed. It may be noted that the three terms ‘switch on’ at times $\tau = 0.15$, $\tau = 0.8$ and $\tau = 0.5$ respectively, corresponding physically to the trolley hitting, the reflected wave arriving and the initial wave arriving. The motion is subsequently periodic, as indicated in the figure.

Figure 9.61 Solution of the moving-load problem for $x = \frac{1}{2}l$ and $\lambda = 0.3$; the supercritical case.



Similarly, the subcritical case $\lambda = 3$ is shown in Figure 9.62. Here the switches are at $\tau = 1.5, 2.5$ and 0.5 respectively for the three terms, with the same interpretation as above. Because of the very odd choice of parameter λ , only one pulse is seen at the centre point, with the terms subsequently cancelling.

Figure 9.62 Solution of the moving-load problem for $x = \frac{1}{2}l$ and $\lambda = 3$; the subcritical case.



While the model illustrates many of the obvious properties of wave propagation, it clearly has its limitations. The discontinuous behaviour in the gradient of the displacement looks unrealistic, and the absence of damping means that oscillations once started continue for ever. It is clear that more subtle modelling of the phenomenon is required to make the solutions realistic, but the general behaviour of the solution would still be followed.

9.10 Engineering application: blood-flow model

A problem of considerable interest is how to deal with the flow of a fluid through a tube with distensible walls and hence variable cross-section. An obvious application is to the flow of blood in a blood vessel. The full Navier–Stokes equations for viscous flow are difficult to solve and the distensible wall, where boundary conditions are not clear,

makes for an impossible problem. An alternative, simpler and more heuristic approach is possible and some useful solutions can be deduced. The work is based on a paper by A. Singer (*Bulletin of Mathematical Biophysics* **31** (1969) 453–70), where more details can be found about the practical application to blood flows.

The assumptions required to set up the model are as follows:

- (1) the flow is one-dimensional;
- (2) the flow is incompressible and laminar;
- (3) the flow is slow, so that all quadratic terms can be neglected;
- (4) the resistance to flow is assumed to be proportional to the velocity;
- (5) there is a leakage through the walls that is proportional to the pressure;
- (6) the cross-sectional area S is a function of the pressure only.

We take the situation illustrated in Figure 9.63, and we denote the pressure by p , the velocity by v , the time by t and the axial distance along the tube by x . The first equation that we derive is a continuity equation, which states that in a time Δt the fluid that comes into the element must leave the element:

$$(S_{t+\Delta t} - S_t)\Delta x = -(vS)_{x+\Delta x}\Delta t + (vS)_x\Delta t - gpS\Delta x\Delta t$$

volume after volume before volume out of right-hand end volume into left-hand end leakage

The proportionality constant g is the leakage per unit volume of tube per unit time. The equation can be rewritten as

$$\frac{S_{t+\Delta t} - S_t}{\Delta t} + \frac{(vS)_{x+\Delta x} - (vS)_x}{\Delta x} + gpS = 0$$

or

$$\frac{\partial S}{\partial t} + \frac{\partial}{\partial x}(vS) + gpS = 0 \quad (9.89)$$

A second equation is required to evaluate v , and this comes from Newton's law that the force is proportional to the rate of change of momentum. The force in the x direction acting on the element in Figure 9.63 is

$$\text{force} = \text{pressure force on left-hand end} - \text{pressure force on right-hand end} - vrS\Delta x$$

pressure force on left-hand end pressure force on right-hand end resistance

where r is the resistance per unit length per unit cross-section per unit time, and is the proportionality constant in assumption (4). The change in the momentum in time Δt is more difficult to compute because of the convection due to the moving fluid. However, these effects only involve second-order terms, and hence can be omitted by assumption (3). The calculation is straightforward under this assumption, so that

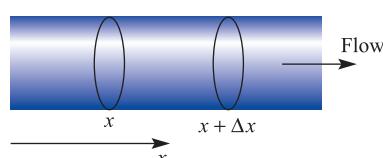


Figure 9.63
An element of the flexible tube in the blood-flow problem.

$$\Delta M = \text{change in momentum} = \rho(v_{t+\Delta t})S\Delta x - \rho(v_t)S\Delta x$$

momentum before momentum after

where ρ is the density of the fluid. Thus

$$\frac{\partial M}{\partial t} = \rho S \frac{\partial v}{\partial t} \Delta x$$

Putting the force equal to the rate of change of momentum, we obtain, on taking the limit as $\Delta x \rightarrow 0$,

$$\rho S \frac{\partial v}{\partial t} = -\frac{\partial(pS)}{\partial x} - vrS \quad (9.90)$$

Now assumption (6) gives $S = S(p)$, so that

$$\frac{1}{S} \frac{\partial S}{\partial x} = \frac{1}{S} \frac{dS}{dp} \frac{\partial p}{\partial x}$$

$$\frac{1}{S} \frac{\partial S}{\partial t} = \frac{1}{S} \frac{dS}{dp} \frac{\partial p}{\partial t}$$

We define $c = (1/S) dS/dp$ as the **distensibility** of the tube, that is the change in S per unit area per unit change in p . Equations (9.89) and (9.90) become

$$cp_t + v_x + cp_x v + gp = 0$$

$$\rho v_t + p_x + cp_x p + rv = 0$$

and since the terms vp_x and pp_x can be neglected by assumption (3), we arrive at our final equations

$$\left. \begin{aligned} cp_t + v_x + gp &= 0 \\ \rho v_t + p_x + rv &= 0 \end{aligned} \right\} \quad (9.91)$$

These are the linearized flow equations, and are identical with the **transmission line** equations describing the flow of electricity down a long, leaky wire such as a trans-atlantic cable (see Exercise 10).

We can now look at special cases that will prove to be very informative about the various terms in the equation.

Case (i): $c = \text{constant}, r = g = 0$

This case corresponds to constant distensibility, which in turn gives $S = A e^{cp}$, since S must satisfy $c = (1/S) dS/dp$. Thus we have made a specific assumption about how S depends on p . The $r = g = 0$ implies the absence of resistance and leakage. Eliminating p between the two equations in (9.91) gives

$$v_{xx} = (c\rho)v_{tt}$$

which is just the wave equation. We know that any pulse will propagate perfectly with a velocity $u = 1/\sqrt{c\rho}$. The assumption in the problem is that the tube is one-dimensional and has no branches. Clearly a heart pulse will propagate to the nearest branch, but there will then be reflection and a complicated behaviour near the branch. In long arteries like the femoral artery the theory can be checked for its validity.

Case (ii): $S = \text{constant}$

Here we are considering a rigid tube where the cross-sectional area does not vary, and hence $c = 0$. Eliminating p between the two equations in (9.91) gives

$$\rho v_t - \frac{1}{g} v_{xx} + rv = 0$$

Substituting $v = V e^{-rt/\rho}$, we have

$$\rho \left(V_t - \frac{rV}{\rho} \right) - \frac{V_{xx}}{g} + rV = 0$$

so that

$$V_t = \frac{1}{\rho g} V_{xx}$$

which is just the diffusion equation. The solution for this rigid-tube case is therefore a damped, diffusion solution. Typically, if we start with a delta-function pulse at the origin then it can be checked that the solution is

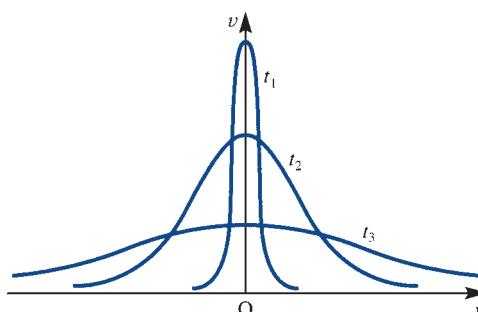
$$v = A \frac{e^{-rt/\rho}}{t^{1/2}} e^{-\rho gx^2/4t}$$

where A is a constant. This solution is plotted in Figure 9.64, and shows the rapid damping. Such a pulse would be most unlikely to propagate far enough for blood to reach the whole of the system.

The two cases considered are extremes, but, just from the analysis performed, some conclusions can be drawn. If there is no distensibility then pulses will not propagate but will just diffuse through the system. We conclude that to move blood through the system with a series of pulses is not possible with rigid blood vessels, and we need flexible walls. Certainly for older people with hardening of the arteries, a major problem is to pump blood round the whole system, and this fact is confirmed by the mathematics.

The actual situation is somewhere between the two cases cited, but there are no simple solutions for such cases except for the ‘balanced line’ case when $cr = g\rho$ (see Exercise 10(c) and Review exercise 20). Singer solves the equations numerically for data appropriate to a dog aorta, and compares his results with experiment. Although the agreement is good, there are problems, since there appears to be a residual pressure after each pulse. The overall pressure would therefore build up to levels that are clearly not acceptable.

Figure 9.64
Development of the solution to the blood-flow problem from a delta function for successive times t_1 , t_2 and t_3 .



9.11 Review exercises (1–21)

- 1 A uniform string is stretched along the x axis and its ends fixed at the points $x = 0$ and $x = a$. The string at the point $x = b$ ($0 < b < a$) is drawn aside through a small displacement ε perpendicular to the x axis, and released from rest at time $t = 0$. By solving the one-dimensional wave equation, show that at any subsequent time t the transverse displacement y is given by

$$y = \frac{2\varepsilon a^2}{\pi^2 b(a-b)} \sum_{n=1}^{\infty} \frac{1}{n^2} \sin\left(\frac{n\pi b}{a}\right) \sin\left(\frac{n\pi x}{a}\right) \times \cos\left(\frac{n\pi c t}{a}\right)$$

where c is the transverse wave velocity in the string.

- 2 The function $\phi(x, t)$ satisfies the wave equation

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{\partial^2 \phi}{\partial t^2} \quad (t > 0, 0 \leq x \leq l)$$

and the conditions

$$\phi(x, 0) = x^2 \quad (0 \leq x \leq l)$$

$$\frac{\partial \phi}{\partial t}(x, 0) = 0 \quad (0 \leq x \leq l)$$

$$\phi(0, t) = 0 \quad (t > 0)$$

$$\frac{\partial \phi}{\partial x}(l, t) = 2l \quad (t > 0)$$

Show that the Laplace transform of the solution is

$$\frac{2}{s^3} + \frac{x^2}{s} - \frac{2}{s^3} \frac{\cosh s(x-l)}{\cosh sl}$$

Using tables of Laplace transforms, deduce that the solution of the wave equation is

$$2xl - \frac{32l^2}{\pi^3} \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^3} \cos\left[\frac{(2n+1)(x-l)\pi}{2l}\right] \times \cos\left[\frac{(2n+1)\pi t}{2l}\right]$$

- 3 The damped vibrations of a stretched string are governed by the equation

$$\frac{1}{c^2} \frac{\partial^2 y}{\partial t^2} + \frac{1}{c^2 \tau} \frac{\partial y}{\partial t} = \frac{\partial^2 y}{\partial x^2} \quad (9.92)$$

where $y(x, t)$ is the transverse deflection, t is the time, x is the position coordinate along the string, and c and τ are positive constants. A taut elastic string, $0 \leq x \leq l$, is fixed at its end points so that $y(0) = y(l) = 0$. Show that separation of variable solutions of (9.92) satisfying these boundary conditions are of the form

$$y_n(x, t) = T_n(t) \sin\left(\frac{n\pi x}{l}\right) \quad (n = 1, 2, \dots)$$

where

$$\frac{1}{c^2} \frac{d^2 T_n}{dt^2} + \frac{1}{c^2 \tau} \frac{dT_n}{dt} + \frac{n^2 \pi^2 T_n}{l^2} = 0$$

Show that if the parameters c , τ and l are such that $2\pi c \tau > l$, the solutions for T_n are all of the form

$$T_n(t) = e^{-t/2\tau} (a_n \cos \omega_n t + b_n \sin \omega_n t)$$

where

$$\omega_n = \frac{n\pi c}{l} \left(1 - \frac{l^2}{4\pi^2 n^2 c^2 \tau^2}\right)^{1/2}$$

and a_n and b_n are constants.

Hence find the general solution of (9.92) satisfying the given boundary conditions.

Given the initial conditions $y(x, 0) = 4 \sin(3\pi x/l)$ and $(\partial y / \partial t)_{t=0} = 0$, find $y(x, t)$.

- 4 A thin uniform beam OA of length l is clamped horizontally at both ends. For small transverse vibrations of the beam the displacement $u(x, t)$ at time t at a distance x from O satisfies the equation

$$\frac{\partial^4 u}{\partial x^4} + \frac{1}{a^2} \frac{\partial^2 u}{\partial t^2} = 0$$

where a is a constant. The restriction that the beam is clamped horizontally gives the boundary conditions

$$u = 0, \quad \frac{\partial u}{\partial x} = 0 \quad (x = 0, l)$$

Show that for periodic solutions of the type

$$u(x, t) = V(x) \sin(\omega t + \varepsilon)$$

where ω and ε are constants, to exist, V must satisfy an equation of the form

$$\frac{d^4 V}{dx^4} = \alpha^4 V \quad (9.93)$$

where $\alpha^4 = (\omega/a)^2$, and the boundary conditions

$$V(0) = V'(0) = V(l) = V'(l) = 0$$

Verify that

$$V = A \cosh \alpha x + B \cos \alpha x + C \sinh \alpha x \\ + D \sin \alpha x$$

where A, B, C and D are constants, satisfies (9.93), and show that this function satisfies the boundary conditions provided that

$$B = -A, \quad D = -C$$

and α is a root of

$$\cos \alpha l \cosh \alpha l = 1$$

- 5** In a uniform bar of length l the temperature $\theta(x, t)$ at a distance x from one end satisfies the equation

$$\frac{\partial^2 \theta}{\partial x^2} = a^2 \frac{\partial \theta}{\partial t}$$

where a is a constant. The end $x = l$ is kept at zero temperature and the other end $x = 0$ is perfectly insulated, so that

$$\theta(l, t) = 0, \quad \frac{\partial \theta}{\partial t}(0, t) = 0 \quad (t > 0)$$

Using the method of separation of variables, show that if initially the temperature in the bar is $\theta(x, 0) = f(x)$ then subsequently the temperature is

$$\theta(x, t) = \sum_{n=0}^{\infty} A_{2n+1} \cos \left[\frac{(2n+1)\pi x}{2l} \right] \\ \times \exp \left[-\frac{(2n+1)^2 \pi^2 t}{4a^2 l^2} \right]$$

where

$$A_{2n+1} = \frac{2}{l} \int_0^l f(x) \cos \left[\frac{(2n+1)\pi x}{2l} \right] dx$$

Given $\theta(x, 0) = \theta_0(l-x)$, where θ_0 is a constant, determine the subsequent temperature in the bar.

- 6** Prove that if $z = x/\sqrt{t}$ and $\phi(x, t) = f(z)$ satisfies the heat-conduction equation

$$\kappa \frac{\partial^2 \phi}{\partial x^2} = \frac{\partial \phi}{\partial t} \quad (9.94)$$

then $f(z)$ must be of the form

$$f(z) = A \operatorname{erf} \left(\frac{z}{2\sqrt{\kappa}} \right) + B$$

where A and B are constants and the **error function** is defined as

$$\operatorname{erf}(\xi) = \frac{2}{\sqrt{\pi}} \int_0^{\xi} e^{-u^2} du$$

A heat-conducting solid occupies the semi-infinite region $x \geq 0$. At time $t = 0$ the temperature everywhere in the solid has the value T_0 . The temperature at the surface, $x = 0$, is suddenly raised, at $t = 0$, to the constant value $T_0 + \phi_0$ and is then maintained at this temperature. Assuming that the temperature field in the solid has the form

$$T = \phi(x, t) + T_0$$

where ϕ satisfies (9.94) in $x > 0$, find the solution of this problem.

- 7** Use the explicit method and the Crank–Nicolson formula to solve the heat-conduction equation

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{\partial \phi}{\partial t}$$

given that ϕ satisfies the conditions

$$\phi = 1 \quad (0 \leq x \leq 1, t = 0)$$

$$\frac{\partial \phi}{\partial x} = \begin{cases} \phi & (x = 0) \\ -\phi & (x = 1) \end{cases} \quad (t \geq 0)$$

Compute $\phi(x, t)$ at $x = 0, 0.2, 0.4, 0.6, 0.8, 1$ when $t = 0.004$ and $t = 0.008$.

- 8** An infinitely long bar of square cross-section has faces $x = 0, x = a, y = 0, y = a$. The bar is made of heat-conducting material, and under steady-state conditions the temperature T satisfies the Laplace equation

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0$$

All the faces except $y = 0$ are kept at zero temperature, while the temperature in the face $y = 0$ is given by $T(x, 0) = x(a - x)$. Show that the temperature distribution in the bar is

$$\frac{8a^2}{\pi^3} \sum_{r=0}^{\infty} \frac{\sin[(2r+1)\pi x/a] \sinh[(2r+1)\pi(a-y)/a]}{(2r+1)^3 \sinh(2r+1)\pi}$$

- 9** (Harder) The function ϕ satisfies the Poisson equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = xy^2$$

inside the area bounded by the parabola $y^2 = x$ and the line $x = 2$. The function ϕ is given at all points on the boundary as $\phi = 1$. By using a square grid of side $\frac{1}{2}$ and making full use of symmetry, formulate a set of finite-difference equations for the unknown values ϕ , and solve.

- 10** A semi-infinite region of incompressible fluid of density ρ and viscosity μ is bounded by a plane wall in the plane $z = 0$ and extends throughout the region $z \geq 0$. The wall executes oscillations in its own plane so that its velocity at time t is $U \cos \omega t$. No pressure gradients or body forces are operative. It can be shown that the velocity of the liquid satisfies the equation

$$\frac{\partial u}{\partial t} = v \frac{\partial^2 u}{\partial z^2}$$

where $v = \mu/\rho$. Establish that an appropriate solution of the equation is

$$u = U e^{-\alpha z} \cos(\omega t - \alpha z)$$

where $\alpha = \sqrt{(\omega/2v)}$.

- 11** Determine the value of the constant k so that

$$U = t^k e^{-r^{3/4}t}$$

satisfies the partial differential equation

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial U}{\partial r} \right) = \frac{\partial U}{\partial t}$$

Sketch the solution for successive values of t .

- 12** The function $z(x, y)$ satisfies

$$\frac{\partial z}{\partial x} + \frac{\partial z}{\partial y} = 0$$

with the boundary conditions

$$z = 2x \quad \text{when } y = -x \quad (x > 0)$$

Find the unique solution for z and the region in which this solution holds. Check the solution using MAPLE.

- 13** The function $\phi(x, y)$ satisfies the Laplace equation

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0$$

in the region $0 < x < \pi$, $0 < y$, and also the boundary conditions

$$\phi \rightarrow 0 \quad \text{as } y \rightarrow \infty$$

$$\phi(0, y) = \phi(\pi, y) = 0$$

Show that an appropriate separation of variables solution is

$$\phi = \sum_{n=1}^{\infty} c_n \sin(nx) e^{-ny}$$

Show that if further

$$\phi(x, 0) = x(\pi - x)$$

then $c_{2m} = 0$ while the odd coefficients are given by

$$c_{2m+1} = \frac{8}{\pi(2m+1)^3}$$

- 14** The boundary-value problem associated with the torsion of a prism of rectangular cross-section $-a \leq x \leq a$, $-b \leq y \leq b$ entails the solution of

$$\frac{\partial^2 \chi}{\partial x^2} + \frac{\partial^2 \chi}{\partial y^2} = -2$$

subject to $\chi = 0$ on the boundary. Show that the differential equation and the boundary conditions on $x = \pm a$ are satisfied by a solution of the form

$$\begin{aligned} \chi &= a^2 - x^2 + \sum_{n=0}^m A_{2n+1} \cosh \left[\frac{2n+1}{2a} \pi y \right] \\ &\quad \times \cos \left[\frac{(2n+1)\pi x}{2a} \right] \end{aligned}$$

From the condition $\chi = 0$ on the boundaries $y = \pm b$, evaluate the coefficients A_{2n+1} .

- 15** When $0 < x < 1$ and $t > 0$ the function $u(x, t)$ satisfies the wave equation

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 u}{\partial t^2}$$

and is also subject to the following boundary conditions:

- (a) $u(0, t) = u(1, t) = 0$ for all $t > 0$
- (b) $\frac{\partial u}{\partial t}(x, 0) = 0$ ($0 < x < 1$)
- (c) $u(x, 0) = 1 - x$ ($0 < x < 1$)

Use the separation method to find the solution for $u(x, t)$ that is valid for $0 < x < 1$ and $t > 0$.

- 16** The excess porewater pressure $u(z, t)$ in an infinite layer of clay satisfies the diffusion equation

$$\frac{\partial u}{\partial t} = c \frac{\partial^2 u}{\partial z^2} \quad (t > 0, 0 < z < h)$$

where t is the time in minutes, z is the vertical height in metres from the base of the clay layer and c is the coefficient of consolidation. There is complete drainage at the top and bottom of the clay layer, which is of thickness h . The distribution of excess porewater pressure $u(z, t)$ is A at $t = 0$ where A is a constant. Show that

$$u(z, t) = \frac{4A}{\pi} \sum_{n=0}^{\infty} \frac{\sin[(2n+1)\pi z/h]}{2n+1} e^{-cr^2(2n+1)^2 th^2}$$

- 17** By seeking a separated solution of the form $\phi = X(x)T(t)$, find a solution to the telegraph equation

$$\frac{\partial^2 \phi}{\partial x^2} = \frac{1}{c^2} \left(\frac{\partial^2 \phi}{\partial t^2} + K \frac{\partial \phi}{\partial t} \right)$$

satisfying the conditions

- (a) $\phi = A \cos px$ for all values of x and for $t = 0$ for the case when $c^2 p^2 > \frac{1}{4} K^2$;
- (b) $\phi = A$ and $\partial \phi / \partial t = -\frac{1}{2} AK$ for $x = 0$ and $t = 0$.

- 18** For the two-dimensional flow of an incompressible fluid the continuity equation may be expressed as

$$\frac{\partial}{\partial r}(r v_r) + \frac{\partial v_\theta}{\partial \theta} = 0$$

where r and θ are polar coordinates in a plane parallel to the flow, and v_r and v_θ are the respective velocity components. Show that a stream function ψ such that

$$v_r = \frac{1}{r} \frac{\partial \psi}{\partial \theta}$$

$$v_\theta = -\frac{\partial \psi}{\partial r}$$

satisfy the continuity equation.

Take

$$\psi = Ur \sin \theta - \frac{Ua^2}{r} \sin \theta$$

and interpret the solution physically.

(An extended problem) Section 9.9 looked at wave propagation caused by moving loads on cables. For loads on beams a similar analysis models such problems as trains going over bridges or loads moving on gantry cranes.

Use a similar analysis for the beam equation

$$a^2 \frac{\partial^4 u}{\partial x^4} + \frac{\partial^2 u}{\partial t^2} = p(x, t)$$

- 20**

(An extended problem) In the blood-flow model in Section 9.10 consider the following cases:

- (a) $S = \text{constant}$, $g = 0$ for a pulsating flow

$$v = v_0 e^{j\omega t} \quad \text{at } x = 0 \text{ for all } t$$

- (b) $S = \text{constant}$, $r = 0$ for a pulsating flow

$$v = v_0 e^{j\omega t} \quad \text{at } x = 0 \text{ for all } t$$

- (c) the balanced-line case when $g\rho = rc$. Show that $v = e^{-g\tau/c} U$ gives

$$\frac{\partial^2 U}{\partial t^2} = \frac{1}{c\rho} \frac{\partial^2 U}{\partial x^2}$$

Solve the equation and interpret your solution.

- 21**

(An extended problem) Fluid flows steadily in the two-dimensional channel shown in Figure 9.65. The temperature $\theta = \theta(x, t)$ depends only on the distance x along the channel and the time t . The

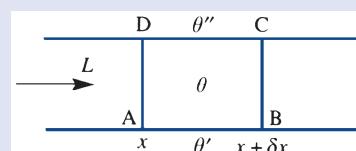


Figure 9.65 An element of the channel in Review exercise 21.

fluid flows at a constant rate so that an amount L crosses any given section in unit time. The specific heat of the fluid is a constant c , and the heat H in a length δx with cross-section S is therefore

$$H = c(S \delta x) \theta$$

Heat is transferred through the walls of the channel, AB and DC, at a rate proportional to the temperature difference between the inside and outside. Heat conduction in the x direction is neglected. Show that the heat balance in the element ABCD leads to the equation

$$cS \frac{\partial \theta}{\partial t} = -Lc \frac{\partial \theta}{\partial x} + k_1(\theta'' - \theta) + k_2(\theta' - \theta)$$

This type of analysis can now be applied to the long heat exchanger illustrated in Figure 9.66. The configuration is considered to be two-dimensional and symmetric with respect to the x axis; in the inner region the flow is to the right, while in the outer regions it is to the left. The regions are separated by metal walls in which similar assumptions to the above are made, except of course there is no fluid flow.

Set up the equations of the system in the form

$$c_1 S_1 \frac{\partial \theta_1}{\partial t} = -L_1 c_1 \frac{\partial \theta_1}{\partial x} + 2k_1(\theta_2 - \theta_1)$$

$$c_2 S_2 \frac{\partial \theta_2}{\partial t} = -k_1(\theta_1 - \theta_2) + k_2(\theta_3 - \theta_2)$$

$$c_3 S_3 \frac{\partial \theta_3}{\partial t} = L_3 c_3 \frac{\partial \theta_3}{\partial x} + k_2(\theta_2 - \theta_3)$$

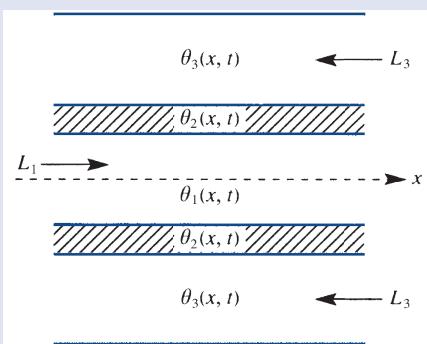


Figure 9.66 Heat-exchanger configuration in Review exercise 21.

where the assumption is made that there is no heat flow through the outside lagged walls. Solve the steady-state equations and fit the arbitrary constants to the conditions that at the inlet ($x = 0$) the fluid enters the inner region at a given temperature $\theta_1 = T_1$, while at the outlet ($x \rightarrow \infty$) the fluid in the outer regions enters at a given temperature $\theta_3 = T_3$. Find flow rates that ensure that this situation is possible, and discuss the implications of any results obtained.

Discuss the assumptions made in setting up this problem, the limitations imposed by the assumptions, possible applications of this type of analysis, and extensions of the work, for example a time-dependent solution.



10 Optimization

Chapter 10 Contents

10.1	Introduction	736
10.2	Linear programming	739
10.3	Lagrange multipliers	764
10.4	Hill climbing	769
10.5	Engineering application: chemical processing plant	790
10.6	Engineering application: heating fin	792
10.7	Review exercises (1–26)	795

10.1**Introduction**

The need to get the ‘best’ out of a system is a very strong motivation in much of engineering. A typical problem may be to obtain the maximum amount of product or to minimize the cost of a process or to find a configuration that gives maximum strength. Sometimes what is ‘best’ is easy to define, but frequently the problem is not so clear cut, and a lot of thought is required to reach an appropriate function to optimize. In most cases there are very severe and natural constraints operating: the problem may be one of maximizing the amount of product, subject to the supply of materials; or it may be minimizing the cost of production, with constraints due to safety standards. Indeed, much of modern optimization is concerned with constraints and how to deal with them.

We have seen in Chapter 9 of *Modern Engineering Mathematics* (MEM) how to obtain the maximum and minimum of a function of many variables. However, the methods described there founder very quickly because most engineering optimization problems are not possible to solve analytically. A simple one-dimensional example soon shows that a numerical solution is required.

Example 10.1

Find the positive x value that maximizes the function

$$y = \frac{\tanh x}{1 + x}$$

Solution Equating the derivative to zero gives

$$\frac{dy}{dx} = 0 = \frac{(1+x) \operatorname{sech}^2 x - \tanh x}{(1+x)^2}$$

so that we need to solve

$$1 + x = \frac{1}{2} \sinh 2x$$

which has no simple positive solutions that can be obtained analytically.

To solve such problems, a set of numerical algorithms was developed during the 1960s as fast computers became available to perform the large amounts of arithmetic required. These algorithms will be described in Section 10.4. Perhaps the main stimulus for this development came from the space industries, where small percentage savings, achieved by doing some mathematics, could save vast amounts of money. The ideas were quickly taken up by ‘expensive’ areas of engineering, such as the chemical and steel industries and aircraft production.

The idea of dealing with constraints is not new: Lagrange developed the theory of **equality**-constrained optimization around the 1800s. However, it was not until the 1940s that **inequality** constraints were studied with any seriousness. The use of Lagrange multipliers for equality constraints was also introduced in Chapter 9 of MEM, and will be looked at again in more detail in Section 10.3 below. The only work on

inequality constraints will be for linear programming problems in Section 10.2. Where inequality constraints are nonlinear, the problems become very difficult, and are the province of specialist books on optimization, for example E. K. P. Chong and S. H. Zak, *An Introduction to Optimization* (fourth edition, New York, Wiley, 2013). Linear programming, however, is much more straightforward, and the basic simplex algorithm has been spectacularly successful – so successful in fact that many workers try to force their problems to be linear when they are clearly not. The computer scientist's maxim GIGO ('garbage in, garbage out') is very applicable to people who try to fit the problem to the mathematics rather than the mathematics to the problem!

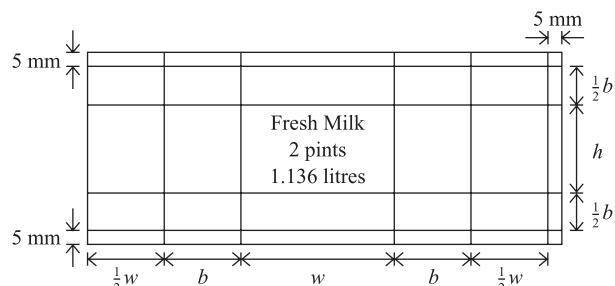
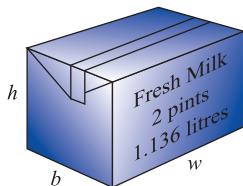
Before considering detailed methods of solution of optimization problems, we shall look at a few examples. Let us first revisit an extended form of the milk carton problem considered in Example 8.34 (and illustrated in Figure 8.38) of MEM.

Example 10.2

A milk carton is designed from a sheet of waxed cardboard as illustrated in Figure 10.1, where a 5 mm overlap has been allowed.

It is to contain 2 pints of milk, and we require the minimum surface area for the carton.

Figure 10.1 Waxed cardboard milk container opened up, with measurements in millimetres and with a 5 mm overlap.



Solution

The only difference between this example and Example 8.34 of MEM is that we no longer insist on a square cross-section. The total area in square millimetres is

$$A = (2b + 2w + 5)(h + b + 10)$$

and the volume of the two-pint container is

$$\text{volume} = hbw = 1\,136\,000 \text{ mm}^3$$

We first note that a constraint, the given volume, occurs naturally in the problem. Because of its simplicity, we can eliminate w from the constraint to give

$$A = (h + b + 10)\left(\frac{2\,272\,000}{hb} + 2b + 5\right)$$

Following the standard minimization procedure and equating partial derivatives to zero gives

$$\frac{\partial A}{\partial h} = \frac{2\ 272\ 000}{hb} + 2b + 5 - (h + b + 10) \frac{2\ 272\ 000}{h^2 b} = 0$$

$$\frac{\partial A}{\partial b} = \frac{2\ 272\ 000}{hb} + 2b + 5 - (h + b + 10) \left(\frac{-2\ 272\ 000}{hb^2} + 2 \right) = 0$$

We therefore have two highly nonlinear equations in the two unknowns h and b , which cannot be solved without resorting to numerical techniques. We shall return to this problem later in Examples 10.12 and 10.18 to see how a practical solution can be obtained.

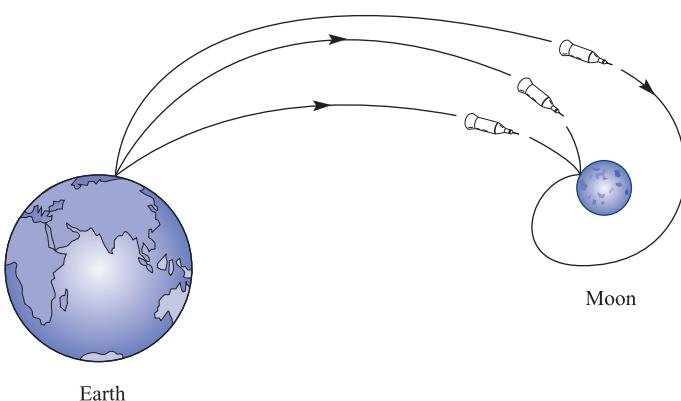
Most practical optimization problems come from very expensive projects where savings of a few per cent can be very significant. Laying natural gas or water pipe networks are typical examples. Without considering the expense of installing compressors, the problem is to minimize the capital cost. This cost is directly related to the weight of the pipe, subject to constraints imposed by pressure-drop limitations, which in turn depend on the pipe diameter in a nonlinear way. Adding the compressors imposes further costs and constraints.

Heat exchangers provide an example of a system where we try to remove heat. We design the flow rates, the pipe sizes and pipe spacing to maximize the heat transferred. A related heating problem might be the design of an industrial furnace. It is required that the energy consumption be minimized subject to constraints on the heat flow and the maintenance of various temperatures.

A final example, the moonshot problem, illustrates a large-scale, very complicated problem that stimulated much of the recent developments in optimization (see Figure 10.2). Which path from a point on the Earth to a point on the Moon should be chosen to minimize the weight of fuel carried by a rocket? The complicated relation between the weight of fuel, the mechanical equations of the rocket and the path must be established before it is possible to proceed to obtain the optimum. The numerous constraints on the strengths of materials, the maximum tolerable acceleration etc. add to the difficulty of the problem.

In the problems discussed above, we have assumed that an optimum exists at a point, and we have asked for the mathematical conditions that must hold. The other way round is much more difficult. Given that the appropriate conditions hold, does an optimum exist, and if so what type of optimum is it? For many simple finite-dimensional

Figure 10.2
The moonshot
problem.



problems these conditions are known, but may not be very simple to apply. To serve as a reminder, the condition $f'(0) = 0$ is a necessary condition for a maximum to exist for the differentiable function $f(x)$ at $x = 0$. It is not *sufficient*, however, as can be seen from the three functions $f_1(x) = x^2$, $f_2(x) = x^3$ and $f_3(x) = -x^2$, which have respectively a minimum, a point of inflection and a maximum at the origin. In many dimensions the difficulties are similar, but much more complicated.

10.2 Linear programming

10.2.1 Introduction

In Section 10.1 it was indicated that constraints are very important in most applications. When all functions are linear, there is an extremely efficient algorithm, developed by Danzig in the 1940s, which will be described for the **linear programming (LP)** problem.

We shall start by posing a particular problem and looking at a simple graphical solution.

Example 10.3

A manufacturing company makes two circuit boards R1 and R2, constructed as follows:

R1 comprises 3 resistors, 1 capacitor, 2 transistors and 2 inductances;

R2 comprises 4 resistors, 2 capacitors and 3 transistors.

The available stocks for a day's production are 2400 resistors, 900 capacitors, 1600 transistors and 1200 inductances. It is required to calculate how many R1 and how many R2 the company should produce daily in order to maximize its overall profits, knowing that it can make a profit on an R1 circuit board of 5p and on an R2 circuit board of 9p.

Solution If the company produces daily x of type R1 and y of type R2 then its stock limitations give

$$3x + 4y \leq 2400 \quad (10.1a)$$

$$x + 2y \leq 900 \quad (10.1b)$$

$$2x + 3y \leq 1600 \quad (10.1c)$$

$$2x \leq 1200 \quad (10.1d)$$

$$x \geq 0, \quad y \geq 0$$

and it makes a profit z given by

$$z = 5x + 9y \quad (10.2)$$

These inequalities are plotted on a diagram as in Figure 10.3(a). The shaded region defines the area for which *all* the inequalities are satisfied, and is called the **feasible region**. The lines of constant profit $z = \text{constant}$, defined by (10.2), are plotted as ‘dashed’ lines in Figure 10.3(b). It is clear from the geometry that the largest possible value of z that intersects the feasible region is at S with $x = 500$, $y = 200$, and this gives the optimal

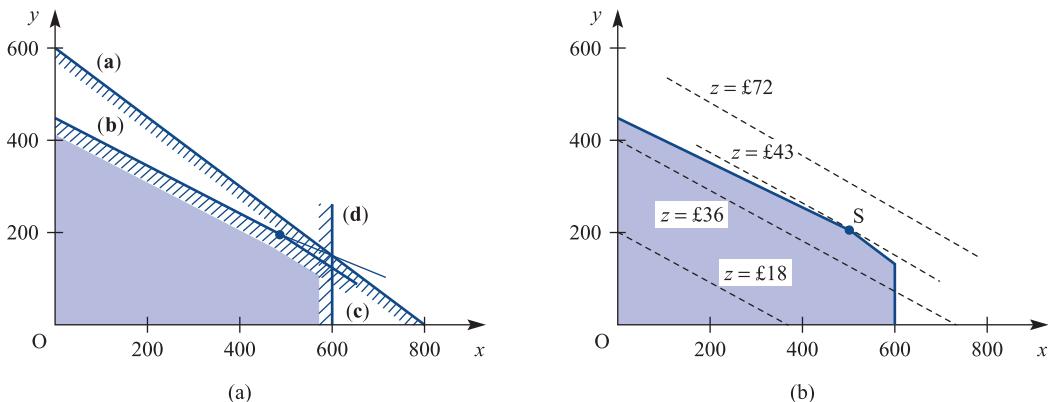


Figure 10.3 (a) Feasible region for the circuit board manufacture problem of Example 10.3. (b) Lines of constant z show that $S(500, 200)$ gives the optimum.

solution. At this point we can analyse the usage of the stocks as in Figure 10.4 and note that a profit of £43 has been made.

Figure 10.4 Table of stock usage.

	Available	Used	Left over
Resistors	2400	2300	100
Capacitors	900	900	0
Transistors	1600	1600	0
Inductances	1200	1000	200

Example 10.3 has encapsulated much of the LP method, and we shall try to extract the maximum amount of information from this example. The function to be optimized, usually denoted by Z , is called the **objective function**. The objective function and the constraints are functions of the **decision variables**. The graphical method will only work if the problem has two decision variables, so we need to consider how to translate the geometry into an algebraic form that will work with any number of variables. Although we shall concentrate in this chapter on small problems in order to illustrate the methods, in practical problems there can be hundreds of variables and constraints. Large problems bring further difficulties that will not be considered here; for instance, how a large amount of information can be input into a computer accurately or how large data sets are handled in the computer. In the MATLAB implementation of LP, there is a specific option to deal with ‘LargeScale’ problems.

From Figure 10.3 it can be seen that the solutions must be at a ‘corner’ of the feasible region, other than in the exceptional case when the profit line $z = \text{constant}$ is parallel to one of the constraints. This follows through into many-dimensional problems, so that it is *only necessary to inspect the corners of the feasible region*. The **simplex method**, described in Section 10.2.3, uses this fact and selects a starting corner, chooses the neighbouring corner that increases z the most, and then repeats the process until no improvement is possible. The method writes the equations into a standard form; it then automates the choice of corner and finally reprocesses the equations back to the standard form again.



Once a solution has been obtained, it may be observed from Figures 10.3 and 10.4 that the **binding constraints** (b) and (c) intersect at S and are satisfied identically, so that all the stocks are used, while the **non-binding constraints** (a) and (d) leave some stock unused. It can also be seen from Figure 10.3 that the constraint (a) is redundant since it does not intersect the feasible region. These might appear obvious comments, but they prove to be useful and relevant observations when a sensitivity analysis is performed. Such an analysis asks whether or not the solution changes as the stocks vary or the costs vary, or the coefficients are changed. In practice, parameters vary over a period, and we wish to know whether a new calculation must be performed or whether the solution that we have already obtained can be used.

10.2.2 Simplex algorithm: an example

We now need to convert the ideas of Section 10.2.1 into a useful algebraic algorithm. There is a whole array of technical terms that are used in LP, and they will be introduced as we reconsider Example 10.3 to develop the solution method. The first step is to introduce **slack variables** r, s, t and u into (10.1) to make the inequality constraints into equality constraints.

If we are given x and y in the feasible region, the variables r, s, t and u provide a measure of how much ‘slack’ is available before all the corresponding resource is used up, so

$$3x + 4y + r = 2400 \quad (10.3a)$$

$$x + 2y + s = 900 \quad (10.3b)$$

$$2x + 3y + t = 1600 \quad (10.3c)$$

$$2x + u = 1200 \quad (10.3d)$$

where x, y, r, s, t and u are now all greater than or equal to zero. We now have more variables than equations, and this enables us to construct a **feasible basic solution** by inspection:

$$\underbrace{x = y = 0; \quad \text{non-basic variables}}_{\text{non-basic variables}} \quad \underbrace{r = 2400, s = 900, t = 1600, u = 1200}_{\text{basic variables}}$$

with 4 **basic variables** (the same number as constraints, which are non-zero) and 2 **non-basic variables** (the remainder of the variables, which are zero). (Note: This corresponds to the origin in Figure 10.3.)

The algebraic equivalent of moving to a neighbouring corner is to increase one of the non-basic variables from zero to its largest possible value. From the profit function given in (10.2), we have

$$z = 5x + 9y$$

Currently z has the value zero, and it seems sensible to change y , since the coefficient of y is larger; this will increase z the most. So keep $x = 0$ in (10.3) and increase y to its maximum value in each case: either

- (a) change y to 600 and reduce r to zero, or
- (b) change y to 450 and reduce s to zero, or
- (c) change y to $533\frac{1}{3}$ and reduce t to zero, or
- (d) note that there is no effect on changing y .

Choose option (b), since increasing y above 450 will make s negative, which would then violate the condition that all variables must be positive. Interchange s and y between the set of basic and non-basic variables and rewrite in the same form as (10.3). This is achieved by solving for y from (10.3b), $y = 450 - \frac{1}{2}x - \frac{1}{2}s$, and substituting to give

$$x - 2s + r = 600 \quad (10.4a)$$

$$\frac{1}{2}x + \frac{1}{2}s + y = 450 \quad (10.4b)$$

$$\frac{1}{2}x - \frac{3}{2}s + t = 250 \quad (10.4c)$$

$$2x + u = 1200 \quad (10.4d)$$

and, from (10.2),

$$z = 4050 + \frac{1}{2}x - \frac{9}{2}s \quad (10.4e)$$

The problem is now reduced to exactly the same form as (10.3), and the same procedure can be applied. The non-basic variables are $x = s = 0$, and the basic variables are $r = 600$, $y = 450$, $t = 250$ and $u = 1200$. z has increased its value from 0 to 4050.

Now only x can be increased, since increasing the other non-basic variable, s , would decrease z . Increasing x to 500 in (10.4c) and reducing t to 0 is the best that can be done. Using (10.4c) to write $x = 3s - 2t + 500$, we now eliminate x from the other equations to give

$$s - 2t + r = 100 \quad (10.5a)$$

$$2s - t + y = 200 \quad (10.5b)$$

$$-3s + 2t + x = 500 \quad (10.5c)$$

$$6s - 4t + u = 200 \quad (10.5d)$$

and

$$z = 4300 - 3s - t \quad (10.5e)$$

We now have the final solution, since increasing s or t can only decrease z . Thus we have $x = 500$, $y = 200$, which is in agreement with the previous graphical solution, the maximum profit is $z = 4300$ as before, and the amounts left over in Figure 10.4 are just the 100 and 200 appearing on the right-hand sides of (10.5a, d).

We have just described the essentials of the **simplex algorithm**, although the method of working may have appeared a little haphazard. It can be tidied up and formalized by writing the whole system in **tableau form**. Equations (10.3) are written with the basic variables in the left-hand column, the coefficients in the equations placed in the appropriate array element and the objective function z placed in the first row with minus signs inserted.

	Non-basic variables		Basic variables				Solution
	x	y	r	s	t	u	
Objective function z	-5	-9	0	0	0	0	0
Basic variables	r	3	4	1	0	0	2400
	s	1	2	0	1	0	900
	t	2	3	0	0	1	1600
	u	2	0	0	0	1	1200

The current solution can easily be read from the tableau. The basic variables in the left-hand column are equal to the values in the solution column, so $r = 2400$, $s = 900$, $t = 1600$ and $u = 1200$. The remaining non-basic variables are zero, namely $x = y = 0$. The profit z is read similarly as the entry in the solution column, namely $z = 0$. The negative signs in the z row ensure that z remains positive in the subsequent manipulation. It should be noted that a 4×4 unit matrix (shown shaded) occurs in the tableau in the basic variable columns, with zeros occurring above in the z row. This standard display is always the starting place for the simplex method, with the only possible complication being that the columns of the unit matrix might be shuffled around. The algorithm can now be performed in a series of steps:

Step 1

Choose the most negative entry in the z row and mark that column (the y column in this case).

Step 2

Evaluate the ratios of the solution column and the *positive* entries in the y column, choose the smallest of these and mark that row (the s row in this case).

	x	y	r	s	t	u	Solution	
z	-5	-9	0	0	0	0	0	Ratios
r	3	4	1	0	0	0	2400	$2400/4 = 600$
s	1	2	0	1	0	0	900	$900/2 = 450$
t	2	3	0	0	1	0	1600	$1600/3 = 533\frac{1}{3}$
u	2	0	0	0	0	1	1200	-

Step 3

Change the marked basic variable in the left-hand column to the marked non-basic variable in the top row (in this case s changes to y in the left-hand column).

Step 4

Make the pivot (the element in the position where the marked row and column cross) 1 by dividing through. In this case we divide the row elements by 2. These series of steps lead to the tableau

	x	y	r	s	t	u	Solution
z	-5	-9	0	0	0	0	0
r	3	4	1	0	0	0	2400
y	$\frac{1}{2}$	1	0	$\frac{1}{2}$	0	0	450
t	2	2	0	0	1	0	1600
u	2	0	0	0	0	1	1200

Step 5

Clear the y column by subtracting an appropriate multiple of the y row (this is just Gaussian elimination); for example, $(z \text{ row}) + 9 \times (y \text{ row})$, $(r \text{ row}) - 4 \times (y \text{ row})$ and so on. This leads to the tableau

	x	y	r	s	t	u	Solution
z	$-\frac{1}{2}$	0	0	$\frac{9}{2}$	0	0	4050
r	1	0	1	-2	0	0	600
y	$\frac{1}{2}$	1	0	$\frac{1}{2}$	0	0	450
t	$\frac{1}{2}$	0	0	$-\frac{3}{2}$	1	0	250
u	2	0	0	0	0	1	1200

This tableau can now easily be recognized as equations (10.4). It may be noted that the unit matrix appears in the tableau again, with the columns permuted, and the z row has zero entries in the basic variable columns.

The tableau is in exactly the standard form, and is ready for reapplication of the five given steps. *Steps 1 and 2* give the tableau

	x	y	r	s	t	u	Solution
z	$-\frac{1}{2}$	0	0	$\frac{9}{2}$	0	0	4050
r		0	1	-2	0	0	600
y		1	0	$\frac{1}{2}$	0	0	450
t	$\frac{1}{2}$	0	0	$-\frac{3}{2}$	1	0	250
u	2	0	0	0	0	1	1200

Steps 3, 4 and 5 then produce a final tableau (compare with equations (10.5))

	x	y	r	s	t	u	Solution
z	0	0	0	3	1	0	4300
r	0	0	1	1	-2	0	100
y	0	1	0	2	-1	0	200
x	1	0	0	-3	2	0	500
u	0	0	0	6	-4	1	200

All the entries in the z row are now positive, so the optimum is achieved. The solution is read from the tableau directly; the left-hand column equals the right-hand column, giving $z = 4300$, $r = 100$, $y = 200$, $x = 500$ and $u = 200$, which is in agreement with the solution obtained in Example 10.3.

10.2.3 Simplex algorithm: general theory

We can now generalize the problem to the standard form of finding the **maximum** of the objective function

$$z = c_1x_1 + c_2x_2 + \dots + c_nx_n$$

subject to the constraints

$$\left. \begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n \leq b_2 \\ \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n \leq b_m \end{array} \right\} \quad (10.6)$$

by the **simplex algorithm**, where the b_1, b_2, \dots, b_m are all positive and all decision variables x_1, x_2, \dots, x_n are non-negative. A linear program written in this form is said to be in **standard form**. By introducing the slack variables $x_{n+1}, \dots, x_{n+m} \geq 0$ we convert (10.6) into the **standard tableau**

	x_1	x_2	\dots	x_n	x_{n+1}	x_{n+2}	\dots	x_{n+m}	Solution
z	$-c_1$	$-c_2$	\dots	$-c_n$	0	0	\dots	0	0
x_{n+1}	a_{11}	a_{12}	\dots	a_{1n}	1	0	\dots	0	b_1
x_{n+2}	a_{21}	a_{22}	\dots	a_{2n}	0	1	\dots	0	b_2
\vdots	\vdots	\vdots		\vdots	\vdots	\vdots		\vdots	\vdots
x_{n+m}	a_{m1}	a_{m2}	\dots	a_{mn}	0	0	\dots	1	b_m

Any subsequent tableau takes this general form, with an $m \times m$ unit matrix in the basic variables columns. As noted in the previous example, the basic variables change, so the left-hand column will have m entries, which can be any of the variables, x_1, \dots, x_{m+n} . The unit-matrix columns are usually not in the above neat form but are permuted and hence the zeros of the ‘ z ’ row can be any of the 1 to $n+m$ entries corresponding to the basic variables.

The five basic steps in the algorithm follow quite generally:

Step 1

Choose the most negative value in the z row, say $-c_i$. (Identify column i)
If all the entries are positive then the maximum has been achieved.

Step 2

Evaluate $b_1/a_{1i}, b_2/a_{2i}, \dots, b_m/a_{mi}$ for all positive a_{ki} . (Identify row j)
Select the minimum of these numbers, say b_j/a_{ji} .

Step 3

Replace x_{n+j} by x_i in the basic variables in the left-hand column. (Change the basis)

Step 4

In row j replace a_{jk} by a_{jk}/a_{ji} for $k = 1, \dots, n+m+1$. (Make pivot = 1)
 (Note that the first row and the final column are treated as part of the tableau for computation purposes, $-c_p = a_{0p}$, $b_q = a_{q(n+m+1)}$)

Step 5

In all other rows, $l \neq j$, replace a_{lk} by $a_{lk} - a_{li}a_{jk}$ for all $k = 1, \dots, m+n+1$ and for each row $l = 0, \dots, m$ ($l \neq j$). (Gaussian elimination)

The algorithm is then repeated until at Step 1 the maximum is achieved. The method provides an extremely efficient way of searching through the corners of the feasible region. To inspect all corners would require the computation of $\binom{m+n}{m}$ points, while the simplex algorithm reduces this very considerably, often down to something of the order of $m+n$.

Several checks should be made at the completion of each cycle, since it may be possible to identify an exceptional case. Perhaps the most complicated of the exceptions is when one of the $b_i = 0$ during the calculation, implying that one of the basic variables is zero. This can be a temporary effect, in which case the problem goes away at the next iteration, or it may be permanent, and that basic variable is indeed zero in the optimal solution. The best that may be said, other than going into sophisticated techniques found in specialist books on LP, is that problems are possible and the computation should be watched carefully. The solution can get into a cycle that cannot be broken.

A second exception, that should be noted carefully, occurs when one of the $c_i = 0$ for a *non-basic* variable in the optimal tableau. The normal simplex algorithm can then change the solution without changing the z row by selecting this i column at Step 1. Because $c_i = 0$, Step 5 is never used on the z row at all. This case corresponds to a degenerate solution with many **alternative solutions** to the problem, and geometrically the profit function is parallel to one of the constraints.

The third exception occurs at Step 2 when all the $a_{1i}, a_{2i}, \dots, a_{mi}$ in the optimal column are zero or negative and it becomes impossible to identify a row to continue the method. The region in this case is **unbounded**, and a careful look at the original problem is required to decide whether this is reasonable, since it may still be possible to get a solution to such a problem.

Example 10.4

Find the maximum of

$$z = 5x_1 + 4x_2 + 6x_3$$

subject to

$$4x_1 + x_2 + x_3 \leq 19$$

$$3x_1 + 4x_2 + 6x_3 \leq 30$$

$$2x_1 + 4x_2 + x_3 \leq 25$$

$$x_1 + x_2 + 2x_3 \leq 15$$

$$x_1, x_2, x_3 \geq 0$$

Solution

The example cannot be solved graphically, since it has three variables, but is in a correct form for the simplex algorithm. The initial tableau gives the solution $x_4 = 19$, $x_5 = 30$, $x_6 = 25$, $x_7 = 15$ and non-basic variables $x_1 = x_2 = x_3 = 0$.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	Solution	
z	-5	-4	-6	0	0	0	0	0	Ratios
x_4	4	1	1	1	0	0	0	19	$19/1 = 19$
x_5	3	4	(6)	0	1	0	0	30	$30/6 = 5$
x_6	2	4	1	0	0	1	0	25	$25/1 = 25$
x_7	1	1	2	0	0	0	1	15	$15/2 = 7.5$

In the initial tableau the pivot is identified, and x_5 is removed from the basic variable column and replaced by x_3 . The pivot is made equal to unity by dividing the x_3 row by 6. The other entries in the x_3 column are then made zero by the Gaussian elimination in Step 5. This gives the tableau

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	Solution	
z	-2	0	0	0	1	0	0	30	Ratios
x_4	($\frac{7}{2}$)	$\frac{1}{3}$	0	1	$-\frac{1}{6}$	0	0	14	4
x_3	$\frac{1}{2}$	$\frac{2}{3}$	1	0	$\frac{1}{6}$	0	0	5	10
x_6	$\frac{3}{2}$	$\frac{10}{3}$	0	0	$-\frac{1}{6}$	1	0	20	13.3
x_7	0	$-\frac{1}{3}$	0	0	$-\frac{1}{3}$	0	1	5	-

The process is then repeated and the pivot is again found, x_4 is replaced by x_1 in the first column, and the next tableau is constructed by following the remaining steps of the simplex algorithm, giving the tableau

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	Solution
z	0	$\frac{4}{21}$	0	$\frac{4}{7}$	$\frac{19}{21}$	0	0	38
x_1	1	$\frac{2}{21}$	0	$\frac{2}{7}$	$-\frac{1}{21}$	0	0	4
x_3	0	$\frac{13}{21}$	1	$-\frac{1}{7}$	$\frac{4}{21}$	0	0	3
x_6	0	$\frac{67}{21}$	0	$-\frac{3}{7}$	$-\frac{2}{21}$	1	0	14
x_7	0	$-\frac{1}{3}$	0	0	$-\frac{1}{3}$	0	1	5

Thus the solution is now optimal, and gives $x_1 = 4$, $x_2 = 0$, $x_3 = 3$, and $z = 38$ as the maximum value. Note that the first two constraints are binding, that is satisfied exactly, while the other two are not. This can easily be deduced by looking at the slack variables in the initial tableau. We have $x_4 = x_5 = 0$, corresponding to the first two constraints, and $x_6 \neq 0$, $x_7 \neq 0$ for the last two constraints.



Computers are particularly helpful when there is an efficient algorithm, such as the simplex algorithm for solving LP problems, since they can perform the arithmetic with speed and accuracy. A typical implementation of the algorithm in MAPLE for Example 10.4 is now given:

```
with(simplex):
constr:={4*x1+x2+x3<=19, 3*x1+4*x2+6*x3<=30,
         2*x1+4*x2+x3<=25, x1+x2+2*x3<=15};
obj:=5*x1+4*x2+6*x3;
maximize(obj, constr, NONNEGATIVE);
```

These few lines of code give $x_1 = 4$, $x_2 = 0$ and $x_3 = 3$ instantly. Similarly in MATLAB, LP problems can be solved but are set up in a slightly different way. It always solves the minimum problem

$$\min_x f^T x \quad \text{such that} \quad \begin{cases} Ax \leq b \\ Aeq = beq \\ lb \leq x \leq ub \end{cases}$$

and the way that the problem is tackled can be controlled in optimset. The following lines of code give the solution to Example 10.4:

```
f=[-5;-4;-6]; A=[4 1 1;3 4 6;2 4 1;1 1 2]; b=[19;30;25;15];
Aeq=[]; beq=[]; lb=zeros(3,1); ub=[]; x0=[];
%[] indicates not used but the lower bound, lb, must be
%set to zero
options=optimset('LargeScale','off','Simplex','on');
[x,fval,exitflag,output,lambda]=linprog(f,A,b,Aeq,beq,lb,
                                         ub,x0,options)
```

Typing lambda.ineqlin gives the values 0.5714, 0.9048, 0, 0 which are the values in the z row of the final tableau and determines whether or not the inequalities are binding.

Clearly this is the quick way to get the ‘answer’, but it does not give any understanding of the method. The package has the facilities to go through the steps of the algorithm one at a time so it can be used to help with the arithmetic while leaving the user to determine the steps of the method.

Example 10.5

A firm has two plants, P1 and P2, that can produce a particular chemical. The product is made from three constituents, A, B and C. In a given period there are 36 000 litres of A, 30 000 litres of B and 12 000 litres of C available. Plant P1 requires the constituents A, B, C to be mixed in the ratio 4:2:1 respectively, and the manufacturer makes a profit of £1.50 per litre of product; plant P2 requires the ratio 3:3:1, and gives a profit of £1 per litre of product.

Determine how production should be allocated to each plant to maximize the profits, and how much of A, B and C remain.

There is a major breakdown in the supply of chemical C, so that only 8000 litres are available in the given period. How should production be changed to maximize the profits, how much has profit been reduced, and how much of A, B and C remain?

Solution

For each 1000 litres produced in plant P1, $\frac{4}{7} \times 1000$ will be constituent A, $\frac{2}{7} \times 1000$ will be B and $\frac{1}{7} \times 1000$ will be C. For each 1000 litres produced in plant P2, $\frac{3}{7} \times 1000$ will be constituent A, $\frac{3}{7} \times 1000$ will be B and $\frac{1}{7} \times 1000$ will be C. Thus, taking the three constituents in turn and letting x_1 and x_2 represent respectively the amount (in 1000 litre units) produced in plants P1 and P2, we obtain

$$\frac{4}{7}x_1 + \frac{3}{7}x_2 \leq 36 \quad 4x_1 + 3x_2 \leq 252$$

$$\frac{2}{7}x_1 + \frac{3}{7}x_2 \leq 30 \quad \text{or} \quad 2x_1 + 3x_2 \leq 210$$

$$\frac{1}{7}x_1 + \frac{1}{7}x_2 \leq 12 \quad x_1 + x_2 \leq 84$$

$$x_1, x_2 \geq 0$$

and the profit

$$z = 1.5x_1 + x_2$$

We can immediately construct the initial tableau

	x_1	x_2	x_3	x_4	x_5	Solution	
z	-1.5	-1	0	0	0	0	Ratios
x_3	4	3	1	0	0	252	63
x_4	2	3	0	1	0	210	105
x_5	(1)	1	0	0	1	84	84

The pivot has been found, and hence we introduce x_1 into the basis and construct the next tableau following the steps of the simplex algorithm:

	x_1	x_2	x_3	x_4	x_5	Solution
z	0	$\frac{1}{8}$	$\frac{3}{8}$	0	0	94.5
x_1	1	$\frac{3}{4}$	$\frac{1}{4}$	0	0	63
x_4	0	$\frac{3}{2}$	$-\frac{1}{2}$	1	0	84
x_5	0	$\frac{1}{4}$	$-\frac{1}{4}$	0	1	21

The z row is all positive, and hence we can immediately read off the solution (multiply by 1000 to re-establish proper costs)

$$x_1 = 63\,000, \quad x_2 = 0, \quad z = £94\,500$$

and only plant P1 is utilized. From the initial tableau we see that since $x_3 = 0$, there are zero litres of A remaining; $x_4 = 84$, so that we have $(84/7) \times 1000 = 12\,000$ litres of B remaining; and $x_5 = 21$, so that $(21/7) \times 1000 = 3000$ litres of C remain.

After the breakdown, the 12 000 litres of C are reduced to 8000 litres, so that the first tableau becomes

	x_1	x_2	x_3	x_4	x_5	Solution	
z	-1.5	-1	0	0	0	0	Ratios
x_3	4	3	1	0	0	252	63
x_4	2	3	0	1	0	210	105
x_5	(1)	1	0	0	1	56	56

We note that we have a different pivot, and hence we expect a different solution. The next tableau is derived in the usual way, giving

	x_1	x_2	x_3	x_4	x_5	Solution
z	0	0.5	0	0	1.5	84
x_3	0	-1	1	0	-4	28
x_4	0	1	0	1	-2	98
x_1	1	1	0	0	1	56

The tableau is again optimal, so

$$x_1 = 56\,000, \quad x_2 = 0, \quad z = £84\,000$$

The profit is thus reduced by £10 500 by the breakdown, but still only plant P1 is used. The remaining amounts of A, B and C can be checked to be 4000, 14 000 and zero litres respectively.

Since this problem has only two variables, it would be instructive to check these results using the graphical method.

Example 10.6

Find the maximum of

$$z = 4x_1 + 2x_2 + 4x_3$$

subject to

$$3x_1 + x_2 + 2x_3 \leq 320$$

$$x_1 + x_2 + x_3 \leq 100$$

$$2x_1 + x_2 + 2x_3 \leq 200$$

Solution

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	-4	-2	-4	0	0	0	0
x_4	3	1	2	1	0	0	320
x_5	1	1	(1)	0	1	0	100
x_6	2	1	2	0	0	1	200

Note that in the above tableau there is some arbitrariness in the choices in both Steps 1 and 2. In Step 1 the column is chosen arbitrarily between the x_1 and x_3 columns. From the ratios, the x_5 row is selected from the x_5 and x_6 rows at Step 2, which both have equal ratios. Steps 3–5 are then followed to give the tableau

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	0	2	0	0	4	0	400
x_4	1	-1	0	1	-2	0	120
x_3	(1)	1	1	0	1	0	100
x_6	0	-1	0	0	-2	1	0

Although this is the optimal solution with $x_1 = x_2 = 0$, $x_3 = 100$ and $z = 400$, we have $c_1 = 0$ in the z row. Since x_1 is a non-basic variable, there is degeneracy. If we follow through the algorithm, choosing the first column at Step 1, we obtain an equally optimal solution in the following tableau. Replace x_3 by x_1 in the basic variables, and subtract the x_1 row from the x_4 row:

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	0	2	0	0	4	0	400
x_4	0	-2	-1	1	-3	0	20
x_1	1	1	1	0	1	0	100
x_6	0	-1	0	0	-2	1	0

This solution gives $x_1 = 100$, $x_2 = x_3 = 0$ and $z = 400$ once more. It can easily be deduced that $x_1 = 100(1 - \alpha)$, $x_2 = 0$, $x_3 = 100\alpha$ is an optimal solution for any $0 \leq \alpha \leq 1$ with $z = 400$. We could have observed this fact geometrically, since z is just a multiple of the left-hand side of the last constraint.

10.2.4 Exercises

- 1 Use the graphical method to find the maximum value of

$$f = 4x + 5y$$

subject to

$$3x + 7y \leq 10$$

$$2x + y \leq 3$$

$$x, y \geq 0$$

- 2 Sketch the constraints

$$2x - y \leq 6$$

$$x + 2y \leq 8$$

$$3x + 2y \leq 18$$

$$y \leq 3$$

and verify that the maximum of the function $x + y$ in the feasible region is at $x = 4$ and $y = 2$. Check the solution with the simplex method. Use a package such as MAPLE or MATLAB to verify the solution.

- 3 A manufacturer produces two types of cupboard, which are constructed from chipboard and oak veneer that both come in standard widths. The first type requires 4 m of chipboard and 5 m of oak veneer, takes 5 h of labour to produce and gives a profit of £24 per unit. The second type requires 5 m of chipboard and 2 m of oak veneer, takes 3 h of labour to produce and gives a £12 profit per unit.

On a weekly basis there are 400 m of chipboard available, 200 m of oak veneer and a maximum of 250 h of labour. Write this problem in a linear programming form. Use the simplex method to determine how many cupboards of each type should be made to maximize profits. How much profit is made? Which of the scarce resources remain unused? Show that the amount of oak veneer available can be reduced to 175 m without affecting the basis. What is the new solution, and by how much is the profit reduced?

- 4 A factory manufactures nails and screws. The profit yield is 2p per kg nails and 3p per kg screws. Three units of labour are required to manufacture 1 kg nails and 6 units to make 1 kg screws. Twenty-four units of labour are available. Two units of raw material are needed to make 1 kg nails and 1 unit for 1 kg screws. Determine the manufacturing policy that yields maximum profit from 10 units of raw material.

- 5 A manufacturer makes two types of cylinder, CYL1 and CYL2. Three materials, M1, M2 and M3, are required for the manufacture of each cylinder. The following information is provided:

	Quantities of materials required		
	M1	M2	M3
CYL1	1	1	2
CYL2	5	2	1

Quantities of materials available		
M1	M2	M3
45	21	24

£4 profit is made on one CYL1 and £3 profit on one CYL2. How many of each cylinder should the manufacturer make in order to maximize profit?

- 6 The Yorkshire Clothing Company makes two styles of jacket, the 'York' and the 'Wetherby'. The York requires 3 m of cloth and 3 h of labour, and makes a profit of £25. The Wetherby needs 4 m of cloth and 2 h of labour, and makes a profit of £30. The Yorkshire has 400 m of cloth available and 300 h of labour available each week. Advise the company on the number of each style it should produce in order to maximize profits.

The company is prepared to buy more cloth to increase its profits, but it will not employ any more labour. Under this revised policy, is there a strategy that will increase its profits?

- 7 Find the optimal solution of the following LP problem: maximize

$$z = kx_1 + 20x_2$$

subject to

$$x_1 + 2x_2 \leq 20$$

$$3x_1 + x_2 \leq 25$$

$$x_1, x_2 \geq 0$$

where k is a positive parameter representing variable profitability. Use both the simplex method and the graphical method, and interpret the results geometrically.

- 8 Use the simplex method to solve the following problem: maximize



$$2x_1 + x_2 + 4x_3 + x_4$$

subject to

$$2x_1 + x_3 \leq 3$$

$$x_1 + 3x_3 + x_4 \leq 4$$

$$4x_2 + x_3 + x_4 \leq 3$$

$$x_1, x_2, x_3, x_4 \geq 0$$

- 9 A publisher has three books available for printing, B1, B2 and B3. The books require varying amounts of paper, and the total paper supplies are limited:

	B1	B2	B3	Total units available
Units of paper required per 1000 copies	3	2	1	60
Profit per 1000 copies	£900	£800	£300	

Books B1 and B2 are similar in content, and the total combined market for these two books is estimated to be at most 15 000 copies. Determine how many copies of each book should be printed to maximize the overall profit.

- 10

Euroflight is considering the purchase of new aircraft. Long-range aircraft cost £4 million each, medium-range £2 million each and short-range £1 million each, and Euroflight has £60 million to invest. The estimated profit from each type of aircraft is £0.4 million, £0.3 million and £0.15 million respectively. The company has trained pilots for at most a total 25 aircraft. Maintenance facilities are limited to a maximum of the equivalent of 30 short-range aircraft. Long-range aircraft need twice as much maintenance as short-range ones, and medium-range 1.5 times as much. Set this up as a linear programming problem, and solve it. Aircraft can only be bought in integer numbers, so estimate how many of each type should be bought.

- 11

Find $x_1, x_2, x_3, x_4 \geq 0$ that maximize

$$f = 6x_1 + x_2 + 2x_3 + 4x_4$$

subject to

$$2x_1 + x_2 + x_4 \leq 3$$

$$x_1 + x_3 + x_4 \leq 4$$

$$x_1 + x_2 + 3x_3 + 2x_4 \leq 10$$

10.2.5 Two-phase method

The previous section only dealt with ' \leq ' constraints and did not consider ' \geq ' constraints. These prove to be much more troublesome, since there is no obvious initial feasible solution, and Phase 1 of the **two-phase method** is solely concerned with getting such a solution. Once this has been obtained, we then move to Phase 2. This is the standard simplex method, starting from the solution just obtained from Phase 1. A simple example will illustrate the problems involved and the basic ideas of the two-phase method.

Example 10.7

Find the maximum of

$$z = x + y$$

subject to

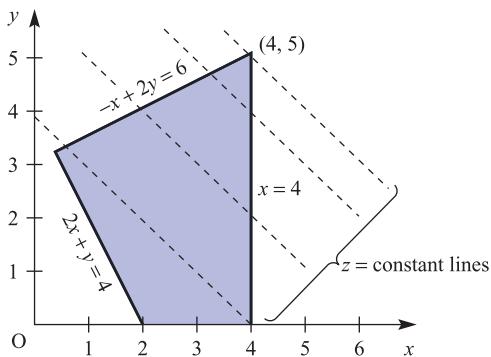
$$-x + 2y \leq 6$$

$$x \leq 4$$

$$2x + y \geq 4$$

$$x, y \geq 0$$

Figure 10.5
Feasible region of Example 10.7.



Solution The region defined by the constraints is shown in Figure 10.5. It is clear from the figure that the origin is *not* in the feasible region and that $x = 4, y = 5$ gives the optimal solution. We have already appreciated that the graphical method is only useful for two-dimensional problems, so we must explore how the simplex method copes with this problem.

Add in the positive variables r, s and t to give

$$\begin{array}{rcl} -x + 2y + r & = 6 \\ x & + s & = 4 \\ 2x + y & - t & = 4 \end{array}$$

The r and s are the usual slack variables. Because we must subtract t to take away the surplus, t is called a **surplus variable**. The obvious solution $x = y = 0, r = 6, s = 4, t = -4$ does *not* satisfy the condition that all variables be positive. The algebra is saying that the origin is not in the feasible region. Because the simplex method works so well, the last equation is forced into standard form by adding in yet another variable, u , called an **artificial variable**, to give

$$2x + y - t + u = 4$$

Now we have a feasible solution $x = y = t = 0, r = 6, s = 4, u = 4$, but not to the problem we originally stated. As the term ‘artificial variable’ implies, we wish to get rid of u and then reduce the problem back to our original one at a feasible corner. The variable u can be eliminated by forcing it to zero and this can be done by entering **Phase 1** with a new cost function

$$z' = -u$$

We see that if we can maximize z' then this is at $u = 0$, and our Phase 1 will be complete. The simplex tableau for Phase 1 then takes the form

	x	y	r	s	t	u	Solution
z	-1	-1	0	0	0	0	0
z'	0	0	0	0	0	1	0
r	-1	2	1	0	0	0	6
s	1	0	0	1	0	0	4
u	2	1	0	0	-1	1	4

where z has been included for the elimination but does *not* enter the optimization: only the z' row is considered in Phase 1. It may be observed that the tableau is not of standard form, since u is a basic variable and the (z', u) entry is non-zero. This must be remedied by subtracting the u row from the z' row to give the standard-form tableau

	x	y	r	s	t	u	Solution
z	-1	-1	0	0	0	0	0
z'	-2	-1	0	0	1	0	-4
r	-1	2	1	0	0	0	6
s	1	0	0	1	0	0	4
u	(2)	1	0	0	-1	1	4

Manipulation using the usual simplex algorithm gives the tableau

	x	y	r	s	t	u	Solution
z	0	$-\frac{1}{2}$	0	0	$-\frac{1}{2}$	$\frac{1}{2}$	2
z'	0	0	0	0	0	1	0
r	0	$\frac{5}{2}$	1	0	$-\frac{1}{2}$	$\frac{1}{2}$	8
s	0	$-\frac{1}{2}$	0	1	$\frac{1}{2}$	$-\frac{1}{2}$	2
x	1	$\frac{1}{2}$	0	0	$-\frac{1}{2}$	$\frac{1}{2}$	2

At this stage $z' = 0$ and $u = 0$, so that we have driven u out of the problem, and the z' row and u column can now be deleted. The Phase 1 solution gives $x = 2$, $y = 0$, which can be observed to be a corner of the feasible region in Figure 10.5.

We now enter **Phase 2**, with the z' row and u column deleted, and perform the usual sequence of steps. The initial tableau is

	x	y	r	s	t	Solution
z	0	$-\frac{1}{2}$	0	0	$-\frac{1}{2}$	2
r	0	($\frac{5}{2}$)	1	0	$-\frac{1}{2}$	8
s	0	$-\frac{1}{2}$	0	1	$\frac{1}{2}$	2
x	1	$\frac{1}{2}$	0	0	$-\frac{1}{2}$	2

Two further cycles are required, leading sequentially to the following two tableaux:

	x	y	r	s	t	Solution
z	0	0	$\frac{1}{5}$	0	$-\frac{3}{5}$	$\frac{18}{5}$
y	0	1	$\frac{2}{5}$	0	$-\frac{1}{5}$	$\frac{16}{5}$
s	0	0	$\frac{1}{5}$	1	($\frac{2}{5}$)	$\frac{18}{5}$
x	1	0	$-\frac{1}{5}$	0	$-\frac{2}{5}$	$\frac{2}{5}$

	x	y	r	s	t	Solution
z	0	0	$\frac{1}{2}$	$\frac{3}{2}$	0	9
y	0	1	$\frac{1}{2}$	$\frac{1}{2}$	0	5
t	0	0	$\frac{1}{2}$	$\frac{5}{2}$	1	9
x	1	0	0	1	0	4

We now have an optimum solution, since all the z row entries are non-negative with $x = 4$, $y = 5$ and objective function $z = 9$ in agreement with the graphical solution.

The general **two-phase strategy** is then as follows:

Phase 1

- (a) Introduce slack and surplus variables.
 - (b) Introduce artificial variables alongside the surplus variables, say x_p, \dots, x_q .
 - (c) Write the artificial cost function
- $$z' = -x_p - x_{p+1} - \dots - x_q$$
- (d) Subtract rows x_p, x_{p+1}, \dots, x_q from the cost function z' to ensure there are zeros in the entries in the z' row corresponding to the basic variables.
 - (e) Use the standard simplex method to maximize z' (keeping the z row as an extra row) until $z' = 0$ and

$$x_p = x_{p+1} = \dots = x_q = 0$$

Phase 2

- (a) Eliminate the z' row and artificial columns x_p, \dots, x_q .
- (b) Use the standard simplex method to maximize the objective function z .

There are other approaches to obtaining an initial feasible basic solution, but Phase 1 of the two-phase method gives an efficient way of obtaining a starting point. Geometrically, it uses the simplex method to search the non-feasible vertices until it is driven to a vertex in the feasible region.

Example 10.8

Use the two-phase method to solve the following LP problem: maximize

$$z = 4x_1 + \frac{1}{2}x_2 + x_3$$

subject to

$$x_1 + 2x_2 + 3x_3 \geq 2$$

$$2x_1 + x_2 + x_3 \leq 5$$

$$x_1, x_2, x_3 \geq 0$$

Solution**Phase 1**

Introduce a surplus variable x_4 and a corresponding artificial variable x_5 into the first inequality. A slack variable x_6 is required for the second inequality. The artificial cost is just

$$z' = -x_5$$

and we can construct the initial tableau

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	-4	$-\frac{1}{2}$	-1	0	0	0	0
z'	0	0	0	0	1	0	0
x_5	1	2	3	-1	1	0	2
x_6	2	1	1	0	0	1	5

We subtract the x_5 row from the z' row to eliminate the 1 from the (z', x_5) element, giving the tableau

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	-4	$-\frac{1}{2}$	-1	0	0	0	0
z'	-1	-2	-3	1	0	0	-2
x_5	1	2	(3)	-1	1	0	2
x_6	2	1	1	0	0	1	5

We now apply the steps of the simplex algorithm to give the tableau

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	$-\frac{11}{3}$	$\frac{1}{6}$	0	$-\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{2}{3}$
z'	0	0	0	0	1	0	0
x_3	$\frac{1}{3}$	$\frac{2}{3}$	1	$-\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{2}{3}$
x_6	$\frac{5}{3}$	$\frac{1}{3}$	0	$\frac{1}{3}$	$-\frac{1}{3}$	1	$\frac{13}{3}$

Since $z' = 0$ and the artificial variable x_5 has been driven into the non-basic variables, phase 1 ends.

Phase 2

The z' row and the x_5 column are now deleted, and the following sequence of tableaux constructed following the rules of the simplex algorithm:

	x_1	x_2	x_3	x_4	x_6	Solution
z	$-\frac{11}{3}$	$\frac{1}{6}$	0	$-\frac{1}{3}$	0	$\frac{2}{3}$
x_3	($\frac{1}{3}$)	$\frac{2}{3}$	1	$-\frac{1}{3}$	0	$\frac{2}{3}$
x_6	$\frac{5}{3}$	$\frac{1}{3}$	0	$\frac{1}{3}$	1	$\frac{13}{3}$

	x_1	x_2	x_3	x_4	x_6	Solution
z	0	$\frac{15}{2}$	11	-4	0	8
x_1	1	2	3	-1	0	2
x_6	0	-3	-5	(2)	1	1

	x_1	x_2	x_3	x_4	x_6	Solution
z	0	$\frac{3}{2}$	1	0	2	10
x_1	1	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{5}{2}$
x_4	0	$-\frac{3}{2}$	$-\frac{5}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$

The solution is now optimal, with $x_1 = \frac{5}{2}$, $x_2 = x_3 = 0$ and $z = 10$. Note that the first inequality is not binding, since $x_4 \neq 0$, while the second inequality is binding.



As indicated in the previous section a computer package such as MAPLE or MATLAB can deal easily with LP problems. For ' \geq ' inequalities the packages are equally efficient and will produce the answer, but there is no indication that the two-phase method has been used. For Example 10.8 the MAPLE code

```
with(simplex):
constr:={x1+2*x2+3*x3>=2, 2*x1+x2+x3<=5};
obj:= 4*x1+.5*x2+x3;
maximize(obj,constr,NONNEGATIVE);
```

produces the result $x_1 = 5/2$, $x_2 = 0$ and $x_3 = 0$ instantly.

The corresponding MATLAB code is

```
f=[-4;-0.5;-1];
A=[-1 -2 -3;2 1 1];
b=[-2,5];
Aeq=[];
beq=[];
lb=zeros(3,1);
ub=[];
x0=[];
options=optimset('LargeScale','off','Simplex','on');
[x,fval,exitflag,output,lambda]=linprog(f,A,b,Aeq,beq,lb,
                                         ub,x0,options)
```

Example 10.9

Three ores, A, B and C, are blended to form 100 kg of alloy; the percentage contents and the costs are as follows:

Ore	A	B	C
Iron	70	60	0
Lead	20	10	40
Copper	10	30	60
Cost (£ kg^{-1})	3000	2000	1000

The alloy must contain at least 20% iron, at least 25% lead but less than 48% copper. Find the blend of ores that minimizes the cost of the alloy.

Solution Let x_1 , x_2 and x_3 be the weights (kg) of ores A, B and C respectively in the 100 kg of alloy. The constraints give

$$\text{iron} \quad 0.7x_1 + 0.6x_2 \geq 20$$

$$\text{lead} \quad 0.2x_1 + 0.1x_2 + 0.4x_3 \geq 25$$

$$\text{copper} \quad 0.1x_1 + 0.3x_2 + 0.6x_3 \leq 48$$

and to make the 100 kg of alloy,

$$x_1 + x_2 + x_3 = 100$$

The cost is

$$3000x_1 + 2000x_2 + 1000x_3$$

which is to be *minimized*.

To reduce the problem to standard form, we change the problem to a *maximization* of

$$z = -3000x_1 - 2000x_2 - 1000x_3$$

For the inequality constraints we use surplus variables x_4 and x_5 and a slack variable x_6 . We require two artificial variables x_7 and x_8 alongside the surplus variables. Thus the inequalities become

$$0.7x_1 + 0.6x_2 - x_4 + x_7 = 20$$

$$0.2x_1 + 0.1x_2 + 0.4x_3 - x_5 + x_8 = 25$$

$$0.1x_1 + 0.3x_2 + 0.6x_3 + x_6 = 48$$

To deal with the equality constraint, we introduce a further artificial variable x_9 :

$$x_1 + x_2 + x_3 + x_9 = 100$$

We must drive x_9 to zero, to ensure that the equality holds, so it is essential to put x_9 into the artificial cost function. (Note that this is the standard way of dealing with an equality constraint.) We first enter *Phase 1*. Steps (a)–(c) of Phase 1 of the two-phase method give the initial tableau

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	Solution
z	3000	2000	1000	0	0	0	0	0	0	0
z'	0	0	0	0	0	0	1	1	1	0
x_7	0.7	0.6	0	-1	0	0	1	0	0	20
x_8	0.2	0.1	0.4	0	-1	0	0	1	0	25
x_6	0.1	0.3	0.6	0	0	1	0	0	0	48
x_9	1	1	1	0	0	0	0	0	1	100

It is necessary to remove the 1s from the z' row in the basic variable columns x_7 , x_8 and x_9 . Following (d) of the general strategy, we replace the z' row by $(z' \text{ row}) - (x_7 \text{ row}) - (x_8 \text{ row}) - (x_9 \text{ row})$ to give the tableau

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	Solution
z	3000	2000	1000	0	0	0	0	0	0	0
z'	-1.9	-1.7	-1.4	1	1	0	0	0	0	-145
x_7	0.7	0.6	0	-1	0	0	1	0	0	20
x_8	0.2	0.1	0.4	0	-1	0	0	1	0	25
x_6	0.1	0.3	0.6	0	0	1	0	0	0	48
x_9	1	1	1	0	0	0	0	0	1	100

Several tableaux need to be completed to drive z' to zero and complete Phase 1, with the final tableau being

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	Solution
z	0	-2000	0	0	-10 000	0	0	10 000	-1000	-250 000
z'	0	0	0	0	0	0	1	1	1	0
x_1	1	1.5	0	0	5	0	0	-5	2	75
x_3	0	-0.5	1	0	-5	0	0	5	-1	25
x_6	0	0.45	0	0	2.5	1	0	-2.5	0.4	25.5
x_4	0	0.45	0	1	3.5	0	-1	-3.5	1.4	32.5

Removing the artificial variables and the z' row gives the tableau

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	0	-2000	0	0	-10 000	0	-250 000
x_1	1	1.5	0	0	5	0	75
x_3	0	-0.5	1	0	-5	0	25
x_6	0	0.45	0	0	2.5	1	25.5
x_4	0	0.45	0	1	3.5	0	32.5

The algorithm is now ready for *Phase 2*, since a sensible feasible basic solution is available. The standard procedure leads, after many cycles, to the final tableau

	x_1	x_2	x_3	x_4	x_5	x_6	Solution
z	333.3	0	0	0	0	3333.3	-140 000
x_4	0.3	0	0	1	0	-2	4
x_3	-0.67	0	1	0	0	3.33	60
x_2	1.67	1	0	0	0	-3.33	40
x_5	-0.3	0	0	0	1	1	3

and the solution can be read off as

$$x_1 = 0, \quad x_2 = 40, \quad x_3 = 60$$

with the cost minimized at £140 000. (Note that the cost in the tableau is negative, since the original problem is a minimization problem.) It may be noted that $x_6 = 0$, so the copper constraint is binding while the iron constraint gives 4% more than the required minimum and the lead constraint gives 3% more than the required minimum.

Except for simple, illustrative examples, the amount of computational work in the two-phase strategy is heavy and requires the use of a computer package. Even for the comparatively simple Example 10.9, many tableaux were required in the solution. The computer packages MAPLE and MATLAB have no difficulty in dealing with the equality constraint that appears in the problem.



10.2.6 Equality constraints and variables that are unrestricted in sign

An equality constraint can be written as two constraints: one ' \leq ' and one ' \geq ' constraint. For example, the equation $6x_1 - x_2 = 10$ can be expressed as

$$6x_1 - x_2 \geq 10$$

$$6x_1 - x_2 \leq 10$$

All the problems considered so far have demanded the decision variables be non-negative. In many applications the decision variables can take any real value. That is, they are **unrestricted in sign (URS)**. To solve such problems with the simplex method the URS variables are reformulated as the difference between two non-negative variables. So if $-\infty < x_i < \infty$ then we write $x_i = x'_i - x''_i$, where $x'_i, x''_i \geq 0$. If $x'_i > x''_i$ then $x_i > 0$ and if $x'_i < x''_i$ then $x_i < 0$.

Example 10.10

Use the simplex method to solve

$$\text{Max } z = 4x_1 + 2x_2$$

subject to

$$2x_1 + x_2 \leq 7$$

$$x_1 + x_2 \leq 3, x_1 \geq 0, -\infty < x_2 < \infty$$

Solution

Decision variable x_2 is URS so we let $x_2 = x'_2 - x''_2$, where the new variables are non-negative and substitute for x_2 into the linear program to obtain

$$\text{Max } z = 4x_1 + 2x'_2 - 2x''_2$$

subject to

$$2x_1 + x'_2 - x''_2 \leq 7$$

$$x_1 + x'_2 - x''_2 \leq 3$$

$$x_1, x'_2, x''_2 \geq 0$$

We can now use the simplex algorithm in the usual way. The initial tableau is (with slack variables r and s for constraints 1 and 2, respectively)

	x_1	x'_2	x''_2	r	s	Solution	
z	-4	-2	2	0	0	0	Ratios
r	2	1	-1	1	0	7	$\frac{7}{2} = 3.5$
s	(1)	1	-1	0	1	3	3

Proceeding as discussed in Section 10.2.3, the next two iterations yield

	x_1	x'_2	x''_2	r	s	Solution	
z	0	2	-2	0	4	12	Ratios
r	0	-1	(1)	1	-2	1	1
x_1	1	1	-1	0	1	3	-3

	x_1	x'_2	x''_2	r	s	Solution
z	0	0	0	2	0	14
x''_2	0	-1	1	1	-2	1
x_1	1	0	0	1	-1	4

At this point the algorithm terminates because all entries in the z row are non-negative. Reading the final table tells us that the maximum value of z is 14 and this occurs when $x_1 = 4$ and $x''_2 = 1$. It is important to give a solution in terms of the variables of the problem, that is x_1 and x_2 . Since $x_2 = x'_2 - x''_2$ and x'_2 is a non-basic variable (equal to zero), we say

$$\text{Max } z = 14 \text{ when } x_1 = 4 \text{ and } x_2 = -1$$

One may ask how to handle linear programming problems that restrict the decision variables to the integers. This is an important question but integer programming is significantly more challenging than standard linear programming and beyond the scope of this book. Suffice to say that most linear programming software packages have in-built routines for integer variable (for example, `intcon` and `intlinprog` in MATLAB).

10.2.7 Exercises

- 12 Use the graphical approach to solve the LP problem
 $\max(x + 2y)$

subject to the constraints

$$1 \leq y \leq 4$$

$$x + y \leq 5$$

Check your solution using the two-phase method and by using a MAPLE or MATLAB implementation.

- 13 Use the simplex method to find positive values of x_1 and x_2 that minimize
 $f = 10x_1 + x_2$

subject to

$$4x_1 + x_2 \leq 32$$

$$2x_1 + x_2 \geq 12$$

$$2x_1 - x_2 \leq 4$$

$$-2x_1 + x_2 \leq 8$$

Sketch the points obtained by the simplex method on a graph, indicating how the points progress through Phases 1 and 2 to the solution.

14

The Footsie company produces boots and shoes. If no boots are made, the company can produce a maximum of 250 pairs of shoes in a day. Each pair of boots takes twice as long to make as each pair of shoes. The maximum daily sales of boots and shoes are 200, but 25 pairs of boots must be produced to satisfy an important customer. The profits per pair of boots and shoes are £8 and £5 respectively. Determine the daily production plan to maximize profits. Use the two-phase method to obtain the solution, and verify your result with a graphical solution.

15

In Exercise 9 there is an additional union agreement that at least 50 000 books must be printed. Does the solution change? If so, calculate the new optimum strategy.

16

Solve the LP problem



$$\max(x + y + z)$$

subject to the constraints

$$x \geq 1$$

$$x + 2y \leq 3$$

$$y + 3z \leq 4$$

by using the two-phase method. Check your result using MAPLE or MATLAB.

17

Solve the following LP problem: minimize



$$2x_1 + 7x_2 + 4x_3 + 5x_4$$

subject to

$$x_1 - x_3 - x_4 \geq 0$$

$$x_2 + x_3 \geq 2$$

$$x_1, x_2, x_3, x_4 \geq 0$$

18



A trucking company requires antifreeze that contains at least 50% of pure glycol and at least 5% of anticorrosive additive. The company can buy three products, A, B and C, whose constituents and costs are as follows:

	A	B	C
% glycol	65	25	80
% additive	10	3	0
Cost (£/litre)	1.8	0.9	1.5

What blend will provide the required antifreeze solution at minimum cost? What is the cost of 100 litres of solution?

19



A builder is constructing three different styles of house on an estate, and is deciding which styles to erect in the next phase of building. There are 40 plots of equal size, and the different styles require 1, 2 and 2 plots respectively. The builder anticipates shortages of two materials, and estimates the requirements and supplies (in appropriate units) to be as follows:

	Requirements			Total supply
	Style 1	Style 2	Style 3	
Facing stone	1	2	5	58
Weather boarding	3	2	1	72

The local authority insists that there be at least 5 more houses of style 2 than style 1. If the profits on the houses are £1000, £1500 and £2500 for styles 1, 2 and 3 respectively, find how many of each style the builder should construct to maximize the total profit.

20



A manufacturer produces three types of carpeting, C1, C2 and C3. Two of the raw materials, M1 and M2, are in short supply. The following table gives the supplies of M1 and M2 available (in 1000s of kg), the quantities of M1 and M2 required for each 1000 m² of carpet, and the profits made from each type of carpet (in £1000s per 1000 m²):

	Quantities required		
	M1	M2	Profit
C1	1	1	2
C2	1	1	3
C3	1	0	0
Quantities available	5	4	

Carpet of type C3 is non-profit-making, but is included in the range in order to enable the

company to satisfy its customers. The company has policies that require that, if x_1 , x_2 and x_3 1000s of m² of C1, C2 and C3 respectively are made then

$$x_1 \geqslant 1$$

and

$$x_1 - x_2 + x_3 \geqslant 2$$

How much carpet of each type should the company manufacture in order to satisfy the constraints and maximize profits?

10.3 Lagrange multipliers

10.3.1 Equality constraints

In Section 10.2 we looked at the situation where all the functions were linear. As soon as functions become nonlinear, the problems become very much more difficult. This is generally the case in most of mathematics, and is certainly true in optimization problems.

In Section 9.7.4 of MEM it was shown how to use Lagrange multipliers to solve the problem of the optimization of a nonlinear function of many variables subject to **equality constraints**. For the general problem it was shown that the necessary conditions for the extremum of

$$f(x_1, x_2, \dots, x_n)$$

subject to

$$g_i(x_1, x_2, \dots, x_n) = 0 \quad (i = 1, 2, \dots, m \ (m < n))$$

are

$$\frac{\partial f}{\partial x_k} + \lambda_1 \frac{\partial g_1}{\partial x_k} + \lambda_2 \frac{\partial g_2}{\partial x_k} + \dots + \lambda_m \frac{\partial g_m}{\partial x_k} = 0 \quad (k = 1, 2, \dots, n) \quad (10.7)$$

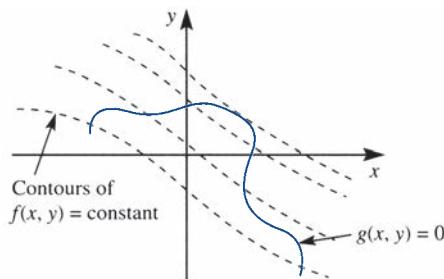
where $\lambda_1, \lambda_2, \dots, \lambda_m$ are the **Lagrange multipliers**. These n equations must be solved together with the m constraints $g_i = 0$. Thus there are $n + m$ equations in the $n + m$ unknowns x_1, x_2, \dots, x_n and $\lambda_1, \lambda_2, \dots, \lambda_m$.

For a two-variable problem of finding the extremum of

$$f(x, y) \text{ subject to the single constraint } g(x, y) = 0$$

the problem is illustrated geometrically in Figure 10.6. We are looking for the maximum, say, of the function $f(x, y)$, but *only* considering those points in the plane that lie on the curve $g(x, y) = 0$. The mathematical conditions look much simpler of course, as

Figure 10.6 Lagrange multiplier problem.



$$\left. \begin{aligned} f_x + \lambda g_x &= 0 \\ f_y + \lambda g_y &= 0 \\ g &= 0 \end{aligned} \right\} \quad (10.8)$$

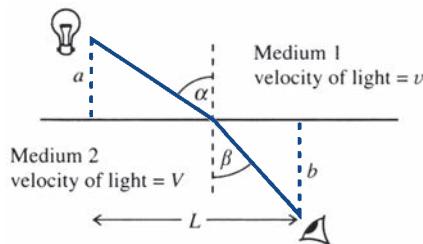
There are two comments that should be noted. First, the method fails if $g_x = g_y = 0$ at the solution point. Such points are called **singular points**: fortunately they are rare, but their existence should be noted. Secondly, sufficient conditions for a maximum, a minimum or a saddle point can be derived, but they are difficult to apply and not very useful. Example 10.13 will illustrate an intuitive approach to sufficiency.

A few examples should be enough to remind the reader of the problems involved and to show the techniques required to solve the equations.

Example 10.11

Fermat stated in 1661 that ‘Light travels along the shortest path’. Find the path that joins the eye to an object when they are in separate media (Figure 10.7).

Figure 10.7 Fermat's shortest-path problem.



Solution

The velocities of light in the two media are v and V . The time of transit of light is given by the geometry as

$$T = \frac{a}{v \cos \alpha} + \frac{b}{V \cos \beta}$$

and is then subject to the geometrical constraint that

$$L = a \tan \alpha + b \tan \beta$$

Applying (10.8),

$$0 = \frac{\partial T}{\partial \alpha} + \lambda \frac{\partial g}{\partial \alpha} = \frac{a}{v} \sec \alpha \tan \alpha + \lambda a \sec^2 \alpha$$

$$0 = \frac{\partial T}{\partial \beta} + \lambda \frac{\partial g}{\partial \beta} = \frac{b}{V} \sec \beta \tan \beta + \lambda b \sec^2 \beta$$

These give as the only solution

$$\sin \alpha = -\lambda v, \quad \sin \beta = -\lambda V$$

or

$$\frac{\sin \alpha}{\sin \beta} = \frac{v}{V} = \mu$$

which is known as **Snell's law**.

In Example 10.11, to obtain Snell's law, the solution of the equation was quite straightforward, but it is rarely so easy. Frequently it is technically the most difficult task, and it is easy to miss solutions. We return to the 'milk carton' problem discussed earlier to illustrate the point.

Example 10.12

Find the minimum area of the milk carton problem stated in Example 10.2 and illustrated in Figure 10.1.

Solution

Taking measurements in millimetres, the basic mathematical problem is to minimize

$$A = (2b + 2w + 5)(h + b + 10)$$

subject to

$$hbw = 1136000$$

Applying the Lagrange multiplier equations (10.7), we obtain

$$0 = (2b + 2w + 5) + \lambda bw$$

$$0 = (2b + 2w + 5) + 2(h + b + 10) + \lambda hw$$

$$0 = 2(h + b + 10) + \lambda hb$$

giving four equations in the four unknowns h , b , w and λ . If we eliminate λ and w from these equations, we are left with the same equations that we derived in Example 10.2, which have no simple analytical solutions.

The only way to proceed further with Example 10.12 is by a numerical solution. Thus, even with simple problems such as this one, we need a numerical algorithm, and in most realistic problems in science and engineering we encounter similar severe computational difficulties. It is often a problem even to write down the function or the constraints explicitly. Such functions frequently emerge as numbers from a complicated computer program. It is therefore essential to look for efficient numerical algorithms to optimize such functions. This will be the substance of Section 10.4.

A final example shows the situation where there are more than two variables involved. An indication will be given of the difficulties of *proving* that the point obtained is a maximum.

Example 10.13

A hopper is to be made from a cylindrical portion connected to a conical portion as indicated in Figure 10.8. It is required to find the maximum volume subject to a given surface area.

Solution

We can compute the volume of the hopper as

$$V = \pi R^2 L + \frac{1}{3} \pi R^3 \tan \alpha$$

and find its maximum subject to the surface area being given as

$$A = 2\pi RL + \pi R^2 \sec \alpha$$

Applying (10.7) with the appropriate variables, we obtain

$$0 = \frac{\partial V}{\partial R} + \lambda \frac{\partial g}{\partial R} = 2\pi RL + \pi R^2 \tan \alpha + \lambda(2\pi L + 2\pi R \sec \alpha)$$

$$0 = \frac{\partial V}{\partial L} + \lambda \frac{\partial g}{\partial L} = \pi R^2 + \lambda 2\pi R$$

$$0 = \frac{\partial V}{\partial \alpha} + \lambda \frac{\partial g}{\partial \alpha} = \frac{1}{3} \pi R^3 \sec^2 \alpha + \lambda \pi R^2 \sec \alpha \tan \alpha$$

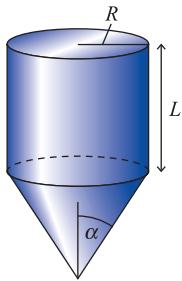


Figure 10.8 Hopper of Example 10.13.

First, λ can be easily evaluated as $\lambda = -\frac{1}{2}R$. The last of the above three equations becomes

$$\pi R^2 \sec^2 \alpha \left(\frac{1}{3} R + \lambda \sin \alpha \right) = 0$$

and hence $\sin \alpha = \frac{2}{3}$. Since $0 \leq \alpha \leq \frac{1}{2}\pi$, we have $\alpha = 0.730$ rad or 41.8° . A little further algebraic manipulation between the constraints and the above equations gives $R^2 = A/\pi\sqrt{5}$, so that $R = 0.377A^{1/2}$ and $V = 0.126A^{3/2}$.

Normally it would be assumed that this is a maximum for the volume on intuitive or geometrical grounds. To prove this rigorously, we take small variations around the suspected maximum and show that the volume is larger than all its neighbours. For simplicity let $R^2 = (A/\pi\sqrt{5})(1 + \delta)$, $\sec \alpha = (3/\sqrt{5})(1 + \varepsilon)$, evaluate L from the constraint $g = 0$ and then calculate V to second order in ε and δ by Taylor's theorem. Some careful algebra gives

$$V = \frac{A^{3/2}}{3\pi^{1/2} 5^{1/4}} \left(1 - \frac{1}{8} \delta^2 - \frac{9}{16} \varepsilon^2 \right)$$

This shows that for any non-zero values of ε and δ we obtain a smaller volume, and hence we have proved that we have found a maximum.

It should be reiterated that the major problem lies in solving the Lagrange multiplier equations and not in writing them down. This is typical, and supports the need for good numerical algorithms to solve such problems; they only have to be marginally more difficult than Example 10.13 to become impossible to manipulate analytically.

10.3.2 Inequality constraints

Although we do not intend to consider them in any detail here, for reference we shall state the basic extension of (10.7) to the case of **inequality constraints**. Kuhn and Tucker proved the following result in the 1940s.

To maximize the function

$$f(x_1, \dots, x_n)$$

subject to

$$g_i(x_1, \dots, x_n) \leq 0 \quad (i = 1, \dots, m)$$

the equivalent conditions to (10.7) are

$$\begin{aligned} \frac{\partial f}{\partial x_k} - \lambda_1 \frac{\partial g_1}{\partial x_k} - \dots - \lambda_m \frac{\partial g_m}{\partial x_k} &= 0 \quad (k = 1, \dots, n) \\ \left. \begin{array}{l} \lambda_i g_i = 0 \\ \lambda_i \geq 0 \\ g_i \leq 0 \end{array} \right\} &\quad (i = 1, \dots, m) \end{aligned}$$

The equation $\lambda_i g_i = 0$ gives two alternative conclusions for each constraint, either

- (a) $g_i = 0$, in which case the constraint is ‘active’ and the corresponding $\lambda_i > 0$, or
- (b) $\lambda_i = 0$ and $g_i < 0$, so that the optimum is away from this constraint and the Lagrange multiplier is not necessary.

Implementation of the Kuhn–Tucker result is not very easy, even though in principle it looks straightforward. There are so many cases to check that it becomes very susceptible to error.

10.3.3 Exercises

- 21 Find the optimum of $f = x^2 + xy + y^2$ subject to the constraint $x + y = 1$ using Lagrange multipliers.
Show that the optimum is a minimum.

- 22 Find the shortest distance from the origin to the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (a < b)$$

- 23 Determine the lengths of the sides of a rectangle with maximum area that can be inscribed within the ellipse

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

- 24 Find the optimum of $f = xy^2z$ subject to $x + 2y + 3z = 6$ using a Lagrange multiplier method.

- 25 Show that the stationary points of $f = x^2 + y^2 + z^2$ subject to $x + y - z = 0$ and $yz + 2zx - 2xy = 1$ are given by the solution of the equations

$$0 = 2x + \lambda + (2z - 2y)\mu$$

$$0 = 2y + \lambda + (z - 2x)\mu$$

$$0 = 2z - \lambda + (y + 2x)\mu$$

Add the last of these two equations, and show that either $\mu = -2$ or $y + z = 0$. Hence deduce the stationary points.

26 A rectangular box without a lid is to be made. It is required to maximize the volume for a given surface area. Find the dimensions of the box when the total surface area is A .

- 27 (Harder) The lowest frequency of vibration, α , of an elastic plate can be computed by minimizing

$$I[\omega] = \iint_R (\nabla^2 \omega)^2 dx dy$$

subject to

$$\iint_R \omega^2 dx dy = 1$$

over all functions $\omega(x, y)$, where R is the region of the plate in the (x, y) plane. If R is the square region $|x| \leq 1, |y| \leq 1$ and the plate is clamped at its edges, use the approximation

$$\omega = A \cos^2 \frac{1}{2} \pi x \cos^2 \frac{1}{2} \pi y$$

to show that $I[\omega_{\min}] = \alpha^2 = \frac{8}{9} \pi^4$.

Use the improved approximation

$$\omega = \cos^2 \frac{1}{2} \pi x \cos^2 \frac{1}{2} \pi y (A + B \cos \frac{1}{2} \pi x \cos \frac{1}{2} \pi y)$$

to get a better estimate of α .

(Note: $\int_{-1}^1 \cos^{2n} \frac{1}{2} \pi z dz = (2n)!/(n!)^2 2^{2n-1}$ for non-negative integers n . Preferably use an algebraic symbolic manipulator, for example MAPLE, to evaluate the differentials and integrals.)

28

Use the Kuhn–Tucker criteria to find the minimum of

$$2x_1^2 + x_2^2 + 2x_1 x_2$$

subject to

$$x_1 - x_2 \leq \alpha$$

where α is a parameter. Find the critical value of α at which the nature of the solution changes. Sketch the situation geometrically to illustrate the change.

10.4 Hill climbing

10.4.1 Single-variable search

Most practical problems give calculations that cannot be performed explicitly, and need a numerical technique. Typical cases are those in Examples 10.1 and 10.12, where the final equations cannot be solved analytically and we need to resort to numerical methods. This is not an uncommon situation, and **hill climbing methods** were devised to cope with just such problems.

In many engineering problems the functions that we are trying to optimize cannot be written down explicitly. Take for example a vibrational problem where the frequencies of vibration are calculated from an eigenvalue problem. These frequencies will depend on the parameters of the physical system, and it may be necessary to make the largest frequency as low as possible. To illustrate this idea, suppose that the eigenvalues come from the equation

$$\begin{vmatrix} a - \lambda & -1 & 0 \\ -1 & -\lambda & -1 \\ 0 & -1 & a^2 - \lambda \end{vmatrix} = 0$$

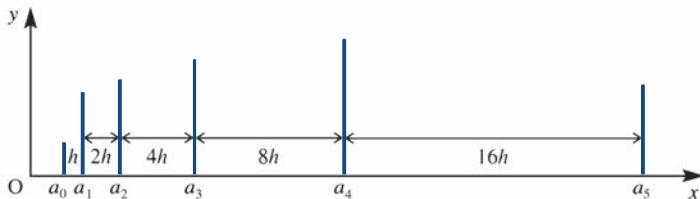
where there is just one parameter a . In this case the mathematical problem is to find $\min_a(\lambda_{\max})$. We note that there is no explicit formula for λ_{\max} as a simple function of a ,

and our function is the result of a solution of the determinantal equation. For some values of a the function can be calculated easily, $\lambda_{\max}(-1) = \sqrt{3}$, $\lambda_{\max}(0) = \sqrt{2}$ and $\lambda_{\max}(1) = 2$, but any other value requires a considerable amount of work. However, a computer package such as MATLAB will perform this hard work very quickly with the instruction



Figure 10.9

Bracketing procedure.



`lambda = max(eig([a -1 0; -1 0 -1; 0 -1 a^2])).` Calculating the derivative of the function with respect to a is too difficult even to contemplate.

In the determinant example we have a function of a single variable, but in most problems there are many variables. One of the commonest methods of attack is to obtain the maximum or minimum as a *sequence of single-variable searches*. We choose a direction and search in this direction until we have found the optimum of the function in the chosen direction. We then select a new direction and repeat the process. For this to be a successful method, we need to be able to perform single-variable searches very efficiently. This section therefore deals with single-variable problems, and only then in Section 10.4.3 are multivariable techniques discussed. In deciding on a strategy for solution, one crucial point is whether derivatives can or cannot be calculated. In the eigenvalue problem, calculation of the derivative is difficult, and would probably not be attempted. If the derivative can be obtained, however, more information is available, and any numerical method can be speeded up considerably. With the increase in sophistication of computers, this is becoming a less important consideration, since a good numerical approximation to the derivative is usually quite satisfactory. This is certainly the case in the MATLAB routine `fminunc`.

The basic problem is to determine the maximum of a function $y=f(x)$ that is difficult to evaluate and for which the derivative may or may not be available. The task is performed in two stages: in Stage 1 we **bracket** the maximum by obtaining x_1 and x_2 such that $x_1 \leq x_{\max} \leq x_2$ as described in Figures 10.9 and 10.10, and in Stage 2 we devise a method that iterates to the maximum to any desired accuracy, as in Figures 10.11 and 10.12.

Figure 10.10MATLAB code for obtaining a bracket for the maximum of $f(x)$.

(a) {Derivative not known}

If the function is defined by the anonymous function
`f=@(x)`

then the following code

```
aold=-1;h=.01;nmax=10; % aold, h, nmax are specified at values appropriate to the problem
zold=f(aold);a=aold+h;h=2*h;z=f(a);n=0;% step 1
while (z>zold)&(n<nmax)
    n=n+1;h=2*h;aoldold=aold;aold=a;a=a+h;zoldold=zold;zold=z;z=f(a); % subsequent steps
end
```

provides the bracket `[aoldold,aold,a]`.

(b) {Derivative known}

If the function and its derivative are defined by the anonymous functions
`f=@(x)`
`fdash=@(x)`

then the following code

```
a=.01;h=.01;nmax=10;
zold=f(a);zdashold=fdash(a);aold=a;a=a+h;z=f(a);zdash=fdash(a);h=2*h;n=0;
while (zdash>0) & (n<nmax)
    n=n+1;zold=z;zdashold=zdash;aold=a;z=f(a);zdash=fdash(a);a=a+h;h=2*h;
end
```

provides the bracket `[aold,a]`.

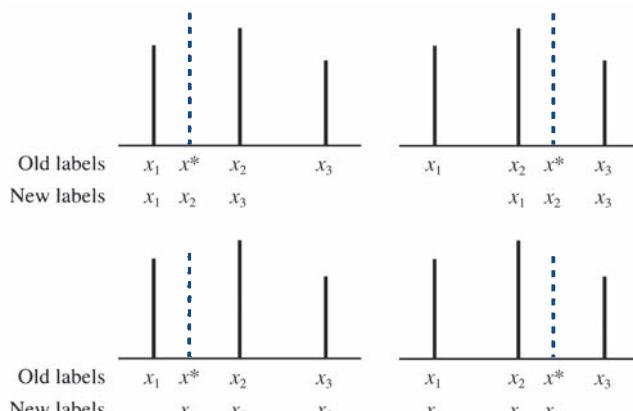
Figure 10.11

(a) MATLAB file *qapp.m* for the quadratic approximation algorithm; the function segment for $f(x)$ is declared in the file *fnn.m*; (b) diagrams corresponding to the four cases considered in the program.



```
function [x,f]=qapp(a,b)
% a=[a1,a2,a3] is the input vector of three points from bracketing
% b=[f(a1),f(a2),f(a3)] is the vector of function values
x=a;f=b;p=polyfit(a,b,2);
xstar=-0.5*p(2)/p(1);fstar=fnn(xstar);
if fstar>b(2)
    if xstar<a(2), x(3)=a(2);f(3)=b(2);
    else x(1)=a(2);f(1)=b(2); end
    x(2)=xstar;f(2)=fstar;
else
    if xstar<a(2), x(1)=xstar;f(1)=fstar;
    else x(3)=xstar;f(3)=fstar; end
end
% x contains the three points of the new bracket and f the function values
```

(a)



(b)

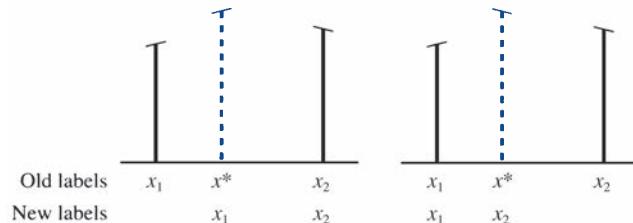
Figure 10.12

(a) MATLAB file *cufit.m* for the cubic approximation algorithm for the maximum of $f(x)$; x and the function segments $f(x)$ and $f'dash(x)$ are declared in the file *cub.m*; (b) diagrams corresponding to the two cases considered in the program.



```
function [an,bn]=cufit(a,b)
% a=[x1 f(x1) fdash(x1)] and b=[x2 f(x2) fdash(x2)] are the input vectors
v=[a(2);b(2);a(3);b(3)];
A=[a(1)^3 a(1)^2 a(1) 1;b(1)^3 b(1)^2 b(1) 1;3*a(1)^2 2*a(1) 1 0;3*b(1)^2 2*b(1) 1 0];
p=A\w;xstar=(-p(2)-sqrt(p(2)^2-3*p(1)*p(3)))/(3*p(1));c=cub(xstar);
if c(3)>0 an=c;bn=b; else bn=c;an=a;end
% an and bn contain the new bracket vectors
```

(a)



(b)

A bracket is most quickly achieved by starting at a given point, taking a step in the increasing direction and proceeding in this direction, doubling the step length at each step until the bracket is obtained (Figure 10.9). If the derivative $f'(x)$ of the function is known and positive when evaluated at x_1 and negative at the next point, x_2 , in the search then $x_1 < x_{\max} < x_2$. If we do not have the derivative of $f(x)$ then we need three points to bracket the maximum of the function. We can say that if the function increases in value between x_1 and x_2 , and then decreases between x_2 and x_3 then $x_1 < x_{\max} < x_3$ and $x_{\max} \in [x_1, x_2, x_3]$. The bracket $[x_1, x_2, x_3]$ is called a **bracketing triple**. The basic idea is summarized in Figure 10.10, which gives a MATLAB procedure for this technique. It is written as a ‘stand-alone’ segment and can be adapted for other packages, but it would normally be incorporated in a more general program. It is assumed that sensible values for a and h have been chosen, but in a working program great care has to be taken. A great deal of effort is required to cope with inappropriate choices and to prevent the program aborting.

The algorithm works very efficiently provided that appropriate safeguards are included, but it is not foolproof. The maximum number of steps chosen, $nmax$, is usually 10, and the initial value of h is small compared with the overall dimension of the problem under consideration.

Example 10.14

Find a bracket for the first maximum of $f(x) = x \sin x$ using the algorithm in Figure 10.10.

Solution

Choose $a = 0.01$ and $h = 0.01$; the algorithm then gives

a	0.01	0.02	0.04	0.08	0.16	0.32	0.64	1.28	2.56	5.12
f	0.000	0.000	0.002	0.006	0.025	0.101	0.382	1.226	1.406	-4.701
f'	0.020	0.040	0.080	0.160	0.317	0.618	1.111	1.325	-1.590	-

If the derivative is *not* used then the bracket is $1.28 \leq x \leq 5.12$.

If the derivative is used then the bracket is $1.28 \leq x \leq 2.56$.

Stage 2 of the calculation is to use the bracket just obtained and then iterate to an accurate maximum. A simple and efficient approach is to use a **polynomial approximation** to estimate the maximum and then choose the ‘best’ points to repeat the calculation. Another, not discussed here, is the **Golden search** algorithm.

If it is assumed that no derivative is available then a bracket is known from the algorithm of Figure 10.10, so that x_1, x_2, x_3 and the corresponding f_1, f_2, f_3 , with $f_1 < f_2$ and $f_3 < f_2$, are given. The **quadratic polynomial** through these points can be written down immediately: it is just the **Lagrange interpolation formula** that was discussed in Section 2.3 of MEM. It can easily be checked that the quadratic which passes through the required points is given by

$$f \approx F = \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} f_3 + \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} f_1 + \frac{(x - x_3)(x - x_1)}{(x_2 - x_3)(x_2 - x_1)} f_2 \quad (10.9)$$

Then $F' = 0$ at the point x^* , which is given, after a little algebra, by

$$x^* = \frac{(x_2^2 - x_3^2)f_1 + (x_3^2 - x_1^2)f_2 + (x_1^2 - x_2^2)f_3}{2[(x_2 - x_3)f_1 + (x_3 - x_1)f_2 + (x_1 - x_2)f_3]} \quad (10.10)$$



A MATLAB procedure for the algorithm that uses this new x^* and f^* is given in Figure 10.11, where the *best* three values are chosen for the next iteration. The code can be easily adapted for other packages. It can be re-written in terms of anonymous functions as in Figure 10.10 but there is great merit in breaking the program into small units that can be checked independently; these are *M-files* in MATLAB as in Figure 10.11.

The method works exceptionally well by repeating the instruction $[a, b] = qapp(a, b)$, but again it is not totally foolproof, and remedial checks need to be put into a working program. The stopping criterion is very problem-dependent, and requires thought and numerical experimentation. The MATLAB procedure `fminbnd` uses the quadratic approximation method (for the minimum problem). It only requires a bound for the minimum and uses another method, the Golden section (see Exercise 34), to obtain three starting values. It then proceeds similarly to the current algorithm. The two lines of code below solve the eigenvalue problem posed at the start of this section

```
options=optimset('display','iter');
[x,fval]=fminbnd('max(eig([x -1 0;-1 0 -1;0 -1 x\2])),  
-1,1,options)
```

It is worth looking at the full code of the MATLAB procedure for `fminbnd` to appreciate the enormous effort required to automate the problem fully, to deal with errors and failures and to make the program ‘user friendly’.

Example 10.15

Find the first maximum of $f(x) = x \sin x$ given the values from Example 10.14, namely $x_1 = 1.28$, $x_2 = 2.56$, $x_3 = 5.12$, $f_1 = 1.226$, $f_2 = 1.406$ and $f_3 = -4.701$.

Solution

From (10.10), $x^* = 2.03$ and $f^* = 1.820$, so for the next iteration choose

$$\begin{aligned} x_1 &= 1.28, & x_2 &= 2.03, & x_3 &= 2.56 \\ f_1 &= 1.226, & f_2 &= 1.820, & f_3 &= 1.406 \end{aligned}$$

From (10.10), $x^* = 1.98$ and $f^* = 1.816$, so for the next iteration choose

$$\begin{aligned} x_1 &= 1.98, & x_2 &= 2.03, & x_3 &= 2.56 \\ f_1 &= 1.816, & f_2 &= 1.820, & f_3 &= 1.406 \end{aligned}$$

From (10.10), $x^* = 2.027$ and $f^* = 1.820$, so the method has almost converged.

When the *derivative* is available, a better approximating polynomial than (10.9) can be used, since $x_1, f_1, f'_1 (> 0)$, x_2, f_2 , and $f'_2 (< 0)$ are known from the bracketing algorithm, and this data can be fitted to a *cubic polynomial*

$$f \simeq F = ax^3 + bx^2 + cx + d$$

$$F' = 3ax^2 + 2bx + c$$

In this case fitting the values just gives the matrix equation

$$\begin{bmatrix} f_1 \\ f_2 \\ f'_1 \\ f'_2 \end{bmatrix} = \begin{bmatrix} x_1^3 & x_1^2 & x_1 & 1 \\ x_2^3 & x_2^2 & x_2 & 1 \\ 3x_1^2 & 2x_1 & 1 & 0 \\ 3x_2^2 & 2x_2 & 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} \quad (10.11)$$

and the maximum of F is given by $F' = 0$, so that

$$x^* = \frac{-b \pm (b^2 - 3ac)^{1/2}}{3a} \quad (10.12)$$

and the negative sign is chosen (the positive sign for a minimum problem) to ensure that $x_1 < x^* < x_2$.

A simple algorithm uses these results to choose the appropriate bracket for the next iteration. The algorithm is illustrated in Figure 10.12, and is an efficient iterative way of evaluating the maximum; just repeat the instruction `[a, b] = cuifit(a, b)`. Unfortunately it is very easy to make errors in a hand computation, and many people prefer the quadratic algorithm for this purpose. On a computer, however, the cubic approximation method is almost universally used.

Example 10.16

Find the first maximum of $f(x) = x \sin x$ given the bracket values from Example 10.14, namely $x_1 = 1.28$, $f_1 = 1.226$, $f'_1 = 1.325$, and $x_2 = 2.56$, $f_2 = 1.406$, $f'_2 = -1.590$.

Solution

Solving (10.11) gives $a = -0.3333$, $b = 0.7814$ and $c = 0.9630$, and hence, from (10.12),

$$x^* = 2.036, \quad f^* = 1.820, \quad f^{*\prime} = -0.0192$$

Thus x_2 , f_2 and f'_2 are replaced by x^* , f^* and $f^{*\prime}$, and x_1 , f_1 and f'_1 are retained. Equation (10.11) is now recomputed and solved to give

$$a = -0.4626, \quad b = 1.412, \quad c = -0.0153$$

Hence, from (10.12),

$$x^* = 2.029, \quad f^* = 1.820, \quad f^{*\prime} = -0.0007$$

Further iterations may be performed to get an even more accurate value. Comparison with Example 10.15 shows that both the quadratic and cubic algorithms work well for this function.



There are ‘built in’ maximizing routines included in most computer packages which can deal with most functions that arise in engineering computations. In MATLAB the single instruction `fminbnd(' -x*sin(x)', 1.28, 5.12)` produces the value $x = 2.0288$ instantly.

10.4.2 Exercises

29

- (a) Find a bracket for the minimum of the function $f(x) = x + 1/x^2$. Start at $x = 0.1$ and $h = 0.2$.
 (b) Use two cycles of the quadratic approximation to obtain an estimate of the minimum.
 (c) Use one cycle of the cubic approximation to obtain an estimate of the minimum.

30

The function

$$f = \frac{\sin x}{1+x^2}$$



has been computed as follows:

x	-0.5	0	1	3
f	-0.3825	0	0.4207	0.0141
f'	0.3952	1	-0.1506	-0.1075

Compare the brackets, obtained from the bracketing procedure, to be used in calculating the maximum of the function: (a) without using the derivatives, and (b) using the derivatives. Use these brackets to perform one iteration of each of the quadratic and the cubic algorithms. Compare the values obtained from the two calculations.

31

- Starting with the bracket $1 < x < 3$, determine an approximation to the maximum of the function



$$f(x) = x(e^{-x} - e^{-2x})$$

- (a) using two iterations of the quadratic algorithm, and
 (b) using two iterations of the cubic algorithm.
 (c) How many iterations are required to obtain three-figure accuracy?

32

Use the quadratic algorithm to obtain an estimate of the value of x that gives the minimum value to the largest root of the eigenvalue equation (that is, $\min_x [\lambda_{\max}(x)]$)

$$\begin{vmatrix} x - \lambda & -1 & 0 \\ -1 & -\lambda & -1 \\ 0 & -1 & x^2 - \lambda \end{vmatrix} = 0$$

Use the bracket given by $x = 1$ and $x = -1$.

33

Show that if $x_1 = x_2 - h$ and $x_3 = x_2 + h$ then (10.10) reduces to

$$x^* = x_2 + \frac{1}{2}h \frac{f_1 - f_3}{f_1 - 2f_2 + f_3}$$

(Note: This provides a better hand computation method than the Lagrange interpolation approach. The best x value is chosen and h is replaced by $\frac{1}{2}h$ after each step.) Show that this formula is a numerical form of the Newton–Raphson method applied to the equation $f'(x) = 0$.

34

An interval AB is divided symmetrically at points C and D. If AC/AD = AD/AB show that C divides AB in the Golden Ratio $\alpha = 1/2(3 - \sqrt{5}) = 0.382\dots$

The function $f(x)$ is known to have a maximum in the interval $a_1 \leq x \leq a_4$. It is evaluated at the points a_1 , a_4 and at the golden section points $a_2 = (1 - \alpha)a_1 + \alpha a_4$ and $a_3 = \alpha a_1 + (1 - \alpha)a_4$. If $f(a_2) > f(a_3)$ then the new bracket is taken as $a_1 \leq x \leq a_3$; if $f(a_2) > f(a_3)$ then the new bracket is taken as $a_2 \leq x \leq a_4$. The method is then repeated. Test the method on the functions

- (a) $f(x) = x \sin x$ with bracket 0, 2.5;
 (b) $f(x) = \frac{1}{(1-x)^2} \left(\ln x + \frac{2(1-x)}{1+x} \right)$ with bracket 1.5, 2.5.

10.4.3 Simple multivariable searches: steepest ascent and Newton's method

As indicated in Section 10.4.1, many multivariable search methods use a sequence of single-variable searches to achieve a maximum. The fundamental question is how to choose a sensible direction in which to search for the top of the hill with a very limited

amount of local information. The problem can be visualized when no derivatives are available as sitting in a dense fog and trying to get to the top of the hill with only an altimeter available. If derivatives are available then the fog has lifted a little, and we can now see a few meters around, so that at least we can see which is the uphill direction. The criteria for choosing a direction are (a) an easy choice of direction and (b) one that gives an efficient climbing method. This is not a simple task. We consider two methods in this section: **the method of steepest ascent** and **Newton's method**. Although they are rarely used for industrial-scale problems, they provide the basis for many more advanced methods. Modern methods are difficult to program, not because the basic method is difficult but because of the vast amount of remedial action that must be taken to prevent the program failing when something goes wrong. The general advice here is to understand the basic idea behind a method and then to implement it using a program from a reliable software library such as NAG (distributed by Numerical Algorithms Group Ltd of Oxford, UK) or a package such as MATLAB.

Perhaps the most obvious method of choosing a search direction is to use the locally steepest direction. For the function $f(x_1, x_2, \dots, x_n)$ this is known to be in the gradient direction $\mathbf{G} = \text{grad } f = [\partial f / \partial x_1, \partial f / \partial x_2, \dots, \partial f / \partial x_n]$. If the derivatives are not available then in current practice they are evaluated numerically.

The gradient direction can be easily proved to give the maximum change. From a given point (a_1, \dots, a_n) , we proceed in the direction (h_1, h_2, \dots, h_n) with given step length h so that $h_1^2 + h_2^2 + \dots + h_n^2 = h^2$. We then require

$$\max F = f(a_1 + h_1, \dots, a_n + h_n) - f(a_1, \dots, a_n)$$

subject to the constraint

$$h_1^2 + \dots + h_n^2 = h^2$$

The problem is one of Lagrange multipliers, so

$$0 = \frac{\partial F}{\partial h_i} - \lambda 2h_i = \frac{\partial f}{\partial h_i} - 2\lambda h_i \quad (i = 1, \dots, n)$$

Thus

$$\frac{\partial f}{\partial h_i} = \frac{\partial f}{\partial x_i} = 2\lambda h_i \quad (i = 1, \dots, n)$$

and hence

$$[h_1, h_2, \dots, h_n] \text{ is proportional to } \left[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right] = \mathbf{G}$$

The method of steepest ascent (or **steepest descent**, for minima) then proceeds from the point \mathbf{a} by choosing the gradient direction \mathbf{G} (or $-\mathbf{G}$ for minima) for a search direction. We therefore need to find

$$\max_t \{g(t) = f(a_1 + tG_1, a_2 + tG_2, \dots, a_n + tG_n)\} \tag{10.13}$$

Since (10.13) is a function of a single variable t , the methods of Section 10.4.1 are appropriate, and we should expect the cubic algorithm to be used in the optimization. Once the best available point in the search direction has been found, $\mathbf{a} + t_{\max} \mathbf{G}$, the new gradient direction is computed and the whole process is repeated. The algorithm is fairly straightforward, and is summarized in Figure 10.13.

Figure 10.13
Steepest-ascent
algorithm.

```

read (keyb, a1, . . . , an)
repeat
    {evaluate f(a) and G1 = ∂f/∂x1, G2 = ∂f/∂x2, . . . }
    {maximize g(t) in (10.13) by the cubic algorithm
     to give a new point a ← a + tmaxG}
until {G = 0}

```

The steepest-ascent (or descent) method has the great advantage of being simple and secure, but it has the disadvantage of being very slow, particularly near to the optimum. It is rarely used nowadays, but does form the basis of the hill climbing methods described in Section 10.4.5.

Example 10.17

Find the maximum of the function

$$f(x_1, x_2) = -(x_1 - 1)^4 - (x_1 - x_2)^2$$

by the steepest-ascent method, starting at the point (0, 0).

Solution

It is clear that (1, 1) gives the maximum, but this example is used to illustrate the basic method. The gradient is easily calculated from the partial derivatives

$$\frac{\partial f}{\partial x_1} = -4(x_1 - 1)^3 - 2(x_1 - x_2), \quad \frac{\partial f}{\partial x_2} = 2(x_1 - x_2)$$

Cycle 1: At the point (0, 0), $f = -1$, $\mathbf{G} = [4, 0]$, the search direction is $x_1 = 4t$, $x_2 = 0$, and we require

$$\max_t \{g(t) = -(4t - 1)^4 - 16t^2\}$$

This can be calculated as $t_{\max} = 0.10256$, so that we can start Cycle 2.

Cycle 2: The new point is (0.41025, 0), $f = -0.259$ and $\mathbf{G} = [0, 0.8205]$. The next search is in the direction $x_1 = 0.41025$, $x_2 = 0.8205t$, and we require

$$\max_t \{g(t) = -0.1210 - (0.41025 - 0.8205t)^2\}$$

This has the obvious solution $t_{\max} = \frac{1}{2}$, so that we can move to Cycle 3.

Cycle 3: The new point is (0.41025, 0.41025), $f = -0.1210$, and $\mathbf{G} = [0.8205, 0]$.

The calculation can be continued until $\mathbf{G} = 0$ to the required accuracy.

There are a couple of points to note from this calculation. The function value is steadily increasing, which is a good feature of the method, but after the first few iterations the method progresses in a large number of very small steps. The successive search directions are parallel to the axes, and hence are perpendicular to each other. This perpendicularity is just a restatement of the known result that the gradient vector is perpendicular to the contours (see Section 3.2.1). In Example 10.17 the function $g(t)$ is written down explicitly for clarity. In practice on a computer this would not be done, since once the search direction has been established, x_1 and x_2 are known functions of t only, and by the chain rule we have

$$\frac{dg}{dt} = \frac{\partial f}{\partial x_1} \frac{dx_1}{dt} + \frac{\partial f}{\partial x_2} \frac{dx_2}{dt} = \frac{\partial f}{\partial x_1} G_1 + \frac{\partial f}{\partial x_2} G_2$$

Since x_1 and x_2 are known, both $\partial f/\partial x_1$ and $\partial f/\partial x_2$ can be calculated, and $\mathbf{G} = [G_1, G_2]$ is the known search direction; therefore dg/dt is computed without explicitly writing down the function g .

The major criticism of the steepest-ascent method is that it is slow to converge, and so the question arises as to how it can be speeded up. One method in use in many programs is to use a fixed number of iterations in the line search, provided the function is increased. Some experimentation is required on how to implement this idea, but it can lead to significant improvement in speed.

It is well known that given a suitable initial condition, Newton–Raphson methods (see Section 9.4.8 of MEM) converge very rapidly, so the same basic idea is tried for these problems. It is convenient to employ matrix notation, and indeed most multi-dimensional optimization methods are written in matrix form. This gives a compact notation, and arrays appear naturally in computer languages.

Taylor's theorem (see Section 9.4.1 of MEM) can be written in matrix form to second order as

$$f(a_1 + h_1, \dots, a_n + h_n) \approx f(a_1, \dots, a_n) + \mathbf{h}^T \mathbf{G} + \frac{1}{2} \mathbf{h}^T \mathbf{J} \mathbf{h} \quad (10.14)$$

where

$$\mathbf{h} = \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

The form (10.14) can be verified by multiplying out the matrices and comparing with the standard form of Taylor's theorem. The vector \mathbf{G} is just the gradient vector, which is now written in matrix form as a column vector, and \mathbf{J} is an $n \times n$ symmetric matrix of second derivatives called the **Hessian** (or Jacobian) **matrix**.

The **Newton method** takes (10.14) as an approximation to f , finds the maximum (or minimum) of this quadratic approximation, and uses the optimal value of a to start the cycle again.

The optimum of (10.14) is given by $\partial f/\partial h_i = 0$ ($i = 1, 2, \dots, n$). The first of these conditions gives

$$0 = \frac{\partial f}{\partial h_1} = [1 \ 0 \ 0 \ \cdots \ 0] \mathbf{G} + \frac{1}{2} [1 \ 0 \ \cdots \ 0] \mathbf{J} \mathbf{h} + \frac{1}{2} \mathbf{h}^T \mathbf{J} \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (10.15)$$

Noting that, since \mathbf{J} is symmetric, for any vectors \mathbf{r} and \mathbf{s} we have

$$\mathbf{r}^T \mathbf{J} \mathbf{s} = (\mathbf{r}^T \mathbf{J} \mathbf{s})^T = \mathbf{s}^T \mathbf{J} \mathbf{r}$$

and hence (10.15) can be written as

$$0 = [1 \ 0 \ 0 \ \cdots \ 0] (\mathbf{G} + \mathbf{J} \mathbf{h})$$

Similarly for the other components we obtain

$$0 = [0 \ 1 \ 0 \ \dots \ 0] (\mathbf{G} + \mathbf{J}\mathbf{h})$$

$$\vdots \qquad \qquad \vdots$$

$$0 = [0 \ 0 \ 0 \ \dots \ 1] (\mathbf{G} + \mathbf{J}\mathbf{h})$$

The only way to satisfy this set of equations is to have

$$\mathbf{G} + \mathbf{J}\mathbf{h} = 0$$

and hence, provided the inverse exists,

$$\mathbf{h} = -\mathbf{J}^{-1}\mathbf{G}$$

The update rule for Newton's method is thus $\mathbf{a}_{i+1} = \mathbf{a}_i - \mathbf{J}_i^{-1}\mathbf{G}_i$ and the basic algorithm is now straightforward, as indicated in Figure 10.14.

Figure 10.14 Newton algorithm.

```
repeat
  { $\mathbf{a}_i$  known, calculate  $\mathbf{G}_i$ ,  $\mathbf{J}_i$ }
  {evaluate  $\mathbf{a}_{i+1} = \mathbf{a}_i - \mathbf{J}_i^{-1}\mathbf{G}_i$ }
until {sufficient accuracy}
```

When Newton's algorithm of Figure 10.14 converges, it does so very rapidly and satisfies our request for a fast method. For a quadratic function it only takes one iteration so, provided the function looks like a quadratic, the method will be expected to converge rapidly.

Example 10.18

Use Newton's method to find the maximum of

$$A = (h + b + 10) \left(\frac{2\,272\,000}{hb} + 2b + 5 \right)$$

Solution Here we have returned to the ‘milk carton’ problem of Example 10.2. We can calculate

$$\frac{\partial A}{\partial h} = \left(\frac{2\,272\,000}{hb} + 2b + 5 \right) - (h + b + 10) \frac{2\,272\,000}{h^2b}$$

$$\frac{\partial A}{\partial b} = \left(\frac{2\,272\,000}{hb} + 2b + 5 \right) + (h + b + 10) \left(-\frac{2\,272\,000}{hb^2} + 2 \right)$$

$$\frac{\partial^2 A}{\partial h^2} = 2(b + 10) \frac{2\,272\,000}{h^3b}$$

$$\frac{\partial^2 A}{\partial b \partial h} = 2 + \frac{10 \times 2\,272\,000}{h^2b^2}$$

$$\frac{\partial^2 A}{\partial b^2} = 4 + \frac{2 \times 2\,272\,000(h + 10)}{hb^3}$$

Iteration 1

$$\mathbf{a} = \begin{bmatrix} 100 \\ 100 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} -44.92 \\ 375.1 \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} 4.998 & 2.227 \\ 2.227 & 8.998 \end{bmatrix}$$

$$\mathbf{a}_{\text{new}} = \begin{bmatrix} 100 \\ 100 \end{bmatrix} - \begin{bmatrix} 0.2249 & -0.0557 \\ -0.0557 & 0.1249 \end{bmatrix} \begin{bmatrix} -44.92 \\ 375.1 \end{bmatrix} = \begin{bmatrix} 131 \\ 50.6 \end{bmatrix}$$

Iteration 2

$$\mathbf{a} = \begin{bmatrix} 131 \\ 50.6 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} -52.3 \\ -465.7 \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} 2.421 & 2.517 \\ 2.517 & 41.75 \end{bmatrix}$$

$$\mathbf{a}_{\text{new}} = \begin{bmatrix} 131 \\ 50.6 \end{bmatrix} - \begin{bmatrix} -0.4407 & -0.0266 \\ -0.0266 & 0.0255 \end{bmatrix} \begin{bmatrix} -52.3 \\ 465.7 \end{bmatrix} = \begin{bmatrix} 141.7 \\ 61.1 \end{bmatrix}$$

Iteration 3

$$\mathbf{a} = \begin{bmatrix} 141.7 \\ 61.1 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} -4.493 \\ -98.76 \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} -1.858 & 2.303 \\ 2.303 & 25.33 \end{bmatrix}$$

$$\mathbf{a}_{\text{new}} = \begin{bmatrix} 141.7 \\ 61.1 \end{bmatrix} - \begin{bmatrix} 0.6067 & -0.0552 \\ -0.0552 & 0.0445 \end{bmatrix} \begin{bmatrix} -4.493 \\ -98.76 \end{bmatrix} = \begin{bmatrix} 139 \\ 65.2 \end{bmatrix}$$

The iterations converge very rapidly; $h = 139$, $b = 65$ is not far from the solution, and gives $A = 827 \text{ cm}^2$.



Since the Newton method involves matrices MATLAB proves to be a very suitable package to perform the manipulations. The simple instructions

```
z=[100;100]
[a,G,J]=newton(z(1),z(2)), z=z-J\G'
```

with the last line repeated, gives the successive iterations of the example. The following listing gives the m-file *newton.m* that is used for Example 10.18.

```
function [a,agrad,ajac]=newton(h,b)
t1=h+b+10;t2=2272000/(h*b)+2*b+5; a=t1*t2;
agrad(1)=t2-t1*2272000/(h^2*b);
agrad(2)=t2+t1*(-2272000/(h*b^2)+2);
ajac(1,1)=2*(b+10)*2272000/(h^3*b);
ajac(1,2)=2+10*2272000/(h^2*b^2);ajac(2,1)=ajac(1,2);
ajac(2,2)=4+2*(h+10)*2272000/(h*b^3);
```

Unfortunately Newton's method is very unreliable, particularly for higher-dimensional problems, unless the starting value is close to the maximum. The reason for this is fairly clear. The analysis given only uses the necessary condition for a maximum, but it would

apply equally well to a minimum or saddle point. In multi-dimensional problems saddle points are abundant, and the usual failure of the Newton method is that it proceeds towards a distant saddle point and then diverges.

Applying the method to the Rosenbrock function, called the ‘banana’ function in MATLAB,

$$f(x, y) = 100(y^2 - x)^2 + (1 - x)^2$$

the unreliability, but rapid convergence, is illustrated in Figure 10.15.

Figure 10.15
Behaviour of Newton’s method for the Rosenbrock function for various starting points.

Starting point	Iterations to convergence	Final point	Comments
(-1.9, 1)	1	(1, 1)	minimum
(-1.9, 0.9)	6	(1, 1)	minimum
(-1.9, 0.5)	6	(1, -1)	minimum
(-1.9, 0) run 1	—	—	aborts, J is singular
(-1.9, 0) run 2	1	(1/101, 0)	saddle point
(-1.9, -0.5)	7	(1, 1)	minimum
(-1.9, -1)	1	(1, -1)	minimum

10.4.4 Exercises

- 35 Follow the first two complete cycles in the steepest-descent algorithm for finding the minimum of the function

$$f(x_1, x_2) = \mathbf{x}^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \frac{1}{2} \mathbf{x}^T \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \mathbf{x}$$

starting at $\mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

- 36 Show that the function

$$f(x, y) = 2(x+y)^2 + (x-y)^2 + 3x + 2y$$

has a minimum at the point $(-\frac{7}{16}, -\frac{3}{16})$.

Starting at the point $(0, 0)$, use one iteration of the steepest-descent algorithm to determine an approximation to the minimum point. Show that one iteration of Newton’s method yields the minimum point from *any* starting point.

- 37 Use the steepest-ascent method to find the maximum of the function



$$f(x, y, z) = -(x-y+z)^2 - (2x+z-2)^2 - (z^2 - 1)^2$$

starting from $(2, 2, 2)$.

- 38 Minimize the function



$$f(x, y, z) = (x-y+z)^2 + (2x+z-2)^2 + (z^2 - 1)^2$$

by Newton’s method. starting at $(2, 2, 2)$.

- 39 A new link road is to be constructed from a city centre to an existing road. In suitable coordinates and units the city centre is at $(0, 0)$ and the existing road has equation $y = 11 - 2x$. The cost of construction is proportional to the length of road, but it is twice as expensive to construct the road in the urban region $|x| \leq 1$ compared with outside the region. The link road consists of two straight sections, inside and outside the urban region. Find the equations of the two sections that minimize the overall cost.

10.4.5 Advanced multivariable searches

To overcome the problems of evaluating second derivatives, which are rarely available, and of the unreliability of the Newton method, but to use its speed of convergence, several methods were produced in the early 1960s. Two have survived and are now the methods currently available in most program libraries. **Conjugate-gradient methods** give one approach, but these will not be described here. We shall look at a method due to Davidon, commonly called **DFP** (after Davidon, Fletcher and Powell, who developed the method) or **quasi-Newton methods**. There is a whole class of such methods, but only one will be studied. Further details and more advanced topics can be found in specialist optimization books such as D. G. Luerberger and Y. Ye, *Linear and Nonlinear Programming* (fourth edition, New York, Springer, 2016).

The basic idea of the DFP method is to look for the *minimum* of the function $f(x_1, \dots, x_n)$ with gradient given by the column vector $\mathbf{G} = [\partial f / \partial x_1 \quad \partial f / \partial x_2 \quad \dots \quad \partial f / \partial x_n]^T$ by iterating with a matrix \mathbf{H}_i , which will be updated at each iteration, so that

$$\mathbf{a}_{i+1} = \mathbf{a}_i - \lambda \mathbf{H}_i \mathbf{G}_i \quad (10.16)$$

The reliable but slow steepest-descent method chooses $\mathbf{H}_i = \mathbf{I}$, the unit matrix, and $\lambda = \lambda_{\min}$, while the less reliable but fast Newton method chooses $\mathbf{H}_i = \mathbf{J}_i^{-1}$ and $\lambda = 1$. Thus the idea is to compute a sequence of \mathbf{H}_i so that $\mathbf{H}_0 = \mathbf{I}$ and $\mathbf{H}_i \rightarrow \mathbf{J}^{-1}$ as the minimum is approached. The basic analysis required to implement this scheme is quite difficult and beyond the scope of this book, so only the briefest of outlines will be given. For a quadratic function

$$f = c + \mathbf{x}_i^T \mathbf{G} + \frac{1}{2} \mathbf{x}_i^T \mathbf{J} \mathbf{x}_i$$

at two successive points the gradient is given by

$$\mathbf{G}_i = \mathbf{G} + \mathbf{J} \mathbf{x}_i$$

$$\mathbf{G}_{i+1} = \mathbf{G} + \mathbf{J} \mathbf{x}_{i+1}$$

so, subtracting,

$$\mathbf{G}_{i+1} - \mathbf{G}_i = \mathbf{J}(\mathbf{x}_{i+1} - \mathbf{x}_i)$$

Writing $\mathbf{h}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ and $\mathbf{y}_i = \mathbf{G}_{i+1} - \mathbf{G}_i$ and working on the assumption that $\mathbf{H}_i \mathbf{J} = \mathbf{I}$ since we require $\mathbf{H}_i \rightarrow \mathbf{J}^{-1}$, we obtain

$$\mathbf{H}_i \mathbf{y}_i = \mathbf{H}_i \mathbf{J} \mathbf{h}_i = \mathbf{h}_i \quad (10.17)$$

It is found to be impossible to satisfy (10.17) unless an exact solution is known, so the next best thing is to satisfy (10.17) one step behind as

$$\mathbf{H}_{i+1} \mathbf{y}_i = \mathbf{h}_i \quad (10.18)$$

There is a whole class of matrices that satisfy the key equation (10.18) but only the original Davidon matrix is quoted here, namely

$$\mathbf{H}_{i+1} = \mathbf{H}_i - \frac{\mathbf{H}_i \mathbf{y}_i \mathbf{y}_i^T \mathbf{H}_i}{\mathbf{y}_i^T \mathbf{H}_i \mathbf{y}_i} + \frac{\mathbf{h}_i \mathbf{h}_i^T}{\mathbf{h}_i^T \mathbf{y}_i} \quad (10.19)$$

It can be shown that for a quadratic function this sequence of \mathbf{H} s produces \mathbf{J}^{-1} in n iterations, where n is the dimension of the problem. The basic algorithm is described in Figure 10.16 for the minimum of a general function $f(x_1, \dots, x_n)$ with gradient $\mathbf{G} = [\partial f / \partial x_1 \quad \dots \quad \partial f / \partial x_n]^T$.

Figure 10.16
DFP algorithm for the minimum of $f(x_1, x_2, \dots, x_n)$.

```

read {initial values  $\mathbf{x}_0, \mathbf{H}_0 = \mathbf{I}$ }
      {calculate the gradient  $\mathbf{G}_0$  and  $f_0$ }
repeat
  {Find  $\min_{\lambda} f(\mathbf{x}_i - \lambda \mathbf{H}_i \mathbf{G}_i)$ , by the cubic algorithm}
  {Put  $\mathbf{x}_{i+1} = \mathbf{x}_i - \lambda_{\min} \mathbf{H}_i \mathbf{G}_i$ }
  {Calculate  $f_{i+1}, \mathbf{G}_{i+1}$  and hence  $\mathbf{h}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$ 
   and  $\mathbf{y}_i = \mathbf{G}_{i+1} - \mathbf{G}_i$ }
  {Update  $\mathbf{H}_i$  to  $\mathbf{H}_{i+1}$  by (10.19)}
until {sufficient accuracy}

```

This method was a major breakthrough in the early 1960s, and is still one of the best and most reliable available. Proofs of convergence and computational experience are available in advanced texts on optimization. To repeat a word of warning: these programs are very long and tedious to write because of the large amount of checking and remedial work that has to be inserted to prevent the program stopping. Such programs are available in software libraries, and these should be used.

Example 10.19

Use the DFP method to find the minimum of

$$f(x, y) = x^4 + y^4 + (2x + y - 5)^2$$

starting at $(0, 0)$.

Solution Note that only the first derivatives

$$\mathbf{G} = \begin{bmatrix} 4x^3 + 4(2x + y - 5) \\ 4y^3 + 2(2x + y - 5) \end{bmatrix}$$

are required in this method.

Iteration 1

$$\mathbf{a}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbf{G}_0 = \begin{bmatrix} -20 \\ -10 \end{bmatrix}, \quad \mathbf{H}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad f_0 = 25$$

and we search in the direction

$$\mathbf{a} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -20 \\ -10 \end{bmatrix} = \begin{bmatrix} 20\lambda \\ 10\lambda \end{bmatrix}$$

for a minimum of f ; that is, we compute

$$\min_{\lambda} \{(20\lambda)^4 + (10\lambda)^4 + (50\lambda - 5)^2\}$$

The cubic algorithm gives $\lambda_{\min} = 0.06414$, so

$$\mathbf{a}_1 = \begin{bmatrix} 1.2828 \\ 0.6414 \end{bmatrix}, \quad \mathbf{G}_1 = \begin{bmatrix} 1.2706 \\ -2.5308 \end{bmatrix}$$

$$\mathbf{h}_0 = \begin{bmatrix} 1.2828 \\ 0.6414 \end{bmatrix}, \quad \mathbf{y}_0 = \begin{bmatrix} -21.2706 \\ 7.4692 \end{bmatrix}$$

Thus \mathbf{H} is calculated from (10.19) as

$$\mathbf{H}_1 = \begin{bmatrix} 0.1611 & -0.2870 \\ -0.2870 & 0.9031 \end{bmatrix}$$

Iteration 2 $\mathbf{a}_1 = \begin{bmatrix} 1.2828 \\ 0.6414 \end{bmatrix}, \quad \mathbf{G}_1 = \begin{bmatrix} 1.2706 \\ -2.5308 \end{bmatrix}, \quad \mathbf{H}_1 = \begin{bmatrix} 0.1611 & -0.2870 \\ -0.2870 & 0.9031 \end{bmatrix}, \quad f_1 = 6.092$

We now search in the direction

$$\begin{aligned} \mathbf{a} &= \begin{bmatrix} 1.2828 \\ 0.6414 \end{bmatrix} - \lambda \begin{bmatrix} 0.1611 & -0.2870 \\ -0.2870 & 0.9031 \end{bmatrix} \begin{bmatrix} 1.2706 \\ -2.5308 \end{bmatrix} \\ &= \begin{bmatrix} 1.2828 - 0.9309\lambda \\ 0.6414 + 2.6501\lambda \end{bmatrix} \end{aligned}$$

and the cubic algorithm gives $\lambda_{\min} = 0.272$. We can now compute the next point and its gradient as

$$\mathbf{a}_2 = \begin{bmatrix} 1.1782 \\ 0.9390 \end{bmatrix}, \quad \mathbf{G}_2 = \begin{bmatrix} -0.2761 \\ -0.0969 \end{bmatrix}, \quad f = 5.6101$$

The new Davidon matrix is calculated from (10.19) as

$$\mathbf{H}_2 = \begin{bmatrix} 0.0597 & -0.0050 \\ -0.0050 & 0.1191 \end{bmatrix}$$

Iteration 3 The next iteration gives

$$\mathbf{a}_3 = \begin{bmatrix} 1.1879 \\ 0.9452 \end{bmatrix}, \quad \mathbf{G}_3 = \begin{bmatrix} -0.0122 \\ 0.0192 \end{bmatrix}, \quad \mathbf{H}_3 = \begin{bmatrix} 0.0461 & -0.0216 \\ -0.0216 & 0.1019 \end{bmatrix}, \quad f = 5.6084$$

The iterations continue until convergence at $x = 1.1886$ and $y = 0.9434$.

In current practice the steepest-descent method is rarely used because it is too slow. Newton's method requires a matrix of second derivatives, which are not usually available, and, although fast, it is too unreliable as illustrated in the previous section, particularly in higher-dimensional problems. The quasi-Newton methods, however, are widely available in most libraries and packages. They are reliable, very competitive and compare favourably with other well used methods, such as conjugate gradients. In MATLAB the optimization routine `fminunc`, DFP and steepest descent are available, but a variant BFGS (Broyden, Fletcher, Goldfarb, Shanno – see Exercise 40) is the default method. In the Optimization Toolbox of MATLAB the ‘unconstrained nonlinear’ option contains a demo of the three methods on the Rosenbrock function (it was also used to illustrate Newton's method in the previous section).



$$f(x, y) = 100(y^2 - x)^2 + (1 - x)^2$$



There are some first class graphics showing the progress of the method and the convergence found from the starting point $(-1.9, 1)$ is

BFGS	34 iterations	50 function evaluations
DFP	40 iterations	64 function evaluations
Steepest descent	exceeds limit	250 function evaluations

This performance is typical of the methods. The complication of the methods and the intimate relation with computers illustrates the need for packages to perform the extensive arithmetic. For the milk carton problem, Example 10.2, the function information is put in the M-file milk.m

```
function [f,g]=milk(x)
f=(x(1)+x(2)+10)*(2272000/(x(1)*x(2))+2*x(2)+5);
if nargout>1
    g(1)=(2272000/(x(1)*x(2))+2*x(2)+5)-(x(1)+x(2)+10)
        *2272000/(x(1)^2*x(2));11
    g(2)=(2272000/(x(1)*x(2))+2*x(2)+5)+(x(1)+x(2)+10)*
        (-2272000/(x(1)*x(2)^2)+2);
end
```

The instructions

```
x0=[100;100];
options=optimset('GradObj','on','display','iter');
[x,fval]=fminunc(@milk,x0,options)
```

produce the minimum of 827 cm^2 with $h = 138.6 \text{ mm}$ and $b = 65.7 \text{ mm}$ in six iterations.

A major development since the 1970s has been in devising modifications that avoid the line searches. It was found that the latter were very time-consuming, so there was great pressure to avoid them. The searches were replaced by one or more steps in the search direction until the function has been reduced ‘sufficiently’ and then the matrix H is updated. This not only reduces the number of function evaluations, which is normally the most expensive part of the routine, but it is found to reduce the number of iterations required. The BFGS variant is found to be very suitable for this approach and it is used in the MATLAB implementation. What is meant by ‘sufficiently’ requires careful consideration and a discussion can be found in R. Fletcher, *Practical Methods of Optimization*. See also D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming* (fourth edition, New York, Springer, 2016).

The early development of numerical optimization algorithms led to very distinct methods for cases when derivatives were or were not available. As computers developed in speed, this difference became less necessary, since derivatives could be calculated very rapidly by a numerical method. Options are now built into packages and library routines which use derivatives when supplied, or use a numerical approximation if the derivatives are not supplied.

Adapting methods to deal with constraints is of intense interest in practice, and has been a major thrust in the subject. For fully nonlinear problems with nonlinear

constraints the practical difficulties are very severe, but, robust programs are now available in most libraries and packages.

10.4.6 Least squares

When minimization problems are in the form of least squares

$$F = f_1^2(x_1, x_2, \dots, x_n) + f_2^2(x_1, x_2, \dots, x_n) + \dots + f_m^2(x_1, x_2, \dots, x_n)$$

then there is a slightly different and potentially more efficient technique that exploits the specific structure of the problem. In matrix form F becomes

$$F = [f_1 \ f_2 \ \dots \ f_m] \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} = f^T f$$

Each of the functions is expanded by Taylor's theorem, in matrix form, to first order about $\mathbf{x} = \mathbf{a}$, for example putting $x_1 = a_1 + h_1, x_2 = a_2 + h_2, \dots$ for f_1 ,

$$f_1 = (x_1, x_2, \dots, x_n) = f_1(a_1, a_2, \dots, a_n) + \left[\frac{\partial f_1}{\partial x_1} \frac{\partial f_1}{\partial x_2} \cdots \frac{\partial f_1}{\partial x_n} \right] \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}$$

Where the partial derivatives are evaluated at $\mathbf{x} = \mathbf{a}$. Repeating for all the m functions gives in matrix form

$$\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}_{x=a} + \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}$$

or in a more compact notation

$$f(\mathbf{x}) = f(\mathbf{a}) + \mathbf{J}\mathbf{h} \quad (10.20)$$

The minimum of F requires

$$0 = \frac{\partial F}{\partial x_1} = 2 \left[\frac{\partial f_1}{\partial x_1} \frac{\partial f_2}{\partial x_1} \cdots \frac{\partial f_m}{\partial x_1} \right] \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix}$$

Repeating this differentiation for each of the variables and putting the equations into matrix form $0 = \mathbf{J}^T \mathbf{f}(\mathbf{x})$ and using (10.20) gives

$$0 = \mathbf{J}^T \mathbf{f}(\mathbf{x}) = \mathbf{J}^T(f(\mathbf{a}) + \mathbf{J}\mathbf{h})$$

Hence we can now compute \mathbf{h} as

$$\mathbf{h} = -(\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T \mathbf{f}(\mathbf{a})$$

provided, of course, that the inverse exists. This is the same result that was quoted in Section 1.8.3 on singular value decomposition. To compute the minimum the result is iterated as

$$\mathbf{a}_{i+1} = \mathbf{a}_i - (\mathbf{J}_i^T \mathbf{J}_i)^{-1} \mathbf{J}_i^T \mathbf{f}_i \quad (10.21)$$

Example 10.20

Find the minimum of the function, starting at $(0, 0)$,

$$F = (x + y - 1)^2 + (x - y + 1)^2 + (2x - y)^2$$

Solution

$$\mathbf{f} = \begin{bmatrix} x + y - 1 \\ x - y + 1 \\ 2x - y \end{bmatrix} \quad \mathbf{J} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 2 & -1 \end{bmatrix}$$

so at the start point

$$\mathbf{f} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \text{ and } (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T = \frac{1}{14} \begin{bmatrix} 5 & 1 & 4 \\ 8 & -4 & -2 \end{bmatrix}$$

From (10.21) the new value of \mathbf{a} is $(2/7, 6/7)$ with $F = 2/7$. Because all the functions are linear the minimum is obtained in one iteration.



MATLAB has a built-in procedure to solve this type of problem; the following instructions obtains the same result quickly

```
C=[1,1;1,-1;2,-1];d=[1;-1;0];% f = Cx - d
x=lsqlin(C,d,[],[]) % Can use linear equality and
                      % inequality constraints, [] indicates empty
```

Example 10.21

The experimental data

X	0	1.2	2	5	10
f	0	0.80	1.15	1.55	1.89

is thought to fit the function

$$f = \frac{aX}{1 + bX}$$

Estimate the values of a and b and compare.

Solution The minimization of the least squares function

$$F(a, b) = \sum_{i=1}^4 \left(f_i - \frac{aX_i}{1+bX_i} \right)^2$$

should give good estimates of a and b . Note that the origin automatically satisfies the function. Take

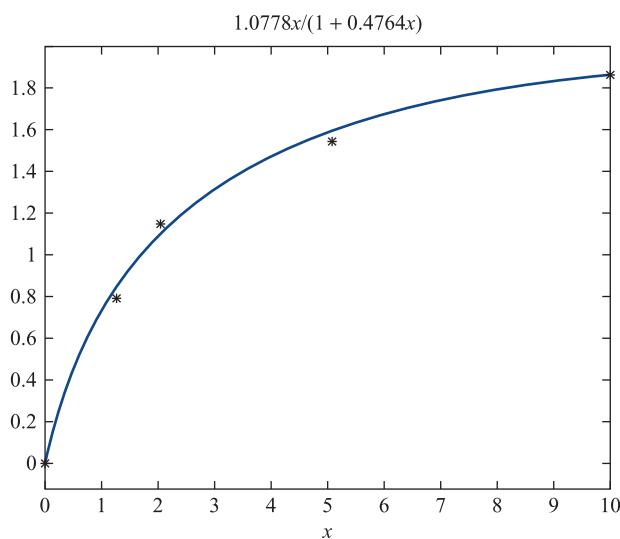
$$\mathbf{f} = \begin{bmatrix} f_1 - \frac{aX_1}{1+bX_1} \\ f_2 - \frac{aX_2}{1+bX_2} \\ f_3 - \frac{aX_3}{1+bX_3} \\ f_4 - \frac{aX_4}{1+bX_4} \end{bmatrix} \text{ and } \mathbf{J} = \begin{bmatrix} \frac{-X_1}{1+bX_1} & \frac{aX_1^2}{(1+bX_1)^2} \\ \frac{-X_2}{1+bX_2} & \frac{aX_2^2}{(1+bX_2)^2} \\ \dots & \dots \end{bmatrix}$$

Starting at $a=1$ and $b=0.5$ the iteration (10.21) was written in MATLAB and quickly gives a converged value of $a=1.0779$ and $b=0.4764$. However starting from $(1, 1)$ the method quickly diverges. It is important that a good starting guess is known. The code

```
ezplot('1.0778*x/(1+0.4764*x)', [0,10]), hold on
plot(0,0,'*',1.2,0.8,'*',2,1.15,'*',5,1.55,'*',10,1.89,'*)
```

produces Figure 10.17, which gives an illustration of how good the fit is. Note that least squares often gives a better fit to experimental data than interpolation methods, such as splines, since rogue points are not dominant in the method.

Figure 10.17
Fit of the function in
Example 10.21.





MATLAB has built-in procedures to deal with exactly this type of curve fitting. The following instructions obtain the same result.

```
xd=[0 1.2 2 5 10];yd=[0 0.80 1.15 1.55 1.89]; % inserts
                                                 the data
gg=@(b,xd)b(1)*xd./(1+b(2)*xd); % sets up the test function
x=lsqcurvefit(gg,[1 1],xd,yd) % returns the values
x = 1.0779 0.4764
```

A much wider range of start values can be used since `lsqcurvefit` can access a variety of methods of solution and contains numerous safeguards.

10.4.7 Exercises

- 40 Minimize the following functions by the DFP method, completing two cycles:

- (a) $f(x, y) = (x - y)^2 + 4(x - 1)^2$, starting at $(2, 2)$;
- (b) $f(x, y, z) = (x - y + z)^2 + (2x + z - 2)^2 + z^4$, starting at $(0, 0, 0)$.

- 41 Show that the updating formulas

$$(i) \quad \mathbf{H}_{i+1} = \mathbf{H}_i + \frac{(\mathbf{h}_i - \mathbf{H}_i \mathbf{y}_i)(\mathbf{h}_i - \mathbf{H}_i \mathbf{y}_i)^T}{(\mathbf{h}_i - \mathbf{H}_i \mathbf{y}_i)^T \mathbf{y}_i} \quad \text{rank (1)}$$

and

$$(ii) \quad \mathbf{H}_{i+1} = \mathbf{H}_i + \left(1 + \frac{\mathbf{y}_i^T \mathbf{H}_i \mathbf{y}_i}{\mathbf{h}_i^T \mathbf{y}_i} \right) \frac{\mathbf{h}_i \mathbf{h}_i^T}{\mathbf{h}_i^T \mathbf{y}_i} - \left(\frac{\mathbf{h}_i \mathbf{y}_i^T \mathbf{H}_i + \mathbf{H}_i \mathbf{y}_i \mathbf{h}_i^T}{\mathbf{h}_i^T \mathbf{y}_i} \right) \quad (\text{BFGS})$$

satisfy (10.18). Follow the DFP method through for two cycles, but using these updates for \mathbf{H} , on the functions

- (a) $x^2 + 2y^2$, starting at $(1, 2)$;
- (b) $x^2 + (x - y + 1)^2 + y^2z^2$, starting at $(0.5, 0.5, 0.5)$.

- 42 Show that the update formula (with suffixes suppressed)

$$\mathbf{H}' = \mathbf{H} + \nu \mathbf{p}^T - \mathbf{H} \mathbf{u} \mathbf{q}^T$$

satisfies the basic quasi-Newton equation (10.18)

$$\mathbf{H}' \mathbf{u} = \mathbf{v}$$

where \mathbf{p} and \mathbf{q} are vectors satisfying

$$\mathbf{p}^T \mathbf{u} = 1, \quad \mathbf{q}^T \mathbf{u} = 1$$

but are otherwise arbitrary.

By making a suitable choice of α, β and α', β' in the expressions

$$\mathbf{p} = \alpha \mathbf{v} + \beta \mathbf{H} \mathbf{u}$$

$$\mathbf{q} = \alpha' \mathbf{v} + \beta' \mathbf{H} \mathbf{u}$$

show that the Davidon formula (10.19) and formula (i) in Exercise 41 can be obtained.

- 43

An alternative algorithm for finding the minimum of a function, $f(\mathbf{x})$ with gradient \mathbf{g} , of several variables is due to Fletcher and Reeves. Starting at the point \mathbf{x}_0 , the first search direction is chosen as $\mathbf{p}_0 = -\mathbf{g}_0$. Successive search directions are given by

$$\mathbf{p}_i = -\mathbf{g}_i + \frac{\mathbf{g}_i^T \mathbf{g}_i}{\mathbf{g}_{i-1}^T \mathbf{g}_{i-1}} \mathbf{p}_{i-1}$$

and successive points satisfy

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \lambda_{i-1} \mathbf{p}_{i-1}$$

where λ_{i-1} is chosen to minimize the function $f(\mathbf{x})$ in this search direction.

Apply the method to the functions

$$(a) f = \frac{1}{2}(3x^2 + y^2) \text{ and}$$

$$(b) f = (x - y + 1)^2 + x^2 y^2 + (z - 1)^2$$

10.5 Engineering application: chemical processing plant

A chemical processing plant consists of a main processing unit and two recovery units. Chemicals A and B are fed into the plant, and produce a maximum output of 100 t day^{-1} of material C. The effluent stream is rich in chemical B, which can be recovered from the primary and secondary recovery units. The total recovered, when at full throughput, is 10 t day^{-1} of pure B, and it is fed back into the incoming stream of chemical B. The process is illustrated in Figure 10.18; the numbers in parentheses indicate the maximum flow (in t day^{-1}) that can be sent down the pipes, and the x_i indicate the actual flow (in t day^{-1}).

The chemistry of the process implies that the chemicals must be mixed in given ratios. For the present system it is found that

$$x_1:x_3 = 1:1, \quad x_1:x_4 = 3:5, \quad x_1:x_5 = 3:1, \quad x_5:x_7 = 5:3$$

$$x_5:x_6 = 5:2, \quad x_7:x_8 = 6:1, \quad x_7:x_{10} = 6:5$$

must be maintained for any flow through the system. Chemical A costs $\text{£}100 \text{ t}^{-1}$, chemical B costs $\text{£}120 \text{ t}^{-1}$ and chemical C sells for $\text{£}220 \text{ t}^{-1}$. The running costs are as follows:

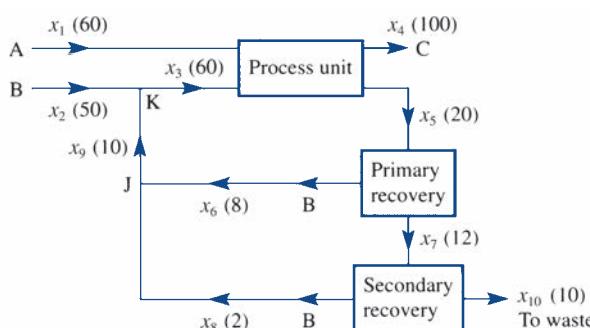
	Variable costs	Fixed costs
Process unit	$\text{£}70 \text{ t}^{-1}$ of product	$\text{£}500 \text{ day}^{-1}$
Primary recovery	$\text{£}30 \text{ t}^{-1}$ of input	$\text{£}200 \text{ day}^{-1}$
Secondary recovery	$\text{£}40 \text{ t}^{-1}$ of input	$\text{£}100 \text{ day}^{-1}$
Disposal of waste	$\text{£}30 \text{ t}^{-1}$	
Indirect cost		$\text{£}400 \text{ day}^{-1}$

It is required to find the most profitable operating policy that can be achieved.

The profit can be written down for a day's production as

$$\begin{aligned} z &= -(\text{fixed costs}) + (\text{profit from sale of C}) - (\text{costs of chemicals A and B}) \\ &\quad - (\text{process unit costs}) - (\text{primary unit costs}) - (\text{secondary unit costs}) \\ &\quad - (\text{waste product costs}) \\ &= -1200 + (220x_4) - (100x_1 + 120x_2) \\ &\quad - (70x_4) - (30x_5) - (40x_7) \\ &\quad - 30x_{10} \\ z &= -1200 - 100x_1 - 120x_2 + 150x_4 - 30x_5 - 40x_7 - 30x_{10} \end{aligned}$$

Figure 10.18
Schematic diagram of a chemical processing plant. The numbers in parentheses give the maximum flow.



The constraints on the flow, given by the maximum throughput, are

$$\begin{aligned}x_1 &\leq 60, \quad x_2 \leq 50, \quad x_3 \leq 60, \quad x_4 \leq 100, \quad x_5 \leq 20, \quad x_6 \leq 8, \quad x_7 \leq 12, \\x_8 &\leq 2, \quad x_9 \leq 10, \quad x_{10} \leq 10\end{aligned}$$

The constraints on the chemistry given by the fixed ratios can be written in a convenient form as

$$\begin{aligned}x_1 - x_3 &= 0, \quad 5x_1 - 3x_4 = 0, \quad x_1 - 3x_5 = 0, \quad 3x_5 - 5x_7 = 0, \quad 2x_5 - 5x_6 = 0, \\x_7 - 6x_8 &= 0, \quad 5x_7 - 6x_{10} = 0\end{aligned}$$

Finally, at the junctions J and K, continuity (what flows in equals what flows out) gives

$$x_6 + x_8 - x_9 = 0, \quad x_2 + x_9 - x_3 = 0$$

The problem thus has 10 variables, 10 inequality constraints and 9 equality constraints. The choice is whether to use the equality constraints to eliminate some of the variables or just to treat the 19 constraints directly by LP. The equations are sufficiently simple to solve for the variables as

$$\begin{aligned}x_2 &= \frac{5}{6}x_1, \quad x_3 = x_1, \quad x_4 = \frac{5}{3}x_1, \quad x_5 = \frac{1}{3}x_1, \quad x_6 = \frac{2}{15}x_1, \quad x_7 = \frac{1}{5}x_1, \quad x_8 = \frac{1}{30}x_1, \\x_9 &= \frac{2}{15}x_1, \quad x_{10} = \frac{1}{6}x_1, \quad z = 27x_1 - 1200\end{aligned}$$

Thus x_1 must be as large as possible, that is, at the value 60, giving a maximum profit of £420 day⁻¹. We must check that all the constraints are satisfied, and indeed this is the case. It is easily seen that each variable reaches its maximum possible value indicated in Figure 10.18.

When we look at variations on the problem, it becomes less clear whether to eliminate or just to use LP directly on the modified equations. For instance a very sensible question is whether it is worth using the primary or secondary recovery units. We can consider this question by allowing a portion to go to waste (at the same cost given previously), as indicated in Figure 10.19.

We add to the previous continuity equations similar equations for the junctions L and H:

$$x_5 - x_{11} - x_{12} = 0$$

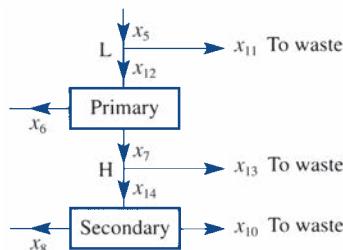
$$x_7 - x_{13} - x_{14} = 0$$

The inputs to the primary and secondary units are now x_{12} and x_{14} , so we need to modify the fixed ratio chemical constraint as follows:

$$\text{replace } 3x_5 - 5x_7 = 0 \text{ by } 3x_{12} - 5x_7 = 0$$

$$\text{replace } 2x_5 - 5x_6 = 0 \text{ by } 2x_{12} - 5x_6 = 0$$

Figure 10.19
Modification to
the chemical
processing plant.



replace $x_7 - 6x_8 = 0$ by $x_{14} - 6x_8 = 0$

replace $5x_7 - 6x_{10} = 0$ by $5x_{14} - 6x_{10} = 0$

In the cost function x_5 is replaced by x_{12} , and x_7 by x_{14} , and the additional waste costs $(-30x_{11} - 30x_{13})$ must be added:

$$\begin{aligned} z = & -1200 + 150x_4 - 100x_1 - 120x_2 \\ & - 30x_{12} - 30x_{11} - 40x_{14} - 30x_{13} - 30x_{10} \end{aligned}$$

We now have three free variables, so we shall certainly need an LP approach. An LP package was used to obtain the solution

$$x_1 = 57.69, \quad x_2 = 50, \quad x_3 = 57.69, \quad x_4 = 96.15$$

$$x_5 = 19.23, \quad x_6 = 7.69, \quad x_7 = 11.54, \quad x_8 = 0$$

$$x_9 = 7.69, \quad x_{10} = 0, \quad x_{11} = 0, \quad x_{12} = 19.23$$

$$x_{13} = 11.54, \quad x_{14} = 0$$

and the profit is £530.77 day⁻¹. It can be seen that $x_{11} = 0$, so nothing is sent to waste before the primary recovery unit; but $x_{14} = 0$, so that all the material from the primary to the secondary goes to waste, and the secondary unit is bypassed. The effect of this strategy is to increase the profit by about 20%.

There are many other variations that can be considered for this model. For instance, pumps often go wrong, so it is important to investigate what happens if the maximum flows are reduced or even cut completely. Once the basic program has been set up, such variations are quite straightforward to implement.

10.6 Engineering application: heating fin

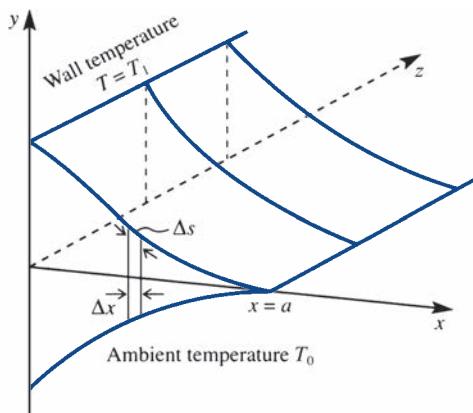
A heating fin is of the shape indicated in Figure 10.20, where the wall temperature is T_1 , the ambient temperature is T_0 and within the fin the value $T = T(x)$ is assumed to depend only on x and to be independent of y and z . Heat is transferred by conduction along the fin, which has thermal conductivity k , and heat is transferred to the outside according to Newton's law of cooling, with surface heat-transfer coefficient h . Considering an area of the fin of unit width in the z direction and height $2y$, the heat transferred by conduction in the x direction is $k2y dT/dx$. The net transfer through the element illustrated is $(d/dx)(k2y dT/dx)$. The total heat lost through a surface element of unit width in the z direction and length $\Delta s = [1 + (dy/dx)^2]^{1/2} \Delta x$ along the surface is $h(T - T_0)\Delta s$. Since there are two surfaces, we can write the heat-transfer equation as

$$2 \frac{d}{dx} \left[ky \frac{d}{dx} (T - T_0) \right] = 2h(T - T_0) \left[1 + \left(\frac{dy}{dx} \right)^2 \right]^{1/2}$$

Provided that dy/dx is not too large, $(dy/dx)^2$ can be neglected, giving

$$\frac{d}{dx} \left[y \frac{d}{dx} (T - T_0) \right] = \frac{h}{k}(T - T_0) \tag{10.22}$$

Figure 10.20
Heating fin.



The mass of the fin is given, so that its cross-sectional area is known, and hence

$$\int_0^a y \, dx = \frac{1}{2} A \quad (10.23)$$

Finally, we wish to maximize the heat transfer, so we require the maximum of

$$I = 2h \int_0^a (T - T_0) \, dx \quad (10.24)$$

over all possible functions $y(x)$.

The problem involves choosing a function $y(x)$ that satisfies (10.23) and then solving (10.22) for $T - T_0$. This is then substituted into (10.24) and the integral evaluated. Out of all possible such functions y , we choose the one that maximizes I . The scheme outlined is extremely difficult and belongs to a class called **variational problems**. An alternative approximate method must be sought. The *assumption* made is that the temperature falls linearly with x as

$$T - T_0 = (T_1 - T_0)(1 - \alpha x)$$

We also assume that $y = 0$ at $x = a$. We thus have two free parameters, α and a , which we can use to give an approximate solution. Given $T - T_0$, y can be computed from (10.22) as

$$y = \frac{h}{k\alpha} (a - \frac{1}{2}\alpha a^2 - x + \frac{1}{2}\alpha x^2)$$

It can be seen that our basic assumption implies that y is quadratic in x . To satisfy the area constraint (10.23), a simple integral is performed to give

$$\frac{1}{2} A = \frac{h}{k\alpha} a^2 (\frac{1}{2} - \frac{1}{3}\alpha a)$$

which gives a relation between α and a . The function I is now integrated as

$$I = 2h(T_1 - T_0) \int_0^a (1 - \alpha x) \, dx$$

or

$$\frac{I}{2h(T_1 - T_0)} = a - \frac{1}{2}\alpha a^2$$

Thus the very difficult problem has been reduced to a Lagrange multiplier problem of maximizing

$$f = a - \frac{1}{2}\alpha a^2$$

subject to

$$g = 0 = \left(\frac{a^2}{2\alpha} - \frac{1}{3}a^3 - S^3 \right) \quad (10.25)$$

where $S^3 = kA/2h$. The Lagrange multiplier analysis gives the equations

$$\frac{\partial(f - \lambda g)}{\partial a} = 1 - \alpha a - \lambda \left(\frac{a}{\alpha} - a^2 \right) = 0$$

$$\frac{\partial(f - \lambda g)}{\partial \alpha} = -\frac{1}{2}a^2 + \lambda \frac{a^2}{2\alpha^2} = 0$$

Hence

$$\lambda = \alpha^2$$

$$1 - 2\alpha a + \alpha^2 a^2 = 0$$

so that $a\alpha = 1$. Substituting back into the constraint (10.25) gives

$$S^3 = \frac{1}{6}a^3, \quad \text{or} \quad a = \left(\frac{3kA}{h} \right)^{1/3}$$

and therefore

$$(T - T_0) = (T_1 - T_0) \left(1 - \frac{x}{a} \right)$$

so that

$$\frac{y}{a} = \frac{ha}{2k} \left(1 - \frac{x}{a} \right)^2$$

Thus, given the physical parameters k , h and A , the ‘best’ shape can be derived. This model shows how a very difficult mathematical problem in optimization can be reduced to a much more straightforward one by an appropriate choice of test functions for $T - T_0$.

10.7 Review exercises (1–26)

- 1 Use the simplex method to find the maximum of the function

$$F = 12x_1 + 8x_2$$

subject to the constraints

$$x_1 + x_2 \leq 350$$

$$2x_1 + x_2 \leq 600$$

$$x_1 + 3x_2 \leq 900$$

$$x_1, x_2 \geq 0$$

Check your results with a graphical solution.

- 2 A manufacturer makes three types of sailboard and is trying to decide how many of each to make in a given week. There are 400 h of labour available, and the three types of sailboard require respectively 10, 20 and 30 h of labour to construct. A shortage of fibreglass and of resin coating is anticipated. The quantities required by each type of sailboard are as follows:

	Type 1	Type 2	Type 3	Total supplies
Fibreglass (kg)	5	10	25	290
Resin coating (litres)	3	2	1	72

If the profits on types 1, 2 and 3 are £10, £15 and £25 respectively, how many of each type should the manufacturer make to maximize profit?

- 3 A motor manufacturer makes a ‘standard’, a ‘super’ and a ‘deluxe’ version of a particular model of car. It is found that each week two of the materials are in limited supply: that of chromium trim being limited to 1600 m per week and that of soundproofing material to 1500 m² per week. The quantity of each of these materials required by a car of each type is as follows:

	Standard	Super	Deluxe
Chromium trim (m)	10	20	30
Soundproofing (m ²)	10	15	20

All other materials are in unlimited supply.

The manufacturer knows that any number of standard models can be sold, but it is estimated that the combined market for the super and deluxe versions is limited to 50 models per week. In addition there is a contractual obligation to supply a total of 70 cars (of any type) each week.

The profits on a standard, super and deluxe model are £100, £300 and £400 respectively. Assuming that the facilities to manufacture any number of cars are available, how many of each model should be made to maximize the weekly profit, and what is that profit?

A poor student lives on bread and cheese. The bread contains 1000 calories and 25 g protein in each kilogram, and the cheese has 2000 calories and 100 g of protein per kg. To maintain a good diet, the student requires at least 3000 calories and 100 g of protein per day. Bread costs 60p per kg and cheese 180p per kg. Find the minimum cost of bread and cheese needed per day to maintain the diet.

5 Use Lagrange multipliers to find the maximum and minimum distances from the origin to the point P lying on the curve

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

6 A solid body of volume V and surface area S is formed by joining together two cubes of different sizes so that every point on one side of the smaller cube is in contact with the larger cube. If $S = 7 \text{ m}^2$, find the maximum and/or minimum values of V for which both cubes have non-zero volumes.

7 Find the maximum distance from the point (1, 0, 0) to the surface represented by

$$2x + y^2 + z = 8$$

8 Find the local extrema of the function

$$F(x, y, z) = x + 2y + 3z$$

subject to the constraint

$$x^2 + y^2 + z^2 = 14$$

Obtain also the *global* maximum and minimum values of F in the region

$$x \geq 0, \quad y \geq 0, \quad z \geq 0, \quad x^2 + y^2 + z^2 \leq 14$$

- 9 A triangle with sides a, b, c has given perimeter $2s$. Recall that the area of the triangle, A , is given by the formula

$$A^2 = s(s - a)(s - b)(s - c)$$

- (i) If a is given, use Lagrange multipliers to find the values of b and c that make the area a maximum.
- (ii) If a, b, c are unrestricted use Lagrange multipliers to find the values that make the area a maximum.

- 10 A nuclear reactor is in the form of a circular cylinder of radius r and height h . According to the theory of nuclear diffusion, the restriction

$$\left(\frac{a}{r}\right)^2 + \left(\frac{\pi}{h}\right)^2 = b$$

applies, where a and b are constants. Use Lagrange multipliers to find the values of r and h that make the volume of the reactor a maximum.

- 11 According to lubrication theory, the lift on a pad bearing, where fluid flows in the narrow gap between a pad and a fixed piece of machinery, is given by

$$F = A \frac{1}{(k-1)^2} \left(\ln k - 2 \frac{k-1}{k+1} \right)$$

where A is a constant, $k = h_1/h_2 > 1$ and h_1 and h_2 are the gap widths at the front and back of the pad. Find the value of k that makes F a maximum by using the bracket/quadratic approximation technique.

- 12 A cylindrical can of radius R (cm) and height H (cm) is to be made with volume 1000 cm^3 . The cost of making the can is proportional to

$$(\text{amount of metal}) \times (\text{machine factor})$$

where the amount of metal is proportional to the surface area of the can (including the two ends) and the machine factor is given by $1 + [1 - (H/4R)]^2$ and reflects the difficulty of machining the can. Show that the cost is

$$2 \left(\frac{1000}{R} + \pi R^2 \right) \left[1 + \left(1 - \frac{1000}{4\pi R^3} \right)^2 \right]$$

Find a bracket for R and use the quadratic algorithm to estimate the radius that minimizes the cost.

- 13 Use the quadratic approximation method to obtain a first estimate of the minimum of the function

$$f(x) = 1 - t + t^2$$

where t is the non-negative root of

$$t^2 + tx - (1 - x^2) = 0$$

Start with the interval $0 \leq x \leq 1$ and note that for $x = 0.5$, $t = 0.6514$ and $f = 0.7729$.

- 14 In Figure 10.21 the disc rotates at a constant angular velocity, so $\theta = \alpha t$. The subsequent movement of the slider P gives $x = x(t)$. If $L/a = \lambda$ show that the velocity, v , of the slider is given by

$$\frac{v}{\alpha a} = -\sin \theta \left[1 + \frac{\cos \theta}{\sqrt{(\lambda^2 - \sin^2 \theta)}} \right]$$

Use the bracket and quadratic approximation technique to evaluate the maximum and minimum velocities of the slider in the case $\lambda = 3$.

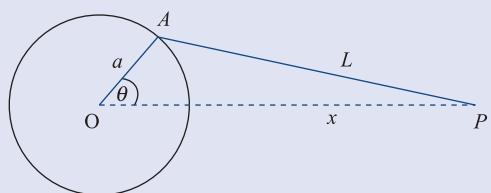


Figure 10.21 Disc and slider in Review exercise 14.

- 15 A trucking company estimates that the cost of running a truck is

$$0.02 \left(2v^4 + \frac{v}{10} \right)$$

pounds per mile at a constant speed v . The driver earns £5 per hour. Find the cost for a journey of D miles. What speed is recommended to minimize the cost?

- 16 Use (a) the steepest-descent method, (b) the Newton method and (c) the DFP method to find the position of the minimum of the function

$$f(x, y) = x^2 + (x - y)^2 + \frac{1}{16}(x + y + 1)^4$$

starting at $(0, 0)$. Perform two cycles of each method, and compare your results.

17

- A compound pendulum consists of a rectangular lamina with a heavy particle embedded in it, as illustrated in Figure 10.22. For small oscillations about the equilibrium, $\theta = \alpha$, and putting $\theta = \alpha + \varepsilon$, the equation of motion is

$$\ddot{\varepsilon} = -\mu^2 \varepsilon$$

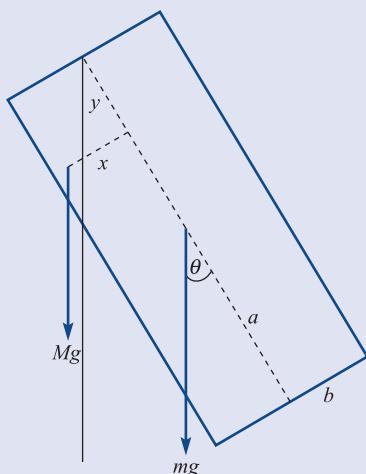


Figure 10.22 Compound pendulum in Review exercise 17.

where the period of the oscillations μ is given, after a substantial calculation, by

$$\frac{a\mu^2}{g} = \frac{\sqrt{[(\lambda + Y)^2 + X^2]}}{X^2 + Y^2 + \lambda(\frac{4}{3} + k^2)}$$

with

$$X = x/a, \quad Y = y/a, \quad k = b/a, \quad \lambda = m/M$$

Explore this expression for maximum and minimum values in the region $|X| \leq k$, $Y \leq 2$; take the case $\lambda = \frac{1}{2}$ and $k = \frac{1}{4}$.

18

- A method called **Partan** uses the notation $\mathbf{D}^{(i)}$ for the gradient of f evaluated at $\mathbf{x} = \mathbf{x}^{(i)}$. The iteration scheme for evaluating the minimum of f using Partan is

$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - \mu_1 \mathbf{D}^{(1)}$$

$$\left. \begin{aligned} \mathbf{z}^{(i)} &= \mathbf{x}^{(i)} - \mu_i \mathbf{D}^{(i)} \\ \mathbf{x}^{(i+1)} &= \mathbf{z}^{(i)} + \lambda_i (\mathbf{z}^{(i)} - \mathbf{x}^{(i-1)}) \end{aligned} \right\} \quad (i = 2, 3, 6)$$

where μ_i and λ_i are chosen by optimum line searches. Sketch the progress of this method up to the point $\mathbf{x}^{(4)}$ for a scalar function of two variables $f(x_1, x_2)$.

Illustrate the use of Partan and the method of steepest descent on the quadratic function

$$f = (x_1 - x_2)^2 + (x_1 - 1)^2$$

starting at $(0, 0)$.

19

- The Newton method described in Section 10.4.3 often fails to converge. One way of overcoming this problem is to restrict the step length at each iteration. Given that $\mathbf{x} = \mathbf{a} + \mathbf{h}$, this can be implemented by constraining \mathbf{h} to have length L , where

$$\mathbf{h}^T \mathbf{h} = L^2$$

Use a Lagrange multiplier λ to show that the result gives

$$\mathbf{x}_{\text{new}} = \mathbf{x}_{\text{old}} - (\mathbf{J} + \lambda \mathbf{I})^{-1} \mathbf{G}$$

The algorithm is then implemented by successively using $\lambda = 0, 1, 10, 100, \dots$ until a reduction in the function f is obtained.

Starting at $\mathbf{x} = [1 \ 1]^T$, perform one complete step of the modified algorithm on the function

$$f = x^2 + y^2 - x^2y$$

20

- Use the method developed in Section 10.4.6 to iterate to the minimum of the functions

$$(a) \quad F = (x - y)^2 + \frac{1}{16}(x + y + 1)^4, \text{ starting at } x = 0, y = 0;$$

$$(b) \quad F = \left(\frac{1}{x+y} \right)^2 + \left(\frac{x}{1+2x+y} \right)^2, \text{ starting at } x = 1, y = 1.$$

21

- It is known from experience that a curve of the form

$$y = 1/(a + bx)$$

should give a good fit to experimental data in the form of a set of points

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

It is required to calculate a and b by a best least-squares fit, and thus to minimize

$$F(a, b) = \sum_{i=1}^p \left(y_i - \frac{1}{a + bx_i} \right)^2$$

Use the least-squares algorithm described in Example 10.21 to fit the function to the data points

$$(0, 1) \quad (1, 0.6), \quad (2, 0.3), \quad (3, 0.2)$$

- 22 A quadratic function $f(x_1, x_2, \dots, x_n)$ with a unique minimum is given in matrix form as

$$f = c + \mathbf{b}^T \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x}$$

Show that a search in the direction

$$\mathbf{x} = \mathbf{a} + \lambda \mathbf{d}$$

produces the minimum at

$$\lambda_{\min} = \frac{-(\mathbf{b} + \mathbf{A}\mathbf{a})^T \mathbf{d}}{\mathbf{d}^T \mathbf{A} \mathbf{d}}$$

- 23 Complete two complete cycles of the steepest-descent algorithm for the function

$$f(x, y) = (1 - x)^2 + (x - y)^2$$

starting at $x = 0, y = 0$. Use Review exercise 22 for the minimization in the search directions.

Show that the minimum is obtained in a single iteration of the Newton method.

- 24 (Harder) It is required to solve the differential equation

$$yy'' - y'^2 + y' = 0$$

with the boundary conditions $y(0) = 1$ and $y(1) = 3$, by a shooting method. The equation is solved for the initial conditions

$$y(0) = 1, \quad y'(0) = \alpha$$

by any suitable method (for example, by a Runge–Kutta method). With this solution, calculate

$$F(\alpha) = y(1)$$

and then try to drive F to the value 3 by minimization of

$$[F(\alpha) - 3]^2$$

In this example illustrate the method by using the exact solution

$$y = (1 + b) e^{x/b} - b$$

for the forward integration.

- 25 (An extended problem) In the chemical processing plant model in Section 10.5 consider the profits when

- (a) the primary pump is faulty and the constraint $x_5 \leq 12$ is imposed;
- (b) the waste pump between primary and secondary fails so that $x_{13} = 0$ (see Figure 10.19).

- 26 (An extended problem) Extend the heating fin analysis in Section 10.6, using a higher approximation to the temperature:

$$T - T_0 = (T_1 - T_0)(1 - \alpha x - \beta x^2)$$

Compare the shape of the fin with that given in the text, and compute the heat transferred in each case.



11

Applied Probability and Statistics

Chapter 11 Contents

11.1	Introduction	800
11.2	Review of basic probability theory	801
11.3	Estimating parameters	810
11.4	Joint distributions and correlation	825
11.5	Regression	841
11.6	Goodness-of-fit tests	863
11.7	Engineering application: analysis of engine performance data	874
11.8	Engineering application: statistical quality control	891
11.9	Poisson processes and the theory of queues	908
11.10	Bayes' theorem and its applications	930
11.11	Review exercises (1–10)	946

11.1

Introduction

Applications of probability and statistics in engineering are very far-reaching. Data from experiments have to be analysed and conclusions drawn, decisions have to be made, production and distribution have to be organized and monitored, and quality has to be controlled. In all of these activities probability and statistics have a central role to play.

The distinction between applied probability and statistics is blurred, but essentially it is this: **applied probability** is about mathematical modelling of a situation that involves random uncertainty, whereas **statistics** is the business of handling data and drawing conclusions, and can be regarded as a branch of applied probability. Most of this chapter is about statistics, but Section 11.9 on queueing theory is applied probability.

When applying statistical methods to a practical problem, the most visible activity is the processing of data, often using a statistical package such as R to apply a formula or standard procedure. The relative ease and obviousness of this activity sometimes leads to a false sense that there is nothing more to it. On the contrary, the handling of the data is quite superficial compared with the essential task of trying to understand both the problem at hand and the assumptions upon which the various statistical procedures are based. If the wrong procedure is chosen, a wrong conclusion may be drawn.

It is, unfortunately, all too easy to use a formula while overlooking its theoretical basis, which largely determines its applicability. It is true that there are some statistical methods that continue to work reliably even where the assumptions upon which they are based do not hold (such methods are called **robust**), but it is unwise to rely too heavily upon this and even worse to be unaware of the assumptions at all.

The conclusions of a statistical analysis are often expressed in a qualified way such as ‘We can be 95% sure that . . .’. At first this seems vague and inadequate. Perhaps a decision has to be made, but the statistical conclusion is not expressed simply as ‘yes’ or ‘no’. A statistical analysis is rather like a legal case in which the witness is required to tell ‘the whole truth and nothing but the truth’. In the present context ‘the whole truth’ means that the statistician must glean as much information from the data as is possible until nothing but pure randomness remains. ‘Nothing but the truth’ means that the statistician must not state the conclusion with any greater degree of certainty or confidence than is justified by the analysis. In fact there is a practical compromise between truth and precision that will be explained in Section 11.3.3. The result of all this is that the decision-maker is aided by the analysis but not pre-empted by it.

In this chapter we shall first review the basic theory of probability and then cover some applications that are beneficial in engineering and many other fields: the statistics of means, proportions and correlation, linear regression and goodness-of-fit testing, quality control and queueing theory.

We will also consider applications of Bayes’ theorem, an important result in probability, and Bayesian statistical inference.



We supply R code for many of the calculations and analyses discussed in this chapter. R is a free software environment for statistical computing and graphics, available from <https://www.r-project.org/>. Unlike some other statistical packages, R is not menu driven, but requires the user to type in and then run commands. Most of these commands are based on functions. We recommend working with R through the easy to use RStudio interface available from <https://www.rstudio.com/>. RStudio provides a customized editor from which commands can be run by R, together with many other features designed to

make R easier to use. A file or R script comprising the R code developed in a session can be saved for future use. An enormous amount of documentation and help about R and RStudio is available from the links just given and from other online sources.

The basic R software, sometimes referred to as base R, is enhanced by thousands of contributed packages which provide specially written R functions to perform a massive range of modern statistical and data science tasks. R can also be used to solve many other problems that arise in engineering. For example, V. A. Bloomfield, *Using R for Numerical Analysis in Science and Engineering* (Boca Raton, FL, Chapman and Hall/CRC, 2014) discusses the use of R in areas including matrix analysis, ordinary and partial differential equations and optimization, and presents a range of interesting engineering case studies. RStudio provides tools for producing reports that integrate R code and the output that it produces with the narrative using a file format called R Markdown. reports of statistical analyses produced in this way are easy to update if the data change and can be readily reproduced by other users. We do not discuss the production of dynamic documents or reproducible research further, except to refer the interested reader to C. Gandrud, *Reproducible Research with R and RStudio* and Y. Xie, *Dynamic Documents with R and knitr* (both second edition, Boca Raton, FL, Chapman and Hall/CRC, 2015). for example.

11.2

Review of basic probability theory

This section contains an overview of the basic theory used in the remainder of this chapter. No attempt is made to explain or justify the ideas or results: a full account can be found in Chapter 13 of *Modern Engineering Mathematics* (MEM) or elsewhere. For the same reason there are no examples or exercises. In the process of reviewing the basic theory, this section also establishes the pattern of notation used throughout the chapter, which follows standard conventions as far as possible. No reader should embark on this chapter without having a fairly thorough understanding of the material in this section.

11.2.1 The rules of probability

We associate a probability $P(A)$ with an **event** A , which in general is a subset of a **sample space** S . The usual set-theoretic operations apply to the events (subsets) in S , and there are corresponding rules that must be satisfied by the probabilities.

Complement rule

$$P(S - A) = 1 - P(A)$$

The **complement** $S - A$ of an event A is often written as \bar{A} .

Addition rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For **disjoint** events, $A \cap B = \emptyset$, and the addition rule takes the simple form

$$P(A \cup B) = P(A) + P(B)$$

Product rule

$$P(A \cap B) = P(A)P(B | A)$$

This is actually the definition of the **conditional probability** $P(B|A)$ of an event B given that an event A has occurred. If A and B are **independent** then the product rule takes the simple form

$$P(A \cap B) = P(A)P(B)$$

since the occurrence of B does not depend on the occurrence of A . A more often used expression for this conditional probability $P(B|A)$ is

$$P(B|A) = \frac{P(A \cap B)}{P(B)}$$

provided $P(B) > 0$.

11.2.2 Random variables

A **random variable** has a sample space of possible numerical values together with a **distribution** of probabilities. Random variables can be either **discrete** or **continuous**. For a discrete random variable (X , say) the possible values can be written as a list $\{v_1, v_2, v_3, \dots\}$ with corresponding probabilities $P(X = v_1), P(X = v_2), P(X = v_3), \dots$. The **mean** of X is then defined as

$$\mu_X = \sum_k v_k P(X = v_k)$$

(sum over all possible values), and is a measure of the central location of the distribution. The **variance** of X is defined as

$$\text{Var}(X) = \sigma_X^2 = \sum_k (v_k - \mu_X)^2 P(X = v_k)$$

and is a measure of dispersion or spread of the distribution about the mean. The symbols μ and σ^2 are conventional for these quantities. In general, the **expected value** of a function $h(X)$ of X is defined as

$$E\{h(X)\} = \sum_k h(v_k) P(X = v_k)$$

of which the mean and variance are special cases obtained by setting $h(x) = x$ and $h(x) = (x - \mu_X)^2$. The **standard deviation** σ_X is the square root of the variance. If the random variable X has units associated with it, then μ_X and σ_X have the same units, while σ_X^2 is measured in the square of the original units.

For a **continuous** random variable (X , say), there is a **probability density function** $f_X(x)$ and a **cumulative distribution function** $F_X(x)$. The cumulative distribution function is defined as

$$F_X(x) = P(X \leq x)$$

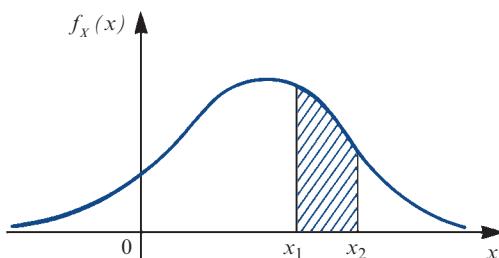
and is the definite integral of the density function:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

This function determines the probabilities of events for the continuous random variable X . The probability that X takes a value within the real interval (x_1, x_2) is the area under the density function over that interval, or equivalently the difference in values of the distribution function at its ends:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f_X(t) dt = F_X(x_2) - F_X(x_1)$$

Figure 11.1
Probability of interval from density function.



(see Figure 11.1). Note that the events $x_1 < X < x_2$, $x_1 \leq X < x_2$, $x_1 < X \leq x_2$ and $x_1 \leq X \leq x_2$ are all equivalent in probability terms for a continuous random variable X because the probability of X being exactly equal to either x_1 or x_2 is zero. The mean and variance of X , and the expected value of a function $h(X)$, are defined in terms of the probability density function by

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x) dx$$

$$\text{Var}(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$$

$$E\{h(X)\} = \int_{-\infty}^{\infty} h(x) f_X(x) dx$$

These definitions assume that the random variable is defined for values of x from $-\infty$ to ∞ . If the random variable is defined in general for values in some real interval, say (a, b) , then the domain of integration can be restricted to that interval, or alternatively the density function can be defined to be zero outside that interval.

Just as events can be independent (and then obey a simple product rule of probabilities), so can random variables be independent. We shall consider this in more detail in Section 11.4. Means and variances of random variables (whether discrete or continuous) have the following important properties (X and Y are random variables, and c is an arbitrary constant):

$$E(cX) = cE(X) = c\mu_X$$

$$\text{Var}(cX) = c^2 \text{Var}(X) = c^2 \sigma_X^2$$

$$E(X + c) = E(X) + c = \mu_X + c$$

$$\text{Var}(X + c) = \text{Var}(X) = \sigma_X^2$$

$$E(X + Y) = E(X) + E(Y) = \mu_X + \mu_Y$$

(this applies whether or not X and Y are independent),

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) = \sigma_X^2 + \sigma_Y^2$$

(this applies only when X and Y are independent).

It is also useful to note that $\text{Var}(X) = E(X^2) - [E(X)]^2$.

Please also note that

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

can provide us with an easy way of finding $\text{Var}(X)$ from $E(X)$ and $E(X^2)$. It is not the definition of $\text{Var}(X)$ but a consequence of the definition

$$\text{Var}(X) = E[(X - E[X])^2].$$

11.2.3 The Bernoulli, binomial and Poisson distributions

The simplest example of a discrete distribution is the **Bernoulli distribution**. This has just two values: $X = 1$ with probability p and $X = 0$ with probability $1 - p$, from which the mean and variance are p and $p(1 - p)$ respectively.

The binomial and Poisson distributions are families of discrete distributions whose probabilities are generated by formulae, and which arise in many real situations. The **binomial distribution** governs the number (X , say) of ‘successes’ in n independent ‘trials’, with a probability p of ‘success’ at each trial:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

where the range of possible values k is $\{0, 1, 2, \dots, n\}$. The binomial distribution can be thought of as the sum of n independent Bernoulli random variables. This distribution (more properly, *family* of distributions) has two **parameters**, n and p . In terms of these parameters, the mean and variance are

$$\mu_X = np$$

$$\sigma_X^2 = np(1 - p)$$

The **Poisson distribution** is defined as

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where the range of possible values k is the set of non-negative integers $\{0, 1, 2, \dots\}$. This has mean and variance both equal to the single parameter λ , and, by setting $\lambda = np$, provides a useful approximation to the binomial distribution that works when n is large and p is small. As a guide, the approximation can be used when $n \geq 25$ and $p \leq 0.1$. The Poisson distribution has many other uses, as will be seen in Section 11.9.



It is easy to calculate binomial and Poisson probabilities in R:

```
dbinom(2, 4, 0.6) # binomial probability with k = 2, n = 4
# and p = 0.6
dpois(4, 2) # Poisson probability with k = 4 and lambda = 2
```

These R commands can be typed into the RStudio’s R script editor window (available initially by means of *File* → *New File* → *R Script*) and then run by R in the Console window by clicking on ‘Run’ or by pressing ‘Ctrl’ and ‘Enter’ together on a Windows machine. Both *dbinom* and *dpois* are examples of R functions. The symbol *#* is used to denote a comment, that will not be processed by R. Here are the results, indicated throughout using *#>*:

```
#> [1] 0.3456
#> [1] 0.09022352
```

The index [1] means that this is the first element of the output; in fact, here there is just one element. Compare this with the output of the `rpois` function that generates 70 random numbers from a Poisson distribution with $\lambda = 2$:

```
rpois(70, 2)
#> [1] 2 0 1 2 2 2 2 1 4 2 1 2 1 1 1 1 1 1 0 4 3 3 0 3 1
1 0 1 3 1 1 1 4 4 2
#> [36] 3 0 3 2 3 2 4 1 3 2 1 3 1 3 1 1 6 2 2 4 2 0 1 1
1 0 2 1 2 1 1 3 1 0 1
```

It is easy to see the help file for an R function:

```
?dbinom
help("dbinom") # Alternative
```

R functions have arguments contained in brackets: (*argument_1*, *argument_2*), for example. It can be seen from the help file (not shown here, but available using R) that `dbinom` has four arguments: *x*, here the value of the number of successes *k*; *size*, the value of the number of trials *n*; *prob*, the probability of success *p*; and *log* which would cause R to report the natural logarithm of $P(X = k)$ if it were set to TRUE (by default *log* is set to FALSE so that this argument can be ignored for our purposes). Hence, the following R commands yield the same results and illustrate the use of arguments:

```
dbinom(2, 4, 0.6) # Arguments specified by order
#> [1] 0.3456
dbinom(x = 2, size = 4, prob = 0.6) # Arguments specified
# by name
#> [1] 0.3456
# Argument order is not important provided arguments are
# specified by name
dbinom(size = 4, prob = 0.6, x = 2)
#> [1] 0.3456
# Do not compute log probability (default)
p <- dbinom(x = 2, size = 4, prob = 0.6, log = FALSE); p
#> [1] 0.3456
# Now compute the natural log of the probability
dbinom(x = 2, size = 4, prob = 0.6, log = TRUE)
#> [1] -1.062473
log(p) # Check
#> [1] -1.062473
```

The term $\binom{n}{k}$ can be computed in R using the `choose` function. Here is an example with *n* = 6 and *k* = 2:

```
choose(6, 2)
#> [1] 15
```

11.2.4 The normal distribution

This is a family of continuous distributions with probability density function given by

$$f_X(x) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x - \mu_X}{\sigma_X}\right)^2\right]$$

for $-\infty < x < +\infty$, where the parameters μ_x and σ_x^2 are the mean and standard deviation of the distribution. It is conventional to denote the fact that a random variable X has a normal distribution by

$$X \sim N(\mu_x, \sigma_x^2)$$

The **standard normal distribution** is a special case with zero mean and unit variance, often denoted by Z :

$$Z \sim N(0, 1)$$

Tables of the standard normal cumulative distribution function

$$\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt$$

are widely available (see, for example, Figure 11.2). These tables can be used for probability calculations involving arbitrary normal random variables. For example, if $X \sim N(\mu_x, \sigma_x^2)$ then

$$P(X \leq a) = P\left(\frac{X - \mu_x}{\sigma_x} \leq \frac{a - \mu_x}{\sigma_x}\right) = \Phi\left(\frac{a - \mu_x}{\sigma_x}\right)$$

We will refer to this method of computing the probability as the standardization approach. R can perform probability calculations as we will see below.

The key result for applications of the normal distribution is the **central limit theorem**: if $\{X_1, X_2, X_3, \dots, X_n\}$ are independent and identically distributed random variables (the distribution being arbitrary), each with mean μ_x and finite variance σ_x^2 , and if

$$W_n = \frac{X_1 + \dots + X_n}{n}, \quad Z_n = \frac{X_1 + \dots + X_n - n\mu_x}{\sigma_x \sqrt{n}}$$

then, as $n \rightarrow \infty$, the distributions of W_n and Z_n tend to $W_n \sim N(\mu_x, \sigma_x^2/n)$ and $Z_n \sim N(0, 1)$ respectively. Loosely speaking, the sum and the mean of independent identical distributed random variables tend to normal distributions.

The central limit theorem is the key to many statistical processes, some of which are described in Section 11.3. One corollary is that the normal distribution can be used to approximate the binomial distribution when n is sufficiently large: if X is binomial with parameters n and p then the approximating distribution (by equating the means and variances) is $Y \sim N(np, np(1-p))$. This is explained (together with the important **continuity correction**) in Section 13.5.5 of MEM, and the approximation is used as follows:

$$\begin{aligned} P(X \leq k) &\approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right) \\ P(X = k) &\approx \Phi\left(\frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$



As a guide, the approximation can be used when $n \geq 25$ and $0.1 \leq p \leq 0.9$. However, as we have seen, binomial probabilities can be computed in R.

Figure 11.2 Table of the standard normal cumulative distribution function $\Phi(z)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
z	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417	
$\Phi(z)$.90	.95	.975	.99	.995	.999	.9995	.99995	.999995	
$2[1 - \Phi(z)]$.20	.10	.05	.02	.01	.002	.001	.0001	.00001	



It is easy to use R to compute the values of $\Phi(z)$ given in Figure 11.2. R provides more decimal places than Figure 11.2. Here are some examples:

```
pnorm(0) # z = 0, giving P(Z <= 0) = Phi(0)
#> [1] 0.5
pnorm(1.52) # z = 1.52, giving P(Z <= 1.52) = Phi(1.52)
#> [1] 0.9357445
pnorm(1.645)
#> [1] 0.9500151
pnorm(1.960)
#> [1] 0.9750021
1 - pnorm(1.645) # P(Z > 1.645) = 1 - Phi(1.645)
#> [1] 0.04998491
pnorm(1.645, lower.tail = FALSE) # easier and better way
# of computing the same
# probability
#> [1] 0.04998491
2 * (1 - pnorm(1.645)) # 2 [1 - Phi(1.645)]
#> [1] 0.09996981
2 * (1 - pnorm(1.960)) # 2 [1 - Phi(1.960)]
#> [1] 0.04999579
```

If the random variable $X \sim N(3,25)$, so that $\mu_X = 3$ and $\sigma_X^2 = 25$ ($\sigma_X = 5$), we can calculate $P(X \leq 2)$ as follows:

```
pnorm((2 - 3) / 5) # using the standardization approach
# formula Phi((x - mu_X) / sigma_X)
#> [1] 0.4207403
```

or more simply as

```
pnorm(2, 3, 5)
```

11.2.5 Sample measures

It is conventional to denote a random variable by an upper-case letter (X , say), and an actual observed value of it by the corresponding lower-case letter (x , say). An observed value x will be one of the set of possible values (sample space) for the random variable, which for a discrete random variable may be written as a list of the form $\{v_1, v_2, v_3, \dots\}$. It is possible to observe a random variable many times (say n times) and obtain a series of values. In this case we assume that the random variable X refers to a **population** (whose characteristics may be unknown), and we refer to the series of random variables $\{X_1, X_2, \dots, X_n\}$ as a **sample**. Each X_i is assumed to have the characteristics of the population, so they all have the same distribution. The actual series of values $\{x_1, x_2, \dots, x_n\}$ consists of data upon which we can work. It is useful to define certain sample measures in terms of the random variables $\{X_1, X_2, \dots, X_n\}$ in order to produce analytical procedures for data. Principal among these measures are the **sample average** and **sample variance**, defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_X^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

respectively. To make calculations shorter it is useful to note that the sample variance is the average of the squares minus the square of the average:

$$S_x^2 = \bar{X}^2 - (\bar{X})^2,$$

in which $\bar{X}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2$. This form of S_x^2 can suffer considerably from numerical problems. We shall also need the following alternative definition of sample variance in Section 11.3.5:

$$S_{x,n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

This is the definition used by R and by many other statistics packages. We can use the properties of means and variances (summarized in Section 11.2.2) to find the mean and variance of the sample average as follows:

$$E(\bar{X}) = \frac{1}{n} E(X_1 + \dots + X_n) = \frac{1}{n} [E(X_1) + \dots + E(X_n)]$$

$$= \frac{n\mu_X}{n} = \mu_X$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \text{Var}(X_1 + \dots + X_n) = \frac{1}{n^2} [\text{Var}(X_1) + \dots + \text{Var}(X_n)]$$

$$= \frac{n\sigma_X^2}{n^2} = \frac{\sigma_X^2}{n}$$

Here we are assuming that the population mean and variance are μ_X and σ_X^2 respectively (which may be unknown values in practice), and that the observations of the random variables X_i are *independent*, a very important requirement in statistics.



We can use R to calculate the sample average \bar{x} and the alternative sample variance $s_{x,n-1}^2$ of a series $\{x_1, x_2, \dots, x_n\}$ of values:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_{x,n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

For example, let us assume that we observe $n = 3$ values $x_1 = 2$, $x_2 = 5$ and $x_3 = 4$. These can be entered into R using:

```
x <- c(2, 5, 4)
# <- can be thought of as the assignment operator
# the function c collects together the data
# so: x is assigned to the collection of data 2, 5, 4
x # To show the contents of an R object just type its name
#> [1] 2 5 4
```

Individual elements can be accessed using [element number(s)]:

```
x[2] # Second element
#> [1] 5
x[2:3] # Elements 2 to 3
#> [1] 5 4
```

```

x[c(1,3)] # Elements 1 and 3
#> [1] 2 4
x[-c(1,3)] # Elements except 1 and 3
#> [1] 5

```

The sample quantities \bar{x} and $s_{X,n-1}^2$ can be computed as:

```

mean(x)
#> [1] 3.666667
var(x)
#> [1] 2.333333

```

These and other R functions have been carefully developed with numerical accuracy in mind.

11.3

Estimating parameters

11.3.1 Interval estimates and hypothesis tests

The first step in statistics is to take some data from an experiment and make inferences about the values of certain parameters. Such parameters could be the mean and variance of a population, or the correlation between two variables for a population. The data are never sufficient to determine the values exactly, but two kinds of inferences can be made:

- (a) a range of values can be computed from the observed data, with intervals computed in this way from repeatedly sampled data containing the population parameter with high probability, or
- (b) a decision can be made as to whether or not the data are compatible with a particular value (or range of values) of the parameter.

The first of these is called **interval estimation**, and provides an assessment of the value that is rather more honest than merely quoting a single number derived from the sample data, which may be more or less uncertain depending upon the sample size. The second approach is called **hypothesis testing** and allows a value of particular interest to be assessed. These two approaches are usually covered in separate chapters in introductory textbooks on statistics, but they are closely related and are often used in conjunction with each other. Tests of simple hypotheses about a specified parameter value will therefore be covered here within the context of interval estimation.

11.3.2 Distribution of the sample average

Suppose that a clearly identified population has a numerical characteristic with an unknown mean value, such as the mean lifetime for a kind of electronic component or the mean salary for a job category. A natural way to estimate this unknown mean is to take a sample from the population, measure the appropriate characteristic, and find the average value. If the sample size is n and the measured values are $\{x_1, x_2, \dots, x_n\}$ then the average value

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

is a reasonable estimate of the population mean μ_x provided that the sample is *representative* and *independent*, and the size n is sufficiently large.

We can be more precise about how useful this estimate is if we treat the sample average as a random variable. Now we have a sample $\{X_1, X_2, \dots, X_n\}$ with average

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and the mean and variance of \bar{X} are given by

$$E(\bar{X}) = \mu_x, \quad \text{Var}(\bar{X}) = \frac{\sigma_x^2}{n}$$

(see Section 11.2.5). This shows that the expected value of the average is indeed equal to the population mean, and that the variance decreases with sample size n and so is smaller for larger samples. However, we can go further. The central limit theorem (Section 11.2.4) tells us that sums of identical random variables tend to have a normal distribution regardless of the distribution of the variables themselves. The only requirement is that a sufficient number of variables contribute to the sum (the actual number required depends very much on the shape of the underlying distribution).

The sample average is a sum of random variables, and therefore has (approximately) a normal distribution for a sufficiently large sample:

$$\tilde{X} \sim N(\mu_x, \sigma_x^2/n)$$

This allows us to use a general method of inference concerning means instead of a separate method for each underlying distribution – even if this were known, which is usually not the case. In practice, a sample size of 25 or more is usually sufficient for the normal approximation.

Example 11.1

For all children taking an examination, the mean mark was 60%, with a standard deviation of 8%. A particular class of 30 children achieved an average of 63%. Is this unusual?

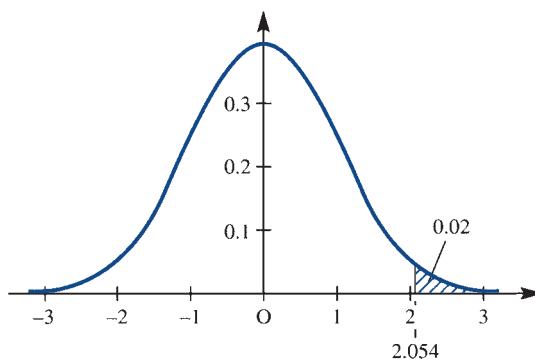
Solution

The average of 63% is higher than the mean, but not by very much. We do not know the true distribution of marks, but the sample average has (approximately) a normal distribution. We can test the idea that this particular class result is a fluke by reducing the sample average to a standard normal in the manner described in Section 11.2.4 and checking its value against the table of the cumulative distribution function $\Phi(z)$ (Figure 11.2):

$$\begin{aligned} P(\bar{X} \geq 63) &= P\left(\frac{\bar{X} - 60}{8/\sqrt{30}} \geq \frac{63 - 60}{8/\sqrt{30}}\right) = P(Z \geq 2.054) = 1 - P(Z < 2.054) \\ &= 1 - \Phi(2.054) = 0.020 \end{aligned}$$

It is unlikely (one chance in 50) that an average as high as this could occur by chance, assuming that the ability of the class is typical. Figure 11.3 illustrates that the result is towards the tail of the distribution. It therefore seems that this class is unusually successful.

Figure 11.3 Normal density function for Example 11.1.



It is useful to see how to perform this probability calculation in a general way in R:

```
mu <- 60 # Mean for all children
sigma <- 8 # Standard deviation for all children
n <- 30 # Sample size
x_bar <- 63 # Sample mean
# Standardization approach
z <- (x_bar - mu) / (sigma / sqrt(n)); z
#> [1] 2.05396
      # Here we separate two R commands using a semicolon to
      # save space
      # In practice, it is usually better to put each R command
      # on a separate line
1 - pnorm(z); pnorm(z, lower.tail = FALSE)
#> [1] 0.0199898
#> [1] 0.0199898
# Direct calculation
pnorm(x_bar, mu, sigma / sqrt(n), lower.tail = FALSE)
#> [1] 0.0199898
```

11.3.3 Confidence interval for the mean

A useful notation will be introduced here. For the standard normal distribution, define z_α to be the point on the z axis for which the area under the density function to its right is equal to α :

$$P(Z > z_\alpha) = \alpha$$

or equivalently

$$\Phi(z_\alpha) = 1 - \alpha$$

(see Figure 11.4a). From the standard normal table we have $z_{0.05} = 1.645$ and $z_{0.025} = 1.96$. By symmetry

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

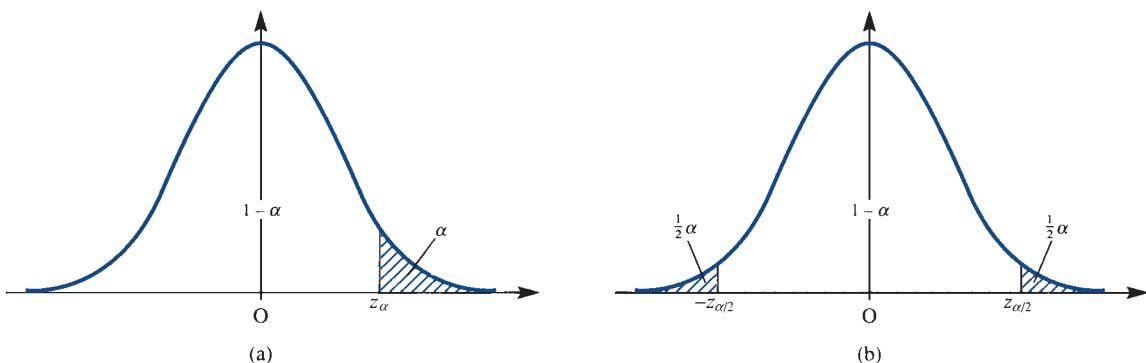


Figure 11.4 Normal density functions with (a) z_α and (b) $z_{\alpha/2}$.

(see Figure 11.4b). Assuming normality of the sample average, we have

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

which, after multiplying through the inequality by σ_X/\sqrt{n} and changing the sign, gives

$$P\left(-z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}} < \mu_X - \bar{X} < z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}\right) = 1 - \alpha$$

so that

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}} < \mu_X < \bar{X} + z_{\alpha/2} \frac{\sigma_X}{\sqrt{n}}\right) = 1 - \alpha$$

Assume for now that the standard deviation of X is known (it is actually very rare for σ_X to be known when μ_X is unknown, but we shall discuss this case first for simplicity and later consider the more general situation where both μ_X and σ_X are unknown).

The interval defined by $(\bar{X} \pm z_{\alpha/2} \sigma_X/\sqrt{n})$ is called a **100(1 - α)% confidence interval for the mean**, with variance known. If a value for α is specified, the upper and lower limits of this interval can be calculated from the sample average. The probability is $1 - \alpha$ that this random interval contains the true mean.

Example 11.2

The temperature (in degrees Celsius) at ten points chosen at random in a large building is measured, giving the following list of readings:

$$\{18^\circ, 16.5^\circ, 17.5^\circ, 18^\circ, 19.5^\circ, 16.5^\circ, 18^\circ, 17^\circ, 19^\circ, 17.5^\circ\}$$

The standard deviation of temperature through the building is known from past experience to be 1°C . Find a 90% confidence interval for the mean temperature in the building.

Solution The average of the ten readings is 17.75°C , and, using $z_{0.05} = 1.645$, the 90% confidence interval is

$$(17.75 \pm 1.645(1/\sqrt{10})) = (17.2, 18.3)$$



This confidence interval can be computed in R as follows:

```
# Input the data
temperature <- c(18, 16.5, 17.5, 18, 19.5, 16.5, 18, 17,
19, 17.5)
sigma <- 1 # Assumed known
n <- length(temperature) # Sample size
x_bar <- mean(temperature) # Sample mean
se <- sigma / sqrt(n) # Standard error of the mean
alpha <- 0.1 # As 90% confidence interval required
z <- qnorm(alpha / 2, lower.tail = FALSE); z
# We want z such that P(Z > z) = alpha / 2
#> [1] 1.644854
# Confidence interval
c(x_bar - z*se, x_bar + z*se)
#> [1] 17.22985 18.27015
```

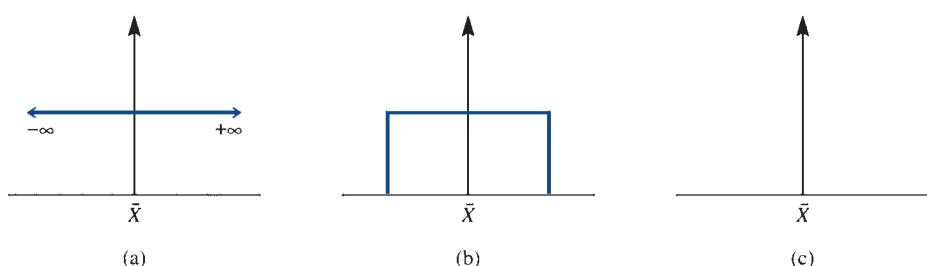
The confidence interval is used to indicate the degree of uncertainty in the sample average. The simplicity of the calculation is deceptive because the idea is very important and easily misunderstood. It is not the population mean that is random but rather the interval that would enclose it $100(1 - \alpha)\%$ of the times the experiment is performed. It is tempting to think of the interval as fixed by the experiment and the mean as a random variable that has a probability $1 - \alpha$ of lying within it, but this is not correct.

Typical values of α are 0.1, 0.05 and 0.01, giving 90%, 95% and 99% confidence intervals respectively. The value chosen is a compromise between truth and precision, as illustrated in Figure 11.5. Loosely speaking, a statement saying that the mean lies within the interval $(-\infty, \infty)$ is 100% true (certain to be the case), but totally uninformative because of its total imprecision. None of the possible values is ruled out. On the other hand, saying that the mean equals the exact value given by the sample average is maximally precise, but again of limited value because the statement is false – or rather the probability of its truth is zero. A statement quoting a finite interval for the mean has a probability of being true, chosen to be quite high, and at the same time it rules out most of the possible values and therefore is highly informative. The higher the probability of truth, the lower the informativeness, and vice versa.

The width of the interval $2z_{\alpha/2}\sigma_x/\sqrt{n}$ also depends on the sample size n . A larger experiment yields a more precise result. If figures for the confidence $1 - \alpha$ and precision (width of the interval) are specified in advance then the sample size can be chosen sufficiently large to satisfy these requirements. In some experimental situations

Figure 11.5

Confidence intervals:
(a) infinite interval;
(b) finite interval;
(c) point value.



(for example, destructive testing) there are incentives to keep sample sizes as small as possible. The experimenter must weigh up these conflicting objectives and design the experiment accordingly.

Example 11.3

A machine fills cartons of liquid; the mean fill is adjustable but the dial on the gauge is not very accurate. The standard deviation of the quantity of fill is 6 ml. A sample of 30 cartons gave a measured average content of 570 ml. Find 90% and 95% confidence intervals for the mean.

Solution Using $\alpha = 0.05$ and $z_{0.025} = 1.96$, the 95% confidence interval is

$$(570 \pm 1.960(6/\sqrt{30})) = (567.8, 572.1)$$

Likewise, using $\alpha = 0.1$ and $z_{0.05} = 1.645$, the 90% confidence interval is

$$(570 \pm 1.645(6/\sqrt{30})) = (568.2, 571.8)$$

As expected, the 95% interval is slightly wider.

11.3.4 Testing simple hypotheses

As explained in Section 11.3.1, the testing of hypotheses about parameter values is complementary to the estimation process involving an interval. A ‘simple’ hypothesis is one that specifies a particular value for the parameter, as opposed to an interval, and it is this type that we shall consider. The following remarks apply generally to parameter hypothesis testing, but will be directed in particular to hypotheses concerning means.

There are two kinds of errors that can occur when testing hypotheses:

- (a) a true hypothesis can be rejected (this is usually referred to as a **type I error**), or
- (b) a false hypothesis can be accepted (this is usually called a **type II error**).

In reality, all simple hypotheses that prescribe particular values for parameters are false, but they may be approximately true and rejection may be the result of an experimental fluke. This is the sense in which a type I error can occur. Any such hypothesis will be rejected if the sample size is large enough. Acceptance really means that there is insufficient evidence to reject the hypothesis, but this is not an entirely negative view because if the hypothesis has survived the test then it has some degree of dependability.

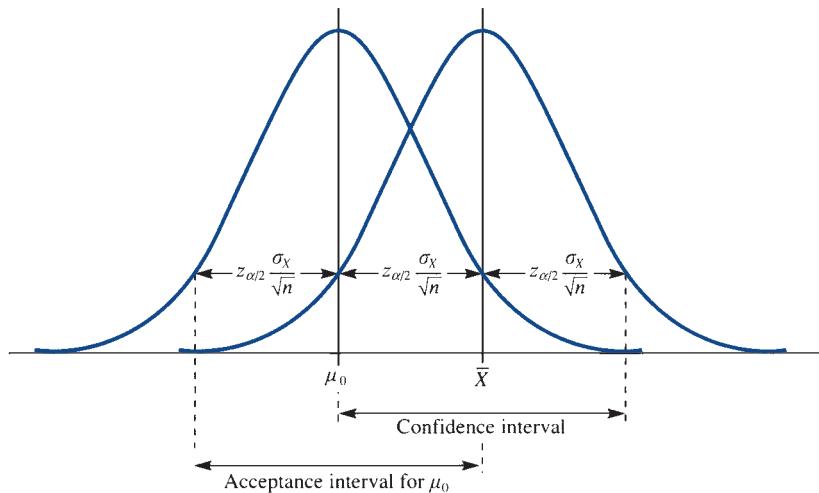
Normally a simple hypothesis is tested by evaluating a **test statistic**, a quantity that depends upon the sample and leads to rejection of the hypothesized parameter value if its magnitude exceeds a certain threshold. If the hypothesized mean is μ_0 then the test statistic for the mean is

$$Z = \frac{X - \mu_0}{\sigma_x / \sqrt{n}}$$

with the hypothesis ‘rejected at significance level α' if $|Z| > z_{\alpha/2}$.

Figure 11.6

Confidence interval
and hypothesis test.



The **significance level** can be regarded as the probability of false rejection, an error of type I. If the hypothesis is true then Z has a standard normal distribution and the probability that it will exceed $z_{\alpha/2}$ in magnitude is α . If Z does exceed this value then either the hypothesis is wrong or else a rare event has occurred. It is easy to show that the test statistic lies on this threshold (for significance level α) exactly when the hypothesized value lies at one or other extreme of the $100(1 - \alpha)\%$ confidence interval (see Figure 11.6). An alternative way to test the hypothesis is therefore to see whether or not the value lies within the confidence interval. For example, the hypothesis that the mean takes the value μ_0 would be rejected at significance level $\alpha = 0.05$ if the corresponding 95% confidence interval does not contain μ_0 .

Example 11.4

For the situation described in Example 11.3 test the hypothesis that the mean fill of liquid is 568 ml (one imperial pint).

Solution

The value of the test statistic is

$$Z = \frac{570 - 568}{6/\sqrt{30}} = 1.83$$

This exceeds $z_{0.05} = 1.645$ (10% significance), but is less than $z_{0.025} = 1.96$ (5% significance). Alternatively, the quoted figure lies within the 95% confidence interval but outside the 90% confidence interval. Either way, the hypothesis is rejected at the 10% significance level but accepted at the 5% level. If the actual mean is 568 ml then there is less than one chance in 10 (but more than one in 20) that a result as extreme as 570 ml will be obtained. It looks as though the true mean is larger than the intended value, but the evidence is not particularly strong. The probability of false rejection (type I error) is somewhere between 5% and 10%, which is small but not negligible.

Examples 11.3 and 11.4 set the pattern for the interpretation and use of confidence intervals. We shall now see how to apply these ideas more generally.

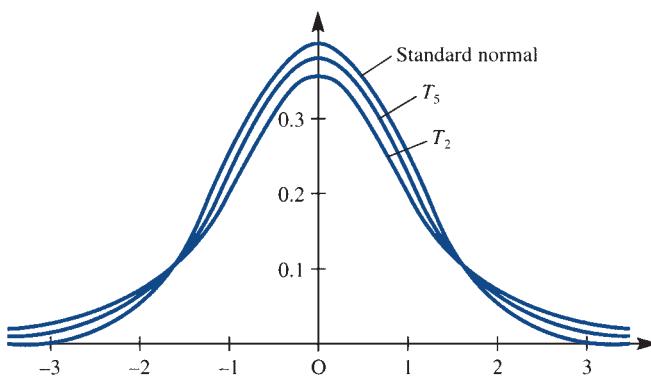
11.3.5 Other confidence intervals and tests concerning means

Mean when variance is unknown

With the basic ideas of interval estimation and hypothesis testing established, it is relatively easy to cover other cases. The first and most obvious is to remove the assumption that the variance is known. If the sample size is large then there is essentially no problem, because the sample standard deviation $S_{X,n-1}$ can be used in place of σ_X in the confidence interval, where

$$S_{X,n-1}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Figure 11.7 Density functions of T_n and z .



This definition was introduced in Section 11.2.5. Note that the sum is divided by $n-1$ rather than n . For a large sample this makes little difference, but for a small sample this form must be used because the ‘*t* distribution’ requires it.

Suppose that the sample size is small, say less than 25. Using $S_{X,n-1}$ in place of σ_X adds an extra uncertainty because this estimate is itself subject to error. Furthermore, the central limit theorem cannot be relied upon to ensure that the sample average has a normal distribution. We have to assume that the data themselves are normal. In this situation the random variable

$$T_n = \frac{\bar{X} - \mu_X}{S_{X,n-1}/\sqrt{n}}$$

has a ***t* distribution** with parameter $n-1$. This distribution resembles the normal distribution, as can be seen in Figure 11.7, which shows the density functions of T_2 and T_5 together with that of the standard normal distribution. In fact T_n tends to the standard normal distribution as $n \rightarrow \infty$. The parameter of the *t* distribution (whose value here is one less than the size of the sample) is usually called the **number of degrees of freedom**.

Defining $t_{\alpha,n-1}$ by

$$P(T_n > t_{\alpha,n-1}) = \alpha$$

(by analogy with z_α), we can derive a $100(1-\alpha)\%$ confidence interval for the mean by the method used in Section 11.3.3:

$$\left(\bar{X} \pm t_{\alpha/2,n-1} \frac{S_{X,n-1}}{\sqrt{n}} \right)$$

This takes explicit account of the uncertainty caused by the use of $S_{X,n-1}$ in place of σ_X . Values of $t_{\alpha,n-1}$ for typical values of α can be read directly from the table of the t distribution, an example of which is shown in Figure 11.8. To obtain a test statistic for an assumed mean μ_0 , simply replace μ_X by μ_0 in the definition of T_n .

Figure 11.8 Table of the t distribution $t_{\alpha,n}$. (Based on Table 12 of *Biometrika Tables for Statisticians*, Volume 1. Cambridge University Press, 1954. By permission of the *Biometrika* trustees.)

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	n
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
∞	1.282	1.645	1.960	2.326	2.576	∞

These tabulated values can be easily calculated in R. Here is an example:

```
alpha <- 0.025
n <- 6 # Number of degrees of freedom
qt(alpha, n, lower.tail = FALSE)
#> [1] 2.446912
```

Example 11.5

The measured lifetimes of a sample of 20 electronic components gave an average of 1250 h, with a sample standard deviation $S_{X,n-1}$ of 96 h. Assuming that the lifetime has a normal distribution, find a 95% confidence interval for the mean lifetime of the population, and test the hypothesis that the mean is 1300 h.

Solution The appropriate figure from the t table is $t_{0.025,19} = 2.093$, so the 95% confidence interval is

$$(1250 \pm 2.093(96)/\sqrt{20}) = (1205, 1295)$$

The claim that the mean lifetime is 1300 h is therefore rejected at the 5% significance level. The same conclusion is reached by evaluating

$$T_n = \frac{1250 - 1300}{96/\sqrt{20}} = -2.33$$

which exceeds $t_{0.025,19}$ in magnitude.



Here are the lifetimes of 5 electronic components in hours: 1106, 1251, 1368, 1101 and 1266. We can test the hypothesis that $\mu_0 = 1300$ h in R as follows:

```
lifetimes <- c(1106, 1251, 1368, 1101, 1266)
t.test(lifetimes, mu = 1300)
#>
#> One Sample t-test
#>
#> data: lifetimes
#> t = -1.5984, df = 4, p-value = 0.1852
#> alternative hypothesis: true mean is not equal to 1300
#> 95 percent confidence interval:
#> 1076.658 1360.142
#> sample estimates:
#> mean of x
#> 1218.4
```

The hypothesis that $\mu_0 = 1300$ h would be rejected at significance level α if the p-value were less than α . Here $\mu_0 = 1300$ h would not be rejected by a test of significance level 0.05 (5%) or 0.10 (10%). R also calculated a 95% confidence interval. A 90% confidence interval can be found using

```
t.test(lifetimes, mu = 1300, conf.level = 0.9)$conf.int
#> [1] 1109.566 1327.234
#> attr(, "conf.level")
#> [1] 0.9
```

Both these intervals contain 1300 h.

Difference between means

Now suppose that we have not just a single sample but two samples from different populations, and that we wish to compare the separate means. Assume also that the variances of the two populations are equal but unknown (the most common situation). Then it can be shown that the $100(1 - \alpha)\%$ confidence interval for the difference $\mu_1 - \mu_2$ between the means is

$$\left(\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2,n} S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right)$$

where \bar{X}_1 and \bar{X}_2 are the respective sample averages, n_1 and n_2 are the respective sample sizes, S_1^2 and S_2^2 are the respective sample variances (using the ' $n - 1$ ' form as above),

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is a pooled estimate of the unknown variance, and

$$n = n_1 + n_2 - 2$$

is the parameter for the t table. The corresponding test statistic for an assumed difference $d_0 = \mu_1 - \mu_2$ is

$$T_n = \frac{\bar{X}_1 - \bar{X}_2 - d_0}{S_p \sqrt{(1/n_1 + 1/n_2)}}$$

For small samples the populations have to be normal, but for larger samples this is not required and the t -table figure can be replaced by $z_{\alpha/2}$.

Example 11.6

Two kinds of a new plastic material are to be compared for strength. From tensile strength measurements of 10 similar pieces of each type, the sample averages and standard deviations were as follows:

$$\bar{X}_1 = 78.3, \quad S_1 = 5.6, \quad \bar{X}_2 = 84.2, \quad S_2 = 6.3$$

Compare the population mean strengths, assuming normal data.

Solution

The pooled estimate of the standard deviation is 5.960, the t table gives $t_{0.025,18} = 2.101$, and the 95% confidence interval for the difference between means is

$$(78.3 - 84.2 \pm 2.101(5.96)/\sqrt{5}) = (-11.5, -0.3)$$

The difference is significant at the 5% level because zero does not lie within the interval. Also, assuming zero difference gives

$$T_n = \frac{78.3 - 84.2}{5.96/\sqrt{5}} = -2.21$$

which confirms the 5% significance.

 The tensile strength measures of 4 pieces of plastic of type 1 were 76.4, 72.9, 78.0 and 87.1, and of 5 pieces of plastic of type 2 were 89.5, 96.7, 94.9, 91.2 and 83.4. We can compare the population mean strengths as follows:

```
type_1 <- c(76.4, 72.9, 78.0, 87.1)
type_2 <- c(89.5, 96.7, 94.9, 91.2, 83.4)
t.test(type_1, type_2, mu = 0, var.equal = TRUE)
# Population variances assumed equal
#>
#> Two Sample t-test
#>
```

```
#> data: type_1 and type_2
#> t = -3.3529, df = 7, p-value = 0.0122
#> alternative hypothesis: true difference in means is not
# equal to 0
#> 95 percent confidence interval:
#> -21.383839 -3.696161
#> sample estimates:
#> mean of x mean of y
#> 78.60 91.14
```

As before, conclusions can be based on the p-value or the confidence interval. Here a test of significance level 0.05 (5%) would reject the hypothesis that $\mu_1 - \mu_2 = 0$ or that $\mu_1 = \mu_2$.

It is also possible to set up confidence intervals and tests for the variance σ_x^2 , or for comparing two variances for different populations. The process of testing means and variances within and between several populations is called **analysis of variance**. This has many applications, and is well covered in statistics textbooks. For a detailed treatment with R implementations, see P. Dalgaard, *Introductory Statistics with R* (second edition, New York, Springer, 2008) or N.J. Harton and K. Kleinman, *Using R and RStudio for Data Management, Statistical Analysis, and Graphics* (second edition, Boca Raton, FL, Chapman and Hall, 2015).

11.3.6 Interval and test for proportion

The ideas of interval estimation do not just apply to means. If probability is interpreted as a long-term proportion (which is one of the common interpretations) then measuring a sample proportion is a way of estimating a probability. The binomial distribution (Section 11.2.3) points the way. We count the number of ‘successes’, say X , in n ‘trials’, and estimate the probability p of success at each trial, or the long-term proportion, by the sample proportion

$$\hat{p} = \frac{X}{n}$$

(it is common in statistics to place the ‘hat’ symbol $\hat{\cdot}$ over a parameter to denote an estimate of that parameter). This only provides a point estimate. To obtain a confidence interval, we can exploit the normal approximation to the binomial (Section 11.2.4)

$$X \sim N(np, np(1-p))$$

approximately, for large n . Dividing by n preserves normality, so

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Following the argument in Section 11.3.3, we have

$$P\left(p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} < \hat{p} < p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

and, after rearranging the inequality,

$$P\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Because p is unknown, we have to make a further approximation by replacing p by \hat{p} inside the square root, to give an approximate $100(1 - \alpha)\%$ confidence interval for p :

$$\left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

The corresponding test statistic for an assumed proportion p_0 is

$$Z = \frac{X - np_0}{\sqrt{[np_0(1-p_0)]}}$$

with $\pm z_{\alpha/2}$ as the rejection points for significance level α .

Example 11.7

In an opinion poll conducted with a sample of 1000 people chosen at random, 30% said that they support a certain political party. Find a 95% confidence interval for the actual proportion of the population who support this party.

Solution The required confidence interval is obtained directly as

$$\left(0.3 \pm 1.96 \sqrt{\frac{(0.3)(0.7)}{1000}}\right) = (0.27, 0.33)$$

A variation of about 3% either way is therefore to be expected when conducting opinion polls with sample sizes of this order, which is fairly typical, and this figure is often quoted in the news media as an indication of maximum likely error.



To obtain this approximate interval in R:

```
prop.test(x = 300, n = 1000)$conf.int
#> [1] 0.2719222 0.3296354
#> attr(,"conf.level")
#> [1] 0.95
```

A similar argument that also exploits the fact that the difference between two independent normal random variables is also normal leads to the following $100(1 - \alpha)\%$ confidence interval for the difference between two proportions, when \hat{p}_1 and \hat{p}_2 are the respective sample proportions:

$$\left(\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\left[\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]}\right)$$

Again it is assumed that n_1 and n_2 are reasonably large. The test statistic for equality of proportions is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{[p(1-\hat{p})(1/n_1 + 1/n_2)]}}$$

where $\hat{p} = (X_1 + X_2)/(n_1 + n_2)$ is a pooled estimate of the proportion.

Example 11.8

One hundred samples of an alloy are tested for resistance to fatigue. Half have been prepared using a new process and the other half by a standard process. Of those prepared by the new process, 35 exhibit good fatigue resistance, whereas only 25 of those prepared in the standard way show the same performance. Is the new process better than the standard one?

Solution

The proportions of good samples are 0.7 for the new process and 0.5 for the standard one, so a 95% confidence interval for the difference between the true proportions is

$$\left(0.7 - 0.5 \pm 1.96 \sqrt{\left[\frac{(0.7)(0.3)}{50} + \frac{(0.5)(0.5)}{50} \right]} \right) = (0.01, 0.39)$$

The pooled estimate of proportion is

$$p = (35 + 25)/(50 + 50) = 0.6$$

so that

$$Z = \frac{0.7 - 0.5}{\sqrt{[(0.6)(0.4)/25]}} = 2.04$$

Both approaches show that the difference is significant at the 5% level. However, it is only just so: if one more sample for the new process had been less fatigue-resistant, the difference would not have been significant at this level. This suggests that the new process is effective – but, despite the apparently large difference in success rates, the evidence is not very strong.



To obtain this approximate interval in R:

```
prop.test(x = c(35, 25), n = c(50, 50), correct = FALSE)$conf.int # No continuity correction is applied
#> [1] 0.01200686 0.38799314
#> attr(, "conf. level")
#> [1] 0.95
```

This method only applies to independent sample proportions. It would not be legitimate to apply it, for instance, to a more elaborate version of the opinion poll (Example 11.7) in which respondents can choose between two (or more) political parties or else support neither. Support for one party usually precludes support for another, so the proportions of those interviewed who support the two parties are not independent. More elaborate

confidence intervals, based on the multinomial distribution, can handle such situations. Chapter 7 of J. J. Faraday, *Extending the Linear Model with R* (second edition, Boca Raton, FL, Chapman and Hall/CRC, 2016) is devoted to modelling multinomial data. This shows how important it is to understand the assumptions upon which statistical methods are based. It would be very easy to look up ‘difference between proportions’ in an index and apply an inappropriate formula.

11.3.7 Exercises



Check your calculations using R whenever possible.

- 1** An electrical firm manufactures light bulbs whose lifetime is approximately normally distributed with a standard deviation of 50 h.

- (a) If a sample of 30 bulbs has an average life of 780 h, find a 95% confidence interval for the mean lifetime of the population.
 (b) How large a sample is needed if we wish to be 95% confident that our sample average will be within 10 h of the population mean?

- 2** Monthly rainfall measurements (in mm) were taken at a certain location for three years, with results as follows:

38 48 50 94 105 53 81 91 110 103 90 84
 115 113 35 130 77 67 72 113 98 37 61 91
 9 112 29 16 56 61 82 132 48 68 114 55

Find the average monthly rainfall for this period. Also find a 95% confidence interval for the mean monthly rainfall, using the measured standard deviation as an estimate of the true value.

- 3** Quantities of a trace impurity in 12 specimens of a new material are measured (in parts per million) as follows:

8.8, 7.1, 7.9, 10.2, 8.9, 7.7, 10.6, 9.4, 9.2, 7.5, 9.0, 8.4

Find a 95% confidence interval for the population mean, assuming that the distribution is normal.

- 4** A sample of 30 pieces of a semiconductor material gave an average resistivity of 73.2 mΩ m, with a sample standard deviation of 5.4 mΩ m. Obtain a 95% confidence interval for the resistivity of

the material, and test the hypothesis that this is 75 mΩ m.

- 5** The mean weight loss of 16 grinding balls after a certain length of time in mill slurry is 3.42 g, with a standard deviation of 0.68 g. Construct a 99% confidence interval for the true mean weight loss of such grinding balls under the stated conditions.

- 6** While performing a certain task under simulated weightlessness, the pulse rate of 32 astronaut trainees increased on average by 26.4 beats per minute, with a standard deviation of 4.28 beats per minute. Construct a 95% confidence interval for the true average increase in the pulse rate of astronaut trainees performing the given task.

- 7** The quality of a liquid being used in an etching process is monitored automatically by measuring the attenuation of a certain wavelength of light passing through it. The criterion is that when the attenuation reaches 58%, the liquid is declared as ‘spent’. Ten samples of the liquid are used until they are judged as ‘spent’ by the experts. The light attenuation is then measured, and gives an average result of 56%, with a standard deviation of 3%. Is the criterion satisfactory?

- 8** A fleet car company has to decide between two brands A and B of tyre for its cars. An experiment is conducted using 12 of each brand, run until they wear out. The sample averages and standard deviations of running distance (in km) are respectively 36 300 and 5000 for A, and 39 100 and 6100 for B. Obtain a 95% confidence interval for the difference in means, assuming the distributions to be normal, and test the hypothesis that brand B tyres outrun brand A tyres.

- 9 A manufacturer claims that the lifetime of a particular electronic component is unaffected by temperature variations within the range 0–60 °C. Two samples of these components were tested, and their measured lifetimes (in hours) recorded as follows:

0 °C: 7250, 6970, 7370, 7910, 6790, 6850,
7280, 7830

60 °C: 7030, 7270, 6510, 6700, 7350, 6770,
6220, 7230

Assuming that the lifetimes have a normal distribution, find 90% and 95% confidence intervals for the difference between the mean lifetimes at the two temperatures, and hence test the manufacturer's claim at the 5% and 10% significance levels.

- 10 Suppose that out of 540 drivers tested at random, 38 were found to have consumed more than the legal limit of alcohol. Find 90% and 95% confidence intervals for the true proportion of drivers who were over the limit during the time of the tests. Are the results compatible with the hypothesis that this proportion is less than 5%?
- 11 It is known that approximately one-quarter of all houses in a certain area have inadequate loft

insulation. How many houses should be inspected if the difference between the estimated and true proportions having inadequate loft insulation is not to exceed 0.05, with probability 90%? If in fact 200 houses are inspected, and 55 of them have inadequate loft insulation, find a 90% confidence interval for the true proportion.

- 12 A drug-manufacturer claims that the proportion of patients exhibiting side-effects to their new anti-arthritis drug is at least 8% lower than for the standard brand X. In a controlled experiment 31 out of 100 patients receiving the new drug exhibited side-effects, as did 74 out of 150 patients receiving brand X. Test the manufacturer's claim using 90% and 95% confidence intervals.

- 13 Suppose that 10 years ago 500 people were working in a factory, and 180 of them were exposed to a material which is now suspected as being carcinogenic. Of those 180, 30 have since developed cancer, whereas 32 of the other workers (who were not exposed) have also since developed cancer. Obtain a 95% confidence interval for the difference between the proportions with cancer among those exposed and not exposed, and assess whether the material should be considered carcinogenic, on this evidence.

11.4 Joint distributions and correlation

Just as it is possible for events to be dependent upon one another in that information that one has occurred changes the probability of the other, so it is possible for random variables to be associated in value. In this section we show how the degree of dependence between two random variables can be defined and measured.

11.4.1 Joint and marginal distributions

The idea that two variables, each of which is random, can be associated in some way might seem mysterious at first, but can be clarified with some familiar examples. For instance, if one chooses a person at random and measures his or her height and weight, each measurement is a random variable – but we know that taller people also tend to be heavier than shorter people, so the outcomes will be related. On the other hand, a person's birthday and telephone number are not likely to be related in any

way. In general, we need a measure of the simultaneous distribution of two random variables.

For two discrete random variables X and Y with possible values $\{u_1, \dots, u_m\}$ and $\{v_1, \dots, v_n\}$ respectively, the **joint distribution** of X and Y is the set of all joint probabilities of the form

$$P(X = u_k \cap Y = v_j) \quad (k = 1, \dots, m; j = 1, \dots, n)$$

The joint distribution contains all relevant information about the random variables separately, as well as their joint behaviour. To obtain the distribution of one variable, we sum over the possible values of the other:

$$P(X = u_k) = \sum_{j=1}^n P(X = u_k \cap Y = v_j) \quad (k = 1, \dots, m)$$

$$P(Y = v_j) = \sum_{k=1}^m P(X = u_k \cap Y = v_j) \quad (j = 1, \dots, n)$$

The distributions obtained in this way are called **marginal distributions** of X and Y .

Example 11.9

Two textbooks are selected at random from a shelf containing three statistics texts, two mathematics texts and three engineering texts. Denoting the number of books selected in each subject by S , M and E respectively, find (a) the joint distribution of S and M , and (b) the marginal distributions of S , M and E .

Solution (a)

Figure 11.9
Joint distribution
for Example 11.9.

		<i>M</i>			Total	
<i>S</i>						
	0	1	2			
0	$\frac{3}{28}$	$\frac{3}{14}$	$\frac{1}{28}$	$\frac{5}{14}$		
1	$\frac{9}{28}$	$\frac{3}{14}$		$\frac{15}{28}$		
2	$\frac{3}{28}$			$\frac{3}{28}$		
Total	$\frac{15}{28}$	$\frac{3}{7}$	$\frac{1}{28}$	1		

The joint distribution (shown in Figure 11.9) is built up element by element using the addition and product rules of probability as follows:

$$P(S = M = 0) = P(E = 2) = \left(\frac{3}{8}\right)\left(\frac{2}{7}\right) = \frac{3}{28}$$

that is, the probability that the first book is an engineering text (three chances out of eight) times the probability that the second book is also (two remaining chances out of seven). Continuing,

$$P(S = 0 \cap M = 1) = P(M = 1 \cap E = 1)$$

$$= \left(\frac{2}{8}\right)\left(\frac{3}{7}\right) + \left(\frac{3}{8}\right)\left(\frac{2}{7}\right) = \frac{3}{14}$$

that is, the probability that the first book is a mathematics text and the second an engineering text, plus the (equal) probability of the books being the other way round. The other probabilities are derived similarly.

- (b) The marginal distributions of S and M are just the row and column totals as shown in Figure 11.9. The marginal distribution of E can also be derived from the table:

$$P(E = 2) = P(S = M = 0) = \frac{3}{28}$$

$$P(E = 1) = P(S = 1 \cap M = 0) + P(S = 0 \cap M = 1) = \frac{15}{28}$$

$$P(E = 0) = P(S = 2) + P(S = 1 \cap M = 1) + P(M = 2) = \frac{5}{14}$$

This is the same as the marginal distribution of S , which is not surprising, because there are the same numbers of engineering and statistics books on the shelf.

Note the way that the joint probabilities and the marginal probabilities sum to 1.

In order to apply these ideas of joint and marginal distributions to continuous random variables, we need to build on the interpretation of the probability density function. The **joint density function** of two continuous random variables X and Y , denoted by $f_{X,Y}(x, y)$, is such that

$$P(x_1 < X < x_2 \text{ and } y_1 < Y < y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{X,Y}(x, y) dy dx$$

for all intervals (x_1, x_2) and (y_1, y_2) . This involves a double integral over the two variables x and y . This is necessary because the joint density function must indicate the relative likelihood of every combination of values of X and Y , just as the joint distribution does for discrete random variables. The joint density function is transformed into a probability by integrating over an interval for both variables. The double integral here can be regarded as a pair of single-variable integrations, with the outer variable (x) held constant during the integration with respect to the inner variable (y). In fact the same answer is obtained if the integration is performed the other way around.

The **marginal density functions** for X and Y are obtained from the joint density function in a manner analogous to the discrete case: by integrating over all values of the unwanted variable:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy \quad (-\infty < x < \infty)$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx \quad (-\infty < y < \infty)$$

Example 11.10

The joint density function of random variables X and Y is

$$f_{X,Y}(x, y) = \begin{cases} 1 & (0 \leq x \leq 1, cx \leq y \leq cx + 1) \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant such that $0 \leq c \leq 1$ (which means that $f_{X,Y}(x, y)$ is unity over the trapezoidal area shown in Figure 11.10 and zero elsewhere). Find the marginal distributions of X and Y . Also find the probability that neither X nor Y exceeds one-half, assuming $c = 1$.

Solution To find the marginal distribution of X , we integrate with respect to y :

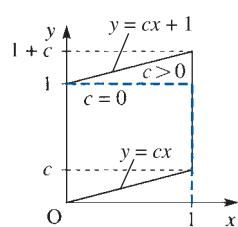


Figure 11.10
Density function for Example 11.10. The dashed unit square indicates the area in which $f_{X,Y}(x,y)$ is non-zero when $c = 0$.

$$f_X(x) = \begin{cases} \int_{cx}^{cx+1} dy = 1 & (0 \leq x \leq 1) \\ 0 & \text{otherwise} \end{cases}$$

The marginal distribution for Y is rather more complicated. Integrating with respect to x and assuming that $0 < c \leq 1$,

$$f_Y(y) = \begin{cases} 1 - \frac{1}{c}(y-1) & (1 \leq y \leq 1+c) \\ 1 & (c \leq y \leq 1) \\ \frac{y}{c} & (0 \leq y \leq c) \end{cases}$$

(Exercise 16). When $c = 0$, the marginal distribution for Y is the same as that for X . Finally, when $c = 1$,

$$P(X \leq \frac{1}{2} \text{ and } Y \leq \frac{1}{2}) = \int_0^{1/2} \int_x^{1/2} 1 dy dx = \int_0^{1/2} \left(\frac{1}{2} - x\right) dx = \frac{1}{8}$$

Here the inner integral (with respect to y) is performed with x treated as constant, and the resulting function of x is integrated to give the answer.

The definitions of joint and marginal distributions can be extended to any number of random variables.

11.4.2 Independence

The idea of independence of events can be extended to random variables to give us the important case in which no information is shared between them. This is important in experiments where essentially the same quantity is measured repeatedly, either within a single experiment involving repetition or between different experiments. As mentioned before, independence within a sample is one of the properties that is assumed by a range of statistical procedures.

Two random variables X and Y are called **independent** if their joint distribution factorizes into the product of their marginal distributions:

$$P(X = u_k \cap Y = v_j) = P(X = u_k)P(Y = v_j) \quad \text{in the discrete case}$$

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \quad \text{in the continuous case}$$

For example, the random variables X and Y in Example 11.10 are independent if and only if $c = 0$ in which case $f_{X,Y}(x,y) = 1$, $f_X(x) = 1$ and $f_Y(y) = 1$ ($0 \leq x \leq 1$, $0 \leq y \leq 1$).

Example 11.11

The assembly of a complex piece of equipment can be divided into two stages. The times (in hours) required for the two stages are random variables (X and Y , say) with density functions e^{-x} and $2e^{-2y}$ respectively. Assuming that the stage assembly times are independent, find the probability that the assembly will be completed within four hours.

Solution The assumption of independence implies that

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) = e^{-x} 2e^{-2y} = 2e^{-(x+2y)}$$

If the time for the first stage is x , the total time will not exceed four hours if

$$Y < 4 - x$$

so the required value is

$$\begin{aligned} P(X + Y < 4) &= \int_0^4 \int_0^{4-x} f_{X,Y}(x, y) dy dx = \int_0^4 \int_0^{4-x} 2e^{-(x+2y)} dy dx \\ &= \int_0^4 (e^{-x} - e^{-(8-x)}) dx = 0.964 \end{aligned}$$

Where random variables are dependent upon one another, it is possible to express this dependence by defining a **conditional distribution** analogous to conditional probability, in terms of the joint distribution (or density function) and the marginal distributions. Examples are

$$\begin{aligned} P(X = u_k \mid Y = v_j) &= P(X = u_k \cap Y = v_j) / P(Y = v_j) && \text{in the discrete case and} \\ f_{X \mid Y}(x \mid y) &= f_{X,Y}(x,y) / f_Y(y) && \text{in the continuous case.} \end{aligned}$$

We shall now consider a numerical measure of dependence that can be estimated from sample data.

11.4.3 Covariance and correlation

The use of mean and variance for a random variable is motivated partly by the difficulty in determining the full probability distribution in many practical cases. The joint distribution of two variables presents even greater difficulties. Since we already have numerical measures of location and dispersion for the variables individually, it seems reasonable to define a measure of association of the two variables that is independent of their separate means and variances so that the new measure provides essentially new information about the variables.

There are four objectives that it seems reasonable for such a measure to satisfy. Its value should

- (a) be zero for independent variables,
- (b) be non-zero for dependent variables,
- (c) indicate the degree of dependence in some well-defined sense, detached from the individual means and variances,
- (d) be easy to estimate from sample data.

It is actually rather difficult to satisfy all of these, but the most popular measure of association gets most of the way.

The **covariance** of random variables X and Y , denoted by $\text{Cov}(X, Y)$, is defined as

$$\text{Cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$$

$$= \begin{cases} \sum_{k=1}^m \sum_{j=1}^n (u_k - \mu_X)(v_j - \mu_Y) P(X = u_k \cap Y = v_j) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) dx dy \end{cases}$$

for discrete and continuous variables respectively. The **correlation** $\rho_{X,Y}$ is the covariance divided by the product of the standard deviations:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

If whenever the random variable X is larger than its mean the random variable Y also tends to be larger than its mean, then the product $(X - \mu_X)(Y - \mu_Y)$ will tend to be positive. The same will be true if both variables tend to be smaller than their means simultaneously. The covariance is then positive. A negative covariance implies that the variables tend to move in opposite directions with respect to their means. Both covariance and correlation therefore measure association relative to the mean values of the variables. It turns out that correlation measures association relative to the standard deviations as well.

It should be noted that the variance of a random variable X is the same as the covariance with itself:

$$\text{Var}(X) = \text{Cov}(X, X)$$

Also, by expanding the product within the integral or sum in the definition of covariance, it is easy to show that an alternative expression is

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Although the sign of the covariance indicates the direction of the dependence, its magnitude depends not only on the degree of dependence but also upon the variances of the random variables, so it fails to satisfy the objective (c). In contrast, the correlation is limited in range

$$-1 \leq \rho_{X,Y} \leq +1$$

and it adopts the limiting values of this range only when the random variables are linearly related:

$$\rho_{X,Y} = \pm 1 \quad \text{if and only if there exist } a, b \text{ such that } Y = aX + b$$

(this is proved in most textbooks on probability theory, such as G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, third edition, Oxford, Clarendon Press, 2000). The magnitude of the correlation indicates the degree of linear relationship, so that objective (c) is satisfied.

Example 11.12

Find the correlation of the random variables S and M in Example 11.9.

Solution

The joint and marginal distributions of S and M are shown in Figure 11.9. First we find the expected values of S and S^2 from the marginal distribution, and hence the variance and standard deviation:

$$E(S) = (1)\frac{15}{28} + (2)\frac{3}{28} = \frac{21}{28}, \quad E(S^2) = (1)^2 \frac{15}{28} + (2^2)(\frac{3}{28}) = \frac{27}{28}$$

$$\text{Var}(S) = \frac{27}{28} - (\frac{21}{28})^2 = \frac{315}{28^2}$$

from which

$$\sigma_S = \frac{3}{28} \sqrt{35}$$

Next we do the same for M :

$$E(M) = (1)\frac{3}{7} + (2)\frac{1}{28} = \frac{1}{2}, \quad E(M^2) = (1)^2 \frac{3}{7} + (2)^2 \frac{1}{28} = \frac{4}{7}$$

$$\text{Var}(M) = \frac{4}{7} - \frac{1}{4} = \frac{9}{28}$$

from which

$$\sigma_M = \frac{3}{2} \sqrt{\frac{1}{7}}$$

All products of S and M are zero except when both are equal to one, so the expected value of the product is

$$E(SM) = (1)\frac{3}{14} = \frac{3}{14}$$

The correlation now follows easily:

$$\rho_{S,M} = \frac{E(SM) - E(S)E(M)}{\sigma_S \sigma_M} = \frac{\frac{3}{14} - (\frac{3}{28})(\frac{1}{2})}{(\frac{3}{28}\sqrt{35})(\frac{3}{2}\sqrt{\frac{1}{7}})} = -\frac{1}{\sqrt{5}} = -0.447.$$

The correlation is negative because if there are more statistics books in the selection then there will tend to be fewer mathematics books, and vice versa, as two textbooks are selected.

Example 11.13

Find the correlation of the random variables X and Y in Example 11.10.

Solution

Proceeding as in Example 11.12, we have for X

$$E(X) = \int_0^1 x \, dx = \frac{1}{2}$$

$$E(X^2) = \int_0^1 x^2 \, dx = \frac{1}{3}$$

so that $\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{12}$. Also, for Y

$$\begin{aligned} E(Y) &= \int_0^c \frac{y^2}{c} \, dy + \int_c^1 y \, dy + \int_1^{1+c} y \left[1 - \frac{1}{c}(y-1) \right] \, dy \\ &= \frac{1}{2}(1+c) \quad \text{after simplification} \end{aligned}$$

$$\begin{aligned} E(Y^2) &= \int_0^c \frac{y^3}{c} \, dy + \int_c^1 y^2 \, dy + \int_1^{1+c} y^2 \left[1 - \frac{1}{c}(y-1) \right] \, dy \\ &= \frac{1}{3}(1+c^2) + \frac{1}{2}c \quad \text{after simplification} \end{aligned}$$

so that $\text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{12}(1+c^2)$. For the expected value of the product we have

$$E(XY) = \int_0^1 \int_{cx}^{cx+1} xy \, dy \, dx = \frac{1}{2} \int_0^1 x(1+2cx) \, dx = \frac{1}{4} + \frac{1}{3}c$$

Finally, the correlation between X and Y is

$$\begin{aligned}\rho_{X,Y} &= \frac{E(XY) - E(X)E(Y)}{\sqrt{[\text{Var}(X)\text{Var}(Y)]}} \\ &= \frac{\frac{1}{4} + \frac{1}{3}c - \frac{1}{4}(1+c)}{\frac{1}{12}\sqrt{(1+c^2)}} = \frac{c}{\sqrt{(1+c^2)}}\end{aligned}$$

Note that in fact the result of Example 11.13 holds for any value of c , and not just for the range $0 \leq c \leq 1$ assumed in Example 11.10. As the value of c increases (positive or negative), the correlation increases also, but its magnitude never exceeds one. It is also clear that if X and Y are independent then $c = 0$ and the correlation is zero. Refer to Figure 11.10 for a geometrical interpretation: when $c = 0$, the sample space is a square within which all points are equally likely, so there is no association between the variables; as c increases (positive or negative), the sample space becomes more elongated as the variables become more tightly coupled to one another.

The general relationship between independence and correlation is expressed as follows: if the random variables X and Y are independent then their correlation is zero. This is easily shown as follows for continuous random variables (or by a similar argument for discrete random variables). First we have

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

and then

$$\begin{aligned}&\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_X(x)f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} (x - \mu_X) f_X(x) dx \int_{-\infty}^{\infty} (y - \mu_Y) f_Y(y) dy = (\mu_X - \mu_X)(\mu_Y - \mu_Y) = 0\end{aligned}$$

since probability density functions integrate to 1.

Unfortunately, the converse does not hold: zero correlation does not imply independence. In general, correlation is a measure of linear dependence, and may be zero or very small for variables that are dependent in a *nonlinear* way (see Exercise 15). Objective (a) is satisfied, therefore, but not objective (b) in general.

Another problem with correlation is that a non-zero value does not imply the presence of a causal relationship between the variables or the phenomena that they measure. Correlation can be ‘spurious’, deriving from some third variable that may be unrecognized at the time. For example, among the economic statistics that are gathered together from many countries, there are figures for the expenditure on luxury goods per head of population, birth rate, and the gross domestic product per capita (GDP). It turns out that there is a large negative correlation between expenditure on luxury goods per head and the birth rate, but no-one would suggest that the expenditure on luxury goods has any direct application in birth control. The GDP is a measure of wealth, and there is a large positive correlation between this and expenditure on luxury goods, and a large negative correlation between GDP and birth rate, both for quite genuine reasons. The correlation between expenditure on luxury goods and birth rate is therefore spurious, and a more sophisticated measure called the **partial correlation** can be used to eliminate the third variable (provided that it is recognized and measured). Indeed, the

partial correlation $\rho_{X,Y|Z}$ between two variables X and Y allowing for variable Z is the correlation between X and Y after the effect of Z has been removed, perhaps using a regression approach (see Section 11.5).

We have considered all the objectives except (d); that this is satisfied is shown in Section 11.4.4.

11.4.4 Sample correlation

There are two kinds of situations where we take samples of values of two random variables X and Y . First we might be interested in the same property for two different populations. Perhaps there is evidence that the mean values are different, so we take samples of each and compare them. This situation was discussed in Section 11.3.5. The second kind involves two different properties for the same population. It is to this situation that correlation applies. We take a single sample from the population and measure the pair of random variables (X_i, Y_i) for each $i = 1, \dots, n$.

For a sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ the **sample correlation coefficient** is defined as

$$r_{X,Y} = \frac{\frac{1}{n} \sum_{i=1}^n [(X_i - \bar{X})(Y_i - \bar{Y})]}{S_X S_Y}$$

Like the true underlying population correlation, the sample correlation is limited in value to the range $[-1, 1]$ and $r_{X,Y} = \pm 1$ when (and only when) all of the points lie along a line. Figure 11.11 contains four typical **scatter diagrams** of samples plotted on the (x, y) plane, with an indication of the correlation for each one. The range of behaviour is shown from independence (a) through imperfect correlation (b) and (c) to a perfect linear relationship (d).

By expanding the product within the outer bracket in the numerator, it is easy to show that an alternative expression is

$$r_{X,Y} = \frac{\bar{X}\bar{Y} - (\bar{X})(\bar{Y})}{S_X S_Y}$$

This expression is quicker to calculate by hand, although it can suffer from considerable numerical problems.

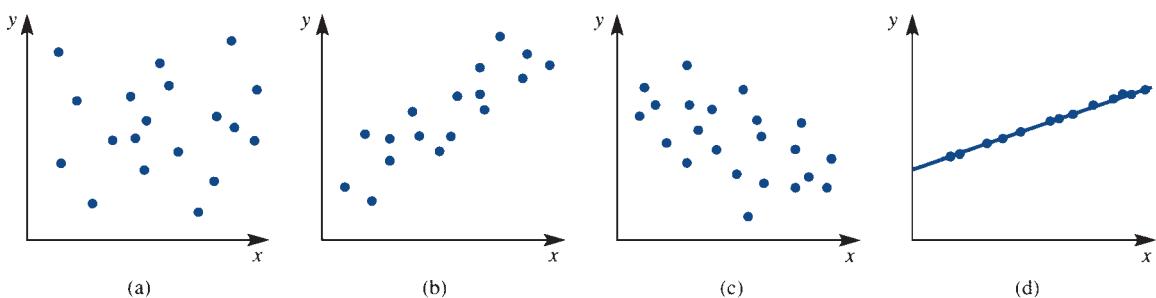


Figure 11.11 Scatter plots for two random variables: (a) $r_{X,Y} = 0$; (b) $r_{X,Y} > 0$; (c) $r_{X,Y} < 0$; (d) $r_{X,Y} = 1$.



The sample correlation coefficient can be computed in R using the function `cor` as we will see in Example 11.14. R functions for calculating summary statistics have been carefully developed with numerical accuracy in mind.

Example 11.14

A material used in the construction industry contains an impurity suspected of having an adverse effect upon the material's performance in resisting long-term operational stresses. Percentages of impurity and performance indexes for 22 specimens of this material are as follows:

% Impurity X_i	4.4	5.5	4.2	3.0	4.5	4.9	4.6	5.0	4.7	5.1	4.4
Performance Y_i	12	14	18	35	23	29	16	12	18	21	27
% Impurity X_i	4.1	4.9	4.7	5.0	4.6	3.6	4.9	5.1	4.8	5.2	5.2
Performance Y_i	13	19	22	20	16	27	21	13	18	17	11

Find the sample correlation coefficient.

Solution

The following quantities are easily obtained from the data:

$$X = 4.6545, \quad S_X = 0.55081, \quad Y = 19.1818, \quad S_Y = 6.0350, \quad \bar{XY} = 87.3591$$

(Note that it is advisable to record these results to several significant digits in order to avoid losing precision when calculating the difference within the numerator of $r_{X,Y}$) The sample correlation is then $r_{X,Y} = \{\bar{XY} - (\bar{X})(\bar{Y})/(S_X S_Y)\} = -0.58$, using the alternative expression. The negative value suggests that the impurity has an adverse effect upon performance. It remains to be seen whether this is statistically significant.



We first show in detail how to work out the sample correlation in R using the alternative expression. We will confirm our result by computing the sample correlation coefficient from the definition and then more easily using R's function `cor`.

```

Impurity_X <- c(4.4, 5.5, 4.2, 3.0, 4.5, 4.9, 4.6,
5.0, 4.7, 5.1, 4.4, 4.1, 4.9, 4.7, 5.0, 4.6, 3.6,
4.9, 5.1, 4.8, 5.2, 5.2)
#
Performance_Y <- c(12, 14, 18, 35, 23, 29, 16, 12, 18, 21,
27, 13, 19, 22, 20, 16, 27, 21, 13, 18, 17, 11)
# Step by step calculations
# Preliminary calculations of sums
X_sum <- sum(Impurity_X); X_sum
Y_sum <- sum(Performance_Y); Y_sum
#> [1] 102.4
#> [1] 422
X_2_sum <- sum(Impurity_X^2); X_2_sum
Y_2_sum <- sum(Performance_Y^2); Y_2_sum

```

```

#> [1] 483.3
#> [1] 8896
XY_sum <- sum(Impurity_X * Performance_Y); XY_sum
#> [1] 1921.9
n <- length(Impurity_X); n # n, sample size
#> [1] 22
# Sample means
X_bar <- X_sum / n; X_bar; Y_bar <- Y_sum / n; Y_bar
#> [1] 4.654545
#> [1] 19.18182
# Sample standard deviations S_X and S_Y
S_2_X <- (X_2_sum / n) - X_bar^2; S_X <- sqrt(S_2_X); S_X
#> [1] 0.5508071
S_2_Y <- (Y_2_sum / n) - Y_bar^2; S_Y <- sqrt(S_2_Y); S_Y
#> [1] 6.035022
# Mean of the products
XY_bar <- XY_sum / n; XY_bar
#> [1] 87.35909
# Sample correlation coefficient using the alternative
# expression
(XY_bar - X_bar * Y_bar) / (S_X * S_Y)
#> [1] -0.5786633
# Sample correlation using the definition
(1 / n) * sum((Impurity_X - X_bar) * (Performance_Y -
Y_bar)) / (S_X * S_Y)
#> [1] -0.5786633
# Sample correlation coefficient using R's function cor
r <- cor(Impurity_X, Performance_Y); r
#> [1] -0.5786633

```

11.4.5 Interval and test for correlation

Correlation is more difficult to deal with than mean and proportion, but for normal random variables X and Y with a true correlation $\rho_{X,Y}$ the sample statistic

$$Z = \frac{\sqrt{(n-3)}}{2} \ln \left[\frac{(1+r_{X,Y})(1-\rho_{X,Y})}{(1-r_{X,Y})(1+\rho_{X,Y})} \right]$$

is approximately standard normal for large n . This can be used directly as a test statistic for an assumed value of $\rho_{X,Y}$. Alternatively, an approximate 100(1 - α)% confidence interval for $\rho_{X,Y}$ can be derived:

$$\left(\frac{1+r-c(1-r)}{1+r+c(1-r)}, \frac{1+r-(1-r)/c}{1+r+(1-r)/c} \right)$$

where

$$c = \exp\left[\frac{2z_{\alpha/2}}{\sqrt{(n-3)}}\right]$$

(the subscripts X and Y have been dropped from $r_{X,Y}$ in this formula).

Example 11.15

For the data in Example 11.14 find 95% and 99% confidence intervals for the true correlation between percentage of impurity and performance index, and test the hypothesis that these are independent.

Solution

The sample correlation (from the 22 specimens) was found in Example 11.14 to be -0.58 . For the 95% confidence interval the constant $c = 2.458$ and the interval itself is $(-0.80, -0.21)$. Similarly, the 99% confidence interval is $(-0.85, -0.07)$. Assuming $\rho_{XY} = 0$, the value of the test statistic is $Z = -2.88$, which exceeds $z_{0.005} = 2.576$ in magnitude. Either way, we can be more than 99% confident that the impurity has an adverse effect upon performance.



Here are the calculations in R:

```
# 95% confidence interval
alpha <- 0.05 # As 95% confidence interval required
z <- qnorm(alpha / 2, lower.tail = FALSE); z
# We want z such that P(Z > z) = alpha / 2
#> [1] 1.959964
constant <- exp(2 * z / sqrt(n - 3)); constant
#> [1] 2.457865
# Confidence interval
c((1 + r - constant*(1 - r)) / (1 + r + constant*(1 - r)),
  (1 + r - (1 - r) / constant) / (1 + r + (1 - r) / constant))
#> [1] -0.8040968 -0.2077362
# 99% confidence interval
alpha <- 0.01 # As 99% confidence interval required
z <- qnorm(alpha / 2, lower.tail = FALSE); z
# We want z such that P(Z > z) = alpha / 2
#> [1] 2.575829
constant <- exp(2 * z / sqrt(n - 3)); constant
#> [1] 3.260471
c((1 + r - constant*(1 - r)) / (1 + r + constant*(1 - r)),
  (1 + r - (1 - r) / constant) / (1 + r + (1 - r) / constant))
#> [1] -0.84867200 -0.06940328
```

```

# Sample statistic
rho <- 0 # Value under the null hypothesis  $H_0$ :  $\rho = 0$ 
Z_statistic <- (sqrt(n - 3) / 2) * log(((1 + r) * (1 - rho)) / ((1 - r) * (1 + rho)))
Z_statistic
#> [1] -2.878838
z # Critical value from above
#> [1] 2.575829

```

The confidence intervals can be obtained in R directly:

```

cor.test(Impurity_X, Performance_Y)
    # Default is a 95% confidence interval
#>
#> Pearson's product-moment correlation
#>
#> data: Impurity_X and Performance_Y
#> t = -3.1731, df = 20, p-value = 0.004781
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#> -0.8040968 -0.2077362
#> sample estimates:
#> cor
#> -0.5786633
cor.test(Impurity_X, Performance_Y, conf.level = 0.99)
    # 99% confidence interval
#>
#> Pearson's product-moment correlation
#>
#> data: Impurity_X and Performance_Y
#> t = -3.1731, df = 20, p-value = 0.004781
#> alternative hypothesis: true correlation is not equal to 0
#> 99 percent confidence interval:
#> -0.84867200 -0.06940328
#> sample estimates:
#> cor
#> -0.5786633

```

R uses a different form of the test statistic. The null hypothesis that $\rho_{XY} = 0$ would be rejected at significance level α if the p-value were less than α . As the p-value of 0.004781 is less than 0.01, we reject the null hypothesis $H_0: \rho_{XY} = 0$ at the 0.01 level of significance.

11.4.6 Rank correlation

As has been previously emphasized the correlation only works as a measure of dependence if

- (1) n is reasonably large,
- (2) X and Y are *numerical* characteristics,
- (3) the dependence is *linear*, and
- (4) X and Y each have a *normal* distribution.

There is an alternative form of sample correlation, which has greater applicability, requiring only that

- (1) n is reasonably large,
- (2) X and Y are *rankable* characteristics, and
- (3) the dependence is *monotonic* (that is, always in the same direction, which may be forward or inverse, but not necessarily linear).

The variables X and Y can have any distribution. For a set of data X_1, \dots, X_n , a **rank** of 1 is assigned to the smallest value, 2 to the next-smallest and so on up to a rank of n assigned to the largest. This applies wherever the values are distinct. Tied values are given the mean of the ranks they would receive if slightly different. The following is an example:

X_i	8	3	5	8	1	9	6	5	3	5	7	2
Rank	10.5	3.5	6	10.5	1	12	8	6	3.5	6	9	2



We can obtain these ranks in R using the `rank` function:

```
x <- c(8, 3, 5, 8, 1, 9, 6, 5, 3, 5, 7, 2)
rank(x)
#> [1] 10.5 3.5 6.0 10.5 1.0 12.0 8.0 6.0 3.5 6.0 9.0 2.0
```

The **Spearman rank correlation coefficient** r_s for data $(X_1, Y_1), \dots, (X_n, Y_n)$ is the correlation of the ranks of X_i and Y_i , where the data X_1, \dots, X_n and Y_1, \dots, Y_n are ranked separately. If the number of tied values is small compared with n then

$$r_s \approx 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n d_i^2$$

where d_i is the difference between the rank of X_i and that of Y_i . The value of r_s always lies in the interval $[-1, 1]$, and adopts its extreme values only when the rankings precisely match (forwards or in reverse).

To test for dependence, special tables must be used for small samples ($n < 20$), but for larger samples the test statistic

$$Z = r_s \sqrt{(n - 1)}$$

is approximately standard normal.

Example 11.16

Find and test the rank correlation for the data in Example 11.14.

Solution

The data with their ranks are as follows:

X_i	4.4	5.5	4.2	3.0	4.5	4.9	4.6	5.0	4.7	5.1	4.4
Rank	5.5	22	4	1	7	14	8.5	16.5	10.5	18.5	5.5
Y_i	12	14	18	35	23	29	16	12	18	21	27
Rank	2.5	6	11	22	18	21	7.5	2.5	11	15.5	19.5
X_i	4.1	4.9	4.7	5.0	4.6	3.6	4.9	5.1	4.8	5.2	5.2
Rank	3	14	10.5	16.5	8.5	2	14	18.5	12	20.5	20.5
Y_i	13	19	22	20	16	27	21	13	18	17	11
Rank	4.5	13	17	14	7.5	19.5	15.5	4.5	11	9	1

From this, the rank correlation is $r_s = -0.361$, and $Z = -1.66$, which exceeds $z_{0.05} = 1.645$ and is therefore just significant at the 10% level. If the approximate formula is used, the sum of squares of differences is 2398, so

$$r_s \approx 1 - \frac{(6)(2398)}{(22)(483)} = -0.354$$

and $Z = -1.62$, which is just short of significance.

These results show that the rank correlation is a more conservative test than the sample correlation $r_{X,Y}$, in that a larger sample tends to be needed before the hypothesis of independence is rejected. A price has to be paid for the wider applicability of the method.



Here are the ranks of the impurity and performance data:

```
rank(Impurity_X)
#> [1] 5.5 22.0 4.0 1.0 7.0 14.0 8.5 16.5 10.5 18.5 5.5
3.0 14.0 10.5
#> [15] 16.5 8.5 2.0 14.0 18.5 12.0 20.5 20.5
rank(Performance_Y)
#> [1] 2.5 6.0 11.0 22.0 18.0 21.0 7.5 2.5 11.0 15.5 19.5
4.5 13.0 17.0
#> [15] 14.0 7.5 19.5 15.5 4.5 11.0 9.0 1.0
```

The Spearman rank correlation coefficient can be found by computing the sample correlation of the ranks:

```
cor(rank(Impurity_X), rank(Performance_Y))
#> [1] -0.3613398
```

The Spearman rank correlation coefficient can be calculated directly in R as:

```
r_S <- cor(Impurity_X, Performance_Y,
            method = "spearman"); r_S
#> [1] -0.3613398
```

The test statistic takes the value

```
Z <- r_S * sqrt(n - 1); Z
#> [1] -1.655867
```

the absolute value 1.655867 of which can be compared with the critical value:

```
alpha <- 0.1
z <- qnorm(alpha / 2, lower.tail = FALSE); z
#> [1] 1.644854
```

A hypothesis test can be performed in R as follows:

```
cor.test(Impurity_X, Performance_Y, method = "spearman")
#>
#> Spearman's rank correlation rho
#>
#> data: Impurity_X and Performance_Y
#> S = 2410.9, p-value = 0.09848
#> alternative hypothesis: true rho is not equal to 0
#> sample estimates:
#> rho
#> -0.3613398
```

Again, R uses a different form of the test statistic. The null hypothesis of zero correlation would be rejected at the significance level α if the p-value were less than α . As the p-value of 0.09848 is less than 0.1, we reject the null hypothesis at the 0.1 level of significance.

The approximate values of the Spearman rank correlation coefficient and the associated test statistic can be calculated as:

```
diff_rank <- rank(Impurity_X) - rank(Performance_Y)
r_S_approx <- 1 - (6 / (n * (n^2 - 1))) * sum(diff_rank^2);
r_S_approx
#> [1] -0.3540373
Z <- r_S_approx * sqrt(n - 1); Z
#> [1] -1.622403
```

11.4.7 Exercises



Check your calculations using R whenever possible.

- 14 Suppose that the random variables X and Y have the following joint distribution:

		X		
		1	2	3
Y	1	0	0.17	0.08
	2	0.20	0.11	0
3	0.14	0.25	0.05	

Find (a) the marginal distributions of X and Y , (b) $P(Y = 3 | X = 2)$, and (c) the mean, variance and correlation coefficient of X and Y .

15

Consider the random variable X with density function

$$f_X(x) = \begin{cases} 1 & (-\frac{1}{2} < x < \frac{1}{2}) \\ 0 & \text{otherwise} \end{cases}$$

Show that the covariance of X and X^2 is zero. (This shows that zero covariance does not imply independence, because obviously X^2 is dependent on X .)

- 16 The joint density function of random variables X and Y is

$$f_{X,Y}(x, y) = \begin{cases} 1 & (0 \leq x \leq 1; cx \leq y \leq cx + 1) \\ 0 & \text{otherwise} \end{cases}$$

where c is a constant such that $0 \leq c \leq 1$.

Find the marginal density function for Y (see Example 11.10).

- 17 Let the random variables X and Y represent the lifetimes (in hundreds of hours) of two types of components used in an electronic system. The joint density function is given by

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{8}x e^{-(x+y)/2} & (x > 0, y > 0) \\ 0 & \text{otherwise} \end{cases}$$

Find (a) the probability that two components (one of each type) will each last longer than 100 h, and (b) the probability that a component of the second type (Y) will have a lifetime in excess of 200 h.

- 18 The following are the measured heights and weights of eight people:

Height (cm)	182.8	162.5	175.2	185.4	170.1	167.6	177.8	172.7
Weight (kg)	86.1	58.3	83.0	92.4	60.2	69.3	83.6	72.7

Find the sample correlation coefficient.

- 19 The number of minutes it took 10 mechanics to assemble a piece of machinery in the morning (X) and in the late afternoon (Y) were measured, with the following results:

X	11.1	10.3	12.0	15.1	13.7	18.5	17.3	14.2	14.8	15.3
Y	10.9	14.2	13.8	21.5	13.2	21.1	16.4	19.3	17.4	19.0

Find the sample correlation coefficient.

- 20 If the sample correlation between resistance and failure time for 30 overloaded resistors is 0.7, find a 95% confidence interval for the true correlation.

- 21 Find a 95% confidence interval for correlation between height and weight using the data in Exercise 18.

- 22 Marks obtained by 20 students taking examinations in mathematics and computer studies were as follows:

Math.	45	77	43	64	58	64	58	54	71	45
	57	52	67	57	54	54	61	58	55	42
Comp.	64	67	47	75	42	65	58	42	70	44
	44	67	49	70	51	58	37	60	42	36

Find the sample correlation coefficient and the 90% and 95% confidence intervals. Hence test the hypothesis that the two marks are independent at the 5% and 10% significance levels. Also find and test the rank correlation.

- 23 Let the random variables X and Y have joint density function given by

$$f_{X,Y}(x, y) = \begin{cases} c(1-y) & (0 \leq x \leq y \leq 1) \\ 0 & \text{otherwise} \end{cases}$$

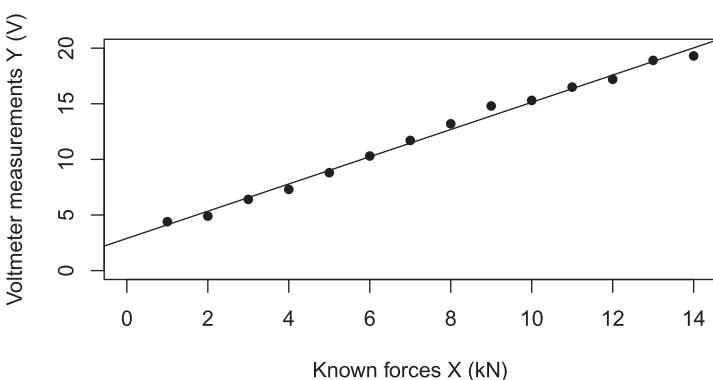
Find (a) the value of the constant c , (b) $P(x < \frac{3}{4}, y > \frac{1}{2})$, and (c) the marginal density functions for X and Y .

- 24 The ball and socket of a joint are separately moulded and then assembled together. The diameter of the ball is a random variable X between 29.8 and 30.3 mm, all values being equally likely. The internal diameter of the socket is a random variable Y between 30.1 and 30.6 mm, again with all values equally likely. The condition for an acceptable fit is that $0 \leq Y - X \leq 0.6$ mm. Find the probability of this condition being satisfied, assuming that the random variables are independent.

11.5 Regression

A procedure that is very familiar to engineers is that of drawing a good straight line through a set of points on a graph. When calibrating a measuring instrument, for example, known inputs are applied, the readings are noted and plotted, a straight line is drawn as close to the points as possible (there are bound to be small errors, so they will not all lie on the line), and the graph is then used to interpret the readings for unknown inputs. It is possible to draw the line by eye, but there is a better way, which involves calculating

Figure 11.12 Scatter plot with regression line (Example 11.17).



the slope and intercept of the line from the data. The given line then minimizes the total squared error for the data points. This procedure (which for historical reasons is called **regression**) can be applied in general to pairs of random variables.

Computer packages such as R are very often used to carry out the regression calculations and display the results. This is of special value when the data tend to follow a curve and various nonlinear models are tried and compared (see Section 11.5.4).

11.5.1 The method of least squares

The correlation was introduced in Section 11.4.3 as a way of measuring the dependence between random variables. Subsequently, we have seen how the correlation can be estimated and the dependence tested using sample data. We can take the idea of correlation between variables (say X and Y) a stage further by assuming that the sample pairs $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ satisfy a linear relationship of the form

$$Y_i = a + bX_i + \varepsilon_i \quad (i = 1, \dots, n)$$

where the intercept a and the slope b are unknown coefficients and the random variables ε_i have zero mean and represent the associated errors, that is the differences between the observed values Y_i and the values on the line $a + bX_i$. This assumption is prompted by the scatter diagrams in Figure 11.11, which illustrate how the points may be concentrated around a line. Figure 11.12 shows a typical scatter diagram again, this time with the line drawn in. If we can estimate the coefficients a and b so as to give the best fit, we shall be able to predict the value of Y when the value of X is known.

The least-squares approach is to choose estimates \hat{a} and \hat{b} to minimize the sum of squares of the values ε_i :

$$Q(a, b) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \{Y_i - (a + bX_i)\}^2$$

Equating to zero the partial derivatives of this sum with respect to the two coefficients gives a pair of equations that determine the minimum:

$$\frac{\partial Q}{\partial a} = 0 \Rightarrow -2 \sum_{i=1}^n \{Y_i - (\hat{a} + \hat{b}X_i)\} = 0$$

$$\frac{\partial Q}{\partial b} = 0 \Rightarrow -2 \sum_{i=1}^n X_i \{Y_i - (\hat{a} + \hat{b}X_i)\} = 0$$

These can be rewritten as

$$\begin{aligned} n\hat{a} + (\sum_i X_i)\hat{b} &= (\sum_i Y_i) \\ (\sum_i X_i)\hat{a} + (\sum_i X_i^2)\hat{b} &= (\sum_i X_i Y_i) \end{aligned}$$

(where $\sum_i = \sum_{i=1}^n$) from which the solution is

$$\hat{b} = \frac{S_{XY}}{S_X^2}, \quad \hat{a} = \bar{Y} - \hat{b}\bar{X}$$

where

$$\bar{X} = \frac{1}{n} \sum_i X_i, \quad \bar{Y} = \frac{1}{n} \sum_i Y_i$$

and

$$S_{XY} = \frac{1}{n} \sum_i (X_i - \bar{X})(Y_i - \bar{Y}) = \bar{XY} - (\bar{X})(\bar{Y})$$

$$S_X^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2 = \bar{X^2} - (\bar{X})^2$$

$$S_Y^2 = \frac{1}{n} \sum_i (Y_i - \bar{Y})^2 = \bar{Y^2} - (\bar{Y})^2$$

are the sample covariance and variances.

This process of fitting a straight line through a set of data of the form $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ is called **linear regression**, and the coefficients are called **regression coefficients**.

Example 11.17

A strain gauge has been bonded to a steel beam, and is being calibrated. The resistance of the strain gauge is converted into a voltage appearing on a meter. Known forces (X , in kN) are applied and voltmeter measurements (Y , in V) are as follows:

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Y	4.4	4.9	6.4	7.3	8.8	10.3	11.7	13.2	14.8	15.3	16.5	17.2	18.9	19.3

Fit a regression line through the data and estimate the tension in the beam when the meter reading is 13.8 V.

Also, estimate the voltmeter measurement when the tension or force is 8.5 kN.

Solution

The following quantities are calculated from the data:

$$\bar{X} = 7.5, \quad S_X = 4.03113, \quad \bar{Y} = 12.0714, \quad S_Y = 4.95068, \quad \bar{XY} = 110.421$$

(When using a hand calculator to solve linear regression problems, it is advisable to work to at least five or six significant digits during intermediate calculations, because the subtraction in the numerator of \hat{b} often results in the loss of some leading digits.) From these results, $\hat{b} = 1.22$ and $\hat{a} = 2.89$ (Figure 11.12). The estimated value of tension for a reading of $Y = 13.8$ V is given by

$$13.8 = 2.89 + 1.22X$$

from which $X = 8.9$ kN.

When the force is 8.5 kN, the voltmeter measurement is estimated to be

$$\hat{Y} = \hat{a} + \hat{b} \times 8.5 = 2.89 + 1.22 \times 8.5 \approx 13.3V$$



We first show in detail how to work out these quantities in R using the formulas. We have seen some of these calculations before. After, we will confirm our result using R's function `lm` for linear modelling. J. J. Faraway, *Linear Models with R* (second edition, Boca Raton, FL, Chapman and Hall/CRC, 2015) provides a book length treatment of linear models.

```
X <- 1:14 # Integer sequence from 1 to 14
Y <- c(4.4, 4.9, 6.4, 7.3, 8.8, 10.3, 11.7, 13.2, 14.8,
15.3, 16.5, 17.2, 18.9, 19.3)
# Step by step calculations, starting with some sums
X_sum <- sum(X); X_sum; Y_sum <- sum(Y); Y_sum
#> [1] 105
#> [1] 169
X_2_sum <- sum(X^2); X_2_sum; Y_2_sum <- sum(Y^2); Y_2_sum
#> [1] 1015
#> [1] 2383.2
XY_sum <- sum(X * Y); XY_sum
#> [1] 1545.9
n <- length(X); n
#> [1] 14
X_bar <- X_sum / n; X_bar
#> [1] 7.5
Y_bar <- Y_sum / n; Y_bar
#> [1] 12.07143
S_2_X <- (X_2_sum / n) - X_bar^2; S_2_X
S_X <- sqrt(S_2_X); S_X
#> [1] 16.25
#> [1] 4.031129
S_2_Y <- (Y_2_sum / n) - Y_bar^2; S_2_Y
S_Y <- sqrt(S_2_Y); S_Y
#> [1] 24.50918
#> [1] 4.950675
XY_bar <- XY_sum / n; XY_bar
#> [1] 110.4214
S_XY <- XY_bar - X_bar * Y_bar; S_XY
#> [1] 19.88571
# Note that S_2_X, S_2_Y and S_XY can be computed
# in a less efficient way according to the definitions as
sum((X - X_bar)^2) / n; sum((Y - Y_bar)^2) / n
sum((X - X_bar) * (Y - Y_bar)) / n
#> [1] 16.25
#> [1] 24.50918
#> [1] 19.88571
```

```

# Estimate of slope
b_hat <- S_XY /S_2_X; b_hat
#> [1] 1.223736
# Estimate of intercept
a_hat <- Y_bar - b_hat * X_bar; a_hat
#> [1] 2.893407

```

R's function `lm` uses a somewhat different and more numerically stable approach based on QR methods to find \hat{a} and \hat{b} . Here's how `lm` can be used:

```

# Define m as the linear model (lm) that results from
# asking: how does Y depend on (~) X?
m <- lm(Y ~ X)
# Look at m and extract the coefficients
m
#>
#> Call:
#> lm(formula = Y ~ X)
#>
#> Coefficients:
#> (Intercept)      X
#>     2.893      1.224
coef(m)
#> (Intercept)      X
#> 2.893407 1.223736

```

To estimate the voltmeter measurement when the tension or force is 8.5 kN we can either perform the calculation directly, or use the `predict` function to make a prediction from the linear model fit:

```

X_new <- 8.5
a_hat + b_hat * X_new
#> [1] 13.29516
# predict requires the newdata to be placed in a data frame
#(see below)
predict(m, newdata = data.frame(X = X_new))
#> 1
#> 13.29516

```

We can add the fitted straight line to a plot of the data using the `abline` function. This code produced Figure 11.12.

```

plot(X, Y,
      xlab = "Known forces X (kN)",
      ylab = "Voltmeter measurements Y (V)",
      xlim = c(0, 14), # Limits on the x axis
      ylim = c(0, 20), # Limits on the y axis
      pch = 16) # Use filled dots as the plotting character
abline(m) # Add the regression line

```

The voltmeter measurements corresponding to the observed values X_1, \dots, X_n are known as the fitted values and can be calculated using $\hat{Y}_i = \hat{a} + \hat{b}X_i, i = 1, \dots, n$.



These values can be easily found in R using the `predict` and `fitted` functions:

```
a_hat + b_hat * X[1:5] # Direct calculation; show the first
# five values
#> [1] 4.117143 5.340879 6.564615 7.788352 9.012088
predict(m) [1:5]
#>      1         2         3         4         5
#> 4.117143 5.340879 6.564615 7.788352 9.012088
fitted(m) [1:5]
#>      1         2         3         4         5
#> 4.117143 5.340879 6.564615 7.788352 9.012088
```



The model $Y_i = a + bX_i + \varepsilon_i, i = 1, \dots, n$, can be written in matrix format as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Let the $n \times 2$ matrix X be defined as

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

which is referred to as the model matrix.

The model matrix can be extracted in R from the fitted model using `model.matrix`:

```
X_mat <- model.matrix(m); head(X_mat) # Show first six
rows, which are labelled
#> (Intercept) x
#> 1          1 1
#> 2          1 2
#> 3          1 3
#> 4          1 4
#> 5          1 5
#> 6          1 6
```

It can be shown mathematically that the estimates $(\hat{a}, \hat{b})^T$ of the coefficients can be found from the model matrix X using matrix multiplication as $(X^T X)^{-1} X^T Y$, in which Y is the vector of observed values $(Y_1, Y_2, \dots, Y_n)^T$. The vector of fitted values $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n)^T$ can be found using

$$\hat{Y} = X \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = X(X^T X)^{-1} X^T Y$$



These fitted values can be calculated from this formula in R using the following steps:

```
X_T_X <- t(X_mat) %*% X_mat
# t means transpose, %*% means matrix multiplication
X_T_X
#>             (Intercept)      X
#> (Intercept)          14 105
#> X                  105 1015
# Note that this is not a diagonal matrix
X_T_X_inverse <- solve(X_T_X) # solve finds the inverse
# Estimates of coefficients
coef_hat <- X_T_X_inverse %*% t(X_mat) %*% Y
coef_hat; coef(m) # Compare with the results from lm
#> [,1]
#> (Intercept) 2.893407
#> X           1.223736
#> (Intercept)      X
#> 2.893407     1.223736
# Fitted values
Y_hat <- X_mat %*% coef_hat
# Confirm by rounding the differences to 5 decimal places
# to take account of numerical inaccuracies; 0.00000
# printed as 0
head(round(Y_hat - fitted(m), 5)) # Differences
#> [,1]
#> 1      0
#> 2      0
#> 3      0
#> 4      0
#> 5      0
#> 6      0
```

There are more efficient ways of performing the above calculations, using `crossprod` for example, but these are not important here.

The columns of the model matrix X are $(1, 1, \dots, 1)^T$ and $(X_1, X_2, \dots, X_n)^T$. These can be thought of as the basis elements of a vector space χ , say, which we can think of as the ‘model vector space’. The vector of fitted values \hat{Y} is the orthogonal or perpendicular projection of the vector of observed values Y into the vector space χ , and in this sense \hat{Y} is the best approximation of Y in the model vector space χ .

There are other ways of defining the basis elements of the vector space χ . Often, the vectors $(1, 1, \dots, 1)^T$ and $(\bar{X}, \bar{X}, \dots, \bar{X})^T$ are taken as basis elements, in which $\bar{X} = \sum_{i=1}^n X_i / n$. These basis vectors are orthogonal because their scalar or dot product \bullet is zero:

$$\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \bullet \begin{pmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{pmatrix}$$

(Calculation developed on the next page)

$$\begin{aligned}
 &= 1 \times (X_1 - X) + 1 \times (X_2 - X) + \cdots + 1 \times (X_n - X) \\
 &= (X_1 - X) + (X_2 - X) + \cdots + 1 \times (X_n - X) \\
 &= X_1 - X_2 + \cdots + X_n - nX \\
 &= nX - nX = 0
 \end{aligned}$$

Because of this we sometimes refer to the functions 1 (constant function) and $X - X$ as ‘orthogonal polynomials’. We also say that the values X_1, X_2, \dots, X_n have been ‘centered’ to have mean 0.



We can fit the ‘centered model’ $Y_i - a' + b'(X_i - X) + \varepsilon_i$ as follows:

```
m_cent <- lm(Y ~ I(X - mean(X)))
# The calculation of X - mean(X) must be enclosed in I
coef(m_cent)
#> (Intercept) I(X - mean(X))
#> 12.071429 1.223736
```

Since $a + bX = (a + bX) + b(X - X)$, it follows that $\hat{a}' = \hat{a} + \bar{X}$ and that $\hat{b}' = \hat{b}$:

```
a_hat + b_hat * mean(X); b_hat
#> [1] 12.07143
#> [1] 1.223736
```

The estimates (\hat{a}', \hat{b}') ^T can also be found from the model matrix

$$X_{\text{cent}} = \begin{pmatrix} 1 & X_1 - \bar{X} \\ 1 & X_2 - \bar{X} \\ \vdots & \vdots \\ 1 & X_n - \bar{X} \end{pmatrix}$$

as

$$(X_{\text{cent}}^T X_{\text{cent}})^{-1} X_{\text{cent}}^T Y$$



They can be calculated from this formula in R using the following code:

```
X - mean(X) # Centered valued
#> [1] -6.5 -5.5 -4.5 -3.5 -2.5 -1.5 -0.5 0.5 1.5 2.5 3.5
4.5 5.5 6.5
mean(X - mean(X))
# The data in X are centered to have mean zero
#> [1] 0
X_center <- model.matrix(m_center)
head(X_center)
#> (Intercept) I(X - mean(X))
#> 1 1 -6.5
#> 2 1 -5.5
#> 3 1 -4.5
#> 4 1 -3.5
#> 5 1 -2.5
#> 6 1 -1.5
```

```

X_cent_T_X_cent <- t(X_cent) %*% X_cent
X_cent_T_X_cent
#> (Intercept) I(X - mean(X))
#> (Intercept) 14 0.0
#> I(X - mean(X)) 0 227.5
# Note that this is a *** diagonal matrix ***
X_cent_T_X_cent_inverse <- solve(X_cent_T_X_cent)
# solve finds the inverse
# Estimates of coefficients
coef_cent_hat <- X_cent_T_X_cent_inverse %*% t(X_cent) %*% Y
coef_cent_hat; coef(m_cent)
# Compare with the results from lm
#> [,1]
#> (Intercept) 12.071429
#> I(X - mean(X)) 1.223736
#> (Intercept) I(X - mean(X))
#> 12.071429 1.223736

```

From the above we see that $X_{\text{cent}}^T X_{\text{cent}}$ is a diagonal matrix and hence is easy to invert. This is because the columns of X_{cent} are orthogonal. In fact, it turns out that under simple assumptions about ε_i that we will meet in the next section, the estimates \hat{a}' and \hat{b}' are independent, whereas \hat{a} and \hat{b} would not be independent. This means that \hat{a}' and \hat{b}' express essentially different information, while \hat{a} and \hat{b} express overlapping information. This has important implications for statistical inference, as discussed in Section 9.4 of J. J. Faraway, *Linear Models with R* (second edition, Boca Raton, FL, Chapman and Hall/CRC, 2015), for example.



The fitted values from the two models are the same as the vector of observed values is projected into the same model vector space χ :

```

round(fitted(m) - fitted(m_cent), 5) # Differences
#> 1 2 3 4 5 6 7 8 9 10 11 12 13 14
#> 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

```

As will be discussed in Exercise 30, sometimes it is required that the regression line passes through the origin, in which case the model is $Y_i = bX_i + \varepsilon_i$ so that the only regression coefficient is the slope b of the line. We can estimate this slope in R as follows:

```

m_through_origin <- lm(Y ~ X - 1)
coef(m_through_origin)
#> X
#> 1.523054
m_through_origin_2 <- lm(Y ~ X + 0)
# Alternative formulation
coef(m_through_origin_2)
#> X
#> 1.523054
sum(Y * X) / sum(X * X) # Direct calculation
#> [1] 1.523054

```

Note that now the model matrix comprises just one column $(X_1, \dots, X_n)^T$:

```
X_through_origin <- model.matrix(m_through_origin)
head(X_through_origin)
#>      X
#> 1   1
#> 2   2
#> 3   3
#> 4   4
#> 5   5
#> 6   6
# Use this to estimate the coefficient b via matrix
# multiplication
solve(crossprod(X_through_origin)) %*%
  t(X_through_origin) %*% Y
#> [,1]
#> X 1.523054
```



In R it is often very useful to keep related data together in a data frame. In general a data frame is a rectangular collection of variables in the columns and observations in the rows. Here is a brief example:

```
df_strain <- data.frame(X, Y)
head(df_strain) # Show the first six data points
#>      X     Y
#> 1   1 4.4
#> 2   2 4.9
#> 3   3 6.4
#> 4   4 7.3
#> 5   5 8.8
#> 6   6 10.3
```

The names of the variables in a data frame can be found using `names`. The `str` function provides a useful, compact display of the internal structure of a data frame. The values of individual variables can be extracted using `$` or `[["variable_name"]]`. Here we illustrate these features in the `df_strain` data frame:

```
names(df.strain)
#> [1] "X"  "Y"
str(df_strain)
#> 'data, frame':14 obs. of 2 variables:
#> $ X: int 1 2 3 4 5 6 7 8 9 10 ...
#> $ Y: num 4.4 4.9 6.4 7.3 8.8 10.3 11.7 13.2 14.8 15.3 ...
df_strain$X # Extract X
#> [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14
df_strain[["Y"]]
# Alternative way of extracting a variable
#> [1] 4.4 4.9 6.4 7.3 8.8 10.3 11.7 13.2 14.8 15.3 16.5
#> 17.2 18.9 19.3
```

```
# Fit a linear model using data in a data frame
m <- lm(Y ~ X, data = df_strain)
# Look for Y and X in data, here the data frame df_strain
coef(m)
#> (Intercept)          X
#> 2.893407      1.223736
```

In practice, data are often supplied as Excel spreadsheets or comma separated variable files. Many functions, for example `read_excel` from the `readxl` package or `read_csv` from the `readr` package, are available to read such data into R as data frames.

Later we will meet a type of R data frame, called a ‘tibble’, specially designed to make life easier in certain circumstances. We will also work with the `dplyr` package which provides very useful tools for data manipulation including `filter` to pick observations by their values, `arrange` to re-order rows, `select` to pick variables by their names, `mutate` to create new variables and `summarize` to collapse many values down to a single summary. For a full discussion of importing data into R, data frames, tibbles and related R structures, and data manipulation see H. Wickham and G. Grolemund, *R for Data Science* (Beijing, O'Reilly, 2016).

Orthogonal polynomials are sometimes used in engineering to model signals observed over time. Let us assume that we observe data Y_t over a time interval $(0, T]$ at n equally spaced time points $t_1 = T/n, t_2 = 2T/n, \dots, t_n = nT/n = T$. Let $\omega = 2\pi/T$. An example of a regression type model that is often used for such data takes the form

$$Y_t = a_0 + a_1 \sin(\omega t) + b_1 \cos(\omega t) + a_2 \sin(2\omega t) + b_2 \cos(2\omega t) + a_3 \sin(3\omega t) + b_3 \cos(3\omega t) + \varepsilon_t, \quad t = t_1, t_2, \dots, t_n$$



The vectors $(1, 1 \dots, 1)^T$, $(\cos(\omega t_1), \cos(\omega t_2), \dots, \cos(\omega t_n))^T$, $(\sin(\omega t_1), \sin(\omega t_2), \dots, \sin(\omega t_n))^T$, $(\cos(2\omega t_1), \cos(2\omega t_2), \dots, \cos(2\omega t_n))^T$, $(\sin(2\omega t_1), \sin(2\omega t_2), \dots, \sin(2\omega t_n))^T$, $(\cos(3\omega t_1), \cos(3\omega t_2), \dots, \cos(3\omega t_n))^T$ and $(\sin(3\omega t_1), \sin(3\omega t_2), \dots, \sin(3\omega t_n))^T$ can be taken as a basis for the vector space χ that we met above. These vectors are orthogonal, as this R example illustrates by showing that the associated model matrix is diagonal:

```
n <- 12; T <- 7; t <- T * (1:n) / n
omega <- 2*pi / T
one <- rep(1, n) # Vector of ones
c_1 <- cos(omega * t)
# Vector of cos(omega t_1), cos(omega t_2), ..., cos(omega t_n)
s_1 <- sin(omega * t)
# Vector of sin(omega t_1), sin(omega t_2), ..., sin(omega t_n)
c_2 <- cos(2 * omega * t)
# Vector of cos(2 omega t_1), cos(2 omega t_2), ..., cos(2 omega t_n)
s_2 <- sin(2 * omega * t)
# Vector of sin(2 omega t_1), sin(2 omega t_2), ..., sin(2 omega t_n)
c_3 <- cos(3 * omega * t)
# Vector of cos(3 omega t_1), cos(3 omega t_2), ..., cos(3 omega t_n)
```

```

s_3 <- sin(3 * omega * t)
# Vector of sin(3 omega t_1), sin(3 omega t_2),...,sin(3 omega t_n)
X_cs <- cbind(one, c_1, s_1, c_2, s_2, c_3, s_3)
# Put all the columns together in a matrix
round(crossprod(X_cs), 2)
# works out X^T X; it is diagonal!
#>      one c_1 s_1 c_2 s_2 c_3 s_3
#> one   12   0   0   0   0   0   0
#> c_1    0   6   0   0   0   0   0
#> s_1    0   0   6   0   0   0   0
#> c_2    0   0   0   6   0   0   0
#> s_2    0   0   0   0   6   0   0
#> c_3    0   0   0   0   0   6   0
#> s_3    0   0   0   0   0   0   6

```

Analysing data that are observed over time is beyond the scope of this chapter and so we do not pursue it further, except to recommend the interested reader to P. S. P. Cowpertwait and A. V. Metcalfe, *Introductory Time Series with R* (New York, Springer, 2009), G. Nason, *Wavelet Methods in Statistics with R* (New York, Springer, 2008) and W. A. Woodward, H. K. Gray and A. C. Elliott, *Applied Time Series Analysis with R* (second edition, Boca Raton, FL, Chapman and Hall/CRC, 2017).

11.5.2 Residuals

The process of fitting a straight line through the data by minimizing the sum of squares of the errors does not involve any statistics as such. However, we often need to test whether the slope of the regression line is significantly different from zero, because this will reveal whether there is any dependence between the random variables. For this purpose we must make the assumption that the unknown errors ε_i , have a normal distribution:

$$\varepsilon_i \sim N(0, \sigma_E^2)$$

The unknown errors can be estimated as

$$\hat{\varepsilon}_i = Y_i - (\hat{a} + \hat{b}X_i)$$

and these quantities, which are often denoted e_i , are known as residuals. Loosely speaking, residuals represent what is left over in Y once the effect of X has been removed. The unknown variance σ_E^2 can be estimated by defining

$$S_E^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n} \sum_{i=1}^n [Y_i - (\hat{a} + \hat{b}X_i)]^2$$

Using the earlier result that $\hat{a} = \bar{Y} - \hat{b}\bar{X}$ gives a more convenient form:

$$\begin{aligned} S_E^2 &= \frac{1}{n} \sum_i [(Y_i - \bar{Y}) - \hat{b}(X_i - \bar{X})]^2 \\ &= \frac{1}{n} \sum_i [(Y_i - \bar{Y})^2 - 2\hat{b}(X_i - \bar{X})(Y_i - \bar{Y}) + \hat{b}^2(X_i - \bar{X})^2] \\ &= S_Y^2 - 2\hat{b}S_{XY} + \hat{b}^2S_X^2 \\ &= S_Y^2 - \hat{b}^2S_X^2 \end{aligned}$$

since $S_{XY} = \hat{b}S_X^2$.

Various confidence intervals are derived in more advanced texts covering linear regression. Here the most useful results will simply be quoted. The $100(1 - \alpha)\%$ confidence interval for the regression slope b is given by

$$\left(\hat{b} \pm t_{\alpha/2, n-2} \frac{S_E}{S_X \sqrt{(n-2)}} \right)$$

It is often useful to have an estimate of the mean value of Y for a given value of X , say $X = x$. The point estimate is $\hat{a} + \hat{b}x$, and the $100(1 - \alpha)\%$ confidence interval for this is

$$\left(\hat{a} + \hat{b}x \pm t_{\alpha/2, n-2} S_E \sqrt{\left[\frac{1 + (x - \bar{X})^2 / S_X^2}{n-2} \right]} \right)$$

Example 11.18

Estimate the residual standard deviation and find a 95% confidence interval for the regression slope for the data in Example 11.17. Also test the hypothesis that the tension in the beam is 10 kN when a voltmeter reading of 15 V is obtained.

Solution Using the results obtained in Example 11.17, the residual standard deviation is

$$S_E = 0.418$$

and, using $t_{0.025, 12} = 2.179$, the 95% confidence interval for b is

$$\left(1.22 \pm 2.179 \frac{0.418}{(4.031) \sqrt{12}} \right) = (1.16, 1.29)$$

Obviously the regression slope is significant – but this is not in doubt. To test the hypothesis that the tension is 10 kN, we can use the 95% confidence interval for the corresponding voltage, which is

$$\begin{aligned} & \left(2.89 + 1.22(10) \pm 2.179(0.418) \sqrt{\left[\frac{1 + (10 - 7.5)^2 / (4.031)^2}{12} \right]} \right) \\ &= (14.8, 15.4) \end{aligned}$$

The measured value of 15 V lies within this interval, so the hypothesis is accepted at the 5% level. A better way to approach this would be to reverse the regression (use force as the Y variable and voltage as the X variable), so that a confidence interval for the tension in the beam for a given voltage could be obtained and the assumed value tested. For the present data this gives (9.6, 10.1) at 95%, so again the hypothesis is accepted (Exercise 27).



We can extract the residuals from the linear model object `m` that we defined above:

```
epsilon_hat <- residuals(m); epsilon_hat
#>           1            2            3            4
#>  0.28285714 -0.44087912 -0.16461538 -0.48835165
#>           5            6            7            8
#> -0.21208791  0.06417582  0.24043956  0.51670330
```

```

#>          9          10          11          12
#> 0.89296703 0.16923077 0.14549451 -0.37824176
#>          13          14
#> 0.09802198 -0.72571129
n <- length(epsilon_hat)

```

To compute S_E :

```

S_2_E <- S_2_Y - b_hat^2 * S_2_X; S_2_E
#> [1] 0.174314
S_E <- sqrt(S_2_E); S_E
#> [1] 0.4175092
# Direct calculation from the residuals
sum(epsilon_hat^2) / n; sqrt(sum(epsilon_hat^2) / n)
#> [1] 0.174314
#> [1] 0.4175092

```

This number can be obtained from m by scaling $\text{summary}(m)\$sigma$ which is defined using a $\sqrt{n-2}$; divisor (there are 2 parameters a and b) instead of a \sqrt{n} divisor:

```

summary(m)\$sigma * sqrt((n - 2) / n)
# sigma has to be scaled
#> [1] 0.4175092

```

For the confidence interval for b :

```

alpha <- 0.05
df <- n - 2 # Number of degrees of freedom
t_value <- qt(alpha / 2, df, lower.tail = FALSE); t_value
#> [1] 2.178813
se_b_hat <- S_E / (S_X * sqrt(n - 2))
# Confidence interval
c(b_hat - t_value * se_b_hat, b_hat + t_value * se_b_hat)
#> [1] 1.158593 1.288879

```

We can obtain 95% confidence intervals for a and b directly using the `confint` function:

```

confint(m)
#>                   2.5 %      97.5 %
#> (Intercept) 2.338733  3.448080
#> X           1.158593  1.288879

```

To work out the 95% confidence interval for the voltage when the tension is 10 kN use the formula:

```

X_new <- 10
Y_hat <- a_hat + b_hat * X_new; Y_hat
# Point (single value) estimate
#> [1] 15.13077
se_Y_hat <- S_E * sqrt((1 + (X_new - X_bar)^2 / S_2_X) /
(n - 2))
c(Y_hat - t_value * se_Y_hat, Y_hat + t_value * se_Y_hat)
#> [1] 14.82177 15.43977

```

This can be obtained using the `predict` function:

```
predict(m, newdata = data.frame(X = X_new), interval =
"confidence")
#>           fit      lwr      upr
#> 1  15.13077 14.82177 15.43977
```

The lower and upper values of the confidence intervals are in the `lwr` and `upr` columns, while the point (single value) estimate of the voltage when the tension is 10 kN is in the `fit` column.

We can obtain p-values for the test of the null hypothesis $H_0 : a = 0$ against the alternative hypothesis $H_1 : a \neq 0$ and for the test of $H_0 : b = 0$ against $H_1 : b \neq 0$ using the `summary` function. The p-values are given in the `Pr(>|t|)` column. Other information is provided:

```
summary(m)
#>
#> Call:
#> lm(formula = Y ~ X, data = df_strain)
#>
#> Residuals:
#>   Min     1Q Median     3Q    Max
#> -0.7257 -0.3367  0.0811  0.2226  0.8930
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  2.8934    0.2546   11.37 8.84e-08 ***
#> X            1.2237    0.0299   40.93 2.93e-14 ***
#> -
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
#>
#> Residual standard error: 0.451 on 12 degrees of
#> freedom
#> Multiple R-squared:  0.9929, Adjusted R-squared:
#> 0.9923
#> F-statistic: 1675 on 1 and 12 DF, p-value: 2.929e-14
```

The reverse regression can be performed as

```
m_rev <- lm(X ~ Y, data = df_strain)
Y_new <- 15
predict(m_rev, newdata = data.frame(Y = Y_new),
       interval = "confidence")
#>           fit      lwr      upr
#> 1  9.876119  9.627683 10.12455
```

The reverse regression approach used here is related to the topic of ‘statistical calibration’, which is discussed briefly in M. Aitkin, B. Francis, J. Hinde and R. Darnell, *Statistical Modelling in R* (Oxford, Oxford University Press, 2009), for example.

11.5.3 Regression and correlation

Both regression and correlation are statistical methods for measuring the linear dependence of one random variable upon another, so it is not surprising that there is a connection between them. From the definition of the sample correlation r_{XY} (Section 11.4.4) and the result for the regression slope \hat{b} , it follows immediately that

$$r_{X,Y} = \frac{S_{XY}}{S_X S_Y} = \frac{\hat{b} S_X}{S_Y}$$

Another expression for the residual variance is then

$$S_E^2 = S_Y^2 - \left(\frac{S_Y r_{X,Y}}{S_X} \right)^2 S_X^2 = S_Y^2 (1 - r_{X,Y}^2)$$

This result has an important interpretation. S_Y^2 is the total variation in the Y values, and S_E^2 is the residual variation after the regression line has been identified, so $r_{X,Y}^2$ is the proportion of the total variation in the Y values that is accounted for by the regression on X : informally, it represents how closely the points are clustered about the line. This is a measure of goodness-of-fit that is especially useful when the dependence between X and Y is nonlinear and different models are to be compared.



We can confirm this result in R:

```
r_XY <- cor(X, Y) # Correlation between X and Y
S_XY / (S_X * S_Y); b_hat * S_X / S_Y # The same
#> [1] 0.9964376
#> [1] 0.9964376
S_2_E
# Residual variation after regression line has been identified
#> [1] 0.174314
S_2_Y * (1 - r_XY^2) # The same
#> [1] 0.174314
r_XY^2
# Proportion of the total variation in Y accounted for by
# the regression
#> [1] 0.9928878
```

This important measure of goodness-of-fit is given by `summary` under `Multiple R-squared`. It can be extracted using:

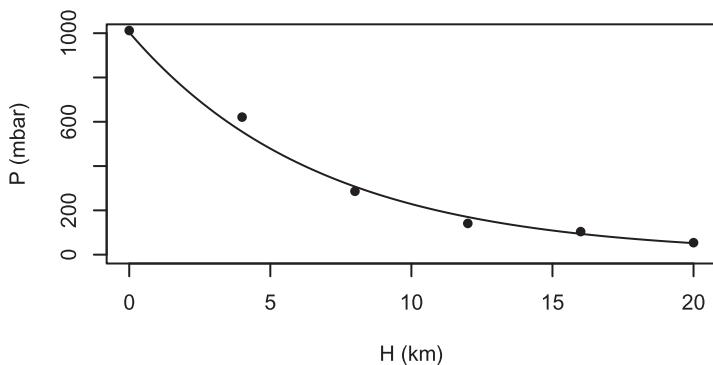
```
summary(m)$r.squared
#> [1] 0.9928878
```

We perform further analysis of the data in Example 11.17 in Exercise 74.

11.5.4 Nonlinear regression

Sometimes the dependence between two random variables is nonlinear, and this shows clearly in the scatter plot; see for instance Figure 11.13. Fitting a straight line through the data would hardly be appropriate. Instead, various models of the dependence can be

Figure 11.13
Nonlinear regression
(Example 11.19).



assumed and tested. In each case the value of $r^2_{X,Y}$ indicates the success of the model in capturing the dependence. One form of nonlinear regression model involves a quadratic or higher-degree polynomial:

$$Y_i = a_0 + a_1 X_i + a_2 X_i^2 + \varepsilon_i \quad (i = 1, \dots, n)$$

The three coefficients a_0 , a_1 and a_2 can be identified by a multivariate regression method that is beyond the scope of this text.



One way of fitting a quadratic model in R is:

```
m_quadratic <- lm(Y ~ X + I(X^2))
# The calculation of X^2 must be enclosed in I
coef(m_quadratic)
#> (Intercept)           X           I(X^2)
#>  2.31373626    1.44111264   -0.01449176
```

Equivalently, we can use the `poly` function:

```
m_quadratic_2 <- lm(Y ~ poly(X, degree = 2, raw = TRUE))
coef(m_quadratic_2)
#> (Intercept)  poly(X, degree = 2, raw = TRUE) 1
#> 2.31373626                           1.44111264
#> poly(X, degree = 2, raw = TRUE) 2
#> -0.01449176
```

If we use orthogonal polynomials, the estimated coefficients are different but the fitted values are the same:

```
# The default is raw = FALSE for orthogonal polynomials,
# so we do not need to specify it
m_quadratic_3 <- lm(Y ~ poly(X, degree = 2))
coef(m_quadratic_3)
#> (Intercept)  poly(X, degree = 2) 1  poly(X, degree = 2) 2
#> 12.071429          18.457740          -0.782018
round(fitted(m_quadratic_3) - fitted(m_quadratic), 5)
#> 1 2 3 4 5 6 7 8 9 10 11 12 13 14
#> 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

As mentioned earlier, there are considerable inferential advantages associated with using orthogonal polynomials; see J. J. Faraway, *Linear Models with R* (second edition, Boca Raton, FL, Chapman and Hall/CRC, 2015), for example, for more details.

A simpler approach is to try models of the form

$$Y_i = aX_i^b\eta_i \quad (i = 1, \dots, n)$$

or

$$Y_i = \exp(a + bX_i)\eta_i \quad (i = 1, \dots, n)$$

where each η_i is a positive multiplicative error. On taking logarithms, these models reduce to the standard linear form:

$$\ln Y_i = \ln a + b \ln X_i + \varepsilon_i \quad (i = 1, \dots, n)$$

or

$$\ln Y_i = a + bX_i + \varepsilon_i \quad (i = 1, \dots, n), \text{ in which } \varepsilon_i = \ln(\eta_i)$$

which can be solved by the usual method.

Example 11.19

The following data for atmospheric pressure (P , in mbar) at various heights (H , in km) have been obtained:

Height H	0	4	8	12	16	20
Pressure P	1012	621	286	141	104	54

These data are plotted in Figure 11.13. The relationship between height and pressure is believed to be of the form

$$P = e^{a+bH}$$

where a and b are constants. Fit and assess a model of this form and predict the atmospheric pressure at a height of 14 km.

Solution

Taking logarithms and setting $Y = \ln P$, leads to a model of the form

$$Y_i = a + bH_i + \varepsilon_i$$

for which the following results are easily obtained:

$$\bar{H} = 10, \quad S_H = 6.83130, \quad Y = 5.43152, \quad S_Y = 1.01638, \quad \bar{HY} = 47.4081$$

Hence

$$\hat{b} = \frac{\bar{HY} - (\bar{H})(\bar{Y})}{S_H^2} = -0.148$$

$$\hat{a} = Y - \hat{b}\bar{H} = 6.91$$

Also, $r_{H,Y}^2 = 0.99$, which implies that the fit is very good (Figure 11.13). In this case there is not much point in trying other models. Finally, the predicted pressure at a height of 14 km is

$$P = e^{6.91 - 0.148(14)} = 126 \text{ mbar}$$



Here are details of the calculations in R, followed by a full analysis using the `lm` function. A plot of the fitted model is also produced.

```

H <- c(0, 4, 8, 12, 16, 20)
P <- c(1012, 621, 286, 141, 104, 54)
Y <- log(P) # Natural log of P
# Step by step calculations, starting with some sums
H_sum <- sum(H); H_sum; Y_sum <- sum(Y); Y_sum
#> [1] 60
#> [1] 32.58914
H_2_sum <- sum(H^2); H_2_sum; Y_2_sum <- sum(Y^2); Y_2_sum
#> [1] 880
#> [1] 183.2069
HY_sum <- sum(H * Y); HY_sum
#> [1] 284.4483
n <- length(H); n
#> [1] 6
H_bar <- H_sum / n; H_bar; Y_bar <- Y_sum / n; Y_bar
#> [1] 10
#> [1] 5.431524
S_2_H <- (H_2_sum / n) - H_bar ^2; S_2_H
S_H <- sqrt(S_2_H); S_H
#> [1] 46.66667
#> [1] 6.831301
S_2_Y <- (Y_2_sum / n) - Y_bar ^2; S_2_Y
S_Y <- sqrt(S_2_Y); S_Y
#> [1] 1.03303
#> [1] 1.016381
HY_bar <- HY_sum / n; HY_bar
#> [1] 47.40805
S_HY <- HY_bar - H_bar * Y_bar; S_HY
#> [1] -6.907184
# Estimate of slope
b_hat <- S_HY /S_2_H; b_hat
#> [1] -0.1480111
# Estimate of intercept
a_hat <- Y_bar - b_hat * H_bar; a_hat
#> [1] 6.911634
# Estimate of r_HY and its square
r_HY <- cor(H, Y) # Correlation between H and Y
S_HY / (S_H * S_Y); b_hat * S_H / S_Y # The same
#> [1] -0.9948122
#> [1] -0.9948122
r_HY^2
#> [1] 0.9896514

```

It's easier to use `lm`, and much more is available:

```

m <- lm(Y ~ H)
coef(m)

```

```

#> (Intercept)      H
#> 6.9116344 -0.1480111
confint(m)
#>               2.5 %    97.5 %
#> (Intercept) 6.6571750 7.1660938
#> H           -0.1690224 -0.1269998
summary(m)
#>
#> Call:
#> lm(formula = Y ~ H)
#>
#> Residuals:
#>       1        2        3        4        5        6
#> 0.008049 0.111741 -0.071554 -0.186742 0.100934 0.037571
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 6.911634 0.091649 75.41 1.85e-07 ***
#> H           -0.148011 0.007568 -19.56 4.03e-05 ***
#> -
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.1266 on 4 degrees of freedom
#> Multiple R-squared: 0.9897, Adjusted R-squared: 0.9871
#> F-statistic: 382.5 on 1 and 4 DF, p-value: 4.03e-05

```

We can perform the prediction in two ways. First, we can use the formula:

```

H_new <- 14
exp(a_hat + b_hat * H_new)
#> [1] 126.4035

```

Secondly, we can use the `predict` function, transforming the result to the scale of P by applying the exponential function:

```

Y_new <- predict(m, newdata = data.frame(H = H_new))
P_new <- exp(Y_new); P_new
# Transform to the scale of P by applying the exp
# function
#> 1
#> 126.4035

```

The following code produces Figure 11.13.

```

# Calculate the fitted values of P given by the model
H_seq <- seq(from = min(H), to = max(H), length = 100)
# 100 equally spaced values along H

Y_pred <- predict(m, newdata = data.frame(H = H_seq))
# predict Y at these values

P_pred <- exp(Y_pred)
# Transform to the scale of P by applying the exp function

```

```
# Plot the data
plot(H, P,
      xlab = "H (km)", ylab = "P (mbar)",
      xlim = c(0,20), ylim = c(0,1000),
      pch = 16)
# Add the fitted curve
lines(H_seq, P_pred)
```

A similar model

$$P_i = \exp(a + bH_i) + \text{error}_i$$

can be fitted using the `nls` function, standing for nonlinear least squares. The algorithm used is iterative and initial values have to be supplied for a and b .

```
m_similar <- nls(P ~ exp(a + b * H), start = list(a = 1,
b = -0.1))
summary(m_similar)
#>
#> Formula: P ~ exp(a + b * H)
#>
#> Parameters:
#> Estimate Std. Error t value Pr(>|t|)
#> a 6.935858 0.032709 212.05 2.97e-09 ***
#> b -0.148539 0.009203 -16.14 8.62e-05 ***
#> -
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 35.2 on 4 degrees of freedom
#>
#> Number of iterations to convergence: 10
#> Achieved convergence tolerance: 8.897e-06
```

11.5.5 Exercises



Check your calculations using R whenever possible.

- 25** Ten files of audio data are annotated by a human labeller. The time (s) this takes per file is a function of the length of the file (s) as follows:

File length (X)	5.4 7.9 10.0 14.2 16.1 16.8 19.6 22.0 25.0 26.7
Annotation time (Y)	13.1 17.3 23.9 30.1 33.5 40.0 43.6 46.5 52.4 60.7

Find the linear regression coefficients.

26

Measured deflections (in mm) of a structure under a load (in kg) were recorded as follows:

Load X	1	2	3	4	5	6	7	8	9	10	11	12
Deflection Y	16	35	45	74	86	96	106	124	134	156	164	182

Draw a scatter plot of the data. Find the linear regression coefficients and predict the deflection for a load of 15 kg.

27

Using the data in Example 11.17, carry out a regression of force against voltage, and obtain a

95% confidence interval for the tension in the beam when the voltmeter reads 15 V, as described in Example 11.18.

- 28 Weekly advertising expenditures X_i and sales Y_i for a company are as follows (in units of £100):

X_i	40	20	25	20	30	50	40	20	50	40	25	50
Y_i	385	400	395	365	475	440	490	420	560	525	480	510

- (a) Fit a regression line and predict the sales for an advertising expenditure of £6000.
- (b) Estimate the residual standard deviation and find a 95% confidence interval for the regression slope. Hence test the hypothesis that the sales do not depend upon advertising expenditure.
- (c) Find a 95% confidence interval for the mean sales when advertising expenditure is £6000.

- 29 A machine that can be run at different speeds produces articles, of which a certain number are defective. The number of defective items produced per hour depends upon machine speed (in rev s⁻¹) as indicated in the following experimental run:

Speed	8	9	10	11	12	13	14	15
Defectives per hour	7	12	13	13	13	16	14	18

Find the regression line for number of defectives against speed, and a 90% confidence interval for the mean number of defectives per hour when the speed is 14 rev s⁻¹.

- 30 Sometimes it is required that the regression line passes through the origin, in which case the only regression coefficient is the slope of the line.

Use the least-squares procedure to show that the estimate of the slope is then

$$\hat{b} = (\sum_i X_i Y_i) / \sum_i X_i^2$$

- 31 A series of measurements of voltage across and current through a resistor produced the following results:

Voltage (V)	1	2	3	4	5	6	7	8	9	10	11	12
Current (mA)	6	18	27	30	42	48	58	69	74	81	94	99

Estimate the resistance, using the result of the previous exercise.

- 32 The pressure P of a gas corresponding to various volumes V was recorded as follows:

V (cm ³)	50	60	70	90	100
P (kg cm ⁻²)	64.7	51.3	40.5	25.9	7.8

The ideal gas law is given by the equation

$$PV^\lambda = C$$

where λ and C are constants. By taking logarithms and using the least-squares method, estimate λ and C from the data and predict P when $V = 80$ cm³.

- 33 The following data show the unit costs of producing certain electronic components and the number of units produced:

Lot size X_i	50	100	250	500	1000	2000	5000
Unit cost Y_i	108	65	21	13	4	2.2	1

Fit a model of the form $Y = aX^b$ and predict the unit cost for a lot size of 300.

11.6 Goodness-of-fit tests

The common classes of distributions, especially the binomial, Poisson and normal distributions, which often govern the data in experimental contexts, are used as the basis for statistical methods of estimation and testing. A question that naturally arises is whether or not a given set of data actually follows an assumed distribution. If it does then the statistical methods can be used with confidence. If not then some alternative should be considered. The general procedure used for testing this can also be used to test for dependence between two variables.

11.6.1 Chi-square distribution and test

No set of data will follow an assumed distribution exactly, but there is a general method for testing a distribution as a statistical hypothesis. If the hypothesis is accepted then it is reasonable to use the distribution as an approximation to reality.

First the data must be partitioned into classes. If the data consist of observations from a discrete distribution then they will be in classes already, but it may be appropriate to combine some classes if the numbers of observations are small. For each class the number of observations that would be expected to occur under the assumed distribution can be worked out. The following quantity acts as a test statistic for comparing the observed and expected class numbers:

$$\chi^2 = \sum_{k=1}^m \frac{(f_k - e_k)^2}{e_k}$$

where f_k is the number of observations in the k th class, e_k is the expected number in the k th class and m is the number of classes. Clearly, χ^2 is a non-negative quantity whose magnitude indicates the extent of the discrepancy between the discretized histogram of data and the assumed distribution. For small samples the histogram is erratic and the comparison invalid, but for large samples the histogram should approximate the true distribution. It can be shown that for a large sample the random variable χ^2 has a '**chi-square distribution**'. This class of distributions is widely used in statistics, and a typical chi-square probability density function is shown in Figure 11.14. We are interested in particular in the value of $\chi_{\alpha,n}^2$ to the right of which the area under the density function curve is α , where n is the (single) parameter of the distribution. These values are extensively tabulated; a typical table is shown in Figure 11.15.

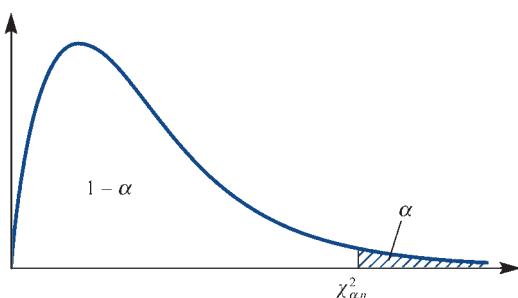


Figure 11.14
The chi-square distribution with $\chi_{\alpha,n}^2$.

Figure 11.15

Table of the chi-square distribution $\chi_{\alpha,n}^2$.

n	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$	n
1	3.841	5.024	6.635	7.879	1
2	5.991	7.378	9.210	10.597	2
3	7.815	9.348	11.345	12.838	3
4	9.488	11.143	13.277	14.860	4
5	11.070	12.832	15.086	16.750	5
6	12.592	14.449	16.812	18.548	6
7	14.067	16.013	18.475	20.278	7
8	15.507	17.535	20.090	21.955	8
9	16.919	19.023	21.666	23.589	9
10	18.307	20.483	23.209	25.188	10
11	19.675	21.920	24.725	26.757	11
12	21.026	23.337	26.217	28.300	12
13	22.362	24.736	27.688	29.819	13
14	23.685	26.119	29.141	31.319	14
15	24.996	27.488	30.578	32.801	15
16	26.296	28.845	32.000	34.267	16
17	27.587	30.191	33.409	35.718	17
18	28.869	31.526	34.805	37.156	18
19	30.144	32.852	36.191	38.582	19
20	31.410	34.170	37.566	39.997	20
21	32.671	35.479	38.932	41.401	21
22	33.924	36.781	40.289	42.796	22
23	35.172	38.076	41.638	44.181	23
24	36.415	39.364	42.980	45.558	24
25	37.652	40.646	44.314	46.928	25
26	38.885	41.923	45.642	48.290	26
27	40.113	43.194	46.963	49.645	27
28	41.337	44.461	48.278	50.993	28
29	42.557	45.722	49.588	52.336	29
30	43.773	46.979	50.892	53.672	30



These tabulated values can be easily calculated in R. Here is an example:

```
alpha <- 0.05
n <- 1 # Number of degrees of freedom
qchisq(alpha, 1, lower.tail = FALSE)
#> [1] 3.841459
```

The hypothesis of the assumed distribution is rejected if

$$\chi^2 > \chi_{\alpha,m-t-1}^2$$

where α is the significance level and t is the number of independent parameters estimated from the data and used for computing the e_k values. The significance level is the probability of false rejection, as discussed in Section 11.3.4. Sometimes the hypothesis is deliberately vague, for example a parameter value may be left unspecified. If the data themselves are used to fix parameter values in the assumed distribution

before testing then the test must be strengthened to allow for this in the form of a correction t in the chi-square parameter.

A useful rule of thumb when using this test is that there should be at most a small number (one or two) of classes with an expected number of observations less than five. If necessary, classes in the tails of the distribution can be merged.

Example 11.20

A die is tossed 600 times, and the numbers of occurrences of the numbers one to six are recorded respectively as 89, 113, 98, 104, 117 and 79. Is the die fair or biased?

Solution

The expected values are $e_k = 100$ for all k , and the test value is $\chi^2 = 10.4$. This is less than $\chi^2_{0.05,5} = 11.07$, so we should expect results as erratic as this at least once in 20 similar experiments. The die may be biased, but the data are insufficient to justify this conclusion.



We can perform the calculation of the χ^2 statistic in R:

```
f <- c(89, 113, 98, 104, 117, 79)
e <- rep(100, 6) # 100 six times
f - e; (f - e)^2; (f - e)^2 / e
#> [1] -11 13 -2 4 17 -21
#> [1] 121 169 4 16 289 441
#> [1] 1.21 1.69 0.04 0.16 2.89 4.41
chisq <- sum((f - e)^2 / e); chisq
#> [1] 10.4
```

We can perform the test directly using the `chisq.test` function. We specify the probabilities of the six outcomes; here the outcomes are equally likely.

```
chisq.test(f, p = rep(1/6, 6))
#>
#> Chi-squared test for given probabilities
#>
#> data: f
#> X-squared = 10.4, df = 5, p-value = 0.06466
```

The expected values e_k , $k = 1, \dots, 6$, can be extracted:

```
chisq.test(f, p = rep(1/6, 6))$expected
#> [1] 100 100 100 100 100 100
```

Example 11.21

The numbers of trucks arriving per hour at a warehouse are counted for each of 500 h. Counts of zero up to eight arrivals are recorded on respectively 52, 151, 130, 102, 45, 12, 5, 1 and 2 occasions. Test the hypothesis that the numbers of arrivals have a Poisson distribution, and estimate how often there will be nine or more arrivals in one hour.

Solution

The hypothesis stipulates a Poisson distribution, but without specifying the parameter λ . Since the mean of the Poisson distribution is λ and the average number of arrivals per

Figure 11.16
Chi-square calculation
for Example 11.21.

Trucks	f_k	p_k	e_k	χ^2
0	52	0.1353	67.7	3.63
1	151	0.2707	135.3	1.81
2	130	0.2707	135.3	0.21
3	102	0.1804	90.2	1.54
4	45	0.0902	45.1	0.00
5	12	0.0361	18.0	2.02
6	5	0.0120	6.0	0.17
7 or more	3	0.0046	2.3	0.24
Totals	500	1.0	500	9.62

hour is 2.02 from the data, it is reasonable to assume that $\lambda = 2$. The columns in the table in Figure 11.16 show the observed counts f_k , the Poisson probabilities p_k , the expected counts $e_k = 500 p_k$ and the individual χ^2 values for each class. The last two classes have been combined because the numbers are so small. One parameter has been estimated from the data, so the total χ^2 value is compared with $\chi^2_{0.05,6} = 12.59$. The Poisson hypothesis is accepted, and on that basis the probability of nine or more trucks arriving in one hour is

$$P(9 \text{ or more}) = 1 - \sum_{k=0}^8 \frac{2^k e^{-2}}{k!} = 0.000\,237$$

This will occur roughly once in every 4200 h of operation.



We can perform this calculation in R (with more accuracy) as follows:

```
number_hours <- 0:8
f <- c(52, 151, 130, 102, 45, 12, 5, 1, 2)
lambda_hat <- sum(number_hours * f) / sum(f); lambda_hat
#> [1] 2.02
# Set lambda_hat to 2, for simplicity
lambda_hat <- 2
# Work out the probabilities
p <- dpois(number_hours, lambda_hat); p
#> [1] 0.1353352832 0.2706705665 0.2706705665 0.1804470443
0.0902235222
#> [6] 0.0360894089 0.0120298030 0.0034370866 0.0008592716
# Eight classes only, with probabilities summing to 1
p_comb <- c(p[1:7], 1 - sum(p[1:7])); p_comb
#> [1] 0.135335283 0.270670566 0.270670566 0.180447044
0.090223522 0.036089409
#> [7] 0.012029803 0.004533806
f_comb <- c(f[1:7], sum(f[8:9])); f_comb
#> [1] 52 151 130 102 45 12 5 3
```

```

# Expected values
e_comb <- 500 * p_comb; e_comb
#> [1] 67.667642 135.335283 135.335283 90.223522 45.111761
18.044704
#> [7] 6.014901 2.266903
(f_comb - e_comb); (f_comb - e_comb)^2
(f_comb - e_comb)^2 / e_comb
#> [1] -15.6676416 15.6647168 -5.3352832 11.7764778
-0.1117611 -6.0447044
#> [7] -1.0149015 0.7330972
#> [1] 245.47499388 245.38335128 28.46524721 138.68543037
0.01249054
#> [6] 36.53845166 1.03002501 0.53743156
#> [1] 3.6276570013 1.8131513483 0.2103313085 1.5371316377
0.0002768799
#> [6] 2.0248850184 0.1712455328 0.2370774642
sum((f_comb - e_comb)^2 / e_comb)
#> [1] 9.621756
qchisq(0.05, 8 - 1 - 1, lower.tail = FALSE)
# t = 1 as lambda is estimated
#> [1] 12.59159
# Probability
1 - ppois(8, lambda_hat); ppois(8, lambda_hat, lower.tail =
= FALSE)
#> [1] 0.0002374473
#> [1] 0.0002374473

```

Because so many statistical methods assume normal data, it is important to have a test for normality, and the chi-square method can be used (Exercise 38 and Section 11.7.4).

11.6.2 Contingency tables

In Section 11.4.3 the correlation was introduced as a measure of dependence between two random variables. The sample correlation (Section 11.4.4) provides an estimate from data. This measure only applies to numerical random variables, and then only works for linear dependence (Exercise 15). The rank correlation (Section 11.4.6) has more general applicability, but still requires that the data be classified in order of rank. The chi-square testing procedure can be adapted to provide at least an indicator of dependence that has the widest applicability of all.

Suppose that each item in a sample of size n can be separately classified as one of A_1, \dots, A_r by one criterion, and as one of B_1, \dots, B_c by another (these may be numerical values, but not necessarily). The number f_{ij} of items in the sample that are classified as ' A_i and B_j ' can be counted for each $i = 1, \dots, r$ and $j = 1, \dots, c$. The table of these numbers (with r rows and c columns) is called a **contingency table** (Figure 11.17). The question is whether the two criteria are independent. If not then some combinations of A_i and B_j will occur significantly more often (and others less often) than would be expected under the assumption of independence. We first have to work out how many would be expected under an assumption of independence.

Figure 11.17

Contingency table.

Class	B_1	...	B_c	Total
A_1	f_{11}	...	f_{1c}	f_{1+}
.	.		.	.
.	.		.	.
.	.		.	.
A_r	f_{r1}	...	f_{rc}	f_{r+}
Total	f_{+1}	...	f_{+c}	n

Let the row and column totals be denoted by

$$f_{i+} = \sum_{j=1}^c f_{ij} \quad (i = 1, \dots, r)$$

$$f_{+j} = \sum_{i=1}^r f_{ij} \quad (j = 1, \dots, c)$$

If the criteria are independent then the joint probability for each combination can be expressed as the product of the separate marginal probabilities:

$$P(A_i \cap B_j) = P(A_i)P(B_j)$$

The chi-square procedure will be used to see how well the data fit this assumption. To test it, we can estimate the marginal probabilities from the row and column totals,

$$P(A_i) \approx \frac{f_{i+}}{n}, \quad P(B_j) \approx \frac{f_{+j}}{n}$$

and multiply the product of these by n to obtain the expected number e_{ij} for each combination:

$$e_{ij} = n \frac{f_{i+}}{n} \frac{f_{+j}}{n} = \frac{f_{i+} f_{+j}}{n}$$

The chi-square goodness-of-fit statistic follows from the actual and expected numbers (f_{ij} and e_{ij}) as a sum over all the rows and columns:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

If the value of this is large then the hypothesis of independence is rejected, because the actual and expected counts differ by more than can be attributed to chance. As explained in Section 11.6.1, the largeness is judged with respect to $\chi^2_{\alpha, m-t-1}$ from the chi-square table. The number of classes, m , is the number of rows times the number of columns, rc . The number of independent parameters estimated from the data, t , is the number of independent marginal probabilities $P(A_i)$ and $P(B_j)$:

$$t = (r - 1) + (c - 1)$$

The number is not $r + c$, because the row and column totals must equal one, so when all but one are specified, the last is determined. Finally,

$$m - t - 1 = rc - (r + c - 2) - 1 = (r - 1)(c - 1)$$

The hypothesis of independence is therefore rejected (at significance level α) if

$$\chi^2 > \chi_{\alpha, (r-1)(c-1)}^2$$

Example 11.22

An accident inspector makes spot checks on working practices during visits to industrial sites chosen at random. At one large construction site the numbers of accidents occurring per week were counted for a period of three years, and each week was also classified as to whether or not the inspector had visited the site during the previous week. The results are shown in bold print in Figure 11.18. Do visits by the inspector tend to reduce the number of accidents, at least in the short term?

Figure 11.18
Contingency table
for Example 11.22.

	Number of accidents				Total
	0	1	2	3	
Visit Residual	20 (13.38) 2.96	3 (7.08) -1.99	1 (2.46) -1.07	0 (1.08) -1.16	24
No visit Residual	67 (73.62) -2.96	43 (38.92) 1.99	15 (13.54) 1.07	7 (5.92) 1.16	132
Total	87	46	16	7	156

Solution

The respective row and column totals are shown in Figure 11.18, together with the expected numbers e_{ij} in parentheses in each cell. For example, the top left cell has observed number 20, row total 24, column total 87, $n = 156$, and hence the expected number

$$e_{11} = (24)(87)/156 = 13.38$$

The chi-square sum is

$$\begin{aligned}\chi^2 &= \frac{(20 - 13.38)^2}{13.38} + \dots + \frac{(7 - 5.92)^2}{5.92} \\ &= 8.94\end{aligned}$$

With two rows, four columns and a significance level of 5%, the appropriate number from the chi-square table is $\chi^2_{0.05,3} = 7.815$. The calculated value exceeds this, and by comparing the observed and expected numbers in the table, it seems clear that the visits by the inspector do tend to reduce the number of accidents. This is not, however, significant at the 2.5% level.



We can perform this analysis using the `chisq.test` function. First, we need to specify the contingency table data as a matrix. We will also compute row, column and overall totals for later use.

```
f <- matrix(c(20, 3, 1, 0, 67, 43, 15, 7),
             nrow = 2,
             # 2 rows (number of columns follows from this)
             byrow = TRUE) # We input the data row by row
# Add row and column names
dimnames(f) <- list("Inspector" = c("Visit", "No Visit"),
                     "Accidents" = 0:3)
f
#>           Accidents
#> Inspector 0 1 2 3
#> Visit      20 3 1 0
#> No Visit   67 43 15 7
#
# Compute the row, column and overall totals for later use
f_row <- rowSums(f); f_col <- colSums(f)
# Row and column totals
n <- sum(f) # Overall total
f_row; f_col; n
#>     Visit No Visit
#>        24    132
#> 0 1 2 3
#> 87 46 16 7
#> [1] 156
# An alternative way of computing row and column totals,
# that will be useful later,
# is by means of the apply function
apply (f, MARGIN = 1, FUN = sum)
# MARGIN 1 corresponds to rows
#>     Visit No Visit
#>        24    132
apply(f, MARGIN = 2, FUN = sum)
# MARGIN 2 corresponds to columns
#> 0 1 2 3
#> 87 46 16 7
```

Now we can perform and confirm the analysis:

```
# Test of independence
chi_ind <- chisq.test(f); chi_ind
#>
#> Pearson's Chi-squared test
#>
#> data: f
#> X-squared = 8.9381, df = 3, p-value = 0.03012
alpha <- 0.05
df <- (nrow(f) - 1) * (ncol(f) - 1); df
#> [1] 3
qchisq(alpha, df, lower.tail = FALSE)
#> [1] 7.814728
```

We can extract the expected values e_{ij} :

```
chi_ind$expected
#>           Accidents
#> Inspector      0       1       2       3
#> Visit     13.38462 7.076923 2.461538 1.076923
#> No Visit  73.61538 38.923077 13.538462 5.923077
```

The values of the so-called Pearson residuals $(f_{ij} - e_{ij})/\sqrt{e_{ij}}$ can be obtained from:

```
chi_ind$residuals
#>           Accidents
#> Inspector      0       1       2       3
#> Visit      1.8082237 -1.5325346 -0.9315516 -1.0377490
#> No Visit -0.7710292 0.6534749 0.3972150 0.4424977
# Confirm by rounding the differences to 5 decimal places
# to take account of numerical inaccuracies; 0.00000
# printed as 0
round(chi_ind$residuals - (f - chi_ind$expected) /
  sqrt(chi_ind$expected), 5)
```

```
#>           Accidents
#> Inspector  0 1 2 3
#> Visit      0 0 0 0
#> No Visit   0 0 0 0
```

The individual contributions $(f_{ij} - e_{ij})^2/e_{ij}$ to the χ^2 statistic can therefore be obtained as:

```
chi_ind$residuals^2
#>           Accidents
#> Inspector      0       1       2       3
#> Visit      3.269673 2.3486622 0.8677885 1.0769231
#> No Visit  0.594486 0.4270295 0.1577797 0.1958042
# Confirm
round(chi_ind$residuals^2 - (f - chi_ind$expected)^2 /
  chi_ind$expected, 5)
```

```
#>           Accidents
#> Inspector  0 1 2 3
#> Visit      0 0 0 0
#> No Visit   0 0 0 0
```

These figures tell us that, after a visit by the inspector, 0 accidents occur more often ($(f_{11} - e_{11})/\sqrt{e_{11}}$ is positive) and 1, 2 and 3 accidents occur less often ($(f_{12} - e_{12})/\sqrt{e_{12}}$, $(f_{13} - e_{13})/\sqrt{e_{13}}$ and $(f_{14} - e_{14})/\sqrt{e_{14}}$ are negative) than would be expected were the number of accidents independent of whether or not the inspector had visited.

A significant chi-square value does not by itself reveal what part or parts of the table are responsible for the lack of independence. A procedure that is often helpful in this respect is to work out the **adjusted residual** for each cell, defined as

$$d_{ij} = \frac{f_{ij} - e_{ij}}{\sqrt{[e_{ij}(1 - f_{i+}/n)(1 - f_{j+}/n)]}}$$

Under the assumption of independence, these are approximately standard normal, so a significant value for a cell suggests that that cell is partly responsible for the dependence overall. The adjusted residuals for the contingency table in Example 11.22 are shown in Figure 11.18, and support the conclusion that visits by the inspector tend to reduce the number of accidents.



The residuals provided by R take the form $(f_{ij} - e_{ij})/\sqrt{e_{ij}}$. These can be transformed to adjusted residuals d_{ij} by first creating a matrix with element (i, j) element $(1 - f_{i+}/n)(1 - f_{j+}/n)$ and then calculating $(f_{ij} - e_{ij})/\sqrt{e_{ij}(1 - f_{i+}/n)(1 - f_{j+}/n)}$ using the outer function:

```
# We have already computed the row, column and overall
# totals
f_row; f_col; n
#> Visit No Visit
#> 24 132
#> 0 1 2 3
#> 87 46 16 7
#> [1] 156
# Required matrix
M <- outer((1 - f_row / n) , (1 - f_col / n))
# Adjusted residuals
chi_ind$residuals / sqrt(M)
#> Accidents
#> Inspector 0 1 2 3
#> Visit 2.955733 -1.984045 -1.069007 -1.154348
#> No Visit -2.955733 1.984045 1.069007 1.154348
# These are available directly
chi_ind$stdres
#> Accidents
#> Inspector 0 1 2 3
#> Visit 2.955733 -1.984045 -1.069007 -1.154348
#> No Visit -2.955733 1.984045 1.069007 1.154348
```

For a useful survey of procedures for analysing contingency tables, see B. S. Everitt, *The Analysis of Contingency Tables*, (second edition, Chapman & Hall, London, 1992). Chapter 6 of J. J. Faraway *Extending the Linear Model with R* (second edition, Boca Raton, Chapman and Hall / CRC, FL, 2016) provides a very detailed modern treatment based on statistical modelling.

11.6.3 Exercises



Check your calculations using R whenever possible.

34

In a genetic experiment, outcome A is expected to occur twice as often as outcome B, which in turn is expected to occur twice as often as outcome C, and exactly one of these three outcomes must occur. In a sample of size 100, outcomes A, B, C occurred 63, 22, 15 times respectively. Test the hypothesis at 5% significance.

35

The number of books borrowed from a library that is open five days a week is as follows: Monday 153, Tuesday 108, Wednesday 120, Thursday 114, Friday 145. Test (at 5% significance) whether the number of books borrowed depends on the day of the week.

36

A new process for manufacturing light fibres is being tested. Out of 50 samples, 32 contained no flaws, 12 contained one flaw and 6 contained two flaws. Test the hypothesis that the number of flaws per sample has a Poisson distribution.

37

In an early experiment on the emission of α -particles from a radioactive source, Rutherford obtained the following data on counts of particles during constant time intervals:

Number of particles	0	1	2	3	4	5	6	7	8	9	10	.10
Number of intervals	57	203	383	525	532	408	273	139	45	27	10	6

Test the hypothesis that the number of particles emitted during an interval has a Poisson distribution.

38

Two samples of 100 data have been grouped into classes as shown in Figure 11.19. The sample average and standard deviation in each case were 10.0 and 2.0 respectively.

- (a) Draw the histogram for each sample.
- (b) Test each sample for normality using the measured parameters.

Class	Sample 1	Sample 2
< 6.5	4	3
6.5–7.5	6	6
7.5–8.5	16	16
8.5–9.5	16	13
9.5–10.5	17	26
10.5–11.5	20	7
11.5–12.5	12	19
12.5–13.5	6	5
> 13.5	3	5

Figure 11.19 Data classification for Exercise 38.

See also Section 11.7.4

39

Shipments of electronic devices have been received by a firm from three sources: A, B and C. Each device is classified as either perfect, intermediate (imperfect but acceptable), or unacceptable. From source A 89 were perfect, 23 intermediate and 12 unacceptable. Corresponding figures for source B were 62, 12 and 8 respectively, and for source C 119, 30 and 21 respectively. Is there any significant difference in quality between the devices received from the three sources?

40

Cars produced at a factory are chosen at random for a thorough inspection. The number inspected and the number of those that were found to be unsuitable for shipment were counted monthly for one year as follows:

Month	Jan.	Feb.	Mar.	Apr.	May	Jun.
Inspected	450	550	550	400	600	450
Defective	8	14	6	3	7	8
Month	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
Inspected	450	200	450	600	600	550
Defective	16	5	12	6	15	9

Is there a significant variation in quality through the year?

- 41 Customers ordering regularly from an on-line clothing catalogue are classed as low, medium and high spenders. Considering four products from the catalogue (a jacket, a shirt, a pair of trousers, and a pair of shoes), the numbers of customers in each class buying these products over a fixed period of time are given in the following table:

Spending level	Jacket	Shirt	Trousers	Shoes
Low	21	94	57	113
Medium	66	157	94	209
High	58	120	41	125

Does this table provide evidence that customers with different spending levels tend to choose different products?

- 42 A quality control engineer takes daily samples of four television sets coming off an assembly line. In a total of 200 working days he found that on 110 days no sets required adjustments, on 73 days one set requires adjustments, on 16 days two sets and on 1 day three sets. Use these results to test the hypothesis that 10% of sets coming off the assembly line required adjustments, at 5% and 1% significance levels. Also test this using the confidence interval for proportion (Section 11.3.6), using the total number of sets requiring adjustments.

11.7 Engineering application: analysis of engine performance data

11.7.1 Introduction

Statistical methods are often used in conjunction with each other. So far in this chapter the examples and exercises have almost always been designed to illustrate the various topics one at a time. This section contains an example of a more extended problem to which several topics are relevant, and correspondingly there are several stages to the analysis.

The background to the problem is this. Suppose that the fuel consumption of a car engine is tested by measuring the time that the engine runs at constant speed on a litre of standard fuel. Two prototype engines, A and B, are being compared for fuel consumption. For each engine a series of tests is performed in various ambient temperatures, which are also recorded. Figure 11.20 contains the data. There are 30 observations each for the four random variables concerned:

- A, running time in minutes for engine A;
- T, ambient temperature in degrees Celsius for engine A;
- B, running time in minutes for engine B;
- U, ambient temperature in degrees Celsius for engine B.

The histograms for the running times are compared in Figure 11.21(a) and those for the temperatures in Figure 11.21(b). The overall profile of temperatures is very similar for the two series of tests, differing only in the number of unusually high or low figures encountered. The profiles of running times appear to differ rather more markedly. It is clear that displaying the data in this way is useful, but some analysis will have to be done in order to determine whether the differences are significant.

Figure 11.20 Data for engine case study.

Engine A				Engine B			
A	T	A	T	B	U	B	U
27.7	24	24.1	7	24.9	13	24.3	17
24.3	25	23.1	14	21.4	19	24.5	16
23.7	18	23.4	16	24.1	18	26.1	18
22.1	15	23.1	9	27.5	19	27.7	14
21.8	19	24.1	14	27.5	21	24.3	19
24.7	16	28.6	23	25.7	17	26.1	5
23.4	17	20.2	14	24.9	17	24.0	17
21.6	14	25.7	18	23.3	19	24.9	18
24.5	18	24.6	18	22.5	21	26.7	23
26.1	20	24.0	12	28.5	12	27.3	28
24.8	15	24.9	18	25.9	17	23.9	18
23.7	15	21.9	20	26.9	13	23.1	10
25.0	22	25.1	16	27.7	17	25.5	25
26.9	18	25.7	16	25.4	23	24.9	22
23.7	19	23.5	11	25.3	30	25.9	16



It is good practice to input data such as these into R using separate variables for the running time, the ambient temperature and the engine:

```
run_time <- c(27.7, 24.3, 23.7, 22.1, 21.8, 24.7, 23.4, 21.6,
            24.5, 26.1, 24.8, 23.7, 25.0, 26.9, 23.7, 24.1,
            23.1, 23.4, 23.1, 24.1, 28.6, 20.2, 25.7, 24.6,
            24.0, 24.9, 21.9, 25.1, 25.7, 23.5, 24.9, 21.4,
            24.1, 27.5, 27.5, 25.7, 24.9, 23.3, 22.5, 28.5,
            25.9, 26.9, 27.7, 25.4, 25.3, 24.3, 24.5, 26.1,
            27.7, 24.3, 26.1, 24.0, 24.9, 26.7, 27.3, 23.9,
            23.1, 25.5, 24.9, 25.9)

amb_temp <- c(24, 25, 18, 15, 19, 16, 17, 14, 18, 20, 15, 15,
              22, 18, 19, 7, 14, 16, 9, 14, 23, 14, 18, 18, 12,
              18, 20, 16, 16, 11, 13, 19, 18, 19, 21, 17, 17,
              19, 21, 12, 17, 13, 17, 23, 30, 17, 16, 18, 14,
              19, 5, 17, 18, 23, 28, 18, 10, 25, 22, 16)

# Generate two levels, each repeated 30 times, to a total
# length of 60
engine <- gl(2, 30, 60, labels = ("engine A", "engine B"))
engine[c(1, 2, 31, 32)] # Show four elements
#> [1] engine A engine A engine B engine B
#> Levels: engine A engine B
levels(engine)
#> [1] "engine A", "engine B")
```

The R object `engine` is a factor with two levels (or possible labels), `engine A` and `engine B`. R uses factors to deal with categorical variables that are represented as descriptions rather than numbers.

We can then store these data together using a data frame:

```
engine_performance <- data.frame(run_time, amb_temp, engine)
head(engine_performance) # Show the first six rows
#>   run_time amb_temp   engine
#> 1    27.7      24 engine A
#> 2    24.3      25 engine A
#> 3    23.7      18 engine A
#> 4    22.1      15 engine A
#> 5    21.8      19 engine A
#> 6    24.7      16 engine A
```

As mentioned in Section 11.1, many useful R functions and interesting data sets can be found in R packages that have been contributed by generous individuals or companies. The `ggplot2` R package by Hadley Wickham provides us with state of the art tools for graphing data such as these. Packages can be installed from RStudio using the Packages tab or, from the command line, use `install.packages`:

```
install.packages("ggplot2",
                 repos ="http://www.stats.bris.ac.uk/R/")
```

If you do not specify an R package repository, a dialogue box will ask you to specify one.

For R to use a package it has to be installed. This can be achieved by using the function `require`:

```
require(ggplot2)
```

`citation("ggplot2")` will provide a citation for the package `ggplot2` so that credit can be given. `ggplot2` forms part of the so-called ‘tidyverse’ (<https://github.com/tidyverse/tidyverse>), a collection of useful R packages that share common philosophies and that are designed to work together. `dplyr` for data manipulation also belongs to the tidyverse, as do `readr` and `readxl` for data input and `tidyverse` for data tidying (not used here).

Here is commented R code to produce Figure 11.21(a), using `ggplot2`:

```
ggplot(engine_performance,
       # Use data from the data frame engine_performance
       aes(x = run_time, fill = engine)) +
       # Specify the aesthetics or features of the graph:
       # On the x axis we plot run_time
       # We fill using a colour determined by engine
       geom_histogram(breaks = 20:30, closed = "left") +
       # Show the data as a histogram
       # with breaks 20, 21,...,30
       # defined as [20, 21), [21, 22),...
       facet_grid(engine ~ .) +
       # A separate panel or facet for each engine
       # arranged in rows, with no columns
       labs(x = "Running time (minutes)", y = "Number")
       # Axes labels
```

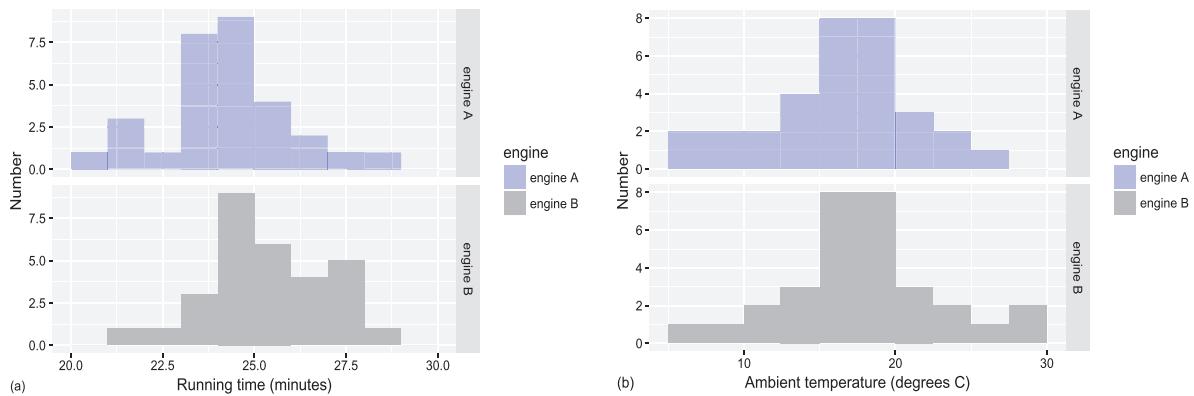


Figure 11.21
Histograms of engine data: (a) running times; (b) temperatures.

Figure 11.21(b) is produced in a similar way:

```
ggplot(engine_performance, aes(x = amb_temp, fill = engine)) +
  geom_histogram(breaks = c(5, 10, 12.4, 15, 17.5, 20, 22.5,
  25, 27.5, 30), closed = "left") +
  facet_grid(engine ~ .) +
  labs(x = "Ambient temperature (degrees C)", y = "Number")
```

In Section 2.3 of H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (second edition, New York, Springer-Verlag, 2016) the three key components of every ggplot2 chart are stated as ‘data, a set of aesthetic mappings between variables in the data and visual properties, and at least one layer which describes how to render each observation. Layers are usually created with a geom function.’ In the above code we can clearly see the specification of the data, the aesthetics and the histogram geometry. We also label the axes.

When planning the analysis, it is as well to consider the questions to which most interest attaches. Do the mean running times for the two engines differ? Does the running time depend on temperature? If so, and if there is a difference in the temperatures for the test series on engines A and B, can this account for any apparent difference in fuel consumption? More particularly, are the data normally distributed? This has a bearing on the methods used, and hence on the conclusions drawn.

11.7.2 Difference in mean running times and temperatures

The sample averages and both versions of standard deviation for the data in Figure 11.20 are as follows:

$$\bar{A} = 24.20, \quad \bar{T} = 16.70, \quad \bar{B} = 25.36, \quad \bar{U} = 18.07$$

$$S_A = 1.761, \quad S_T = 4.001, \quad S_B = 1.657, \quad S_U = 4.932$$

$$S_{A,n-1} = 1.791, \quad S_{T,n-1} = 4.070, \quad S_{B,n-1} = 1.685, \quad S_{U,n-1} = 5.017$$

The average running time for engine B is slightly higher than for engine A. The sample standard deviations are very similar, so we can assume that the true standard deviations

are equal and use the method for comparing means discussed in Section 11.3.5. The pooled estimate of the variance is

$$S_p^2 = \frac{(n-1)(S_{A,n-1}^2 + S_{B,n-1}^2)}{2(n-1)} = \frac{3.208 + 2.839}{2} = 3.023$$

and the relevant value from the t -table is $t_{0.025,58} \approx 1.960$. In fact the sample is large enough for the value for the normal distribution to be taken.

The 95% confidence interval for the difference $\mu_B - \mu_A$ is approximately

$$(\bar{B} - \bar{A} \pm 1.96S_p/\sqrt{15}) = (0.28, 2.04)$$

We may conclude that the difference in mean running times is significant.

Following the same procedure for the temperatures gives the 95% confidence interval for the difference $\mu_U - \mu_T$ to be approximately $(-0.94, 3.68)$. Superficially, this is not significant – and even if the running times do depend on temperature, the similarity of the two test series in this respect enables this factor to be discounted. If so, the fuel performance for engine B is superior to that for engine A. However, if the temperature sensitivity were very high then even a difference in the average too small to give a significant result by this method could create a misleading difference in the fuel consumption figures. This possibility needs to be examined.



We can use the `summarize` function from the `dplyr` package to compute the above summary statistics for each engine. You will probably have to install the `dplyr` package. We also show an example of how we can write a function in R to compute the sample standard deviation s of a sample of data x_1, x_2, \dots, x_n from the alternative sample standard deviation s_{n-1} , using the fact that $s^2 = (s_{n-1}^2 \times (n-1)/n)$ with the consequence that $s = s_{n-1} \times \sqrt{(n-1)/n}$. Here, $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$ (version with divisor n) and $s_{n-1}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$ (version with divisor $n-1$).

```
# This function takes in data x_1, x_2,...,x_n using the argument x
# and returns the sample standard deviation s (defined using
# the divisor n)
# The body of an R function is contained in {...}
sd_n <- function(x){sd(x) * sqrt((length(x) - 1) / length(x))}
# sd uses the n - 1 divisor version of the standard deviation
# The function sd_n that we have created uses
# the n divisor version of the standard deviation
#
# Now for some summary statistics
# In the following %>% can be thought of as a 'pipe',
# with the left side being sent to the right side
require(dplyr)
engine_performance %>% # The data gets
  group_by(engine) %>% # grouped by engine, and
  summarize(m_r_t = mean(run_time), # summarized
            sd_r_t = sd_n(run_time),
            sd_r_t_n_1 = sd(run_time),
```

```

m_a_t = mean(amb_temp),
sd_a_t = sd_n(amb_temp),
sd_a_t_n_1 = sd(amb_temp))
#> # A tibble: 2 × 7
#>   engine      m_r_t    sd_r_t sd_r_t_n_1     m_a_t    sd_a_t sd_a_t_n_1
#>   <fctr>     <dbl>    <dbl>      <dbl>     <dbl>    <dbl>      <dbl>
#> 1 engine A 24.20000 1.76106  1.791166 16.70000 4.001250  4.069652
#> 2 engine B 25.35667 1.65664  1.684961 18.06667 4.932432  5.016754

```

We can perform a t -test to see whether there is a difference in the underlying mean running time between the two engines. The factor `engine` has levels (or labels) `engine A` followed by `engine B`. We need to create a new factor with levels `engine B` followed by `engine A` so that we can get confidence intervals for population $\text{mean}_B - \text{population mean}_A$.

```

engine_performance_2 <- engine_performance %>% # Take the data and
  mutate(engine_rev = factor(engine,
                                levels = c("engine B", "engine A")))
# add the required variable
#
# t-test and confidence interval for running time mean_B - mean_A,
# testing mean_B - mean_A = 0 or mean_B = mean_A
t.test(run_time ~ engine_rev,
       data = engine_performance_2,
       var.equal = TRUE,
       conf.level = 0.95) # 95% confidence interval
#>
#> Two Sample t-test
#>
#> data: run_time by engine_rev
#> t = 2.5762, df = 58, p-value = 0.01256
#> alternative hypothesis: true difference in means is not
#> equal to 0
#> 95 percent confidence interval:
#> 0.2579447 2.0553886
#> sample estimates:
#> mean in group engine B mean in group engine A
#>                25.35667                 24.20000
#
# t-test and confidence interval for ambient temperature
# mean_B - mean_A,
# testing mean_B - mean_A = 0 or mean_B = mean_A
t.test(amb_temp ~ engine_rev,
       data = engine_performance_2,
       var.equal = TRUE,
       conf.level = 0.95) # 95% confidence interval
#>
#> Two Sample t-test
#>

```

```
#> data: amb_temp by engine_rev
#> t = 1.1588, df = 58, p-value = 0.2513
#> alternative hypothesis: true difference in means is not
#> equal to 0
#> 95 percent confidence interval:
#> -0.994169 3.727602
#> sample estimates:
#> mean in group engine B mean in group engine A
#> 18.06667 16.70000
```

The difference between the intervals produced by R and the ones that we calculated above is due to the normal approximation to the t distribution ($t_{0.025,58} = 2.0017 \approx 1.960$) that we have adopted.

11.7.3 Dependence of running time on temperature

The simplest way to test for dependence is to correlate times and temperatures for each engine. To compute the sample correlations, we need the following additional results from the data:

$$\bar{AT} = 407.28, \quad \bar{BU} = 457.871$$

The sample correlations (Section 11.4.4) of A and T and of B and U are then

$$r_{A,T} = \frac{\bar{AT} - (\bar{A})(\bar{T})}{S_A S_T} = 0.445, \quad r_{B,U} = \frac{\bar{BU} - (\bar{B})(\bar{U})}{S_B S_U} = -0.030$$

and the respective 95% confidence intervals (Section 11.4.5) are

$$(0.10, 0.69), \quad (-0.39, 0.33)$$

This is a quite definitive result: the running time for engine A depends positively upon the ambient temperature, but that for engine B does not. The confidence intervals are based on the assumption that all the variables A , T , B and U are normal. The histograms have this character, and a test for normality will be covered later.



These calculations and confidence intervals can be easily obtained in R:

```
engine_performance %>% # The data gets
  group_by(engine) %>% # grouped by engine, and
  summarize(mean_of_prod = mean(run_time * amb_temp))
    # summarized
#> # A tibble: 2 × 2
#>   engine mean_of_prod
#>   <fctr>     <dbl>
#> 1 engine A      407.2767
#> 2 engine B      457.8700
# Sample correlations and 95% confidence intervals
```

```

# Extract the data from engine A
engine_performance_A <- engine_performance %>%
  filter(engine == "engine A")
# Filter out the rows for which engine takes the value engine A
# Note the use of == for 'does it equal?'
# Use the function cor.test with just these data:
# here, cor.test is applied to the variables run_time and
# amb_temp that are found
# in the data frame engine_performance_A
with(engine_performance_A, cor.test(run_time , amb_temp))
#>
#> Pearson's product-moment correlation
#>
#> data: run_time and amb_temp
#> t = 2.6305, df = 28, p-value = 0.0137
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#> 0.1010835 0.6940980
#> sample estimates:
#> cor
#> 0.4451419
# Extract the data from engine B and use the function cor.test
# with just these data
engine_performance_B <- engine_performance %>%
  filter(engine == "engine B")
with(engine_performance_B, cor.test(run_time , amb_temp))
#>
#> Pearson's product-moment correlation
#>
#> data: run_time and amb_temp
#> t = -0.15577, df = 28, p-value = 0.8773
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#> -0.3856070 0.3343885
#> sample estimates:
#> cor
#> -0.02942561

```

Linear regression also reveals the dependence of running time on temperature. Here we assume that the variables are related by a linear model as follows:

$$\left. \begin{array}{l} A_i = c_A + d_A T_i + \varepsilon_i \\ B_i = c_B + d_B U_i + \eta_i \end{array} \right\} \quad (i = 1, \dots, n)$$

where c_A , d_A , c_B and d_B are constants, and the random variables ε_i and η_i represent errors. Using the results of the least-squares analysis in Section 11.5.1, for engine A we have

$$\hat{d}_A = \frac{\bar{AT} - (\bar{A})(\bar{T})}{S_T^2} = 0.196, \quad \hat{c}_A = \bar{A} - \hat{d}_A \bar{T} = 20.9$$

Likewise, for engine B

$$\hat{d}_B = -0.010, \quad \hat{c}_B = 25.5$$



We can obtain these intercepts and slopes, together with the other visual output, using the `lm` function:

```
m_A <- lm(run_time ~ amb_temp,
            # How does run_time depend on amb_temp?
            data = engine_performance,
            # Use the data in the data frame engine_performance
            subset = engine == "engine A")
            # Work with just the data corresponding to engine A
summary(m_A)
#>
#> Call:
#> lm(formula = run_time ~ amb_temp, data = engine_performance,
#>      subset = engine == "engine A")
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -3.4710 -0.8328  0.2769  1.0111  3.1657
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept) 20.92815 1.27904 16.36 7.28e-16 ***
#> amb_temp     0.19592 0.07448   2.63   0.0137 *
#> -
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.632 on 28 degrees of freedom
#> Multiple R-squared: 0.1982, Adjusted R-squared: 0.1695
#> F-statistic: 6.919 on 1 and 28 DF, p-value: 0.0137
m_B <- lm(run_time ~ amb_temp,
            data = engine_performance,
            subset = engine == "engine B")
summary(m_B)
#>
#> Call:
#> lm(formula = run_time ~ amb_temp,
#>      data = engine_performance,
#>      subset = engine == "engine B")
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -3.9474 -1.0623  0.0767  1.2297  3.0834
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
```

```

#> (Intercept) 25.535221 1.188197 21.491 <2e-16 ***
#> amb_temp     -0.009883 0.063445 -0.156 0.877
#> -
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.714 on 28 degrees of freedom
#> Multiple R-squared: 0.0008659, Adjusted R-squared:
#> -0.03482
#> F-statistic: 0.02427 on 1 and 28 DF, p-value: 0.8773

```

Figure 11.22 contains scatter plots of the data with these regression lines drawn. The points are well scattered about the lines. The residual variances, using the results in Sections 11.5.2 and 11.5.3, are

$$S_E^2 = S_A^2 - \hat{d}_A S_T^2 = S_A^2(1 - r_{A,T}^2) = 2.49$$

$$S_F^2 = 2.74$$

As explained in Section 11.5.3, the respective values of r^2 indicate the extent to which the variation in running times is due to the dependence on temperature. For engine B there is virtually no such dependence. For engine A we have $r_{A,T}^2 = 0.198$, so nearly 20% of the variation in running times is accounted for in this way.

If we assume that the errors ε_i and η_i are normal, we can obtain confidence intervals for the regression slopes. The appropriate value from the t table is $t_{\alpha/2,n-2} = t_{0.025,28} = 2.048$, so the 95% confidence interval for d_A is

$$\left(\hat{d}_A \pm 2.048 \frac{S_E}{S_T \sqrt{28}} \right) = (0.04, 0.35)$$

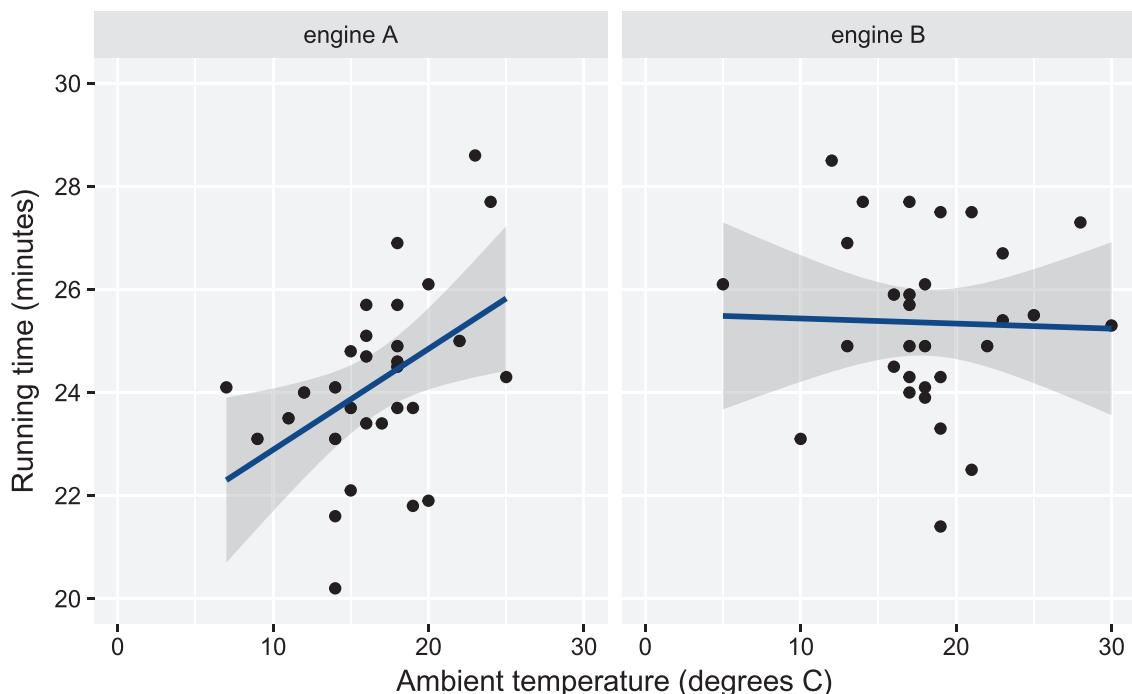


Figure 11.22 Regression of running time against temperature for engine A and engine B. 95% confidence limits (Section 11.5.2) are also shown.

The significance shown here confirms that found for the correlation. The 95% confidence interval for d_B is $(-0.14, 0.12)$ that contains 0.



Figure 11.22 can be produced as follows:

```
ggplot(engine_performance, aes(x = amb_temp, y = run_time)) +
  geom_point() + # Show the data as points
  geom_smooth(method = "lm") +
    # Add a regression line with confidence intervals
    # Specify the scale on the y axis
  scale_y_continuous(breaks = c(20, 22, 24, 26, 28, 30),
                     # Where the breaks should be
                     minor_breaks = NULL,
                     limits = c(20,30)) + # no breaks in-between
  facet_grid(. ~ engine) +
    # A separate facet for each engine arranged in columns, with
    # no rows
  labs(x = "Ambient temperature (degrees C)"
       y = "Running time (minutes)") +
  xlim(0, 30) # x axis limits
```

The values of S_E^2 and S_F^2 can be found from the linear model objects, as can 95% confidence intervals:

```
n_A <- nrow(engine_performance_A) # Number of data points
n_B <- nrow(engine_performance_B)
S_2_E <- summary(m_A)$sigma^2 * (n_A - 2) / n_A
  # sigma has to be scaled
S_2_E
#> [1] 2.4868
S_2_F <- summary(m_B)$sigma^2 * (n_B - 2) / n_B
S_2_F
#> [1] 2.742079
confint(m_A)
#>           2.5 %      97.5 %
#> (Intercept) 18.30816036 23.5481378
#> amb_temp     0.04335172  0.3484867
confint(m_B)
#>           2.5 %      97.5 %
#> (Intercept) 23.101310 27.9691318
#> amb_temp     -0.139845  0.1200788
```



Figure 11.23 illustrates the different relationships between running time and ambient temperature for the two engines. It can be generated using the following code:

```
ggplot(engine_performance,
  aes(x = amb_temp, y = run_time, col = engine)) +
  geom_point() +
  geom_smooth(method = "lm", fullrange = TRUE) +
    # Extend lines to the full range
  scale_y_continuous(breaks = c(20, 22, 24, 26, 28, 30),
                     minor_breaks = NULL, limits c(20,30)) +
  labs(x = "Ambient temperature (degrees C)",
       y = "Running time (minutes)") +
  xlim(0, 30)
```

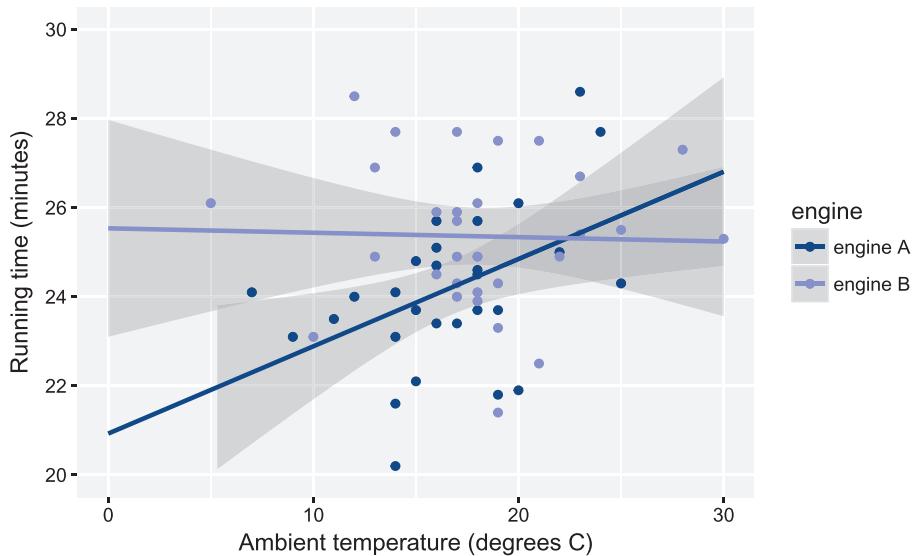


Figure 11.23 Regression of running time against temperature for both engines on the same graph.

In order to quantify these differences we use all the data to fit both regression lines simultaneously. The following R code provides estimates of the intercept c_A ($c_A = 20.9$) and slope d_A ($\hat{d}_A = 0.196$) for engine A and estimates of the differences $c_B - c_A$ ($c_B - c_A = 25.5 - 20.9 = 4.6$) and $d_B - d_A$ ($\hat{d}_B - \hat{d}_A = -0.010 - 0.196 = -0.206$). Since $c_B = c_A + (c_B - c_A)$ and $d_B = d_A + (d_B - d_A)$, these differences represent the additional intercept and slope for engine B:

```
m_interaction <- lm(run_time ~ amb_temp * engine, data =
  engine_performance)
# * allows for an interaction between amb_temp and engine
# That is, * allows different intercepts and slopes for the
# two engines
summary(m_interaction)
#>
#> Call:
#> lm(formula = run_time ~ amb_temp * engine, data =
  engine_performance)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -3.9474  -0.9748  0.1187  1.0892  3.1657
#>
#> Coefficients:
#>                               Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                20.92815   1.31145 15.958   <2e-16 ***
#> amb_temp                   0.19592   0.07637  2.565   0.0130 *
#> engineengine B              4.60707   1.75100  2.631   0.0110 *
#> amb_temp:engineengine B -0.20580   0.09834 -2.093   0.0409 *
```

```

#> ---
#> Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 1.674 on 56 degrees of freedom
#> Multiple R-squared: 0.1974, Adjusted R-squared: 0.1544
#> F-statistic: 4.59 on 3 and 56 DF, p-value: 0.006075

95% confidence intervals for  $c_A$ ,  $d_A$ ,  $c_B - c_A$  and  $d_B - d_A$  are obtained in the usual way:
confint(m_interaction)

```

	2.5 %	97.5 %
#> (Intercept)	18.3009974	23.555300790
#> amb_temp	0.0429346	0.348903836
#> engineengine B	1.0993996	8.114744294
#> amb_temp:engineengine B	-0.4027943	-0.008810298

Because these last two confidence intervals do not contain 0, we can conclude that, for the linear relationship between running time and ambient temperature, there is a significant difference in intercept and a significant difference in slope between the two engines. The same conclusion follows by noting that the corresponding p-values are less than 0.05.

We will now use this model to produce point estimates and 95% confidence intervals for the running time for both engines at 10 °C and 20 °C:

```

predict(m_interaction,
        newdata = data.frame(amb_temp = 10, engine = "engine A"),
        interval = "confidence")
#>           fit      lwr      upr
#> 1 22.88734 21.69347 24.08121

predict(m_interaction,
        newdata = data.frame(amb_temp = 10, engine = "engine B"),
        interval = "confidence")
#>           fit      lwr      upr
#> 1 25.43639 24.26298 26.6098

predict(m_interaction,
        newdata = data.frame(amb_temp = 20, engine = "engine A"),
        interval = "confidence")
#>           fit      lwr      upr
#> 1 24.84653 24.05308 25.63999

predict(m_interaction,
        newdata = data.frame(amb_temp = 20, engine = "engine B"),
        interval = "confidence")
#>           fit      lwr      upr
#> 1 25.33756 24.68009 25.99503

```

The 95% confidence intervals do not overlap at 10 °C suggesting that there is a significant difference between the engines at this temperature. The intervals do overlap at 20 °C.

This can be properly quantified by providing 95% confidence intervals for $(c_B + d_B t) - (c_A + d_A t) = (c_B - c_A) + (d_B - d_A)t$ at $t = 10$ and $t = 20$ using the `multcomp` package, which you will probably have to install:

```
require(multcomp)
K_10 <- matrix(c(0, 0, 1, 10), nrow = 1)
# c_B - c_A is the third and d_B - d_A is the fourth parameter
# of the model
# We are interested in (c_B - c_A) + 10 (d_B - d_A) =
# 1 (c_B - c_A) + 10 (d_B - d_A)
interval_10 <- glht(m_interaction, linfct = K_10)
confint(interval_10)
#>
#> Simultaneous Confidence Intervals
#>
#> Fit: lm(formula = run_time ~ amb_temp * engine, data =
#> engine_performance)
#>
#> Quantile = 2.0032
#> 95% family-wise confidence level
#>
#>
#> Linear Hypotheses:
#>             Estimate    lwr      upr
#> 1 == 0    2.5490   0.8751  4.2230
K_20 <- matrix(c(0, 0, 1, 20), nrow = 1)
interval_20 <- glht(m_interaction, linfct = K_20)
confint(interval_20)
#>
#> Simultaneous Confidence Intervals
#>
#> Fit: lm(formula = run_time ~ amb_temp * engine, data =
#> engine_performance)
#>
#> Quantile = 2.0032
#> 95% family-wise confidence level
#>
#>
#> Linear Hypotheses:
#>             Estimate    lwr      upr
#> 1 == 0    0.4910  -0.5394  1.5215
```

The first interval is entirely positive, while the second contains zero. Hence, we may be confident that the difference in running times between engine B and engine A is positive for a low ambient temperature, such as 10°C, but that there is no significant difference when $t = 20^\circ\text{C}$.

11.7.4 Test for normality

The confidence interval statistics are all based on the assumption of normality of the data. Although the sample sizes are reasonably large, so that the central limit theorem can be relied upon to weaken this requirement, it is worth applying a test for normality to see whether there is any clear evidence to the contrary. Here the regression residuals ε_i and η_i will be tested using the method described in Section 11.6.1.

Figure 11.24 shows the histogram of all 60 ‘standardized’ residuals. The residuals have zero mean in any case, and are standardized by dividing by the alternative form of the standard deviation so that they can be compared with a standard normal distribution. It is convenient to use intervals of width 0.4, and the comparison is developed in Figure 11.25.

The normal probabilities for each interval are obtained from the standard normal table of the cumulative distribution function $\Phi(z)$, Figure 11.2, taking successive differences:

$$P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1)$$

These probabilities are multiplied by 60 to obtain the expected number in each interval, and the difference between the observed and expected number for each interval is squared and then divided by the expected number to give the contribution to the total chi-square:

$$\chi^2 = \sum_{k=1}^m \frac{(f_k - e_k)^2}{e_k} \approx 4.8$$

Figure 11.24
Histogram of
standardized residuals.

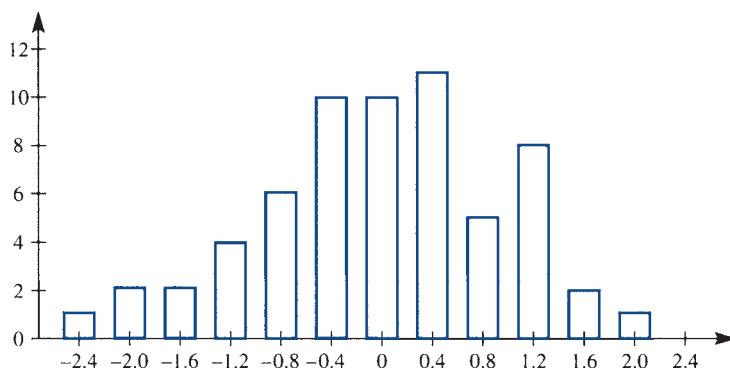


Figure 11.25 Table of
the test for normality.

Interval	Observed (f_k)	Probability	Expected (e_k)	Chi-square
($-\infty, -1.4$)	5	0.0808	4.845	0.005
($-1.4, -1.0$)	4	0.0779	4.674	0.097
($-1.0, -0.6$)	6	0.1156	6.936	0.126
($-0.6, -0.2$)	10	0.1465	8.789	0.167
($-0.2, +0.2$)	8	0.1585	9.511	0.240
($+0.2, +0.6$)	11	0.1465	8.789	0.556
($+0.6, +1.0$)	5	0.1156	6.936	0.540
($+1.0, +1.4$)	8	0.0779	4.674	2.367
($+1.4, +\infty$)	3	0.0808	4.845	0.703
Totals	60	1.0	60	4.801

This is small compared with $\chi^2_{0.05,8} = 15.507$, so the hypothesis of normality is accepted.

It is unwise when applying this test in general to have many classes with expected numbers less than five, so the intervals in the tails of the histogram have been merged.



We can reproduce this analysis in R as follows:

```
r_A <- residuals(m_A) # epsilon_i_hat
r_A <- r_A / sd(r_A)
  # Divide by the alternative form of the standard deviation
r_B <- residuals(m_B)
r_B <- r_B / sd(r_B) # eta_i_hat
r_AB <- c(r_A, r_B) # Collect all the residuals together
# Breaks; use -10 and 10 as end points
breaks_AB <- c(-10, seq(from = -1.4, to = 1.4, by = 0.4), 10)
# Cut at these breaks
r_AB_in_intervals <- cut(r_AB, breaks = breaks_AB)
head(r_AB_in_intervals) # Residuals are assigned to an interval
#> [1] (1,1.4] (-1,-0.6] (-0.6,-0.2] (-1.4,-1] (-10,-1.4] (0.2,0.6]
#> 9 Levels: (-10,-1.4] (-1.4,-1] (-1,-0.6] (-0.6,-0.2] ... (1.4,10]
head(r_AB) # To check
#>           1           2           3           4           5           6
#> 1.2904589 -0.9515012 -0.4705323 -1.1016384 -1.7772822 0.3972420
# Tabulate
f <- table(r_AB_in_intervals); f
#> r_AB_in_intervals
#> (-10,-1.4] (-1.4,-1] (-1,-0.6] (-0.6,-0.2] (-0.2,0.2] (0.2,0.6]
#>      5          4          6          10          8          11
#> (0.6,1]   (1,1.4] (1.4,10]
#>      5          8          3
prob <- diff(pnorm(breaks_AB)); prob # Probabilities
#> [1] 0.08075666 0.07789859 0.11559786 0.14648717 0.15851942
  0.14648717
#> [7] 0.11559786 0.07789859 0.08075666
  # pnorm provides Phi, then diff takes lag 1 differences
e <- 60 * prob; e # Expected values
#> [1] 4.845400 4.673916 6.935872 8.789230 9.511165 8.789230
  6.935872 4.673916
#> [9] 4.845400
(f - e)^2 / e # Chi-square values
#> r_AB_in_intervals
#> (-10,-1.4] (-1.4,-1] (-1,-0.6] (-0.6,-0.2] (-0.2,0.2] (0.2,0.6]
#> 0.004932782 0.097169563 0.126279162 0.166790838 0.240098876 0.556078537
#> (0.6,1]   (1,1.4] (1.4,10]
#> 0.540321366 2.366931208 0.702831516
```

```

sum((f - e)^2 / e) # Chi-square statistic
#> [1] 4.801434
qchisq(0.05, length(f) - 1, lower.tail = FALSE)
#> [1] 15.50731
# Alternatively, using chisq.test
chisq.test(f, p = prob)
#>
#> Chi-squared test for given probabilities
#>
#> data: f
#> X-squared = 4.8014, df = 8, p-value = 0.7786

```

A quantile-comparison plot (see Figure 11.26) provides a graphical way of checking normality. We do not go into the details, except to say that if most of the points lie in the confidence envelope, then we would accept the hypothesis of normality. Here all the points lie in the confidence envelope.

Figure 11.26 can be generated using the `car` package (which you will probably need to install):

```

require(car)
qqPlot(r_AB, ylab = "Standardized residuals")

```

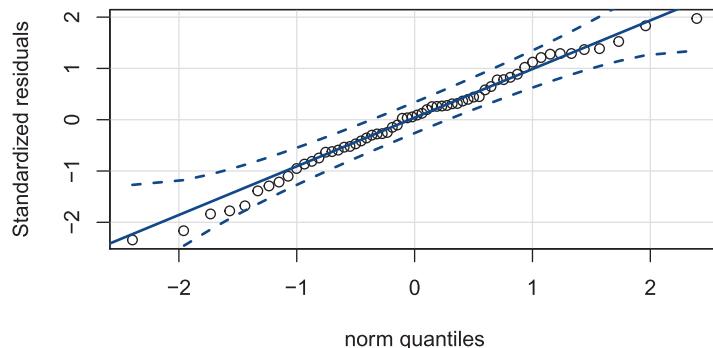


Figure 11.26 Quantile-comparison plot of the standardized residuals.

11.7.5 Conclusions

All the questions posed in Section 11.7.1 have now been answered. Engine B has an average running time that is significantly higher than that for engine A, showing that it has the advantage in fuel consumption. However, this statement requires qualification. The running time for engine A depends upon ambient temperature. The temperature difference between the two test series was not significant, and does not account for the difference in average running times. However, engine B will only maintain its fuel advantage up to a certain point. This point cannot be identified very precisely because of considerable residual scatter in the data. There are many potential sources of this scatter, such as errors in measuring out the fuel, or variations in the quantity and

consistency of the engine oil. The scatter has a normal distribution, which justifies the statistics behind the conclusions reached. We will discuss this example further in Section 11.10.3.

11.8 Engineering application: statistical quality control

11.8.1 Introduction

Every manufacturer recognizes the importance of quality, and every manufacturing process involves some variation in the quality of its output, however that is to be measured. Experience shows that tolerating a lack of quality tends to be more costly in the end than promoting a quality approach. It follows that quality control is a major and increasing concern, and methods of statistical quality control are more important than ever. The domain of these methods now extends to the construction and service industries as well as to manufacturing – wherever there is a process that can be monitored in quantitative terms. Internet traffic can also be monitored by statistical quality control methods to ensure stable performance.

Traditionally, quality control involved the accumulation of batches of manufactured items, the testing of samples extracted from these batches, and the acceptance or rejection (with appropriate rectifying action) of these batches depending upon the outcome. The essential problem with this is that it is too late within the process: it is impossible to inspect or test quality into a product. More recently the main concern has been to design the quality into the product or service and to monitor the process to ensure that the standard is maintained, in order to prevent any deficiency. Assurance can then be formally given to the customer that proper procedures are in place.

Control charts play an important role in the implementation of quality. The idea of these is introduced in Section 13.6 in MEM, where Shewhart charts for counts of defectives are described. In order for this section to be as self-contained as possible, some of that material is repeated here. This section then covers more powerful control charts and extends the scope of what they monitor.

First note that there are two main alternative measures of quality: **attribute** and **variable**. In attribute measure, regular samples from the process are inspected and for each sample the number that fail according to some criterion is plotted on a chart. In variable measure, regular samples are again taken, but this time the sample average for some numerical measure (such as dimension or lifetime) is plotted.

11.8.2 Shewhart attribute control charts

A Shewhart control chart provides a plot of the count of ‘defectives’ (the number in the sample failing according to some chosen criterion) against sample number. It is assumed that a small (specified) proportion of ‘defective’ items in the process is permitted. It will also show the two limits on the count of defectives, corresponding to probabilities of one in 40 and one in 1000 of a sample count falling outside if the process is ‘in control’; that is, conforming to the specification. These are called **warning** and **action limits** respectively, and are denoted by c_w and c_A .

Any sample point falling outside the action limit would normally result in the process being suspended and the problem corrected. Roughly one in 40 sample points will fall outside the warning limit purely by chance, but if this occurs repeatedly or if there is a clear trend upwards in the counts of defectives then action may well be taken before the action limit itself is crossed.

To obtain the warning and action limits, we use the Poisson approximation to the binomial. If the acceptable proportion of defective items is p , usually small, and the sample size is n then for a process in control the defective count C , say, will be a binomial random variable with parameters n and p . Provided that n is not too small, this binomial random variable can be approximated by a Poisson random variable with mean parameter np . From this approximation we can obtain:

$$P(C \geq c) = \sum_{k=c}^{\infty} \frac{(np)^k e^{-np}}{k!}$$

Equating this to $\frac{1}{40}$ and then to $\frac{1}{1000}$ gives equations that can be solved for the warning limit c_w and the action limit c_A respectively, in terms of the product np .

Example 11.23

Regular samples of 50 are taken from a process making electronic components, for which an acceptable proportion of defectives is 5%. Successive counts of defectives in each sample are as follows:

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Count	3	5	2	2	1	6	4	4	2	6	7	4	5	5	8	6	5	9	7	8

At what point would the decision be taken to stop and correct the process?

Solution

The control chart is shown in Figure 11.27. Taking np to be 2.5, the warning limit $c_w = 5.5$ and action limit $c_A = 8.5$ can be read from the table in Figure 11.28. The half-integer values are to avoid ambiguity when the count lies on a limit. There are warnings at samples 6, 10, 11, 15 and 16 before the action limit is crossed at sample 18. Strictly, the decision should be taken at that point, but the probability of two consecutive warnings is less than one in 1600 by the product rule of probabilities assuming independence, which would justify taking action after sample 11.

Figure 11.27
Attribute control chart
for Example 11.23.

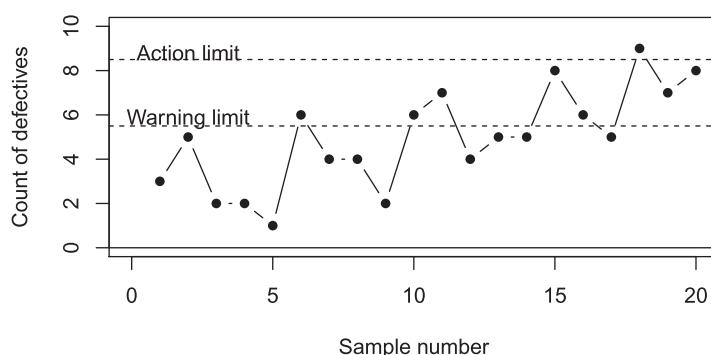


Figure 11.28

Shewhart attribute control limits: n is sample size, p is probability of defect, c_w is warning limit and c_a is action limit.

c_w or c_a	np for c_w	np for c_a
1.5	<0.44	<0.13
2.5	0.44–0.87	0.13–0.32
3.5	0.87–1.38	0.32–0.60
4.5	1.38–1.94	0.60–0.94
5.5	1.94–2.53	0.94–1.33
6.5	2.53–3.16	1.33–1.77
7.5	3.16–3.81	1.77–2.23
8.5	3.81–4.48	2.23–2.73
9.5	4.48–5.17	2.73–3.25
10.5	5.17–5.87	3.25–3.79
11.5	5.87–6.59	3.79–4.35
12.5	6.59–7.31	4.35–4.93
13.5	7.31–8.05	4.93–5.52
14.5	8.05–8.80	5.52–6.12
15.5	8.80–9.55	6.12–6.74
16.5	9.55–10.31	6.74–7.37
17.5	10.31–11.08	7.37–8.01
18.5	11.08–11.85	8.01–8.66
19.5	11.85–12.63	8.66–9.31
20.5	12.63–13.42	9.31–9.98
21.5	13.42–14.21	9.98–10.65
22.5	14.21–15.00	10.65–11.33
23.5	15.00–15.80	11.33–12.02
24.5	15.80–16.61	12.02–12.71
25.5	16.61–17.41	12.71–13.41
26.5	17.41–18.23	13.41–14.11
27.5	18.23–19.04	14.11–14.82
28.5	19.04–19.86	14.82–15.53
29.5	19.86–20.68	15.53–16.25
30.5		16.25–16.98
31.5		16.98–17.70
32.5		17.70–18.44
33.5		18.44–19.17
34.5		19.17–19.91
35.5		19.91–20.66



We can perform the calculations and produce Figure 11.27 in R:

```
count <- c(3, 5, 2, 2, 1, 6, 4, 4, 2, 6, 7, 4, 5, 5, 8,
       6, 5, 9, 7, 8)
N <- length(count)
n <- 50; p <- 0.05; n * p
#> [1] 2.5
c_W <- 5.5; c_A <- 8.5 # Warning and Action limits
plot(1:N, count,
     type = "b", # Both points and lines
     pch = 16, # Filled dots as plotting characters
     xlab = "Sample number", ylab = "Count of defectives",
     xlim = c(0, 20), ylim = c(0, 10))
abline(h = c(c_W, c_A), lty = "dashed")
# Dashed horizontal lines at Warning and Action limits
```

```
abline(h = 0) # Horizontal line at 0
text(2, c_W + 0.3, "Warning limit") # Text at (2, c_W + 0.3)
text(2, c_A + 0.3, "Action limit")
```

An alternative practice (especially popular in the USA) is to dispense with the warning limit and to set the action limit (called the **upper control limit, UCL**) at three standard deviations above the mean. Because the count of defectives is binomial with mean np and variance $np(1 - p)$, this means that

$$\text{UCL} = np + 3\sqrt{[np(1 - p)]}$$

Example 11.24

Find the UCL and apply it to the data in Example 11.23.

Solution

From $n = 50$ and $p = 0.05$ we infer that $\text{UCL} = 7.1$, which is between the warning limit c_W and the action limit c_A in Example 11.23. The decision to correct the process would be taken after the 15th sample, the first to exceed the UCL.



The R packages `qcc` and `spc` provide functions to produce a range of control charts. We shall mainly use functions from `qcc`; see L. Seruca, ‘`qcc`: An R package for quality control charting and statistical process control’, *R News*, 4, 11–17, 2004 (https://www.r-project.org/doc/Rnews/Rnews_2004-1.pdf), for a brief description.

Figure 11.29 shows the UCL and LCL (lower control limit, here is zero) with coloured points used to highlight problems. It is generated using the following code:

```
sample_sizes <- rep(50, N) # 20 samples of size 50
require(qcc)
# Load the package, which you will probably have to install
np_chart <- qcc(count, type = "np",
                  sizes = sample_sizes,
                  center = 0.05, # Target value
                  confidence.level = 1 - 1/1000, #
                  Corresponds to 1 / 1000
                  ylim = c(0, 10),
                  xlab = "Sample number",
                  ylab = "Count of defectives")
```

11.8.3 Shewhart variable control charts

Suppose now that the appropriate assessment of quality involves measurement on a continuous scale rather than success or failure under some criterion. This arises whenever some dimension of the output is critical for applications. Again we take samples, but this time we measure this critical dimension and average the results. The Shewhart chart for this variable measure is a plot of successive sample averages against sample number.

The warning and action limits c_W and c_A on a Shewhart chart are those points for which the probabilities of a false alarm (where the result exceeds the limit even though

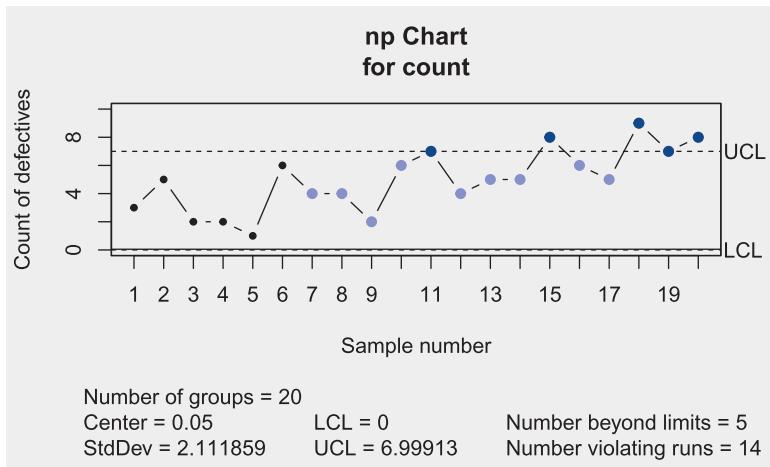


Figure 11.29 Attribute control chart for Example 11.23 produced by the R package qcc.

the process is in control) are one in 40 and one in 1000 respectively. For variable measure the critical quantity can be either too high or too low, so the sample average must be tested in each direction with the stated probability of exceedance for each limit. It follows that the limits are determined by

$$P(\bar{X} > \mu_x + c_w) = P(\bar{X} < \mu_x - c_w) = \frac{1}{40}$$

$$P(\bar{X} > \mu_x + c_a) = P(\bar{X} < \mu_x - c_a) = \frac{1}{1000}$$

where \bar{X} is the sample average and μ_x the design mean.

Provided that the sample size n is not too small, the central limit theorem allows the sample average to be assumed normal (Section 11.3.2),

$$\bar{X} \sim N(\mu_x, \sigma_x^2/n)$$

and the normal distribution table (Figure 11.2) then gives

$$c_w = \frac{1.96 \sigma_x}{\sqrt{n}}, \quad c_a = \frac{3.09 \sigma_x}{\sqrt{n}}$$

Example 11.25

Measurements of sulphur dioxide concentration (in $\mu\text{g m}^{-3}$) in the air are taken daily at five locations, and successive average readings are as follows:

$$\begin{aligned} & 64.2, 56.9, 57.7, 67.9, 61.7, 59.7, 55.6, 63.7 \\ & 58.3, 66.4, 67.2, 65.2, 63.1, 67.6, 64.1, 66.7 \end{aligned}$$

It is suspected that the mean increased during that time. Assuming normal data with a long-term mean of 60.0 and standard deviation of 8.0, investigate whether an increase occurred.

Solution

From $n = 5$ and $\sigma_x = 8$ we have $c_w = 7.0$ and $c_a = 11.1$ (Figure 11.30). The warning limit is 67.0, which is exceeded by sample numbers 4, 11 and 14. The action limit is 71.1, which is not exceeded. The readings are suspiciously high – but not sufficiently so for the conclusion to be justified.

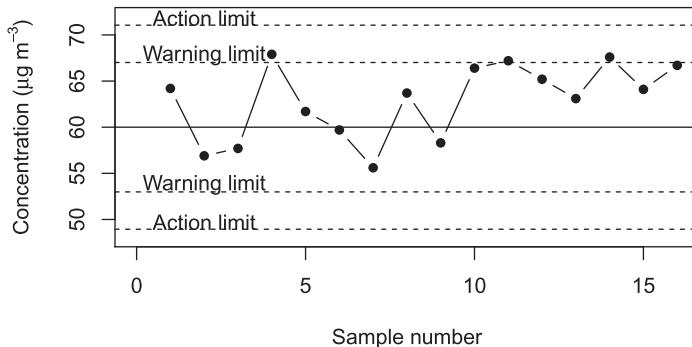


Figure 11.30 Variable control chart for Example 11.25.



We can perform the calculations and produce Figure 11.30 in R:

```
# Sulphur dioxide
SO2 <- c(64.2, 56.9, 57.7, 67.9, 61.7, 59.7, 55.6, 63.7,
      58.3, 66.4, 67.2, 65.2, 63.1, 67.6, 64.1, 66.7)
N <- length(SO2)
n <- 5; mu_X <- 60; sigma_X <- 8
# Warning and Action limits
c_W <- qnorm(1 / 40, lower.tail = FALSE) * sigma_X /
sqrt(n); c_W
#> [1] 7.01218
c_A <- qnorm(1 / 1000, lower.tail = FALSE) * sigma_X /
sqrt(n); c_A
#> [1] 11.05595
# Plot; the plot needs more space on left side for axis
# label superscript
p <- par(mar = c(5.1, 4.1 + 1, 4.1, 2.1))
# original parameters can be restored using par(p)
plot(1:N, SO2,
     type = "b", pch = 16,
     xlab = "Sample number",
     ylab = expression(paste("Concentration (", mu * g ~
       m^-3, ")")),
     # Axis label using Greek letter and power
     xlim = c(0,16), ylim = c(48, 72))
abline(h = mu_X) # Horizontal line at the long-term mean
# Dashed horizontal lines at Warning and Action limits
abline(h = c(mu_X - c_W, mu_X + c_W, mu_X - c_A, mu_X +
c_A), lty = "dashed")
abline(h = 0) # Horizontal line at 0
text(2, mu_X - c_W + 0.75, "Warning limit")
text(2, mu_X + c_W + 0.75, "Warning limit")
```

```
text(2, mu_X - c_A + 0.75, "Action limit")
text(2, mu_X + c_A + 0.75, "Action limit")
```

The following code uses the qcc package to produce Figure 11.31 on which the UCL and LCL are shown:

```
sample_sizes <- rep(5, N) # 16 samples of size 5
# Plot needs more space on left side for axis label
# superscript
p <- par(mar = c(5.1, 4.1 + 1, 4.1, 2.1))
xbar_chart <- qcc(SO2, type = "xbar",
                    sizes = sample_sizes,
                    center = 60.0, # Target value
                    std.dev = 8, # Assumed standard deviation
                    confidence.level = 1 - 2 / 1000,
                    # Corresponds to 1 / 1000 (two limits);
                    # alternatively specify nsigmas
                    xlim = c(0,16), ylim = c(48, 72),
                    xlab = "Sample number",
                    ylab = expression(paste("Concentration (",
                    mu * g ~ m^-3, ")")))
```

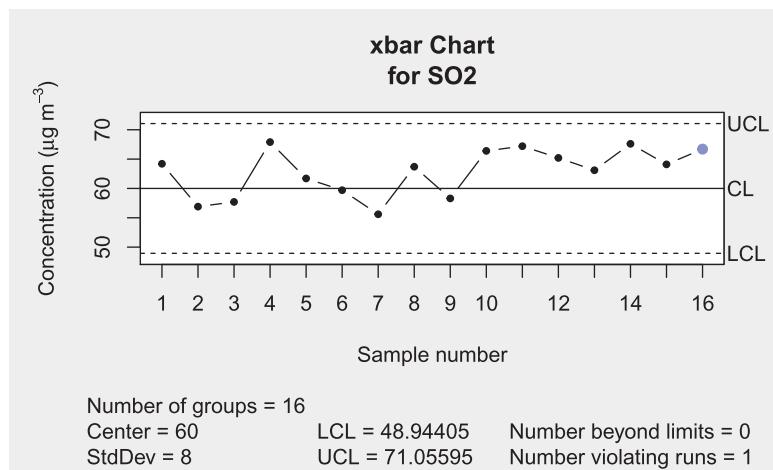


Figure 11.31 Variable control chart for Example 11.25 produced by the R package qcc.

As discussed in Section 11.8.2, the practice in the USA is somewhat different: there are no warning limits, only action limits at three standard deviations on either side of the mean. For a variable chart this allows a deviation from the mean of at most $3\sigma_x/\sqrt{n}$, which is very close to the action limit usually used in the UK.

11.8.4 Cusum control charts

The main concern in designing a control chart is to achieve the best compromise between speedy detection of a fault on the one hand and avoidance of a proliferation of false alarms on the other. If the chart is too sensitive, it will lead to a large number of unnecessary shutdowns. The Shewhart charts, on the other hand, are rather conservative in that they are slow to indicate a slight but genuine shift in performance away from the design level. This derives from the fact that each sample point is judged independently and may well lie inside the action limits, whereas the cumulative evidence over several samples might justify an earlier decision. Rather informal methods involving repeated warnings and trends are used, but it is preferable to employ a more powerful control chart. The **cumulative sum (cusum)** chart achieves this.

Suppose that we have a sequence $\{Y_1, Y_2, \dots\}$ of observations, which may be either counts of defectives or sample averages. From this a new sequence $\{S_0, S_1, \dots\}$ is obtained by setting

$$S_0 = 0,$$

$$S_m = \max\{0, S_{m-1} + Y_m - r\} \quad (m = 1, 2, \dots)$$

where r is a constant ‘reference value’. This gives a cumulative sum of values of $Y_m - r$, which is reset to zero whenever it goes negative. The out-of-control decision is made when

$$S_m > h$$

where h is a constant ‘decision interval’. This will detect an increasing mean; a separate but similar procedure can be used to detect a decreasing mean. Values of r and h for both attribute and variable types of control can be obtained from tables such as those in J. Murdoch, *Control Charts* (London, Macmillan, 1979), from which the attribute table in Figure 11.32 has been extracted. For variable measure (with process design mean μ_x and standard deviation σ_x) the following are often used:

$$r = \mu_x + \frac{\sigma_x}{2\sqrt{n}}, \quad h = 5\frac{\sigma_x}{\sqrt{n}}$$

Figure 11.32 Cusum attribute chart control data.

np	r	h	np	r	h
0.22	1	1.5	2.35	4	4.5
0.39	1	2.5	2.60	4	5.5
0.51	2	1.5	2.95	5	4.5
0.62	1	4.5	3.24	5	5.5
0.69	1	5.5	3.89	6	5.5
0.79	2	2.5	4.16	6	6.5
0.86	3	1.5	5.32	7	8.5
1.05	2	3.5	6.07	8	8.5
1.21	3	2.5	7.04	9	9.5
1.52	3	3.5	8.01	10	10.5
1.96	3	5.5	9.00	11	11.5
2.16	5	2.5	10.00	12	12.5

Example 11.26

Regular samples of 50 are taken from a process making electronic components, for which an acceptable proportion of defectives is 5%. Successive counts of defectives in each sample are as follows:

Sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Count	3	5	2	2	1	6	4	4	2	6	7	4	5	5	8	6	5	9	7	8

At what point would the decision be taken to stop and correct the process?

Solution

The acceptable proportion of defectives is $p = 0.05$ and the regular sample size is $n = 50$. From the table in Figure 11.32, with $np = 2.5$ the nearest figures for reference value and decision interval are $r = 4$ and $h = 5.5$. The following shows the cusum S_m for $1 \leq m \leq 20$ below each count of defectives Y_m , and the cusum is also plotted in Figure 11.33:

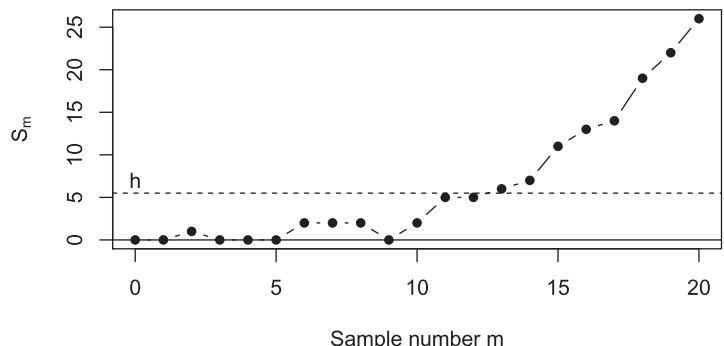
Count	3	5	2	2	1	6	4	4	2	6	7	4	5	5	8	6	5	9	7	8
Cusum	0	1	0	0	0	2	2	2	0	2	5	5	6	7	11	13	14	19	22	26

For example,

$$S_{13} = S_{12} + Y_{13} - r = 5 + 5 - 4 = 6$$

and because this exceeds $h = 5.5$, the decision to take action would be made after the 13th sample. This result can be compared with that of a Shewhart chart applied to the same data (Example 11.23), which suggests that action should be taken after 18 samples.

Figure 11.33
Cusum control chart
for Example 11.26.



We can perform the calculations and produce Figure 11.33 in R:

```

n <- 50; p <- 0.05; n * p
#> [1] 2.5
r <- 4; h <- 5.5
# Set-up space for S
N <- length(count)
S <- rep(NA, N + 1)
S[1] <- 0 # R starts indexing at 1, not 0

```

```

# Note the indexing on count, due to different range of m
# (1 to N + 1, not 0 to N)
for(m in 2:(N + 1)) {
  S[m] <- max(0, S[m-1] + count[m-1] - r)}
S # First value was set to 0
#> [1] 0 0 1 0 0 0 2 2 2 0 2 5 5 6 7 11 13 14 19 22 26
plot(0:N, S,
     type = "b", pch = 16,
     xlab = "Sample number m", ylab = expression(S[m]),
     xlim = c(0, 20), ylim = c(0, 26))
abline(h = h, lty = "dashed")
abline(h = 0)
text(0, h + 1.5, "h")

```

Example 11.27

Construct a cusum chart for the sulphur dioxide data in Example 11.25.

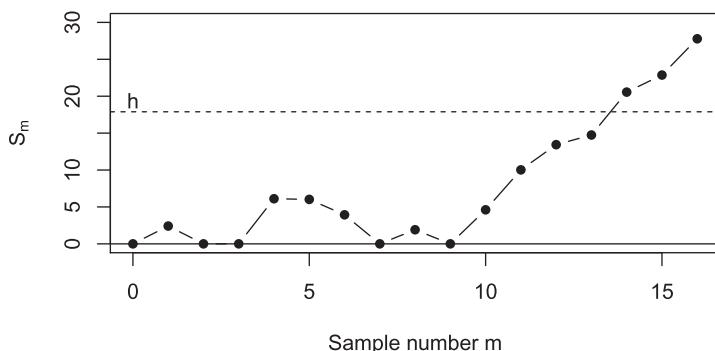
Solution From $\mu_x = 60$, $\sigma_x = 8$ and $n = 5$ we have

$$r = 60 + \frac{8}{2\sqrt{5}} = 61.8, \quad h = 5 \frac{8}{\sqrt{5}} = 17.9$$

The following table shows the sample average X_m and cusum S_m for $1 \leq m \leq 16$, and the cusum is also plotted in Figure 11.34:

Average	64.2	56.9	57.7	67.9	61.7	59.7	55.6	63.7
Cusum	2.4	0	0	6.1	6.0	3.9	0	1.9
Average	58.3	66.4	67.2	65.2	63.1	67.6	64.1	66.7
Cusum	0	4.6	10.0	13.4	14.7	20.6	22.9	27.8

Figure 11.34
Cusum control chart
for Example 11.27.



Because $S_{14} = 20.5$ exceeds $h = 17.9$, this chart suggests that the SO_2 concentration did increase during the experiment, a stronger result than that obtained from the Shewhart chart in Example 11.25.



We can perform the calculations and produce Figure 11.34 in R:

```

mu_X <- 60; sigma_X <- 8; n <- 5
r <- mu_X + sigma_X / (2 * sqrt(n)); r
#> [1] 61.78885
h <- 5 * sigma_X / sqrt(n); h
#> [1] 17.88854
# Set-up space for S
N <- length(SO2)
S <- rep(NA, N + 1)
S[1] <- 0
for(m in 2:(N + 1)) {S[m] <- max(0, S[m-1] + SO2[m-1] - r)}
S # First value was set to 0
#> [1] 0.000000 2.411146 0.000000 0.000000 6.111146
6.022291 3.933437
#> [8] 0.000000 1.911146 0.000000 4.611146 10.022291
13.433437 14.744582
#> [15] 20.555728 22.866874 27.778019
plot(0:N, S,
      type = "b", pch = 16,
      xlab = "Sample number m", ylab = expression(S[m]),
      xlim = c(0, 16), ylim = c(0, 30))
abline(h = h, lty = "dashed")
# A dashed horizontal line at h
abline(h = 0) # Horizontal line at 0
text(0, h + 1.5, "h")

```

Cusum control charts for measurements on a continuous scale can be produced in R from raw data using the `qcc` package for example.

It can be shown that the cusum method will usually detect an out-of-control condition (involving a slight process shift) much sooner than the strict Shewhart method, but with essentially the same risk of a false alarm. For instance, the cusum method leads to a decision after 13 samples in Example 11.26 compared with 18 samples in Example 11.23 for the same data. The measure used to compare the two methods is the **average run length (ARL)**, which is the mean number of samples required to detect an increase in proportion of defectives (or process average) to some specified level. It has been shown that the ARL for the Shewhart chart can be up to four times that for the cusum chart (J. Murdoch, *Control Charts*, London, Macmillan, 1979).

11.8.5 Moving-average control charts

The cusum chart shows that the way to avoid the relative insensitivity of the Shewhart chart is to allow the evidence of a shift in performance to accumulate over several samples. There are also **moving-average control charts**, which are based upon a weighted sum of a number of observations. The best of these, which is very similar to the cusum chart in operation, is the **geometric moving-average (GMA) chart**. This will be described here for variable measure, but it also works for attribute measure (Exercise 51).

Suppose that the successive sample averages are $\bar{X}_1, \bar{X}_2, \dots$, each from a sample of size n . Also suppose that the design mean and variance are μ_x and σ_x^2 . Then the GMA is the new sequence given by

$$\begin{aligned} S_0 &= \mu_x \\ S_m &= r\bar{X}_m + (1 - r)S_{m-1} \quad (m = 1, 2, \dots) \end{aligned}$$

where $0 < r \leq 1$ is a constant. The smaller the value of r the smaller is the contribution of new data points to S_m . The statistical properties of this sequence are simpler than for the cusum sequence. First, by successively substituting for S_{m-1}, S_{m-2} and so on, we can express S_m directly in terms of the sample averages:

$$S_m = r \sum_{i=0}^{m-1} [(1 - r)^i \bar{X}_{m-i}] + (1 - r)^m \mu_x$$

Then, using the summation formula

$$\sum_{i=0}^{m-1} x^i = \frac{1 - x^m}{1 - x}$$

it is easy to show (Exercise 52) that the mean and variance of S_m are

$$\begin{aligned} \mu_{S_m} &= E(S_m) = \mu_x \\ \sigma_{S_m}^2 &= \text{Var}(S_m) = \frac{r}{2 - r} [1 - (1 - r)^{2m}] \frac{\sigma_x^2}{n} \end{aligned}$$

After the first few samples the variance of S_m tends to a constant value:

$$\sigma_{S_m}^2 \rightarrow \left(\frac{r}{2 - r} \right) \frac{\sigma_x^2}{n} \quad \text{as } m \rightarrow \infty$$

as $0 \leq 1 - r < 1$

If US practice is followed then the upper and lower control limits can be set at $(\mu_x \pm 3\sigma_{S_m})$. If UK practice is followed then, from the approximate normality of the sample averages and the fact that sums of normal random variables are also normal, it follows that S_m is approximately normal, so the warning and action limits can be set at $(\mu_x \pm 1.96\sigma_{S_m})$ and $(\mu_x \pm 3.09\sigma_{S_m})$ respectively (although the warning limits now have less significance).

It remains to choose a value for r . If we set $r = 1$, the whole approach reduces to the standard Shewhart charts. Small values of r (say around 0.2) lead to early recognition of small shifts of process mean, but if r is too small, a large shift may remain undetected for some time.

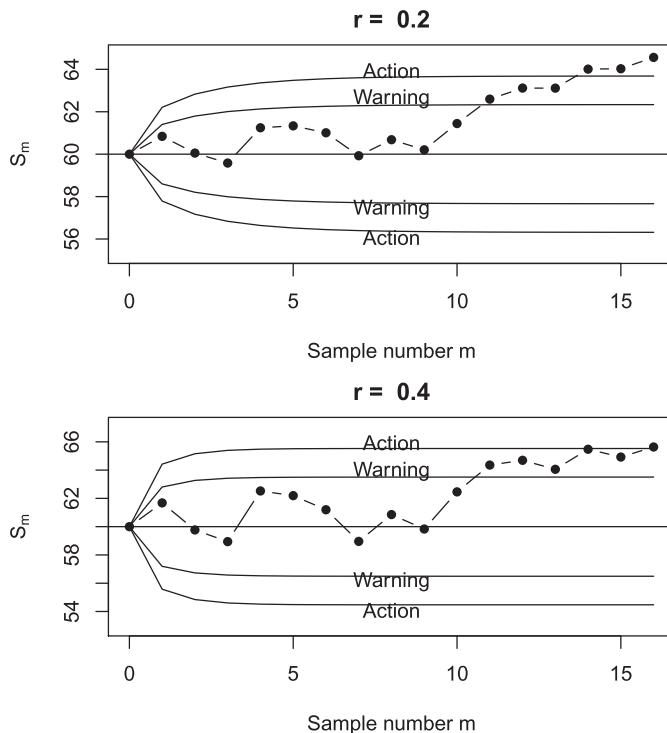
Example 11.28

Construct GMA charts for the sulphur dioxide data in Example 11.25, using $r = 0.2$ and $r = 0.4$.

Solution

The control charts can be seen in Figure 11.35. Clearly the warning and action limits converge fairly quickly to constant values, so little is lost by using those values in practice. The warning limit is exceeded from sample 11 for both values of r . The action limit is exceeded from sample 14 for $r = 0.2$ (as for the cusum chart in Example 11.27) and at sample 16 for $r = 0.4$.

Figure 11.35 Moving-average control chart for Example 11.28: for $r=0.2$ (top) and $r=0.4$ (bottom).



We can perform the calculations and produce Figure 11.35 in R. To experiment with different values of r , we create an R function `GMA_chart`, the construction and use of which we describe in detail.

```
GMA_chart <- function(sample_averages, n, mu_X, sigma_X, r,
                      x_lims = c(0, length(sample_averages)),
                      y_lims = c(mu_X - 4 * sqrt(r / (2 - r)) *
                                sigma_X / sqrt(n), mu_X + 4 * sqrt(r / (2 - r)) * sigma_X /
                                sqrt(n)),
                      text_shift = 0.3) {
  #
  # Function to construct a GMA chart
  # sample_averages: vector of sample averages
  # n: sample size, mu_X: design mean
  # sigma_X: design standard deviation
  # r: constant, r and (1 - r) provide the weights for the
  # sample average and moving average
  # x_lims: limits for the x axis, default: 0 to length of
  # sample_averages
  # y_lims: limits for the y axis, default: 4 standard
  # deviations of  $S_m$  from the center
  # text_shift: how much to shift the text on the chart,
  # default: 0.3
  #
```

```

# Set-up space for S
N <- length(sample_averages)
S <- rep(NA, N + 1)
S[1] <- mu_X
    # Assign S_0, remembering that in R we start indexing at 1
# Iteratively compute S
for(m in 2:(N + 1)) {S[m] <- r * sample_averages[m - 1] +
(1 - r) * S[m-1]}
# Variance of S_m
m <- 0:N # Values of m from 0 to N
sigma_2_S_m <- (r / (2 - r)) * (1 - (1 - r)^(2 * m)) *
sigma_X^2 / n
sigma_S_m <- sqrt(sigma_2_S_m) # Standard deviation
# Multiplicative constants
q_W <- qnorm(1 / 40, lower.tail = FALSE)
q_A <- qnorm(1 / 1000, lower.tail = FALSE)
plot(0:N, S,
      type = "b", pch = 16,
      main = paste("r = ", r),
      # Main title giving the value of r
      xlab = "Sample number m", ylab = expression(S[m]),
      xlim = x_lims, ylim = y_lims)
      # Limits on the x and y axis
abline(h = mu_X) # Central line
action_upper <- mu_X + q_A * sigma_S_m # Action limits
action_lower <- mu_X - q_A * sigma_S_m
warning_upper <- mu_X + q_W * sigma_S_m # warning limits
warning_lower <- mu_X - q_W * sigma_S_m
lines(0:N, action_upper); lines(0:N, action_lower)
    # Draw the lines
lines(0:N, warning_upper); lines(0:N, warning_lower)
text_position <- floor(N/2) # Text approximately half way
text(text_position, action_upper[text_position + 1] +
text_shift, "Action")
text(text_position, action_lower[text_position + 1] -
text_shift, "Action")
text(text_position, warning_upper[text_position + 1] +
text_shift, "Warning")
text(text_position, warning_lower[text_position + 1] -
text_shift, "Warning")
}
#
# Use the function
# Plot in matrix format, two rows, one column; reduce space
# below and above
p <- par(mfrow = c(2,1), mar = c(5.1 - 1, 5.1, 4.1 - 2,
2.1))
#
GMA_chart(sample_averages = SO2,
n = 5, mu_X = 60, sigma_X = 8, r = 0.2)

```

```
GMA_chart(sample_averages = SO2,
n = 5, mu_X = 60, sigma_X = 8, r = 0.4,
text_shift = 0.4)
# Change text_shift for better text placement
```

Try also running the code that produced Figure 11.35 for $r = 1$ and compare the result with Figure 11.33. What happens when $r > 0$ is set to a very small number? Moving-average control charts can be produced in R from raw data using the `qcc` package for example.

11.8.6 Range charts

The **sample range** is defined as the difference between the largest and smallest values in the sample. The range has two functions in quality control where the quality is of the variable rather than the attribute type. First, if the data are normal then the range (R , say) provides an estimate $\hat{\sigma}$ of the standard deviation σ by

$$\hat{\sigma} = R/d$$

where d is a constant that depends upon the sample size n as follows:

n	2	3	4	5	6	7	8	9	10	11	12
d	1.128	1.693	2.059	2.326	2.534	2.704	2.847	2.970	3.078	3.173	3.258

It is clearly quicker to evaluate this than the sample standard deviation S , and for the small samples typically used in quality control the estimate is almost as good.

The other reason why the range is important is because the quality of production can vary in dispersion as well as (or instead of) in mean. Control charts for the range R are more commonly used than charts for the sample standard deviation S when monitoring variability within the manufacturing process, and all three types of chart discussed above (Shewhart, cusum and moving-average) can be applied to the range. **Range charts** (or **R charts**) are designed using tables that can be found in specialized books on quality control, for example D. C. Montgomery, *Introduction to Statistical Quality Control* (seventh edition, New York, Wiley, 2012).



Range charts can be produced in R from raw data using the `qcc` package for example.



We conclude our discussion of statistical quality control by briefly illustrating some of the above techniques using real data supplied with the `qcc` package. In particular, we will work with the `pistonrings` data frame:

Piston rings for an automotive engine are produced by a forging process. The inside diameter [stored in the variable `diameter`] of the rings manufactured by the process is measured on 25 samples [stored in the variable `sample`], each of size 5, for the control Phase I, when preliminary samples from a process being considered ‘in control’ are used to construct control charts. Then, a further 15 samples, again each of size 5, are obtained for Phase II.

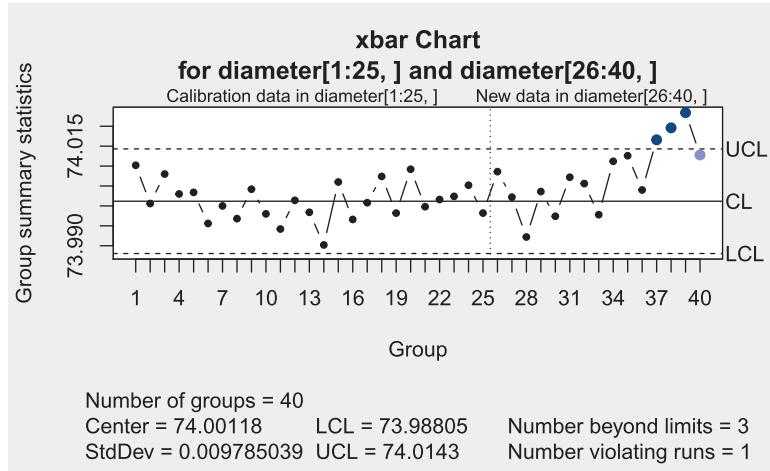


Figure 11.36 Variable control chart for the pistonrings data produced by the qcc package. The chart computations are performed using the Phase I ‘in control’ data. The chart can then be used to monitor the Phase II data.

The variable `trial` takes the value `TRUE` for Phase I data and the value `FALSE` for Phase II data.

We now reproduce two of the many examples in the qcc help file. This code produces the variable control chart shown in Figure 11.36:

```
data("pistonrings") # To access the data
str(pistonrings)
#> 'data.frame': 200 obs. of 3 variables:
#> $ diameter: num 74 74 74 74 74 ...
#> $ sample : int 1 1 1 1 1 2 2 2 2 ...
#> $ trial : logi TRUE TRUE TRUE TRUE TRUE ...
# Arrange the data in samples of size 5
diameter <- with(pistonrings, qcc.groups(diameter, sample))
head(diameter)
#>      [,1]     [,2]     [,3]     [,4]     [,5]
#> 1 74.030 74.002 74.019 73.992 74.008
#> 2 73.995 73.992 74.001 74.011 74.004
#> 3 73.988 74.024 74.021 74.005 74.002
#> 4 74.002 73.996 73.993 74.015 74.009
#> 5 73.992 74.007 74.015 73.989 74.014
#> 6 74.009 73.994 73.997 73.985 73.993
# Variable control chart based on Phase I data (25 samples,
# samples 1 to 25),
# also showing Phase II data (15 samples, samples 26 to 40)
c1 <- qcc(diameter[1:25,], type = "xbar",
           newdata = diameter[26:40,])
```

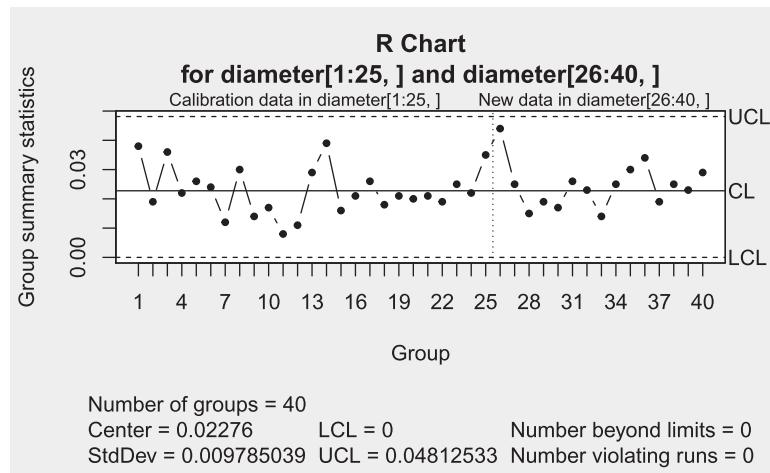


Figure 11.37 Range chart for the pistonrings data produced by the qcc package.

This code produces the range chart shown in Figure 11.37:

```
# R chart based on Phase I data, also showing Phase II data
c2 <- qcc(diameter[1:25,], type = "R",
            newdata = diameter[26:40,])
```

We finish by noting that a recent book-length treatment of statistical quality control using R is provided by P. Qiu, *Introduction to Statistical Process Control* (Boca Raton, FL, Chapman and Hall/CRC, 2014). In addition to a deep coverage of the Shewhart, cusum, moving-average and range charts described in this section, this book also reviews change-point detection and multivariate statistical process control. Non-parameter process control methodology for when underlying assumptions about normality do not hold is also discussed, as is profile monitoring when the stability of the relationship between variables over time is important. Five R packages are mentioned, and data sets and R code are supplied on a website.

11.8.7 Exercises



Confirm your work using R wherever possible.

43

It is intended that 90% of electronic devices emerging from a machine should pass a simple on-the-spot quality test. The numbers of defectives among samples of 50 taken by successive shifts are as follows:

5, 8, 11, 5, 6, 4, 9, 7, 12, 9, 10, 14

Find the action and warning limits, and the sample number at which an out-of-control decision is taken. Also find the UCL (US practice) and the sample number for action.

44

Thirty-two successive samples of 100 castings each, taken from a production line, contained the following numbers of defectives

3, 3, 5, 3, 5, 0, 3, 1, 3, 5, 4, 2, 4, 3, 5, 4
3, 4, 5, 6, 5, 6, 4, 4, 7, 5, 4, 8, 5, 6, 6, 7

If the proportion that are defective is to be maintained at 0.02, use the Shewhart method (both UK and US standards) to indicate whether this proportion is being maintained, and if not then give the number of samples after which action should be taken.

- 45** A bottling plant is supposed to fill bottles with 568 ml (one imperial pint) of liquid. The standard deviation of the quantity of fill is 3 ml. Regular samples of 10 bottles are taken and their contents measured. After subtracting 568 from the sample averages, the results are as follows:

−0.2, 1.3, 2.1, 0.3, −0.8, 1.7, 1.3, 0.6, 2.5, 1.4, 1.6, 3.0

Using a Shewhart control chart, determine whether the mean fill requires readjustment.

- 46** Average reverse-current readings (in nA) for samples of 10 transistors taken at half-hour intervals are as follows:

12.8, 11.2, 13.4, 12.1, 13.6, 13.9, 12.3, 12.9, 13.8, 13.1, 12.9, 14.0, 13.7, 13.4, 14.2, 13.1, 14.0, 14.0, 15.1, 14.3

The standard deviation is 3 nA. At what point, if any, does the Shewhart control method indicate that the reverse current has increased from its design value of 12 nA?

- 47** Using the data in Exercise 45, apply (a) a cusum control chart and (b) a moving-average control chart with $r = 0.3$.

- 48** Using the data in Exercise 46, apply (a) a cusum control chart and (b) a moving-average control chart with $r = 0.3$.

- 49** Apply a cusum control chart to the data in Exercise 43.

- 50** Apply a cusum control chart to the data in Exercise 44.

- 51** The diameters of the castings in Exercise 44 are also important. Twelve of each sample of 100 were taken, and their diameters measured and averaged.

The differences (in mm) between the successive averages and the design mean diameter of 125 mm were as follows:

0.1, 0.3, −0.2, 0.4, 0.1, 0.0, 0.2, −0.1, 0.2, 0.4, 0.5, 0.1, 0.4, 0.6, 0.3, 0.4, 0.3, 0.6, 0.5, 0.4, 0.2, 0.3, 0.5, 0.7, 0.3, 0.1, 0.6, 0.5, 0.6, 0.7, 0.4, 0.5

Use (a) Shewhart, (b) cusum and (c) moving-average (with $r = 0.2$) control methods to test for an increase in actual mean diameter, assuming a standard deviation of 1 mm.

- 52** Prove that the mean and variance of the geometric moving-average S_m defined in Section 11.8.5 for variable measure are given by

$$E(S_m) = \mu_x$$

$$\sigma_{S_m}^2 = \text{Var}(S_m) = \frac{r}{2-r} [1 - (1-r)^{2m}] \frac{\sigma_x^2}{n}$$

- 53** Suppose that the moving-average control chart is to be applied to the counts of defectives in attribute quality control. Find the mean and variance of S_m in terms of the sample size n , the design proportion of defectives p and the coefficient r . Following US practice, set the upper control limit at three standard deviations above the mean, and apply the method to the data in Example 11.26, using $r = 0.2$.

- 54** The design diameter of a moulded plastic component is 6.00 cm, with a standard deviation of 0.2 cm. The following data consist of successive averages of samples of 10 components:

6.04, 6.12, 5.99, 6.02, 6.04, 6.11, 5.97, 6.06, 6.05, 6.06, 6.17, 6.03, 6.13, 6.05, 6.17, 5.97, 6.07, 6.14, 6.03, 5.99, 6.10, 6.01, 5.96, 6.12, 6.02, 6.20, 6.11, 5.98, 6.02, 6.12

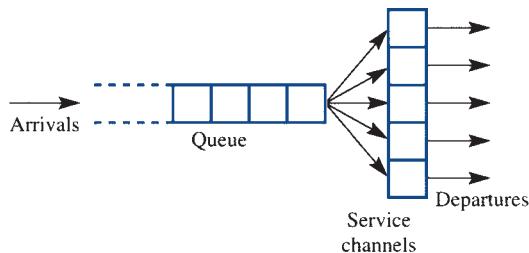
After how many samples do the Shewhart, cusum and moving-average (with $r = 0.2$) control methods indicate that action is needed?

11.9

Poisson processes and the theory of queues

Probability theory is often applied to the analysis and simulation of systems, and this can be a valuable aid to design and control. This section, which is therefore applied probability rather than statistics, will illustrate how understanding of systems can be gained from an initial mathematical model using probability-based analysis and computer simulation.

Figure 11.38
A typical queueing system.



11.9.1 Typical queueing problems

Queues are everywhere: in banks and shops, at airports and seaports, traffic intersections and hospitals, and in computer and communication networks. Somebody has to decide on the level of service facilities. The problem, in essence, is that it is costly to keep customers waiting for a long time, but it is also costly to provide enough service facilities so that no customer ever has to wait at all. Queues of trucks, aeroplanes or ships may be costly because of the space they occupy or the lost earnings during the idle time. Queues of people may be costly because of lost productivity or because people will often go elsewhere in preference to joining a long queue. Queues of jobs or packets of data in computer networks are costly in loss of time-efficiency. Service facilities are costly in capital, staffing and maintenance. Probabilistic modelling, combined with simulation, allows performance evaluation for queues and networks, which can be of great value in preparing the ground for design decisions.

The mathematical model of a simple queueing system is based on the situation shown in Figure 11.38. **Customers** join the queue at random times that are independent of each other – the **inter-arrival time** (between successive arrivals) is a random variable. When a **service channel** is free, the next customer to be served is selected from the queue in a manner determined by the **service discipline**. After being served the customer departs from the queueing system. The **service time** for each customer is another random variable. The distributions of inter-arrival time and service time are usually assumed to take one of a number of standard patterns. The commonest assumption about service discipline is that the next customer to be served is the one who has been queueing the longest time (first in, first out).

The queueing system may be regarded from either a static or a dynamic viewpoint. Dynamically, the system might start from an initial state of emptiness and build up with varying rates of arrivals and varying numbers of service channels depending upon queue length. This is hard to deal with mathematically, but can be treated by computer simulation. Useful information about queues can, however, be obtained from the static viewpoint, in which the rate at which arrivals occur is constant, as is the number of service channels, and the system is assumed to have been in operation sufficiently long to have reached a steady state. At any time the queue length will be a random variable, but the distribution of queue length is then independent of time.

We need to find the distributions of queue length and of waiting time for the customer, and how these vary with the number of service channels. Costs can be worked out from these results.

11.9.2 Poisson processes

Consider the arrivals process for a queueing system. We shall assume that the customers join the queue at random times that are independent of each other. Other assumptions

Figure 11.39

Random events (\times)
on a time axis.



about the pattern of arrivals would give different results, but this is the most common one. We can therefore think of the arrivals as a stream of events occurring at random along a time axis, as depicted in Figure 11.39. The inter-arrival time T , say, will be a continuous random variable with probability density function $f_T(t)$ and cumulative distribution function $F_T(t)$.

One way to formulate the assumption of independent random arrivals is to assert that at any moment the distribution of the time until the next arrival is independent of the time elapsed since the previous arrival (because arrivals are ‘blind’ to each other). This is known as the **memoryless property**, and can be expressed as

$$P(T \leq t + h | T > t) = P(T \leq h) \quad (t, h \geq 0)$$

where t denotes the actual time since the previous arrival and h denotes a possible time until the next arrival. Using the definition of conditional probability (Section 11.2.1), we can write this in terms of the distribution function $F_T(t)$ as

$$\frac{P(t < T \leq t + h)}{1 - P(T \leq t)} = \frac{F_T(t + h) - F_T(t)}{1 - F_T(t)} = F_T(h)$$

Rearranging at the second equality and then dividing through by h gives

$$\frac{1}{h}[F_T(t + h) - F_T(t)] = \frac{F_T(h)}{h}[1 - F_T(t)]$$

Letting $h \rightarrow 0$, we obtain a first-order linear differential equation for $F_T(t)$:

$$\frac{d}{dt}F_T(t) = \lambda[1 - F_T(t)]$$

where

$$\lambda = \lim_{h \rightarrow 0} \frac{F_T(h)}{h}$$

With the initial condition $F_T(0) = 0$ (because inter-event times must be positive), the solution is

$$F_T(t) = 1 - e^{-\lambda t} \quad (t \geq 0)$$

and hence the probability density function is

$$f_T(t) = \frac{d}{dt}F_T(t) = \lambda e^{-\lambda t} \quad (t \geq 0)$$

This is the density function of an **exponential distribution with parameter λ** , and it follows by integrating $E[T] = \int_0^\infty t \lambda e^{-\lambda t} dt$ that the mean time between arrivals is $1/\lambda$. The parameter λ is the **rate of arrivals** (number per unit time).

Example 11.29

A factory contains 30 machines of a particular type, each of which breaks down every 100 operating hours on average. It is suspected that the breakdowns are not independent. The operating time intervals between 10 consecutive breakdowns (of any machine) are measured and the shortest such interval is only six minutes. Does this lend support to the suspicion of non-independent breakdowns?

Solution

Collectively, the machines break down at the rate of 30/100 or 0.3 per hour. If the breakdowns are independent then the interval between successive breakdowns will have an exponential distribution with parameter 0.3. The probability that such an interval will exceed six minutes is

$$P(\text{interval} > 0.1) = \int_{0.1}^{\infty} 0.3 e^{-0.3t} dt = e^{-0.3(0.1)} = 0.9704$$

and the probability that all nine intervals (between 10 breakdowns) will exceed this time is $(0.9704)^9 = 0.763$. Hence the probability that the shortest interval will be six minutes or less is one minus this, or 0.237. This is quite likely to have happened by chance, so it does not support the suspicion of non-independent intervals.



We may perform these calculations in R:

```
p_interval <- pexp(0.1, rate = 0.3, lower.tail = FALSE)
p_interval
#> [1] 0.9704455
p_interval^9; 1 - p_interval^9
#> [1] 0.7633795
#> [1] 0.2366205
```

The assumption of independent random arrivals therefore leads to a particular distribution of inter-arrival time, parametrized by the rate of arrivals. Two further conclusions also emerge. First, the number of arrivals that occur during a fixed interval of length H has a Poisson distribution with parameter λH :

$$P(k \text{ arrivals during interval of length } H) = \frac{(\lambda H)^k e^{-\lambda H}}{k!} \quad (k = 0, 1, 2, \dots)$$

This will not be proved here, but is easily seen to be consistent with an exponential distribution of inter-arrival time T because

$$\begin{aligned} F_T(t) &= P(T \leq t) = 1 - P(T > t) \\ &= 1 - P(\text{no event during interval of length } t) \\ &= 1 - e^{-\lambda t} \end{aligned}$$

using the Poisson distribution. Because of this distribution, events conforming to these assumptions are known as a **Poisson process**.

The other conclusion is that the probability that an arrival occurs during a short interval of length h is equal to $\lambda h + O(h^2)$, regardless of the history of the process. Suppose that a time t has elapsed since the previous arrival, and consider a short interval of length h starting from that point:

$$\begin{aligned} P(\text{arrival during } (t, t+h)) &= P(T \leq t+h | T > t) = F_T(h) \\ &= 1 - e^{-\lambda h} = \lambda h + O(h^2) \end{aligned}$$

using the memoryless property and the expansion of $e^{-\lambda h}$ to first order. Furthermore, the probability of more than one arrival during a short interval of length h is $O(h^2)$.

Example 11.30

A computer receives on average 60 batch jobs per day. They arrive at a constant rate throughout the day and independently of each other. Find the probability that more than four jobs will arrive in any one hour.

Solution

The assumptions for a Poisson process hold, so the number of jobs arriving in an interval of one hour ($H = 1$) is a Poisson random variable with parameter $\lambda H = 60/24$. Hence

$$P(\text{more than four jobs}) = 1 - P(0 \text{ or } 1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ jobs})$$

$$= 1 - e^{-\lambda H} \left[1 + \lambda H + \frac{(\lambda H)^2}{2!} + \frac{(\lambda H)^3}{3!} + \frac{(\lambda H)^4}{4!} \right] = 0.109$$



We may perform these calculations in R:

```
lambda_H <- 60 / 24
1 - sum(dpois(0:4, lambda = lambda_H));
ppois(4, lambda = lambda_H, lower.tail = FALSE)
#> [1] 0.108822
#> [1] 0.108822
```



We will now simulate and display realizations of the Poisson process described in Example 11.30. Please note that in general every time simulation code is run a different simulated process results. We can make the results from simulations reproducible by using `set.seed`; please read the help file. Here is a very simple example based on generating five random numbers from a standard normal distribution:

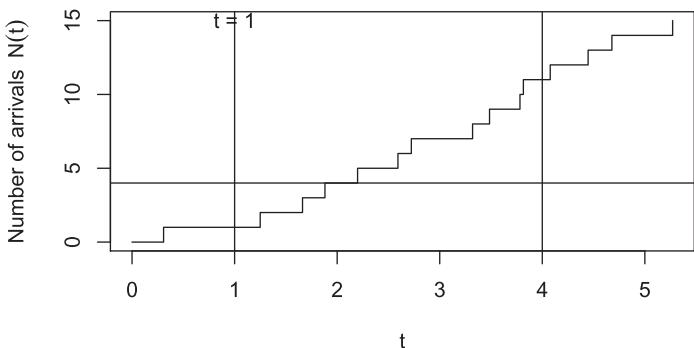
```
set.seed(23); rnorm(5)
#> [1] 0.1932123 -0.4346821 0.9132671 1.7933881 0.9966051
set.seed(23); rnorm(5) # The same
#> [1] 0.1932123 -0.4346821 0.9132671 1.7933881 0.9966051
rnorm(5) # Different
#> [1] 1.10749049 -0.27808628 1.01920549 0.04543718 1.57577959
set.seed(seed = NULL)
# re-initializes as if no seed had yet been set
```

We now simulate a Poisson process with $\lambda = 60 / 24$ over a 4 hour period. Let $N(t)$ be the number of arrivals up to time t . We know that $N(t) \sim \text{Po}(\lambda t)$, where \sim means ‘is distributed as’. Our simulation strategy will be to generate n arrivals where n is sufficiently large to cover a 4 hour period with high probability; that is, n is such that $P(N(4) \leq n) = 0.95$ say. This number n is said to be the 0.95-quantile of a $\text{Po}(\lambda \times 4)$ random variable. Let us find n :

```
lambda <- 60 / 24
# qpois is the quantile function for the Poisson distribution
n <- qpois(0.95, lambda = lambda * 4); n
#> [1] 15
```

Figure 11.40

A simulated Poisson process. The number of arrivals $N(t)$ is plotted against time t .



We will now simulate n inter-arrival times, which we know follow an exponential distribution with parameter λ . We use `set.seed` to allow the results to be reproduced:

```
set.seed(39)
# Set the seed of the random number generator for reproducibility
inter_arrival_times <- rexp(n = n, rate = lambda)
```

The arrival times themselves are the cumulative sums of these inter-arrival times, where the cumulative sums of x_1, x_2, x_3, \dots are $x_1, x_1 + x_2, x_1 + x_2 + x_3, \dots$. Cumulative sums can be found using the `cumsum` function. Here is a simple example:

```
x <- c(1, 3, 2, -7)
cumsum(x)
#> [1] 1 4 6 -1
```

We now work out the arrival times:

```
arrival_times <- cumsum(inter_arrival_times)
```

The corresponding values of the number of arrivals $N(t)$ are $1, 2, \dots, n$:

```
N_t <- 1:n
```

We can also include the fact that $N(0) = 0$:

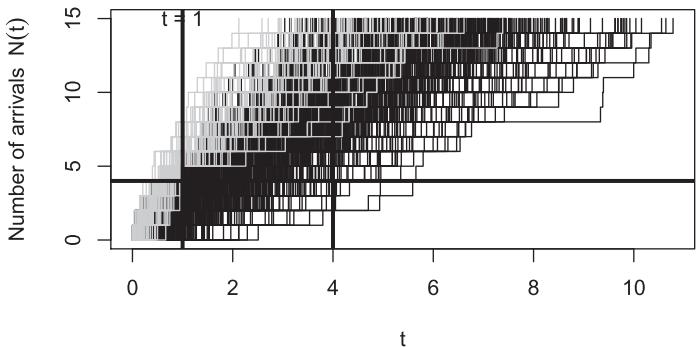
```
arrival_times_0 <- c(0, arrival_times); N_t_0 <- c(0, N_t)
```

We can display this Poisson process by plotting $N(t)$ against t , as shown in Figure 11.40, for which the R code is:

```
p <- par(mar = c(5.1, 4.1 + 1, 4.1, 2.1))
# Plot need more space on left side for axis label
plot(arrival_times_0, N_t_0,
      type = "s", # We plot using stair steps
      xlab = "t", ylab = expression(paste("Number of arrivals",
~ N(t))))
abline(v = c(1, 4)) # A vertical line at 1 and 4 hours
abline(h = 4)
# A horizontal line at 4, corresponding to 4 jobs as in
# Example 11.30
text(1, 15, "t = 1") # Text "t = 1" at (1, 15)
```

Figure 11.41

Five hundred simulated Poisson processes. Processes for which there have been five arrivals before $t = 1$ are shaded in grey.



We will now generate 500 Poisson processes, using the replicate function. We will plot the resulting matrix of arrival times as shown in Figure 11.41, in which we shade Poisson processes that have five arrivals by time $t = 1$ in grey:

```
N_paths <- 500 # Number of Poisson process paths
# Replicate (repeat) the code in {} N_paths times
arrival_times_0_mat <- replicate(N_paths,
  {c(0, cumsum(rexp(n = n, rate = lambda))))})
# Plot need more space on left side for axis label
p <- par(mar = c(5.1, 4.1 + 1, 4.1, 2.1))
# We will plot the Poisson processes with five arrivals by time
# t = 1 after the other processes to make them stand out better
# To achieve this, we first define the plotting axes, without
# plotting anything
matplot(arrival_times_0_mat, N_t_0,
  # matplot plots matrices (here, a matrix and a vector)
  type = "n", # Plot nothing
  xlab = "t", ylab = expression(paste("Number of arrivals",
    ~ N(t))))
#
# Identify the Poisson processes with five arrivals before time
# t = 1
# That is, the time of the 5th arrival is before 1
five_arrivals_before_1 <- arrival_times_0_mat[N_t_0 == 5,] < 1
  # Logical vector five_arrivals_before_1[1:10]
#> [1] FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE
#> FALSE
all_processes <- seq_len(N_paths)
  # Sequence of path labels 1, 2,...,N_paths
# Indicies of Poisson processes with five arrivals before time
# t = 1
# Think of [] as 'such that'
processes_five_arrivals_before_1 <-
  all_processes[five_arrivals_before_1]
```

```

# Indicies of the other Poisson processes; ! ('not') turns
# TRUE/FALSE to FALSE/TRUE
processes_not_five_arrivals_before_1 <-
  all_processes[!five_arrivals_before_1]
# First, add these Poisson processes to the plot
matlines(arrival_times_0_mat[,,
  processes_not_five_arrivals_before_1],
  # Select the required columns; matrices are indexed using
  # [row, column]
  # all rows are required
  N_t_0,
  type = "s", lty = 1,
  # We plot using stair steps and line type 1 (continuous)
  col = "black") # Use black
# Now, add on top the Poisson processes with five arrivals
# before time t = 1
matlines(arrival_times_0_mat[,,
  processes_five_arrivals_before_1], # Select columns
  N_t_0,
  type = "s", lty = 1,
  col = "grey") # Use grey
# Lines and text as before
abline(v = c(1, 4), h = 4, lwd = 3) # Increased line width
text(1, 15, "t = 1")

```

Let us confirm the results of Example 11.30 by extracting the number of events at time $t = 1$. We will use a much larger number of simulated Poisson processes for this. We will again use the function `apply` which will allow us to apply a function to every column of a matrix:

```

N_paths <- 10000 # 10000 Poisson process paths
arrival_times_0_mat <- replicate(N_paths, {c(0, cumsum(rexp(n
  = n, rate = lambda))))})
# Show the arrival times of the first five Poisson processes
arrival_times_0_mat[1:10,1:5]
#>           [,1]      [,2]      [,3]      [,4]      [,5]
#> [1,] 0.0000000 0.0000000 0.000000 0.0000000 0.0000000
#> [2,] 1.010416 0.1326435 1.021045 0.3315623 0.5446702
#> [3,] 1.053304 0.1473250 1.042809 0.4019004 0.8778008
#> [4,] 1.604397 0.1693694 1.278456 0.4834921 0.9221729
#> [5,] 1.631006 0.3054541 2.535763 0.5296068 1.4048160
#> [6,] 1.639120 0.5941504 2.571398 1.2126070 1.6758765
#> [7,] 1.968432 0.9050178 2.970899 1.4489347 1.9070222
#> [8,] 2.108633 1.4639267 3.899057 1.7042519 1.9995854
#> [9,] 2.662620 1.5898577 4.125454 1.8247135 2.1912819
#> [10,] 2.964361 1.8224266 4.362871 2.0112748 2.2430705
# Find the subscript of the largest arrival_time no greater
# than t = 1
subscript_t <- function(arrival_times_0, t = 1){ # Default
  value of t is 1

```

```
# The coordinate of the first value of arrival_times_0 that is
# greater than t
first_greater <- which(arrival_times_0 > t)[1]
# Return the previous coordinate
first_greater - 1
}
# Apply subscript_t to the columns (MARGIN = 2) of
# arrival_times_0_mat
subscripts_at_1 <- apply(arrival_times_0_mat,
                           MARGIN = 2, FUN = subscript_t, t = 1)
# t = 1 not needed as default value
# Subscripts of the largest arrival_time no greater than t = 1,
# shown for first five Poisson processes
subscripts_at_1[1:5]
#> [1] 1 7 1 5 4
# Find the corresponding values of N(t), the number of arrivals
# up to time t
N_t_at_1 <- N_t_0[subscripts_at_1]
N_t_at_1[1:5] # For the first five Poisson processes
#> [1] 0 6 0 4 3
# Find the proportions
proportions <- prop.table(table(N_t_at_1)) ; proportions
#> N_t_at_1
#>      0      1      2      3      4      5      6      7
#> 0.0760 0.2068 0.2574 0.2107 0.1356 0.0692 0.0306 0.0099
#>      8      9     10     11
#> 0.0028 0.0006 0.0003 0.0001
# Compare with Po(lambda) probabilities
round(dpois(0:max(N_t_at_1), lambda = lambda * 1), 4) # Rounded
#> [1] 0.0821 0.2052 0.2565 0.2138 0.1336 0.0668 0.0278
#>      0.0099 0.0031 0.0009
#> [11] 0.0002 0.0000
# Approximate probability in Example 11.30; R starts indexing
# from 1
1 - sum(proportions[1 + (0:4)]) ; ppois(4, lambda = lambda * 1,
    lower.tail = FALSE)
#> [1] 0.1135
#> [1] 0.108822
```

11.9.3 Single service channel queue

Consider a queueing system with a Poisson arrival process with mean rate λ per unit time, and a single service channel. The behaviour of the queueing system depends not only on the arrival process but also upon the distribution of service times. A common

assumption here is that the service time distribution (like that of inter-arrival time) is exponential. Thus the probability density function of service time S is

$$f_S(s) = \mu e^{-\mu s} \quad (s \geq 0)$$

Unlike the inter-arrival time distribution in Section 11.9.2, this is not based on an assumption of independence or the memoryless property, but simply on the fact that in many queueing situations most customers are served quickly but a few take a lot longer, and the form of the distribution conforms with this fact. This assumption is therefore on much weaker ground than that for the arrival time distribution. The parameter μ is the mean number of customers served in unit time (with no idle periods), and the mean service time is $1/\mu$. With this service distribution, the probability that a customer in the service channel will have departed after a short time h is equal to $\mu h + O(h^2)$, independent of the time already spent in the service channel.

Distribution of the number of customers in the system

We can now derive the distribution of the number of customers in the queueing system. Considering the system as a whole (queue plus service channel), the number of customers in the system at time t is a random variable. Let $p_n(t)$ be the distribution of this random variable:

$$p_n(t) = P(n \text{ customers in the system at time } t) \quad (n = 0, 1, 2, \dots)$$

Consider the time $t + h$, where h is small. The probability of more than one arrival or more than one departure during this time is $O(h^2)$, and will be ignored. There are four ways in which there can be n (assumed greater than zero) customers in the system at that time:

- (1) there are n in the system at t , and no arrival or departure by $t + h$; the probability of this is given by

$$p_n(t)(1 - \lambda h)(1 - \mu h) + O(h^2) = p_n(t)(1 - \lambda h - \mu h) + O(h^2)$$

- (2) there are n in the system at t , and one arrival and one departure by $t + h$; the probability is given by

$$p_n(t)(\lambda h)(\mu h) + O(h^2) = O(h^2)$$

- (3) there are $n - 1$ in the system at t , and one arrival but no departure by $t + h$; the probability is given by

$$p_{n-1}(t)(\lambda h)(1 - \mu h) + O(h^2) = p_{n-1}(t)(\lambda h) + O(h^2)$$

- (4) there are $n + 1$ in the system at t , and no arrivals but one departure by $t + h$; the probability is given by

$$p_{n+1}(t)(1 - \lambda h)(\mu h) + O(h^2) = p_{n+1}(t)(\mu h) + O(h^2)$$

Summing the probabilities of these mutually exclusive events gives the probability of n customers in the system at time $t + h$ as

$$\begin{aligned} p_n(t+h) &= p_n(t)(1 - \lambda h - \mu h) + p_{n-1}(t)(\lambda h) \\ &\quad + p_{n+1}(t)(\mu h) + O(h^2) \quad (n = 1, 2, \dots) \end{aligned} \tag{11.1}$$

Similarly, there are two ways in which the system can be empty ($n = 0$) at time $t + h$: empty at t and no arrival before $t + h$, or one customer at t who departs before $t + h$. This gives

$$p_0(t+h) = p_0(t)(1 - \lambda h) + p_1(t)(\mu h) + O(h^2) \quad (11.2)$$

Rearranging equations (11.1) and (11.2) and taking the limit as $h \rightarrow 0$, we obtain

$$\begin{aligned} \frac{d}{dt} p_n(t) &= \lim_{h \rightarrow 0} \frac{1}{h} [p_n(t+h) - p_n(t)] \\ &= -(\lambda + \mu)p_n(t) + \lambda p_{n-1}(t) + \mu p_{n+1}(t) \quad (n = 1, 2, \dots) \\ \frac{d}{dt} p_0(t) &= -\lambda p_0(t) + \mu p_1(t) \end{aligned}$$

This is a rather complex set of recursive differential equations for the probabilities $p_n(t)$. If we assume that the arrival and service parameters λ and μ are constant and that the system has been in operation for a long time then the distribution will not depend upon t ; the derivatives therefore vanish, and we are left with the following algebraic equations for the *steady-state* distribution p_n :

$$\begin{aligned} 0 &= -(\lambda + \mu)p_n + \lambda p_{n-1} + \mu p_{n+1} \quad (n = 1, 2, \dots) \\ 0 &= -\lambda p_0 + \mu p_1 \end{aligned}$$

Defining the ratio of arrival and service parameters λ and μ as $\rho = \lambda/\mu$ and dividing through by μ , we have

$$\begin{aligned} p_{n+1} &= (1 + \rho)p_n - \rho p_{n-1} \quad (n = 1, 2, \dots) \\ p_1 &= \rho p_0 \end{aligned}$$

To solve these, we first assume that $p_n = \rho^n p_0$. Clearly this works for $n = 0$ and $n = 1$. Substituting,

$$p_{n+1} = (1 + \rho)\rho^n p_0 - \rho\rho^{n-1}p_0 = \rho^{n+1}p_0$$

so the assumed form holds for $n + 1$, and therefore for all n by induction. It remains only to identify p_0 from the fact that the distribution must sum to unity over $n = 0, 1, 2, \dots$:

$$1 = p_0 \sum_{n=0}^{\infty} \rho^n = \frac{p_0}{1 - \rho} \text{ provided } \rho < 1$$

Hence $p_0 = 1 - \rho$ and

$$p_n = (1 - \rho)\rho^n \quad (n = 0, 1, 2, \dots)$$

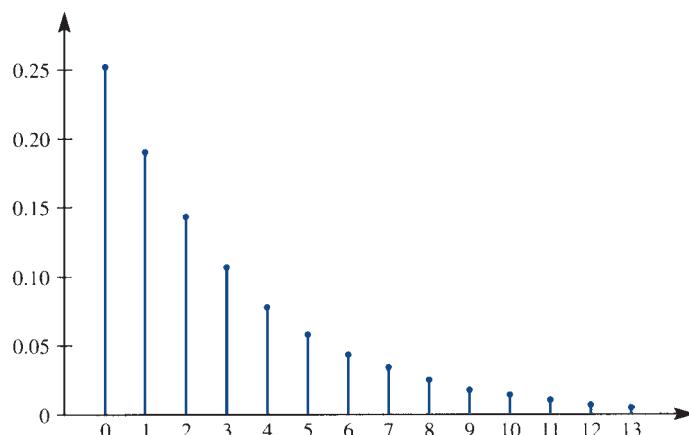
This is known as the **geometric distribution**, and is a discrete version of the exponential distribution (Figure 11.42). Note that this result requires that $\rho < 1$, or equivalently $\lambda < \mu$. If this condition fails to hold, the arrival rate swamps the capacity of the service channel, the queue gets longer and longer, and no steady-state condition exists.

Queue length and waiting time

The queue length distribution now follows easily:

$$\begin{aligned} P(\text{queue empty}) &= p_0 + p_1 \\ &= 1 - \rho^2 \end{aligned}$$

Figure 11.42
Geometric distribution
(with $\rho = 0.75$).



$$\begin{aligned} P(n \text{ in queue}) &= P(n+1 \text{ in system}) \\ &= (1 - \rho)\rho^{n+1} \quad (n = 1, 2, \dots) \end{aligned}$$

Denoting the mean numbers of customers in the system and in the queue by N_S and N_Q respectively,

$$N_S = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1 - \rho}, \quad N_Q = \sum_{n=1}^{\infty} (n - 1)p_n = \frac{\rho^2}{1 - \rho}$$

(Exercise 57). From this it follows that $N_S = N_Q + \rho_0$. Since in the steady state the mean time between departures must equal the mean time between arrivals ($1/\lambda$), it is plausible that the mean total time in the system for each customer, W_S say, is given by

$$\begin{aligned} W_S &= \text{mean number in system} \times \text{mean time between departures} \\ &= \frac{\rho/\lambda}{1 - \rho} = \frac{1}{\mu - \lambda} \end{aligned}$$

The mean waiting time in the queue, W_Q say, is then

$$\begin{aligned} W_Q &= \text{mean time in system} - \text{mean service time} \\ &= W_S - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda} \end{aligned}$$

These results for W_S and W_Q can be derived more formally from the respective waiting time distributions. For example, the distribution of total time in the system can be shown to be exponential with parameter $\mu - \lambda$, and the waiting time in the queue can be expressed as

$$P(\text{waiting time in queue} \geq t) = \rho e^{-(\mu - \lambda)t} \quad (t > 0)$$

Example 11.31

If customers in a shop arrive at a single check-out point at the rate of 30 per hour and if the service times have an exponential distribution, what mean service time will ensure that 80% of customers do not have to wait more than five minutes in the queue and what will be the mean queue length?

Solution With $\lambda = \frac{30}{60} = 0.5$ and $t = 5$, the queue waiting time gives

$$0.2 = \rho e^{-(\mu-\lambda)t}$$

that is,

$$0.2\mu = 0.5 e^{2.5-5\mu}$$

This is a nonlinear equation for μ , which may be solved by standard methods to give $\mu = 0.743$. The mean service time is therefore $1/\mu$ or 1.35 min, and the mean queue length is

$$N_Q = \frac{\rho^2}{1 - \rho} = 1.39, \quad \text{using } \rho = \frac{\lambda}{\mu} = 0.673$$



We can perform these calculations in R. This code illustrates how to solve the equation $f(\mu) = 0.2\mu - 0.5e^{2.5-5\mu} = 0$.

```
lambda <- 30/60; percent <- 80; time_min <- 5
f <- function(mu, lambda, percent, time_min){(1 - percent
    / 100) * mu - lambda * exp(lambda * time_min -
    time_min * mu)
}
# Solve f(mu) = 0; we need to supply an interval of
# possible mu values
solution_info <- uniroot(f, interval = c(0, 1), lambda =
lambda, percent = percent, time_min = time_min)
# Specify the other values
mu <- solution_info$root; mu; 1 / mu
#> [1] 0.7427287
#> [1] 1.346387
rho <- lambda / mu; rho
#> [1] 0.6731933
N_Q <- rho^2 / (1 - rho); N_Q
#> [1] 1.38672
```

Example 11.32

Handling equipment is to be installed at an unloading bay in a factory. An average of 20 trucks arrive during each 10 h working day, and these must be unloaded. The following three schemes are being considered:

Scheme	Fixed cost/ £ per day	Operating cost/ £ per hour	Mean handling rate/ trucks per hour
A	90	45	3
B	190	50	4
C	450	60	6

Truck waiting time is costed at £30 per hour. Assuming an exponential distribution of truck unloading time, find the best scheme.

Solution Viewing this as a queueing problem, we have

$$\lambda = \text{arrival rate per hour} = 2.0$$

$$\mu = \text{unloading rate per hour}$$

$$\text{mean waiting time for each truck} = 1/(\mu - \lambda)$$

Hence the mean delay cost per truck is

$$\frac{30}{\mu - 2}$$

and the mean delay cost per day is

$$\frac{20 \times 30}{\mu - 2} = \frac{600}{\mu - 2}$$

The proportion of time that the equipment is running is equal to the probability that the system is not empty (the **utilization**), which is

$$1 - p_0 = 1 - (1 - \rho) = \rho$$

Hence the mean operating cost per day is 10ρ times operating cost per hour. The total cost per day (in £) is the sum of the fixed, operating and delay costs, as follows:

Scheme	μ	ρ	Fixed	Operating	Delay	Total
A	3	0.6667	90	300	600	990
B	4	0.5	190	250	300	740
C	6	0.3333	450	200	150	800

Hence scheme B minimizes the total cost.

11.9.4 Queues with multiple service channels

For the case where there are c service channels, all with an exponential service time distribution with parameter μ , a line of argument similar to that in Section 11.9.3 can be found in many textbooks on queueing theory. In particular, it can be shown that the distribution p_n of the number of customers in the system is

$$p_n = \begin{cases} \frac{\rho^n}{n!} p_0 & (0 \leq n \leq c) \\ \frac{\rho^n}{c^{n-c} c!} p_0 & (n > c) \end{cases}$$

where $\rho = \lambda/\mu$ and

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{(c-1)!(c-\rho)} \right]^{-1}$$

The mean numbers in the queue and in the system are

$$N_Q = \frac{\rho^{c+1}}{(c-1)!(c-\rho)^2} p_0, \quad N_S = N_Q + \rho$$

and the mean waiting times in the queue and in the system are

$$W_Q = \frac{N_Q}{\lambda}, \quad W_S = W_Q + \frac{1}{\mu}$$

Example 11.33

For the unloading bay problem in Example 11.32 a fourth option would be to install two sets of equipment under scheme A (there is space available to do this). The fixed costs would then double but the operating costs per bay would be the same. Evaluate this possibility.

Solution

With two bays under scheme A, we have $\lambda = 2$, $\mu = 3$ and $c = 2$, so that $\rho = \frac{2}{3}$, and the probability that the system is empty at any time is

$$p_0 = \left(1 + \rho + \frac{\rho^2}{2 - \rho}\right)^{-1} = \frac{1}{2}$$

The probabilities of one truck (one bay occupied) and of two or more trucks (both bays occupied) are then

$$p_1 = \rho p_0 = \frac{1}{3}$$

$$P(\text{two or more trucks}) = 1 - \frac{1}{2} - \frac{1}{3} = \frac{1}{6}$$

The total operating cost per day is the operating cost for when one or other bay is working (£45 per hour) plus that for when both bays are working (£90 per hour), which is

$$10[\frac{1}{3}(45) + \frac{1}{6}(90)] = 300$$

The mean number in the queue is

$$\frac{\rho^3}{(2 - \rho)^2} p_0 = 0.08333$$

so that the mean total time in the system for each truck is

$$\frac{1}{2}(0.08333) + \frac{1}{3} = 0.375$$

Multiplying by the cost per hour and the number of trucks gives the delay cost per day:

$$20(30)(0.375) = 225$$

The total cost per day of this scheme is therefore

$$2(90) + 300 + 225 = £705$$

This is less than the £740 under scheme B, the best of the single-bay options.

O. Jones, R. Maillardet and A. Robinson, *Introduction to Scientific Programming and Simulation Using R* (second edition, Boca Raton, FL, Chapman and Hall/CRC, 2014) provide a detailed treatment of a class of random processes called Markov chains, the future behaviour of which depends only on the present state and not on the past

behaviour before the present. They discuss a range of continuous-time Markov chains, including engineering and queueing examples, using mathematical analysis based on matrices and simulation performed in R. An in-depth mathematical analyses of queues and many related random processes is supplied by G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes* (third edition, Oxford, Clarendon Press, 2001). In their Chapter 11 these authors discuss a notation for queueing systems, due to D. G. Kendall. The notation takes the form $A/B/s$ in which A and B describe the distribution of the inter-arrival and the service times, respectively, and s is the number of service channels. The multiple service channel queue described in this section would be denoted $M/M/c$ or, more fully, $M(\lambda)/M(\mu)/c$. Here M stands for *Markovian* or *memoryless*, the use of this notation being related to the memoryless property of the exponential distribution.

11.9.5 Queueing system simulation

The assumption that the service time distribution is exponential, which underlies the results in Sections 11.9.3 and 11.9.4, is often unrealistic. It is known that it leads to predicted waiting times that tend to be pessimistic, as a result of which costs based on these predictions are often overestimated. Theoretical results for other service distributions exist (see, for example, E. Page, *Queueing Theory in OR*, London, Butterworth, 1972 or the book by Grimmett and Stirzaker mentioned above), but it is often instructive to simulate a queueing system and find the various answers numerically. It is then easy to vary the arrival and service distributions, and the transient (non-steady-state) behaviour of the system also reveals itself.

Annotated R code to simulate a single-channel queueing system is provided in Figure 11.43. We think of the states of the system as being the number of customers in the queue. Each event consists of either an arrival or a departure. The variables `next_arrival` and `next_departure` are used to represent the times until the next arrival and the next departure, and the type of the next event is determined by whichever is smaller. A limit is placed on the number of customers in the system, as often happens in practice. A different simulated queueing system will result every time that the code is run, unless `set.seed` is used, which it is here for reproducibility.



Figure 11.43 R code listing for the queuing system simulation.

```
set.seed(78)
  # Set the seed of the random number generator for
  # reproducibility
# Arrival and departure rates based on Example 11.31
lambda <- 0.5; mu <- 0.743; rho <- lambda / mu; rho
#> [1] 0.6729475
# Maximum number of customers in system
number_max <- 20
# Number of events to be simulated
N <- 500000
# Space to save event times and number of customers (state of
# the system)
```

```

event_times <- rep(NA, N)
number_of_customers <- rep(NA, N)
# System starts with *** no customers *** (an initial arrival must
# be forced)
# number holds current number of customers in system
initial_number <- 0; number <- initial_number
# Define big arrival or departure times to force arrivals or
# departures
big_arrival <- 1000; big_departure <- 1000
# Initialize time
initial_time <- 0; time <- initial_time
# time holds the current time
# Randomly generate the time until the next arrival (system
# starts with no customers)
# Assume distribution is exponential, rate lambda
next_arrival <- rexp(1, rate = lambda)
# Set the time until the next departure to a large number to
# force an arrival next_departure <- big_departure
# Set space to monitor the generated arrival and departure times
arrival_intervals <- rep(NA, N); departure_intervals <- rep(NA, N)
# Loop to generate events
for(i in 1:N){ # Number of events to be simulated
  if(next_arrival < next_departure){
    # The next arrival event occurs before the next
    # departure event
    time <- time + next_arrival
    # Time of the next (arrival) event
    if(number == 0){
      # If system is empty, there is no existing time until
      # next departure
      # and so one needs to be randomly generated
      # Assume distribution is exponential, rate mu
      next_departure <- rexp(1, rate = mu)
      departure_intervals[i] <- next_departure # Monitor it
    } else {
      # Existing time until next departure needs to be reduced
      # after the arrival
      next_departure <- next_departure - next_arrival
    }
    # Number in the system increases by 1 because of the arrival
    number <- number + 1
    if(number == number_max){
      # If system is now at its limit, a departure has to be
      # forced
      next_arrival <- big_arrival
    }
  }
}

```

Figure 11.43 (Continued)

```

} else{
    # A new time until the next arrival is randomly generated
    next_arrival <- rexp(1, rate = lambda)
    arrival_intervals[i] <- next arrival # Monitor it
}
} else {
# The next departure event occurs before the next arrival
# event
time <- time + next_departure
# Time of the next (departure) event
if(number == number_max){
    # If system was at its limit, there is no existing time
    # until the next arrival
    # and so one needs to be randomly generated
    next_arrival <- rexp(1, rate = lambda)
    arrival_intervals[i] <- next_arrival # Monitor it
} else {
    # Existing time until the next arrival needs to be
    # reduced after the departure
    next arrival <- next arrival - next departure
}
# Number in the system decreased by 1 because of the
# departure
number <- number - 1
if(number == 0){
    # If the system is empty, the next event must be forced
    # to be an arrival
    next_departure <- big_departure
} else {
    # Randomly generate a new time until the next departure
    next_departure <- rexp(1, rate = mu)
    departure_intervals[i] <- next departure # Monitor it
}
}
# Save event times and associated number of customers
event_times[i] <- time
number_of_customers[i] <- number
}
# Include initial time (0) and number of customers (0)
event_times <- c(initial_time, event_times)
number_of_customers <- c(initial_number, number_of_customers)
# Record state of system (number of customers) and compute time
# between states
state <- number_of_customers[-(N + 1)]
# Time in final state not known

```

Figure 11.43 (Continued)

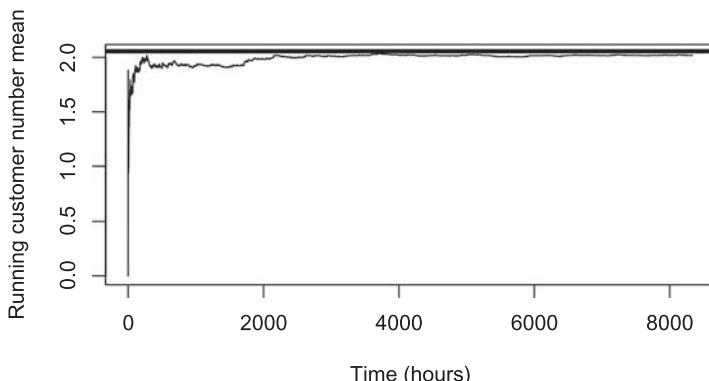
```

time_between_state <- diff(event_times)
# Compute at each time the mean state of (mean number of
# customers in) the system
running_mean_state <- cumsum(time_between_state * state) /
  cumsum(time_between_state)
# Plot this against time
plot(event_times[(-(N + 1)] / 60, # No result for final state,
     time converted to hours
     running_mean_state, type = "l",
     xlab = "Time (hours)", ylab = "Running customer number mean")
# Show the long-term or steady-state limit
abline(h = rho / (1 - rho), lwd = 3)

```

Figure 11.43 (Continued)

Figure 11.44
The running mean of the number of customers in the simulated system against time. The steady-state value $\rho/(1 - \rho)$ is also shown by the horizontal line.



This code can be modified to use other service time distributions by changing `rexp(1, rate = mu)` to generate a realization from the required distribution. An example could be to use the gamma distribution instead of this exponential $\text{Exp}(\text{rate} = \mu)$ distribution. A gamma distribution has two parameters, shape and rate; sometimes scale is used instead of rate where $\text{scale} = 1/\text{rate}$. The exponential distribution is a special case of the gamma distribution with shape parameter set to 1: if $X \sim \text{Gamma}(\text{shape} = 1, \text{rate} = \mu)$, then X follows an $\text{Exp}(\text{rate} = \mu)$ distribution and has mean $E[X] = 1/\mu$ and variance $\text{Var}[X] = 1/\mu^2$. If, for example, $X \sim \text{Gamma}(\text{shape} = 2, \text{rate} = 2\mu)$, then $E[X] = 2/(2\mu) = 1/\mu$ and variance $\text{Var}[X] = 2/(2\mu)^2 = 1/(2\mu^2)$, so that the mean is maintained, but the variance is reduced (since high values occur less often). The R code `rgamma(1, shape = 2, rate = 2 * mu)` would generate a realization from this distribution.

We can use the output from the above code to check some of the theoretical results presented in Section 11.9.3. First, we work with the second half of the simulated values in the hope that these are consistent with the steady-state form of the distribution.

```
# Work with the second half of the simulation
event_times_second_part <- event_times[(N / 2 + 1):(N + 1)]
number_of_customers_second_part <- number_of_customers[(N / 2 + 1):(N + 1)]
length_second_part <- length(event_times_second_part)
# Time between each state
time_between_state_second_part <-
diff(event_times_second_part)
# Omit final value
state_second_part <- number_of_customers_second_part[-length_second_part]
```

Now we can check the steady-state results that the eventual probability of being in state n (n customers) is $p_n = (1 - \rho)\rho^n$, that the mean number in the system is $N_s = \rho/(1 - \rho)$, and that the mean number in the queue is $N_Q = \rho^2/(1 - \rho)$. We can make use of the dplyr package to find the proportion of time spent in each state. You may have to install the dplyr package.

```
require(dplyr) # Load package
# Put state and time data into a data frame
state_time_second_part <- data.frame(state_second_part,
time_between_state_second_part)
# Work out total time in each state
summary_by_state <- state_time_second_part %>%
group_by(state_second_part) %>%
# Separate, results for each state
summarize(total_time =
sum(time_between_state_second_part))
total_time_by_state <- summary_by_state$total_time
# Turn this into a proportion
proportion_time_by_state <- total_time_by_state /
sum(total_time_by_state)
# Compare with theoretical values  $p_n = (1 - \rho)\rho^n$ ,
n = 0, 1, 2
proportion_time_by_state[1:6]
#> [1] 0.32848925 0.22168018 0.14906167 0.09949540
0.06705024 0.04613191
(1 - rho) * rho^(0:5)
#> [1] 0.32705249 0.22008916 0.14810845 0.09966921
0.06707215 0.04513604
#> Mean number in system  $N_s$ 
sum(time_between_state_second_part * state_second_part) /
sum(time_between_state_second_part)
#> [1] 2.015726
# Compare with theoretical value  $\rho / (1 - \rho)$ 
```

```

rho / (1 - rho)
#> [1] 2.057613
#> Mean number in queue N_Q
# If no customer is in the system, there is no customer in
# the queue!
# If there are n > = 1 customers in the system, all but
# one is in the queue
# ifelse(test, yes, no) returns yes if test (which can be
# a vector) is TRUE, and no otherwise
number_in_queue <- ifelse(state_second_part == 0, 0,
                           state_second_part - 1)
sum(time_between_state_second_part * number_in_queue) /
sum(time_between_state_second_part)
#> [1] 1.344216
# Compare with theoretical value rho^2 / (1 - rho)
rho^2 / (1 - rho)
#> [1] 1.384666

```

Checking that the mean total time in the system for each customer is $W_S = 1/(\mu - \lambda)$ and that the mean waiting time in the queue is $W_Q = \rho/(\mu - \lambda)$ is considerably more difficult. We work with the largest possible set of simulated values for which the queue is empty at the beginning and will be straight after the end. If customer i arrives at time t_i^a and departs at time t_i^d , then customer i will be in the system for time $t_i^{in} = t_i^d - t_i^a$. From this it follows that average $t^{in} = \text{average } t^d - \text{average } t^a$. The mean waiting time in the queue $W_Q = W_S - \text{mean service time}$, which can be found from the estimate of W_S and the values of `arrival_intervals` monitored above. Sometimes these approximations can be poor.

```

# Start and finish with an empty queue; identify the
# indicies corresponding to zero
zero_indices <- which(state_second_part == 0)
first_zero <- min(zero_indices); last_zero <-
max(zero_indices)
# Select out the corresponding states and event times,
# omitting the final zero state point
state_0_0 <- state_second_part[first_zero:(last_zero - 1)]
event_times_0_0 <-
event_times_second_part [first_zero:(last_zero - 1)]
# Identify the arrivals (+1) and departures (-1)
# Last state is a departure as queue empty after
arrival_departure <- c(diff(state_0_0), -1)
# Identify times of arrivals and departures
t_a <- event_times_0_0[arrival_departure == 1]
# Arrival times
t_d <- event_times_0_0[arrival_departure == -1]
# Departure times
# Estimate mean time in system w_S
t_bar_in <- mean(t_d) - mean(t_a); t_bar_in

```

```

#> [1] 4.427089
# Compare with theoretical value 1 / (mu - lambda)
1 / (mu - lambda)
#> [1] 4.115226
# Estimate mean service time
mean_service_time <- mean(departure_intervals, na.rm =
TRUE) # Remove NAs
mean_service_time; 1 / mu # Theoretical value
#> [1] 1.343691
#> [1] 1.345895
# Estimate W_Q mean time in queue
t_bar_in - mean_service_time
#> [1] 3.083398
# Compare with theoretical value rho / (mu - lambda)
rho / (mu - lambda)
#> [1] 2.769331

```

The complexity of the above code indicates that writing more general queueing system simulators may be difficult. The R package `simmer` for example may be used to simulate a range of queueing situations; see I. Ucar and B. Smeets, `simmer`; Discrete-Event Simulation for R, R package version 3.6.1, <https://CRAN.R-project.org/package=simmer> (2017). The `Simul8` simulation software <https://www.simul8.com/> is frequently used to simulate queues and many other engineering processes from organization schemes that are drawn using the computer.

11.9.6 Exercises

- 55 A sea area has on average 15 gales annually, evenly distributed throughout the year. Assuming that the gales occur independently, find the probability that more than two gales will occur in any one month.
- 56 Suppose that the average number of telephone calls arriving at a call centre is 30 per hour, and that they arrive independently. What is the probability that no calls will arrive in a three-minute period? What is the probability that more than five calls will arrive in a five-minute period?
- 57 Show that for a single-channel queue with Poisson arrivals and exponential service time distribution the mean numbers of customers in the system and in the queue are
- $$N_S = \frac{\rho}{1-\rho}, \quad N_Q = \frac{\rho^2}{1-\rho}$$
- where ρ is the ratio of arrival and service rates.
(Hint: Differentiate the equation
- $$\sum_{n=0}^{\infty} \rho^n = \frac{1}{1-\rho}$$
- with respect to ρ)
- 58 Patients arrive at the casualty department of a hospital at random, with a mean arrival rate of three per hour. The department is served by one doctor, who spends on average 15 minutes with each patient, actual consulting times being exponentially distributed. Find
- the proportion of time that the doctor is idle;
 - the mean number of patients waiting to see the doctor;
 - the probability of there being more than three patients waiting;

- (d) the mean waiting time for patients;
 (e) the probability of a patient having to wait longer than one hour.

59

A small company operates a cleaning and re-catering service for passenger aircraft at an international airport. Aircraft arrive requiring this service at a mean rate of λ per hour, and arrive independently of each other. They are serviced one at a time, with an exponential distribution of service time. The cost for each aircraft on the ground is put at c_1 per hour, and the cost of servicing the planes at a rate μ is $c_2\mu$ per hour. Prove that the service rate that minimizes the total cost per hour is

$$\mu = \lambda + \sqrt{\left(\frac{c_1\lambda}{c_2}\right)}$$

60

The machines in a factory break down in a Poisson pattern at an average rate of three per hour during the eight-hour working day. The company has two service options, each involving an exponential service time distribution. Option A would cost £20 per hour, and the mean repair time would

61

be 15 min. Option B would cost £40 per hour, with a mean repair time of 12 min. If machine idle time is costed at £60 per hour, which option should be adopted?

Ships arrive independently at a port at a mean rate of one every three hours. The time a ship occupies a berth for unloading and loading has an exponential distribution with a mean of 12 hours. If the mean delay to ships waiting for berths is to be kept below six hours, how many berths should there be at the port?

62

In a self-service store the arrival process is Poisson, with on average one customer arriving every 30 s. A single cashier can serve customers every 48 s on average, with an exponential distribution of service time. The store managers wish to minimize the mean waiting time for customers. To do this, they can either double the service rate by providing an additional server to pack the customer's goods (at a single cash desk) or else provide a second cash desk. Which option is preferable?

11.10 Bayes' theorem and its applications

To end this chapter, we return to the foundations of probability and inference. The definition of conditional probability is fundamental to the subject, and from it there follows the theorem of Bayes, which has far-reaching implications. This Bayesian approach to statistical inference has become very popular. It will be discussed in Section 11.10.3.

11.10.1 Derivation and simple examples

The definition in Section 11.2.1 of the conditional probability $P(B|A)$ of an event B given that another event A occurs can be written as

$$P(A \cap B) = P(B|A)P(A)$$

If A and B are interchanged then this becomes

$$P(A \cap B) = P(A|B)P(B)$$

The left-hand sides are equal, so we can equate the right-hand sides and rearrange, giving

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Now suppose that B is known to have occurred, and that this can only happen if one of the mutually exclusive events

$$\{A_1, \dots, A_n\}, \quad A_i \cap A_j = \emptyset \quad (i \neq j)$$

has also occurred, but which one is not known. The relevance of the various events A_i to the occurrence of B is expressed by the conditional probabilities $P(B|A_i)$. Suppose that the probabilities $P(A_i)$ are also known. The examples below will show that this is a common situation, and we should like to work out the conditional probabilities

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

To find the denominator, we sum from 1 to n :

$$\sum_{i=1}^n P(A_i|B) = 1 = \frac{1}{P(B)} \sum_{i=1}^n P(B|A_i)P(A_i)$$

The sum is equal to 1 by virtue of the assumption that B could not have occurred without one of the A_i occurring. We therefore obtain a formula for $P(B)$:

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

which is sometimes called the **rule of total probability**. Hence we have the following theorem.

Theorem 11.1

Bayes' theorem

If $\{A_1, \dots, A_n\}$ are mutually exclusive events, one of which must occur given that another event B occurs, then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)} \quad (i = 1, \dots, n)$$

[end of theorem](#)

Effectively, Bayes' theorem allows us to reverse the conditioning: from $P(B|A_i)$ (and other probabilities) we may find $P(A_i|B)$. In some applications it is only necessary to know $P(A_i|B)$ up to a multiplicative constant. This means that a common statement of Bayes' theorem is

$$P(A_i|B) \propto P(B|A_i)P(A_i)$$

in which \propto means proportional to: if $f(x) \propto g(x)$, then $f(x) = cg(x)$ for a constant c . For the expression $P(A_i|B) \propto P(B|A_i)P(A_i)$, c would equal $1/P(B)$. Often in Bayesian statistical inference, which we will meet in Section 11.10.3, we will only know probabilities (or probability density functions) up to a multiplicative constant because the so-called 'normalization constant' $P(B)$ cannot be calculated.

Example 11.34

Three machines produce similar car parts. Machine A produces 40% of the total output, and machines B and C produce 25% and 35% respectively. The proportions of the output from each machine that do not conform to the specification are 10% for A, 5%

for B and 1% for C. What proportion of those parts that do not conform to the specification are produced by machine A?

Solution Let D represent the event that a particular part is defective. Then, by the rule of total probability, the overall proportion of defective parts is

$$\begin{aligned} P(D) &= P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) \\ &= (0.1)(0.4) + (0.05)(0.25) + (0.01)(0.35) = 0.056 \end{aligned}$$

Using Bayes' theorem,

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{(0.1)(0.4)}{0.056} = 0.714$$

so that machine A produces 71.4% of the defective parts.

Example 11.35

Suppose that 0.1% of the people in a certain area have a disease D and that a mass screening test is used to detect cases. The test gives either a positive or a negative result for each person. Ideally, the test would always give a positive result for a person who has D, and would never do so for a person who has not. In practice the test gives a positive result with probability 99.9% for a person who has D, and with probability 0.2% for a person who has not. What is the probability that a person for whom the test is positive actually has the disease?

Solution Let T represent the event that the test gives a positive result. Then the proportion of positives is

$$\begin{aligned} P(T) &= P(T|D)P(D) + P(T|D^c)P(D^c) \\ &= (0.999)(0.001) + (0.002)(0.999) \approx 0.003 \end{aligned}$$

and the desired result is

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)} = \frac{(0.999)(0.001)}{0.003} = \frac{1}{3}$$

Despite the high basic reliability of the test, only one-third of those people receiving a positive result actually have the disease. This is because of the low incidence of the disease in the population, which means that a positive result is twice as likely to be a false alarm as it is to be correct.

In connection with Example 11.35, it might be wondered why the reliability of the test was quoted in the problem in terms of

$$P(\text{positive result} | \text{disease}) \quad \text{and} \quad P(\text{positive result} | \text{no disease})$$

instead of the seemingly more useful

$$P(\text{disease} | \text{positive result}) \quad \text{and} \quad P(\text{disease} | \text{negative result})$$

The reason is that the latter figures are contaminated, in a sense, by the incidence of the disease in the population. The figures quoted for reliability are intrinsic to the test can be found experimentally, and may be used anywhere the disease occurs, regardless of the level of incidence.

11.10.2 Applications in probabilistic inference

The scope for applications of Bayes' theorem can be widened considerably if we assume that the calculus of probability can be applied not just to events as subsets of a sample space but also to more general statements about the world. Events are essentially statements about facts that may be true on some occasions and false on others. Scientific theories and hypotheses are much deeper statements, which have great explanatory and predictive power, and which are not so much true or false as gaining or lacking in evidence. One way to assess the extent to which some evidence E supports a hypothesis H is in terms of the conditional probability $P(H|E)$. The relative frequency interpretation of probability does not normally apply in this situation, so a subjective interpretation is adopted. The quantity $P(H|E)$ is regarded as a **degree of belief** in hypothesis H on the basis of evidence E . In an attempt to render the theory as objective as possible, the rules of probability are strictly applied, and an inference mechanism based on Bayes' theorem is employed.

Suppose that there are in fact two competing hypotheses H_1 and H_2 . Let X represent all background information and evidence relevant to the two hypotheses. The probabilities $P(H_1|X)$ and $P(H_1|X \cap E)$ are called the **prior** and **posterior probabilities** of H_1 , where E is a new piece of evidence. The probabilities $P(H_1|X)$ and $P(H_1|X \cap E)$ represent our degrees of belief about the hypothesis H_1 before and after seeing the new evidence E . Similarly, there are prior and posterior probabilities of H_2 . Applying Bayes' theorem to both H_1 and H_2 and cancelling the common denominator $P(E)$ gives

$$\frac{P(H_1|X \cap E)}{P(H_2|X \cap E)} = \frac{P(E|H_1 \cap X)}{P(E|H_2 \cap X)} \frac{P(H_1|X)}{P(H_2|X)}$$

The left-hand side and the second factor on the right-hand side are called the **posterior odds** and **prior odds** respectively, favouring H_1 over H_2 . The first factor on the right-hand side is called the **likelihood ratio**, and it measures how much more likely it is that the evidence event E would occur if the hypothesis H_1 were true than if H_2 were true. The new evidence E therefore 'updates' the prior odds, and the process can be repeated as often as desired whenever new evidence become available, provided that the likelihood ratios can be calculated.

Example 11.36

From experience it is known that when a particular type of computer fails, this is twice as likely to be caused by a short on the serial interface (H_1) as by a faulty memory circuit (H_2). The standard diagnostic test is to measure the voltage at a certain point on the board, and from experience it is also known that a drop in voltage there occurs nine times out of ten when the memory circuit is faulty but only once in six occasions of an interface short. How does the observed drop in voltage (E) affect the assessment of the cause of failure?

Solution

Here we do not need to be concerned about background information. The prior odds are two to one in favour of H_1 , and the likelihood ratio is $(1/6)/(9/10)$, so the posterior odds are given by

$$\frac{P(H_1|E)}{P(H_2|E)} = \left(\frac{10}{54}\right)(2) = 0.370$$

The evidence turns the odds around to about 2.7 to one in favour of H_2 , since $1 / 0.370 \approx 2.7$

Example 11.37

An oil company is prospecting for oil in a certain area, and is conducting a series of seismic experiments. It is known from past experience that if oil is present in the rock strata below, then there is on average one chance in three that a characteristic pattern will appear on the trace recorded by the seismic detector after a test. If oil is absent then the pattern can still appear, but is less likely, appearing only once in four tests on average. After 150 tests in the area the pattern has been seen on 48 occasions. Assuming prior odds of 3:1 against the presence of oil, find the updated odds. Also find the 90% confidence interval for the true probability of the pattern appearing after a test, and hence consider whether oil is present or not.

Solution

Let H_1 and H_2 represent the hypotheses that oil is present and that it is absent respectively. There were effectively 150 pieces of evidence gathered, and the odds need to be multiplied by the likelihood ratio for each. Each time the pattern is present the likelihood ratio is

$$\frac{P(\text{pattern}|H_1)}{P(\text{pattern}|H_2)} = \frac{1}{3}/\frac{1}{4} = \frac{4}{3}$$

and each time it is absent the likelihood ratio is

$$\frac{P(\text{no pattern}|H_1)}{P(\text{no pattern}|H_2)} = \frac{2}{3}/\frac{3}{4} = \frac{8}{9}$$

The updated odds, letting E represent the total evidence, become

$$\frac{P(H_1|E)}{P(H_2|E)} = \left(\frac{4}{3}\right)^{48} \left(\frac{8}{9}\right)^{102} \left(\frac{1}{3}\right) = 2.01$$

The odds that there is oil present are therefore raised to 2:1 in favour.

Confidence intervals for proportions were covered in Section 11.3.6. The proportion of tests for which the pattern was observed is 48/150 or 0.32, so the 90% confidence interval for the probability of appearance is

$$0.32 \pm 1.645 \sqrt{\left[\frac{(0.32)(0.68)}{150}\right]} = (0.26, 0.38)$$

The hypothesis that oil is absent is not compatible with this, because the pattern should then appear with probability 0.25, whereas the hypothesis that oil is present is fully compatible.

For the problem in Example 11.36 it is conceivable that there could be enough repetitions for the relative frequency interpretation to be placed on the probabilities of the two hypotheses. In contrast, in Example 11.37 the probability of the presence or absence of oil is not well suited to a frequency interpretation, but the subjective interpretation is available.

Example 11.37 also provides a contrast between the ‘Bayesian’ and ‘classical’ inference approaches. The classical confidence interval appears to lead to a definite result: H_1 is true and H_2 is false. This definiteness is misleading, because it is possible (although not likely) that the opposite is the case, but the evidence supports one hypothesis more

than the other. The Bayesian approach has the merit of indicating this relative support quantitatively.

In Section 11.10.3 we discuss briefly the Bayesian approach to statistical inference. This approach now enjoys considerable popularity. For a recent, excellent discussion of the Bayesian and frequentist inference approaches see B. Efron and T. Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* (Cambridge, Cambridge University Press, 2016).

One area where Bayesian inference is very important is in decision support and expert systems. In classical decision theory Bayesian inference is used to update the probabilities of various possible outcomes of a decision, as further information becomes available. This allows an entire programme of decisions and their consequences to be planned (see D. V. Lindley, *Making Decisions*, second edition, London, Wiley, 1985). Expert systems often involve a process of reasoning from evidence to hypothesis with a Bayesian treatment of uncertainty (see, for example, R. Forsyth, ed., *Expert Systems, Principles and Case Studies*, London, Chapman & Hall, 1984, and R. G. Cowell, P. Dawid, S. L. Lauritzen and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, Berlin, Springer, 1999). See, also M. Seutari and J.-B. Denis, *Bayesian Networks: With Examples in R* (Boca Raton, FL, Chapman and Hall/CRC, 2015) for a modern R-based treatment of Bayesian networks.

11.10.3 Bayesian statistical inference

The Bayesian approach to statistical inference has become mainstream due to the increasing availability of computational power. It is discussed in many books; see, for example, I. Ntzoufras, *Bayesian Modeling Using WinBUGS* (Hoboken, NJ, Wiley, 2009) and D. Lunn, C. Jackson, N. Best, A. Thomas and D. Spiegelhalter, *The BUGS Book: A Practical Introduction to Bayesian Analysis* (Boca Raton, FL, CRC Press, 2013). Our aim here is to provide a flavour of what can be done and an illustration of the flexibility of the Bayesian approach. We will revisit the analysis of engine performance data discussed in Section 11.7. In Section 11.7.3 we worked with two regression models, which we will state again here using somewhat different notation more suited to the present purpose. For the data from engine A, we adopted the model

$$\text{running time}_i = c_1 + d_1 \text{ ambient temperature}_i + \text{error}_i \quad (i = 1, \dots, 30)$$

while for engine B, the model was

$$\text{running time}_i = c_2 + d_2 \text{ ambient temperature}_i + \text{error}_i \quad (i = 31, \dots, 60)$$

Here, we use 1 for engine A and 2 for engine B because R and related code indexes objects using numbers rather than characters. We also assumed that the errors are independent and follow a $N(0, \sigma^2)$ distribution. In Bayesian statistical inference we tend to work with precision τ instead of variance σ^2 , where $\tau = 1/\sigma^2$. This does not matter as it is just a reparameterization of the model. So, we can now write our data model as

$$\begin{aligned} \text{running time}_i &\sim N(c_1 + d_1 \text{ ambient temperature}_i, \text{precision} = \tau) \text{ for engine A} \\ \text{running time}_i &\sim N(c_2 + d_2 \text{ ambient temperature}_i, \text{precision} = \tau) \text{ for engine B} \end{aligned}$$

in which \sim means ‘is distributed as’. We wish to learn or make inference about the unknown parameters c_1, d_1, c_2, d_2 and τ from the data. We will see that conclusions that we make about τ can be transformed into conclusions about $\sigma^2 = 1/\tau$ and $\sigma = 1/\sqrt{\tau}$.

In the Bayesian approach to statistical inference, the parameters are considered to be random variables and inference is based on the distribution of the parameters given the observed data, which is called the posterior distribution (as it is defined *after* observing the data). This is similar to the approach taken in Section 11.10.2, where we assigned probabilities to hypotheses. The posterior probability density function of the parameters c_1, d_1, c_2, d_2 and τ given the data is often written as $\pi(c_1, d_1, c_2, d_2, \tau | \text{data})$. The posterior probability density function can be found mathematically using Bayes' theorem as

$$\pi(c_1, d_1, c_2, d_2, \tau | \text{data}) \propto \pi(\text{data} | c_1, d_1, c_2, d_2, \tau) \pi(c_1, d_1, c_2, d_2, \tau)$$

in which the ‘likelihood’ $\pi(\text{data} | c_1, d_1, c_2, d_2, \tau)$ can be found from the above data model using independence by multiplying all the normal probability density functions for running times. We also have to specify the prior probability density function $\pi(c_1, d_1, c_2, d_2, \tau)$ which expresses our degree of belief about the parameters before seeing the data. It is common practice to assume that the parameters are independent before seeing the data so that $\pi(c_1, d_1, c_2, d_2, \tau) = \pi(c_1) \pi(d_1) \pi(c_2) \pi(d_2) \pi(\tau)$. Often, our prior beliefs are very vague and so we choose $\pi(c_1), \pi(d_1), \pi(c_2), \pi(d_2)$ and $\pi(\tau)$ to reflect this. For example, $\pi(c_1)$ may be chosen to be a normal probability density function with very low precision, that is very high variance.

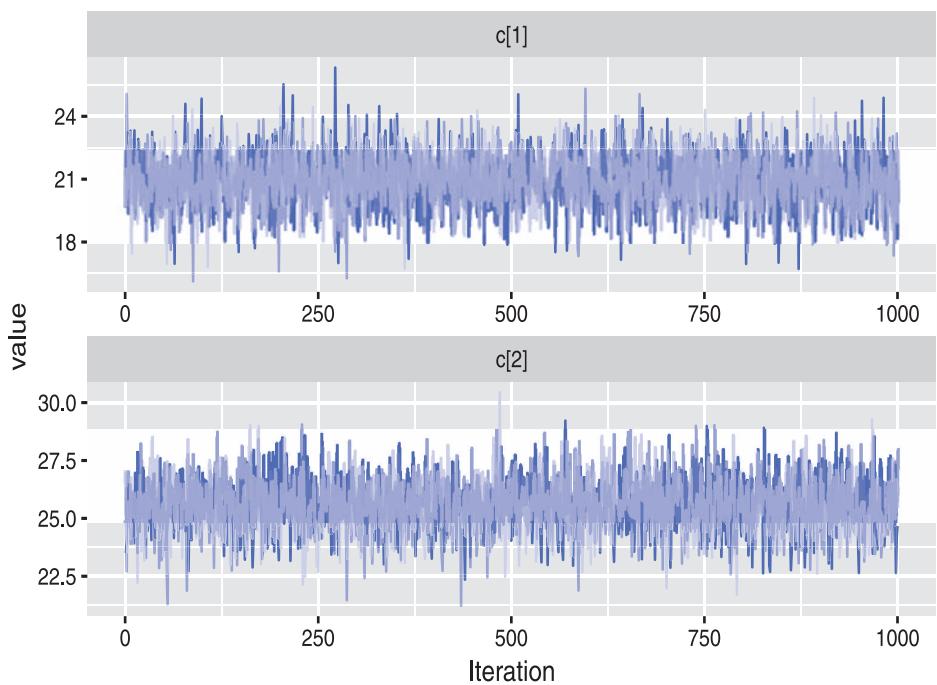
Usually, the posterior distribution is too complicated to be handled mathematically. Here it is five-dimensional and so a mathematical expression for the normalization constant $\pi(\text{data})$ cannot be found. It is, therefore, common practice to understand and summarize the posterior distribution by randomly sampling values from it. Markov chain Monte Carlo algorithms such as the Gibbs Sampler have been developed to perform this sampling and these iterative algorithms can be used even though $\pi(c_1, d_1, c_2, d_2, \tau | \text{data})$ is only known up to a multiplicative constant. They produce sets of values

$$\begin{aligned} & (c_1^{(0)}, d_1^{(0)}, c_2^{(0)}, d_2^{(0)}, \tau^{(0)}) \rightarrow \\ & (c_1^{(1)}, d_1^{(1)}, c_2^{(1)}, d_2^{(1)}, \tau^{(1)}) \rightarrow \dots \rightarrow (c_1^{(j-1)}, d_1^{(j-1)}, c_2^{(j-1)}, d_2^{(j-1)}, \tau^{(j-1)}) \rightarrow \\ & (c_1^{(j)}, d_1^{(j)}, c_2^{(j)}, d_2^{(j)}, \tau^{(j)}) \rightarrow \dots \end{aligned}$$

which are eventually distributed according to the posterior distribution $\pi(c_1, d_1, c_2, d_2, \tau | \text{data})$. $(c_1^{(0)}, d_1^{(0)}, c_2^{(0)}, d_2^{(0)}, \tau^{(0)})$ is called the initial set of values. These sets of values are realizations of a Markov chain because $(c_1^{(j)}, d_1^{(j)}, c_2^{(j)}, d_2^{(j)}, \tau^{(j)})$ is produced using only knowledge of $(c_1^{(j-1)}, d_1^{(j-1)}, c_2^{(j-1)}, d_2^{(j-1)}, \tau^{(j-1)})$. It is common practice to run a sampling algorithm several times to assess whether the memory of the initial set of values has been lost. This is illustrated in Figure 11.45, in which the sequences $c_1^{(j)}$ and $c_2^{(j)}$ are plotted against iteration number $j = 1, \dots, 1000$. In each panel, there are three traceplots, each of which comes from a different initial set of values. These traceplots seem indistinguishable even for small iteration numbers, suggesting that, there is no effective influence due to the initial set of Markov chain values. In practice, far more than 1000 iterations would be used, for example 100 000 iterations, with the samples being produced by the first 50 000 (say) iterations being discarded so that the initial value has no influence. The discarded values are said to come from the ‘burn-in’ phase. Values after burn-in are then considered to be distributed according to the posterior distribution. Because Markov chain Monte Carlo algorithms are iterative, they produce a correlated or dependent sequence of values, since $(c_1^{(j)}, d_1^{(j)}, c_2^{(j)}, d_2^{(j)}, \tau^{(j)})$ depends on $(c_1^{(j-1)}, d_1^{(j-1)}, c_2^{(j-1)}, d_2^{(j-1)}, \tau^{(j-1)})$, which depends on $(c_1^{(j-2)}, d_1^{(j-2)}, c_2^{(j-2)}, d_2^{(j-2)}, \tau^{(j-2)})$ etc. To reduce this dependence,

Figure 11.45

Traceplots of simulated values of c_1 (labelled $c[1]$) and c_2 (labelled $c[2]$). The colours correspond to the three sets of initial Markov chain values.



'thinning' is often applied to the output. Thinning may involve taking every 50th sampled value, for example. Although thinning reduces computational costs, there is no theoretical requirement to use it.

Several programs are available for implementing Gibbs Sampler-type algorithms including `jags` (Just Another Gibbs Sampler), with which we will work. You will have to install the `jags` computational engine from <http://meme-jags.sourceforge.net/>. The R package `R2jags`, which you also have to install, allows `jags` to be run from R. We will specify our model in the BUGS (Bayesian inference Using Gibbs Sampling) language; see D. Lunn, C. Jackson, N. Best, A. Thomas and D. Spiegelhalter, *The BUGS Book: A Practical Introduction to Bayesian Analysis* (Boca Raton, FL, Chapman and Hall/CRC, 2013) for a very detailed treatment.

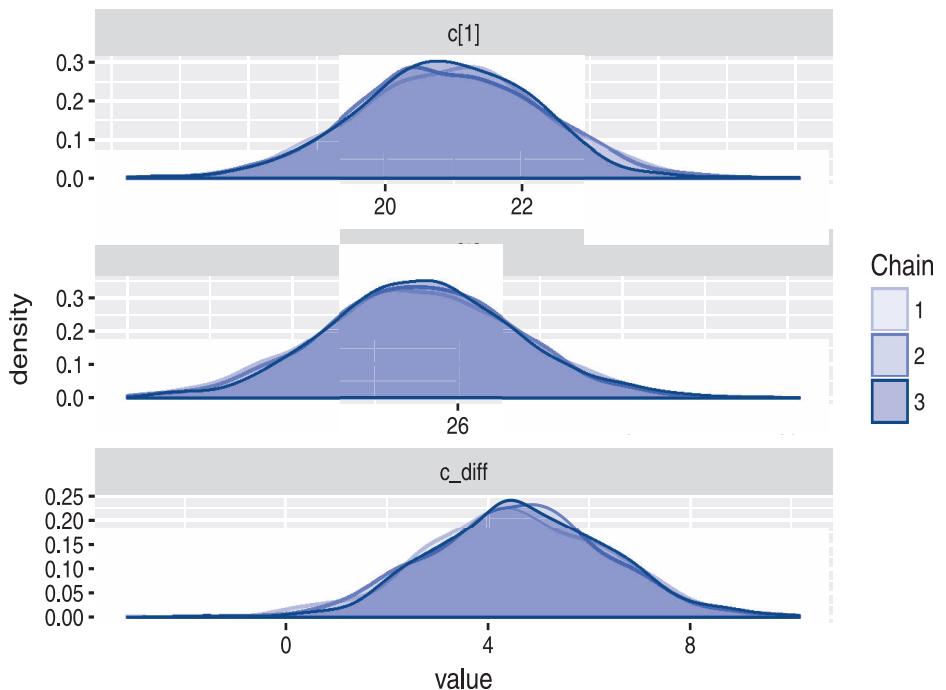
Mathematically, the marginal posterior probability density functions $\pi(c_1 | \text{data})$, $\pi(d_1 | \text{data})$, $\pi(c_2 | \text{data})$, $\pi(d_2 | \text{data})$ and $\pi(\tau | \text{data})$ can be derived from the joint posterior probability density function $\pi(c_1, d_1, c_2, d_2, \tau | \text{data})$ by multi-dimensional integration. These marginal posterior distributions tell us about the individual parameters on their own, while the joint posterior distribution tells us about all the parameters together. In practice, the individual elements of the sampled values $(c_1^{(j)}, d_1^{(j)}, c_2^{(j)}, d_2^{(j)}, \tau^{(j)})$ from the joint posterior distribution provide us with samples from the marginal posterior distributions. For example, the sampled c_1 values, some of which are shown in the top panel of Figure 11.45, would eventually be distributed according to the marginal posterior distribution $\pi(c_1 | \text{data})$. Moreover, a sample from the marginal posterior distribution $\pi(\sigma | \text{data})$ can be easily obtained by transforming the sampled τ values using $\sigma = 1/\sqrt{\tau}$.

In the Bayesian framework we work with credible intervals rather than confidence intervals. A 95% credible interval $(c_{1,b}, c_{1,u})$ for c_1 would have the property that

$$P(c_{1,l} \leq c_1 \leq c_{1,u} | \text{data}) = 0.95$$

Figure 11.46

Approximations of $\pi(c_1 \mid \text{data})$ (labelled $c[1]$), $\pi(c_2 \mid \text{data})$ (labelled $c[2]$) and $\pi(c_2 - c_1 \mid \text{data})$ (labelled c_diff). The colours correspond to the three sets of initial Markov chain values.



so that c_1 lies in the interval $(c_{1,l}, c_{1,u})$ with high probability. Many people find the definition of a credible interval to be more natural than that of a confidence interval. Such a credible interval can be approximated from the sample of c_1 values.



We now explain code that performs Bayesian inference about the parameters of the above model. Running this code provides us with 95% credible intervals for the model parameters and other quantities of interest such as $c_2 - c_1$ (difference in intercepts), $d_2 - d_1$ (difference in slopes) and $(c_2 + d_2t) - (c_1 + d_1t)$ (difference in regression line values) for $t = 10^\circ\text{C}$ and $t = 20^\circ\text{C}$. In addition, we can obtain a 95% credible interval for the point at which the two lines intersect $t_{\text{intersect}} = -(c_2 - c_1)/(d_2 - d_1)$.

In Section 11.7 we saved the data in R objects called `run_time`, `amb_temp` and `engine`. We need to recode `engine` using 1 for engine A and 2 for engine B.

```
engine_12 <- ifelse(engine == "engine A", 1, 2)
# If the condition is TRUE, return 1, else return 2
engine_12
#> [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
#> [26] 1 1 1 1 1 2 2 2 2 2
#> [36] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

Now we write our model using the BUGS language:

```
engine_model <- function() {
  # Define the data model (and hence the likelihood) as
  # stated above
  for(i in 1:n) {
    run_time[i] ~ dnorm(mu[i], tau)
```

```

# Parametrized by the precision tau = 1 / sigma^2
mu[i] <- c[engine_12[i]] + d[engine_12[i]] *
amb_temp[i]

# Different intercept and slope for each of the two
engine
}

# Prior (i.e. before seeing the data) beliefs about the
# unknown parameter
c[1] ~ dnorm(0.0, 1.0E-4)
# Low precision, so vague belief
d[1] ~ dnorm(0.0, 1.0E-4)
c[2] ~ dnorm(0.0, 1.0E-4)
d[2] ~ dnorm(0.0, 1.0E-4)
tau ~ dgamma(1. 0E-3, 1.0E-3)
# We allow tau to take a large range of possible values

#
sigma <- 1.0 / sqrt(tau)
# Definition of sigma, a transformation of tau

# Other quantities to monitor
c_diff <- c[2] - c[1] # Difference in intercepts
d_diff <- d[2] - d[1] # Difference in slopes
run_diff_1 <- (c[2] + d[2] * t_new_1) - (c[1] + d[1] *
t_new_1)
# Difference in regression line values at t_new_1
run_diff_2 <- (c[2] + d[2] * t_new_2) - (c[1] + d[1] *
t_new_2)
# Difference in regression line values at t_new_2
t_intersect <- -(c[2] - c[1]) / (d[2] - d[1])
# Point at which the regression lines intersect
# There would be a problem if d[2] - d[1] were zero!
# This could be overcome using an ifelse construction
# to ensure that the divisor is never actually zero
}

```

We need to specify the data that this BUGS code uses:

```

n <- length(run_time)
t_new_1 <- 10
# Value of t at which to evaluate the difference in
# regression line values
t_new_2 <- 20

```

```

# All the required data
engine_data <- list("run_time", "amb_temp", "engine_12",
                     "n", "t_new_1", "t_new_2")

```

Now we use the `jags` function to randomly sample from the posterior $\pi(c_1, d_1, c_2, d_2, \tau | \text{data})$ in the way described above. You should read the `jags` function help file carefully. Here is the code. We use `set.seed` for reproducibility. Three sets of initial Markov chain values ($c_1^{(0)} d_1^{(0)} c_2^{(0)} d_2^{(0)} \tau_i^{(0)}$) are used.

```

require(R2jags)
set.seed(14)
# Set the seed of the random number generator for
# reproducibility
# Specify arbitrary initial points for the unknown
# parameters c, d and tau
# Here we specify three sets, each provided in a list
# d[2] - d[1] must not be zero to prevent division by zero
initial_points <- list(list(c = c(15, 20), d = c(0.01,
-0.1), tau = 0.1), list(c = c(25, 30), d = c(0.01, -0.1),
tau = 0.2), list(c = c(25, 30), d = c(0.4, 0.2), tau =
0.6))

# Obtain samples from the posterior distribution
engine_posterior <- jags(data = engine_data,
                           # Specify the initial values
                           inits = initial_points,
                           # Parameters of interest to be monitored
                           parameters.to.save = c("c", "d", "sigma", "c_diff",
"d_diff", "run_diff_1", "run_diff_2", "t_intersect"),
                           # Number of samples (some are not used)
                           n.iter = 100000,
                           # Repeat sampling algorithm three times from the three
                           # initial points
                           n.chains = 3,
                           # Function containing the BUGS code
                           model.file = engine_model)

```

We can display the results, part of which we will discuss below:

```

print(engine_posterior, intervals = c(0.025, 0.5, 0.975))
#> fit using jags,
#> 3 chains, each with 1e+05 iterations (first 50000
#> discarded), n.thin = 50
#> n.sims = 3000 iterations saved

```

```

#>          mu.vect  sd.vect   2.5%    50%   97.5% Rhat n.eff
#> c[1]      20.932   1.351  18.238  20.932  23.532 1.0013000
#> c[2]      25.540   1.176  23.274  25.537  27.917 1.0013000
#> c_diff     4.608   1.771   1.086   4.584   8.025 1.0021800
#> d[1]       0.196   0.079   0.043   0.195   0.351 1.0013000
#> d[2]      -0.010   0.063  -0.135  -0.010   0.113 1.0013000
#> d_diff    -0.206   0.100  -0.399  -0.207  -0.013 1.0012000
#> run_diff_1 2.549   0.843   0.918   2.530   4.156 1.0021900
#> run_diff_2  0.490   0.522  -0.522   0.484   1.561 1.0013000
#> sigma      1.693   0.164   1.415   1.681   2.055 1.0021000
#> t_intersect 32.931 518.858 17.015 22.188 49.711 1.2913000
#> deviance   233.148   3.293 228.829 232.439 241.233 1.0031400
#>
#> For each parameter, n.eff is a crude measure of
#> effective sample size,
#> and Rhat is the potential scale reduction factor (at
#> convergence, Rhat=1).
#>
#> DIC info (using the rule, pD = var(deviance)/2)
#> pD = 5.4 and DIC = 238.6
#> DIC is an estimate of expected predictive error (lower
#> deviance is better).

```

We can obtain a 95% credible interval for each parameter by looking at the values in the 2.5% and 97.5% columns. These can be extracted for future use, for example:

```

engine_posterior$BUGSoutput$summary[c("c[1]", "c[2]",
                                         c("2.5%", "97.5%")]

#>          2.5%    97.5%
#> c[1] 18.23815 23.53192
#> c[2] 23.27416 27.91698

```

The mean of the marginal posterior distribution for each parameter (available from the column `mu.vect`) and the median (available from the column `50%`) provide one number posterior summaries. These results will be slightly different each time the code is run as they are based on randomly sampled values from the posterior distribution.

The results here are in line with those that we saw in Section 11.7. We are also able to perform inference about $t_{\text{intersect}}$, which we did not do before, although some extreme values may be generated when the divisor $d_2 - d_1$ is close to zero. This illustrates the flexibility of the simulation-based Bayesian approach to statistical inference.

A range of R packages including `ggmcmc` allow us to provide graphical visualizations of `jags` output. For example, traceplots, such as the one shown in Figure 11.46, can be produced using the following procedure:

First, the `jags` output has to be turned into a so-called ‘`mcmc`’ object:

```
engine_posterior.mcmc <- as.mcmc(engine_posterior)
```

The sampled values can be extracted from an ‘`mcmc`’ object, if required:

```
head(engine_posterior.mcmc[[1]] [, "c[1]"])
#> Markov Chain Monte Carlo (MCMC) output:
#> Start = 50001
#> End = 50301
#> Thinning interval = 50
#> [1] 21.79822 19.94469 19.21991 20.71384 21.12518
#> 20.21288 23.18128
```

Next, this ‘`mcmc`’ object has to be turned into a so-called ‘`ggs`’ object. You will probably need to install the `ggmcmc` package:

```
require(ggmcmc)
engine_posterior.ggs <- ggs(engine_posterior.mcmc)
```

Finally, traceplots (not shown because of space considerations) can be produced. These can be used to assess Markov chain convergence.

```
ggs_traceplot(engine_posterior.ggs) # A lot of plots
ggs_traceplot(engine_posterior.ggs, family = "c")
# Parameters starting with c
ggs_density(engine_posterior.ggs, family = "c")
# Parameters starting with c
```

Density plots provide approximations of marginal posterior density functions. The following code produces Figure 11.46 which shows approximations of $\pi(c_1 \mid \text{data})$, $\pi(c_2 \mid \text{data})$ and $\pi(c_2 - c_1 \mid \text{data})$.

These posterior probability density functions express our beliefs about c_1 , c_2 and $c_2 - c_1$ after seeing the data. The colours correspond to the three sets of initial Markov chain values provided to `jags`. The similarity of density functions suggests that the memory of these initial values has been lost and that the algorithm is indeed producing samples from the posterior.

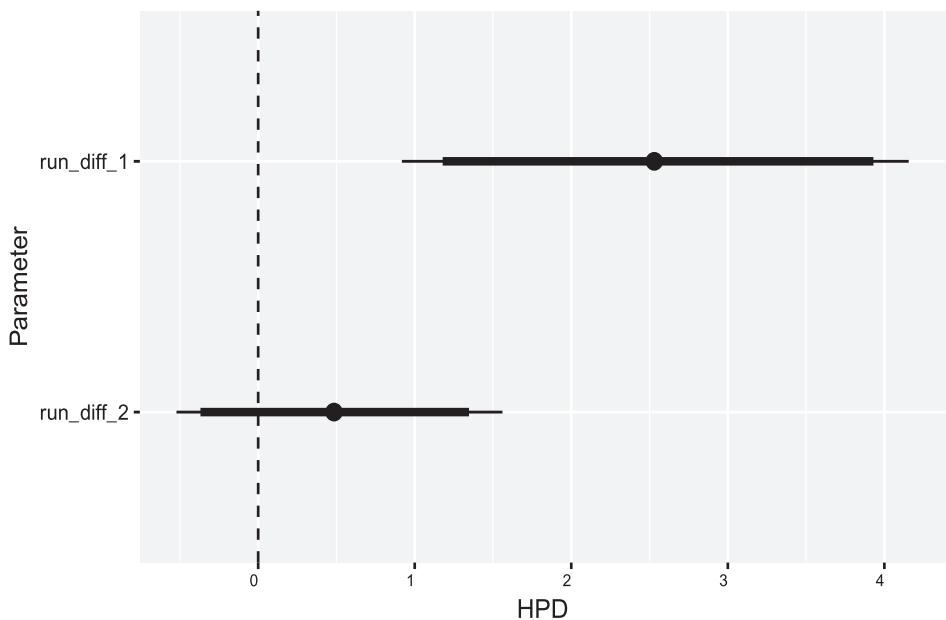
Caterpillar plots show posterior medians, and 90% and 95% narrowest credible intervals. Figure 11.47 presents caterpillar plots for $(c_2 + d_2t) - (c_1 + d_1t)$ (difference in regression line values) at $t = 10^\circ\text{C}$ and $t = 20^\circ\text{C}$. As before, the first interval is entirely positive, while the second provides posterior support for zero. Hence, we may conclude that the underlying difference in running times between engine B and engine A is probably positive for a low ambient temperature, such as 10°C , while there is probably no effective difference when $t = 20^\circ\text{C}$.

```
ggs_caterpillar(engine_posterior.ggs, family = "r") +
geom_vline(xintercept = 0, lty = "dashed")
# Dashed vertical line at zero
```

In addition to the references already mentioned in this section, A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, *Bayesian Data Analysis* (second edition, Boca Raton, FL, Chapman and Hall/CRC, 2013), R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (Boca Raton, FL, Chapman and Hall/CRC, 2016), R. Christensen, W. Johnson, A. Branseum and T. E. Hanson, *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians* (Boca Raton, FL, Chapman and Hall/CRC, 2011), and many others, provide excellent treatments. The `MCMCpack` package contains easy to use functions to perform Bayesian inference using posterior simulation for a number of statistical models. Finally, techniques for determining the number of iterations and the burn-in length to use, and for assessing whether the sampled values really do come from the posterior distribution are referred to as ‘convergence diagnostics’ and are discussed in detail in I. Ntzoufras, *Bayesian Modeling Using WinBUGS* (Hoboken, NJ, Wiley, 2009), for example.

Figure 11.47

Caterpillar plots for $(c_2 + d_2t) - (c_1 + d_1t)$ (difference in regression line values) at $t = 10^\circ\text{C}$ (upper caterpillar) and $t = 20^\circ\text{C}$ (lower caterpillar).



11.10.4 Exercises

63 A telephone-based automated customer care system has three main menu options: 45% of customers choose option 1, 32% choose option 2, and 23% choose option 3. Of those who choose option 1, 28% eventually get routed to a service agent, as do 41% of those who choose option 2 and 16% of those who choose option 3. What is the overall proportion of customers who eventually get routed to a service agent?

64 An explosion at a construction site could have occurred as a result of (a) static electricity, (b) malfunctioning of equipment, (c) carelessness or (d) sabotage. It is estimated that such an explosion would occur with probability 0.25 as a result of (a), 0.20 as a result of (b), 0.40 as a result of (c) and 0.75 as a result of (d). It is also judged that the prior probabilities of the four causes of the explosion are (a) 0.20, (b) 0.40, (c) 0.25, (d) 0.15. Find the posterior probabilities and hence the most likely cause of the explosion.

65 Three marksmen (A, B and C) fire at a target. Their success rates at hitting the target are 60% for A, 50% for B and 40% for C. If each marksman fires one shot at the target and two bullets hit it, then which is more probable: that C hit the target, or did not?

66 An accident has occurred on a busy highway between city A, of 100 000 people, and city B, of 200 000 people. It is known only that the victim is from one of the two cities and that his name is Smith. A check of the records reveals that 10% of city A's population is named Smith and 5% of city B's population has that name. The police want to know where to start looking for relatives of the victim. What is the probability that the victim is from city A?

67 In a certain community, 8% of all adults over 50 have diabetes. If a health service in this community correctly diagnoses 95% of all persons with diabetes as having the disease, and incorrectly diagnoses 2% of all persons without diabetes as having the disease, find the probabilities that

- the community health service will diagnose an adult over 50 as having diabetes,
- a person over 50 diagnosed by the health service as having diabetes actually has the disease.

68 A stockbroker correctly identifies a stock as being a good one 60% of the time and correctly identifies a stock as being a bad one 80% of the time. A stock has a 50% chance of being good. Find the probability that a stock is good if

- the stockbroker identifies it as good,
- k out of n stockbrokers of equal ability independently identify it as good.

69 On a communications channel, one of three sequences of letters can be transmitted: AAAA, BBBB and CCCC, where the prior probabilities of the sequences are 0.3, 0.4 and 0.3 respectively. It is known from the noise in the channel that the probability of correct reception of a transmitted letter is 0.6, and the probability of incorrect reception of the other two letters is 0.2 for each. It is assumed that the letters are distorted independently of each other. Find the most probable transmitted sequence if ABCA is received.

70 The number of accidents per day occurring at a road junction was recorded over a period of 100 days. There were no accidents on 84 days, one accident on 12 days, and two accidents on four days. One hypothesis is that the number of accidents per day has a Poisson distribution with parameter λ (unspecified), and another is that the distribution is binomial with parameters $n = 3$ and p (unspecified). Use the average number of accidents per day to identify the unspecified parameters and compare the hypotheses assuming that the binomial is initially thought to be twice as likely as the Poisson.

71 The following **multinomial distribution** is a generalization of the binomial distribution. Suppose that there are k distinct possible outcomes of an experiment, with probabilities p_1, \dots, p_k , and that the experiment is repeated n times. The probability of obtaining a number n_1 of occurrences of the first possible outcome, n_2 of the second, and so on up to n_k of the k th is

$$P(n_1, \dots, n_k) = \frac{n!}{n_1! \dots n_k!} (p_1)^{n_1} \dots (p_k)^{n_k}$$

Suppose now that there are two competing hypotheses H_1 and H_2 . H_1 asserts that the

probabilities are p_1, \dots, p_k as above, and H_2 asserts that they are q_1, \dots, q_k . Prove that the logarithm of the likelihood ratio is

$$\ln \left[\frac{P(n_1, \dots, n_k | H_1)}{P(n_1, \dots, n_k | H_2)} \right] = \sum_{i=1}^k n_i \ln \left(\frac{p_i}{q_i} \right)$$

72

According to the design specification, of the components produced by a machine, 92% should have no defect, 5% should have defect A alone, 2% should have defect B alone and 1% should have both defects. Call this hypothesis H_1 . The user suspects that the machine is producing more components (say a proportion p_B) with defect B alone, and also more components (say a proportion p_{AB}) with both defects, but is satisfied that 5% have defect A alone. Call this hypothesis H_2 . Of a sample of 1000 components, 912 had no defects, 45 had A alone, 27 had B alone and 16 had both. Using the multinomial distribution (as in Exercise 71), maximize $\ln P(912, 45, 27, 16 | H_2)$ with respect to p_B and p_{AB} , and find the posterior odds assuming prior odds of 5:1 in favour of H_1 .

73

It is suggested that higher-priced cars are assembled with greater care than lower-priced cars. To investigate this, a large luxury model A and a compact hatchback B were compared for defects when they arrived at the dealer's showroom. All

cars were manufactured by the same company. The numbers of defects for several of each model were recorded:

A: {5, 4, 3, 5, 3, 4}

B: {8, 6, 8, 9, 5}

The number of defects in each car can be assumed to be governed by a Poisson distribution with parameter λ . Compare the hypothesis H_1 that $\lambda_A \neq \lambda_B$ with H_2 that $\lambda_A = \lambda_B = \lambda$, using the average numbers of defects to identify the λ values and assuming no initial preference between the hypotheses.

Perform Bayesian inference about the parameters of the linear regression model

$$Y_i = a + bX_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

for the strain gauge data of Example 11.17. Estimate $(Y - a)/b$ when Y is 13.8 V. Estimate the voltmeter measurement when the tension or force is 8.5 kN and 10 kN. Also, monitor the standardized residual quantities $r_{s,i} = (Y_i - a - bX_i)/\sigma = \sqrt{\tau}(Y_i - a - bX_i)$, in which the precision $\tau = 1/\sigma^2$, and produce a caterpillar plot of corresponding credible intervals. (Hint: Construct your BUGS code step by step so that it performs one inference task at a time.)

11.11 Review exercises (1–10)

1

Eight cases each of 12 bottles of wine from a vineyard were tested for evidence of oxidation in the wine. Five of the cases were bottled using standard corks and, of these, six bottles were found to have oxidized. The remaining cases were bottled using plastic bungs and, of these, three bottles were found to have oxidized. Test the hypothesis that there is no difference in the proportion of bottles oxidized for the different types of cork.

2

The amplitude d of vibration of a damped pendulum is expected to diminish by

$$d = d_0 e^{-\lambda t}$$

Successive amplitudes are measured from a trace as follows:

t	1.01	2.04	3.12	4.09	5.22	6.30	7.35	8.39	9.44	10.50
d	2.46	1.75	1.26	0.94	0.90	0.79	0.52	0.49	0.31	0.21

Find a 95% confidence interval for the damping coefficient λ .

3

Successive masses of 1 kg were hung from a wire, and the position of a mark at its lower end was measured as follows:

Load/kg	0	1	2	3	4	5	6	7
Position/cm	6.12	6.20	6.26	6.32	6.37	6.44	6.50	6.57

It is expected that the extension Y is related to the force X by

$$Y = LX/EA$$

where $L = 101.4$ cm is the length, $A = 1.62 \times 10^{-5}$ cm² is the area and E is the Young's modulus of the material. Find a 95% confidence interval for the Young's modulus.

6.8	2.1	1.0	28.1	5.8	19.7	2.9	16.3	10.7	25.3	12.5	1.6	3.0	9.9	15.9
21.3	9.1	6.9	5.6	2.0	2.2	10.2	6.5	6.8	42.5	2.9	7.3	3.1	2.6	1.0
3.8	14.7	3.8	13.9	2.9	4.1	22.7	5.8	7.6	6.4	11.3	51.6	15.6	2.6	7.6
1.2	0.7	1.9	1.8	0.7	0.4	72.0	10.7	8.3	15.1	3.6	6.0	0.1	3.1	12.9
2.2	17.6	3.6	2.4	3.2	0.4	4.4	17.1	7.1	10.1	18.8	3.4	0.2	4.9	12.9
1.8	22.4	11.6	4.2	18.0	3.0	16.2	6.8	3.7	13.6	15.7	0.7	2.7	18.8	29.8
4.9	6.8	10.7	0.9	2.4	3.8	9.0	8.8	4.8	0.3	4.6	4.9	6.1	33.0	6.5

Figure 11.48 Time interval data for Review exercise 4.

- 4 The table in Figure 11.48 gives the intervals, in hours, between arrivals of cargo ships at a port during a period of six weeks. It is helpful to the port authorities to know whether the times of arrival are random or whether they show any regularity. Fit an exponential distribution to the data and test for goodness-of-fit.
- 5 When large amounts of data are processed, there is a danger of transcription errors occurring (for example, a decimal point in the wrong place), which could bias the results. One way to avoid this is to test for **outliers** in the data. Suppose that X_1, \dots, X_n are independent exponential random variables, each with a common parameter λ . Let the random variable Y be the largest of these divided by the sum:

$$Y = X_{\max} / \sum_i X_i$$

It can be shown (V. Barnett and T. Lewis, *Outliers in Statistical Data*, Wiley, Chichester, 1978) that the distribution function of Y is given by

$$F_Y(y) = \sum_{k=0}^{\lfloor 1/y \rfloor} (-1)^k \binom{n}{k} (1-ky)^{n-1} \left(\frac{1}{n} \leq y \leq 1 \right)$$

where $\lfloor 1/y \rfloor$ denotes the integer part of $1/y$. For the data in the Review exercise 4 (Figure 11.48) test the largest value to see whether it is reasonable to expect such a value if the data truly have an exponential distribution. Find 95% confidence intervals for the mean inter-arrival time with this value respectively included and excluded from the data.

- 6 Language courses in French, German and Spanish are offered by an adult learning institute. At the end of each course, the students are asked to grade their response to the course as either very satisfied,

fairly satisfied, neutral, fairly dissatisfied, or very dissatisfied. After gathering data for several terms the results are as follows:

Grade	French	German	Spanish
Very satisfied	16	6	22
Fairly satisfied	63	13	76
Neutral	40	27	60
Fairly dissatisfied	10	13	32
Very dissatisfied	3	5	12

Is there evidence of different levels of satisfaction with the different courses?

- 7 A surgeon has to decide whether or not to perform an operation on a patient suspected of suffering from a rare disease. If the patient has the disease, he has a 50:50 chance of recovering after the operation but only a one in 20 chance of survival if the operation is not performed. On the other hand, there is a one in five chance that a patient who has not got the disease would die as a result of the operation. How will the decision depend upon the surgeon's assessment of the probability p that the patient has the disease? (Hint: Use $P(B|A) = P(B|A \cap C)P(C) + P(B|A \cap C')P(C')$, where A and C are independent.)

- 8 A factory contains 200 machines, each of which becomes misaligned on average every 200 h of operation, the misalignments occurring at random and independently of each other and of other machines. To detect the misalignments, a quality control chart will be followed for each machine, based on one sample of output per machine per hour. Two options have been worked out: option A would cost £1 per hour per machine, whereas

option B would cost £1.50 per hour per machine. The control charts differ in their average run lengths (ARLs) to a signal of action required. Option A (Shewhart) has an ARL of 20 for a misaligned machine, but will also generate false alarms with an ARL of 1000 for a well-adjusted machine. Option B (cusum) has an ARL of four for a misaligned machine and an ARL of 750 for a well-adjusted machine.

When a control chart signals action required, the machine will be shut down and will join a queue of machines awaiting servicing. A single server will operate, with a mean service time of 30 min and standard deviation of 15 min, regardless of whether the machine was actually misaligned. This is all that is known of the service time distribution, but use can be made of the **Pollaczek–Khintchine formula**, which applies to single-channel queues with arbitrary service distributions:

$$N_s = \rho + \frac{(\lambda\sigma_s)^2 + \rho^2}{2(1 - \rho)}$$

(the notation is as in Section 11.9.3, with σ_s the standard deviation of service time).

During the time that a machine is in the queue and being serviced, its lost production is costed at £200 per hour. In addition, if the machine is found to have been misaligned then its output for the previous several hours (given on average by the ARL) must be examined and if necessary rectified, at a cost of £10 per production hour.

Find the total cost per hour for each option, and hence decide which control scheme should be implemented.

- 9** A transmission channel for binary data connects a source to a receiver. The source emits a 0 with

probability α and a 1 with probability $1 - \alpha$, each symbol independent of every other. The noise in the channel causes some bits to be interpreted incorrectly. The probability that a bit will be inverted is p (whether a 0 or a 1, the channel is ‘symmetric’).

- (a) Using Bayes’ theorem, express the four probabilities that the source symbol is a 0 or a 1 given that the received symbol is a 0 or a 1.
- (b) If p is small and the receiver chooses to deliver whichever source symbol is the more likely given the received symbol, find the conditions on α such that the source symbol is assumed to be the same as the received symbol.

10

If discrete random variables X and Y can take possible values $\{u_1, \dots, u_m\}$ and $\{v_1, \dots, v_n\}$ respectively, with joint distribution $P(u_k, v_j)$ (see Section 11.4.1), the **mutual information** between X and Y is defined as

$$I(X; Y) = \sum_{k=1}^m \sum_{j=1}^n P(u_k, v_j) \log_2 \frac{P(u_k, v_j)}{P(X=u_k)P(Y=v_j)}$$

Show that for the binary symmetric transmission channel referred to in Review exercise 9, if X is the source symbol, Y the received symbol and $\alpha = \frac{1}{2}$ then

$$I(X; Y) = 1 + p \log_2 p + (1 - p) \log_2 (1 - p)$$

The interpretation of this quantity is that it measures (in ‘bits’) the average amount of information received for each bit of data transmitted. Show that $I(X; Y) = 0$ when $p = \frac{1}{2}$ and that $I(X; Y) \rightarrow 1$ as $p \rightarrow 0$ and as $p \rightarrow 1$. Interpret this result.

Answers to Exercises

CHAPTER 1

Exercises

1 (a)

$$\mathbf{2} \quad \mathbf{A} = \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & 0 \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The transformation rotates the e_1, e_2 plane through $\pi/4$ about the e_3 axis.

3 (a), (c) and (d)

4 The set of all odd quintic polynomials; it has dimension 3.

5 (a) $\lambda^3 - 12\lambda^2 + 40\lambda - 35$

(b) $\lambda^4 - 4\lambda^3 + 2\lambda^2 + 5\lambda + 2$

6 (a) 2, 0; $[1 \ 1]^T, [1 \ -1]^T$

(b) 4, -1; $[2 \ 3]^T, [1 \ -1]^T$

(c) 9, 3, -3; $[-1 \ 2 \ 2]^T, [2 \ 2 \ -1]^T, [2 \ -1 \ 2]^T$

(d) 3, 2, 1; $[2 \ 2 \ 1]^T, [1 \ 1 \ 0]^T, [0 \ -2 \ 1]^T$

(e) 14, 7, -7; $[2 \ 6 \ 3]^T, [6 \ -3 \ 2]^T, [3 \ 2 \ -6]^T$

(f) 2, 1, -1; $[-1 \ 1 \ 1]^T, [1 \ 0 \ -1]^T, [1 \ 2 \ -7]^T$

(g) 5, 3, 1; $[2 \ 3 \ -1]^T, [1 \ -1 \ 0]^T, [0 \ -1 \ 1]^T$

(h) 4, 3, 1; $[2 \ -1 \ -1]^T, [2 \ -1 \ 0]^T, [4 \ 1 \ -2]^T$

7 (a) 5, $[1 \ 1 \ 1]^T$; 1 (repeated) with two linearly independent eigenvectors, e.g. $[0 \ 1 \ 2]^T, [1 \ 0 \ -1]^T$

(b) -1, $[8 \ 1 \ 3]^T$; 2 (repeated) with one linearly independent eigenvector, e.g. $[1 \ -1 \ 0]^T$

(c) 1, $[4 \ 1 \ -3]^T$; 2 (repeated) with one linearly independent eigenvector, e.g. $[3 \ 1 \ -2]^T$

(d) 2, $[2 \ 1 \ 2]^T$; 1 (repeated) with two linearly independent eigenvectors, e.g. $[0 \ 2 \ -1]^T, [2 \ 0 \ 3]^T$

8 1, $[-3 \ 1 \ 1]^T$

9 2, e.g. $[1 \ 0 \ 1]^T, [0 \ 1 \ 1]^T$

12 -6, 3, 2; $[2 \ 1 \ 1]^T, [-1 \ 1 \ 1]^T, [0 \ 1 \ -1]^T$

13 $[1 \ -1 \ 0]^T$

14 8.59, $[0.61 \ 0.71 \ 0.35]^T$

15 (a) 3.62, $[0.62 \ 1 \ 1]^T$

(b) 7, $[0.25 \ 0.5 \ 1]^T$

(c) 2.62, $[1 \ -0.62 \ -0.62 \ 1]^T$

16 $\lambda_1 = 6; \lambda_2 = 3, \mathbf{e}_2 = [1 \ 1 \ -1]^T;$
 $\lambda_3 = 2, \mathbf{e}_3 = [1 \ -1 \ 0]^T$

17 10.132, 4.491, 0.373

18 (b) 0.59

19 5, 2, -1; $[-1 \ 5 \ 3]^T, [0 \ 2 \ 1]^T, [1 \ 0 \ 0]^T$

20 6, 3, 1; $[1 \ 2 \ 0]^T, [0 \ 0 \ 1]^T, [2 \ -1 \ 0]^T$

21 18, 9, -9; $[2 \ 1 \ 2]^T, [1 \ 2 \ -2]^T, [-2 \ 2 \ 1]^T$

22 2, 1, -1; $[1 \ 3 \ 1]^T, [3 \ 2 \ 1]^T, [1 \ 0 \ 1]^T$

23 -9, 6, 3; $[1 \ 2 \ -2]^T, [2 \ 1 \ 2]^T, [-2 \ 2 \ 1]^T$

$\mathbf{L} = \frac{1}{3} \mathbf{M}, \mathbf{M}$ modal matrix

$$\mathbf{24} \quad [0 \ 0 \ 1]^T, \begin{bmatrix} 2 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

$$\mathbf{25} \quad \mathbf{J} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

26 $\lambda = -2: [0 \ 1 \ 1 \ 0]^T, [1 \ 0 \ 0 \ 1]^T$
 $\lambda = 4: [0 \ 1 \ -1 \ 0]^T, [6 \ -1 \ 0 \ -6]^T$

$$\mathbf{J} = \begin{bmatrix} -2 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 4 & 1 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

27 $y_1^2 + y_2^2 + y_3^2$

28 (a) Positive-definite

(b) Positive-semidefinite

(c) Indefinite

29 (a) $2a > 1$, (b) $2b^2 < 6a - 3$

30 Positive-semidefinite, eigenvalues 3, 3, 0

 31 $k > 2$; when $k = 2$ Q is positive-semidefinite

32 $a > 2$

33 $\lambda > 5$

35 (a) $\begin{bmatrix} 3 & 4 \\ 2 & 3 \end{bmatrix}$ (b) $\begin{bmatrix} 7 & 10 \\ 5 & 7 \end{bmatrix}$ (c) $\begin{bmatrix} 17 & 24 \\ 12 & 17 \end{bmatrix}$

36 (a) $\frac{1}{3} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$ (b) $\frac{1}{11} \begin{bmatrix} -2 & 5 & -1 \\ -1 & -3 & 5 \\ 7 & -1 & -2 \end{bmatrix}$

37 $\begin{bmatrix} 47231 & 47342 & 47270 \\ 47342 & 47195 & 47306 \\ 47270 & 47306 & 47267 \end{bmatrix}$

38 (a) $\begin{bmatrix} e^t & 0 \\ t e^t & e^t \end{bmatrix}$ (b) $\begin{bmatrix} e^t & 0 \\ e^{2t} - e^t & e^{2t} \end{bmatrix}$

40 (a) $\begin{bmatrix} 2t & 2 \\ -1 & 2t-1 \end{bmatrix}$ (b) $\begin{bmatrix} \frac{10}{3} & 0 \\ \frac{7}{2} & \frac{23}{6} \end{bmatrix}$

41 $\begin{bmatrix} t^4 + 2t^2 + t - 4 & t^3 - t^2 + t - 1 \\ 5t^2 + 5 & 5t - 5 \end{bmatrix}$

42 (a) 3, 3 (b) Yes

43 (a) $\begin{bmatrix} \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} & -\frac{3}{\sqrt{10}} \end{bmatrix} \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ -\frac{2}{3} & -\frac{1}{3} & \frac{2}{3} \\ \frac{2}{3} & -\frac{2}{3} & \frac{1}{3} \end{bmatrix}$

(b) $\frac{1}{180} \begin{bmatrix} -1 & 13 \\ 4 & 8 \\ 10 & -10 \end{bmatrix}$

44 $\frac{1}{141} \begin{bmatrix} 13 & 30 & -17 & 6 & -4 \\ 18 & 9 & 9 & 30 & 27 \end{bmatrix}$

45 (a) 1

(b) $\begin{bmatrix} \frac{1}{3} & 0 & \frac{2}{\sqrt{5}} \\ -\frac{2}{3} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} \\ \frac{2}{3} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$

(c) $\frac{1}{18} \begin{bmatrix} 1 & -2 & 2 \\ -1 & 2 & -2 \end{bmatrix}$

(d) $x = \frac{1}{6}, y = -\frac{1}{6}$

46 (a) $x = y = \frac{2}{3}$

47 (b) $\frac{1}{15} \begin{bmatrix} 4 & 5 & 1 & 6 \\ 2 & 10 & 8 & 3 \\ -3 & 0 & 3 & 3 \end{bmatrix}$

48 (a) (i) $x = y = 1$ (ii) $x = y = 1.0909$

(b) (i) $x = y = 1$ (ii) $x = y = 1.4785$

(c) (i) $x = y = 1$ (ii) $x = y = 1.4998$

49 $m = 0.5, c = 0.8$

50 (a) $\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -4 & -5 & -4 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u,$

$y = [1 \ 0 \ 0]x$

(b) $\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & -4 & -2 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 5 \end{bmatrix} u,$

$y = [1 \ 0 \ 0 \ 0]x$

51 (a) $\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -7 & -5 & -6 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u,$

$y = [5 \ 3 \ 1]x$

(b) $\dot{x} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -3 & -4 \end{bmatrix} x + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u, \quad y = [2 \ 3 \ 1]x$

52 $\dot{x} = \begin{bmatrix} -R_1/L_1 & -R_1/L_1 & -1/L_1 \\ -R_1/L_1 & -(R_1 + R_2)/L_2 & -1/L_2 \\ 1/C & 1/C & 1/C \end{bmatrix} x + \begin{bmatrix} 1/L_1 \\ 1/L_2 \\ 0 \end{bmatrix} u,$

$y = [0 \ R_2 \ 0]x$

53 A possible model is

$$\dot{x} = \begin{bmatrix} B(M_1 + M)/MM_1 & 1 & 0 & 0 \\ -(K_1M + KM_1 + KM)/MM_1 & 0 & 1 & 0 \\ -K_1B/MM_1 & 0 & 0 & 1 \\ -K_1K/MM_1 & 0 & 0 & 0 \end{bmatrix} x$$

$$+ \begin{bmatrix} 0 \\ 0 \\ K_1B/MM_1 \\ K_1K_2/MM_1 \end{bmatrix} u, \quad y = [1 \ 0 \ 0 \ 0]x$$

54 $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -\frac{(R_1+R_2+R_4)}{\alpha C_1} & \frac{R_1}{\alpha C_1} \\ \frac{R_1}{\alpha C_2} & -\frac{R_1+R_3}{\alpha C_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$+ \begin{bmatrix} \frac{R_2+R_4}{\alpha C_1} \\ \frac{R_3}{\alpha C_2} \end{bmatrix} u$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{R_1}{\alpha} & -\frac{R_1+R_3}{\alpha} \\ -\frac{R_3}{\alpha}(R_1+R_2+R_4) & \frac{R_1 R_3}{\alpha} \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{R_3}{\alpha} \\ \frac{R_3(R_4+R_2)}{\alpha} \end{bmatrix} u,$$

$\alpha = R_1 R_3 + (R_1 + R_3)(R_2 + R_4)$
 $-2.6 \times 10^2, -1.1 \times 10^2$

55 $\begin{bmatrix} e^t & 0 \\ t e^t & e^t \end{bmatrix}$

56 $\begin{bmatrix} e^{-t}(1+t) & t e^{-t} \\ -t e^{-t} & e^{-t}(1-t) \end{bmatrix}, \quad y = e^{-t}(1+2t)$

57 $[e^t \quad e^t(t+1)]^T$

58 $x_1 = 2 - 4e^{-2t} + 3e^{-3t}, x_2 = 8e^{-2t} - 9e^{-3t}$

59 $x_1 = 4t e^{-2t} + e^{-2t} - e^{-t}, x_2 = 3 e^{-t} - 2 e^{-2t} - 4t e^{-2t}$

60 $x(t) = \begin{bmatrix} -5 + \frac{8}{3}e^{-t} + \frac{10}{3}e^{5t} \\ 3 - \frac{8}{3}e^{-t} + \frac{5}{3}e^{5t} \end{bmatrix}$

61 $x(t) = e^t [3 \quad 2]^T - e^{-3t} [1 \quad -2]^T$

62 $x(t) = \frac{1}{5} \begin{bmatrix} 14 e^{-t} - 4 e^{-6t} \\ 7 e^{-t} + 8 e^{-6t} \end{bmatrix}$

63 $x(t) = e^{-2t} \{ (\cos 2t - \sin 2t)[2 \quad 1]^T - (\cos 2t + \sin 2t)[0 \quad 1]^T \}$

64 $\dot{z} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} z + \begin{bmatrix} \frac{1}{3} \\ 0 \\ -\frac{4}{3} \end{bmatrix} u,$
 $y = [1 \quad -4 \quad -2]z$

65 $\alpha_0 = \frac{1}{2}, \alpha_1 = -\frac{1}{2}, \alpha_2 = \frac{1}{2}$

66 $6, 1, [4 \quad 1]^T, [1 \quad -1]^T$
 $x_1 = 4e^{6t} - 3e^t, x_2 = e^{6t} + 3e^t$

67 Same as Exercise 60

68 Asymptotically stable

69 Asymptotically stable

70 $a > 0, b > 0$

1.10 Review exercises

1 (a) $5, 2, -1; [1 \ 1 \ 5 \ 3]^T, [8 \ 2 \ 1]^T, [1 \ 0 \ 0]^T$

(b) $3, 2, 1; [1 \ 2 \ 1]^T, [2 \ 1 \ 0]^T, [1 \ 0 \ -1]^T$

(c) $3, 1, 0; [1 \ -2 \ 1]^T, [1 \ 0 \ -1]^T, [1 \ 1 \ 1]^T$

2 $6, 3, 1; [1 \ 1 \ 1]^T, [1 \ 1 \ -2]^T, [1 \ -1 \ 0]^T$

3 $b = 1, c = 2; \lambda = 2, 4, 1;$

$[1 \ -2 \ -1]^T, [1 \ 1 \ -1]^T$

4 (a) 4.56; $[0.72 \ 0.84 \ 1]^T$ (b) 1.75

(c) (i) 1.19 (ii) 1.75

5 $[0 \ -1 \ 1]^T$

6 $3, 2, 1; [2 \ 1 \ 1]^T, [3 \ 2 \ 1]^T, [4 \ 3 \ 2]^T$

7 $[\frac{1}{3} \ -\frac{2}{3} \ -\frac{2}{3}]^T, [\frac{2}{3} \ \frac{2}{3} \ -\frac{1}{3}]^T, [\frac{2}{3} \ -\frac{1}{3} \ \frac{2}{3}]^T$

8 $-6, -4, -2, 0; [1 \ -3 \ 3 \ -1]^T, [0 \ 1 \ -2 \ 1]^T,$
 $[0 \ 0 \ 1 \ -1]^T, [0 \ 0 \ 0 \ 1]^T;$
 $C - Ce^{-6t} + 3Ce^{-4t} - 3Ce^{-2t}$

9 (a) (i) $\begin{bmatrix} -29 & 0 \\ -32 & 3 \end{bmatrix}$ (ii) $\begin{bmatrix} 2^k & 0 \\ 2^k - 1 & 1 \end{bmatrix}$

(b) $\begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}$

10 $e_1 = [1 \ 0 \ 0]^T, e_2 = [\frac{3}{8} \ \frac{1}{2} \ 0]^T, e_3 = [0 \ 0 \ \frac{1}{8}]^T$

11 $2, 2 \pm \sqrt{2}; 1; 0; -1, 1; -\sqrt{2}; 1, 1; \sqrt{2}; 1$

12 (a) Positive-semidefinite (b) Positive-definite

(c) Indefinite (d) Negative-semidefinite

(e) Negative-definite

13 $1; 3, [1 \ 1 \ 0]^T; -1, [0 \ -1 \ 1]^T$

14 (a) $\begin{bmatrix} -0.8 & 0.6 \\ 0.6 & 0.8 \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 2.5 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0.8 & 0.6 & 0 \\ -0.6 & 0.8 & 0 \end{bmatrix}$

(b) $\frac{i}{125} \begin{bmatrix} 24 & 32 \\ 18 & 24 \\ -20 & 15 \end{bmatrix} = \begin{bmatrix} 0.192 & 0.256 \\ 0.144 & 0.192 \\ -0.16 & 0.12 \end{bmatrix}$

15 (c) $\begin{bmatrix} \frac{1}{3} \\ -\frac{2}{3} \\ \frac{2}{3} \end{bmatrix} \sqrt{18} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}, \frac{1}{18} \begin{bmatrix} 1 & -2 & 2 \\ -1 & 2 & -2 \end{bmatrix}$

16 $A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, b_1 = [\frac{1}{3} \quad 0 \quad -\frac{4}{3}]^T,$
 $c = [1 \quad -4 \quad -2]^T$

The system is uncontrollable but observable; it is stable

17 $M = \begin{bmatrix} -2 & 1 & -2 \\ 1 & 0 & -4 \\ -1 & 0 & -1 \end{bmatrix}$
 $x(t) = \begin{bmatrix} -14e^{-t} + \frac{127}{4}e^{-2t} - \frac{58}{9}e^{-3t} + \frac{1}{6}t - \frac{47}{36} \\ 7e^{-t} - \frac{29}{9}e^{-3t} + \frac{1}{3}t + \frac{11}{9} \\ -7e^{-t} + \frac{29}{3}e^{-3t} - \frac{2}{3} \end{bmatrix}$

18 (a) 6, $[3 \quad 2 \quad 1]^T$; 3, $[1 \quad -1 \quad 0]^T$, $[\frac{1}{3} \quad \frac{1}{3} \quad \frac{1}{3}]^T$
(c) $\frac{1}{3} \begin{bmatrix} (-3+3t)e^{3t} + 3e^{6t} \\ (1+3t)e^{3t} + 2e^{6t} \\ -e^{3t} + e^{6t} \end{bmatrix}$

19 $x_1 = \frac{5}{3} \cos t + 2 \sin t - \frac{2}{3} \cos 2t$
 $x_2 = \frac{5}{3} \cos t + 2 \sin t + \frac{1}{3} \cos 2t$

CHAPTER 2

Exercises

1 $X(0.3) = 0.985\ 05$

2 $X(1.1) = 0.094\ 913$

3 $X(1) = 1.1571$

4 $X(0.5) = 2.1250$

5 $X_a(2) = 2.811\ 489, X_b(2) = 2.819\ 944,$
 $x(t) = 2\sqrt{(2+t^2)}/\sqrt{3}$

6 $X_a(2) = 1.573\ 065, X_b(2) = 1.558\ 541,$
 $x(t) = \sqrt{1+2\ln t}$

7 $X_a(1.5) = 2.241\ 257, X_b(1.5) = 2.206\ 232,$
 $x(t)\ln x(t) - x(t) = t - 1.981\ 214$

8 (a) $X(0.5) = 0.1238$ (b) $X(1.2) = 1.3740$

9 $X(0.5) = 1.7460$

10 (a) $X(0.5) = 0.7948$ (b) $X(1) = -1.3511$

14 $X(0.5) = 0.1353$

15 (a) $X(0.75) = 3.2345$
(b) $X(2) = 2.2771$

16 (a) $X_{0,2}(2) = 2.242\ 408, X_{0,1}(2) = 2.613\ 104$
Richardson extrapolation estimates the error as 0.123 565 so a step less than 0.0064 should be used.

- (b) $X_{0,2}(2) = 2.788\ 158, X_{0,1}(2) = 2.863\ 456$
Richardson extrapolation estimates the error as 0.025 099 so a step less than 0.014 should be used.
(c) $X_{0,4}(2) = 2.884\ 046, X_{0,2}(2) = 2.897\ 402$
Richardson extrapolation estimates the error as 0.000 890 so a step less than 0.057 should be used.
 $x(2) = 2.898\ 51$ to 5 dp

17 $X(3) = 1.466\ 47$

- 18 (a) $\frac{dx}{dt} = v, x(0) = 1$
 $\frac{dv}{dt} = 4xt - 6(x^2 - t)v, v(0) = 2$
(b) $\frac{dx}{dt} = v, x(1) = 2$
 $\frac{dv}{dt} = -4(x^2 - t^2), v(1) = 0.5$
(c) $\frac{dx}{dt} = v, x(0) = 0$
 $\frac{dv}{dt} = -\sin v - 4x, v(0) = 0$
(d) $\frac{dx}{dt} = v, x(0) = 1$
 $\frac{dv}{dt} = w, v(0) = 2$
 $\frac{dw}{dt} = e^{2t} + x^2t - 6e^t v - tw, w(0) = 0$
(e) $\frac{dx}{dt} = v, x(1) = 1$
 $\frac{dv}{dt} = w, v(1) = 0$
 $\frac{dw}{dt} = \sin t - x^2 - tw, w(1) = -2$
(f) $\frac{dx}{dt} = v, x(2) = 0$
 $\frac{dv}{dt} = w, v(2) = 0$
 $\frac{dw}{dt} = (x^2t^2 + tw)^2, w(2) = 2$
(g) $\frac{dx}{dt} = v, x(0) = 0$
 $\frac{dv}{dt} = w, v(0) = 0$
 $\frac{dw}{dt} = u, w(0) = 4$
 $\frac{du}{dt} = \ln t - x^2 - xw, u(0) = -3$
(h) $\frac{dx}{dt} = v, x(0) = a$
 $\frac{dv}{dt} = w, v(0) = 0$
 $\frac{dw}{dt} = u, w(0) = b$
 $\frac{du}{dt} = t^2 + 4t - 5 + \sqrt{(xt)} - v - (v-1)tu, u(0) = 0$

19 $X(0.3) = 0.299\ 90$

20 $X(0.3) = 0.299\ 64$

21 $X(0.65) = -0.826\ 03$

- 22 $X_{0,4}(1.6) = 1.220\ 254, X_{0,2}(1.6) = 1.220\ 055$
Richardson extrapolation estimates the error as 0.000 013 so, to obtain an error less than 5×10^{-7} , a step less than 0.088 should be used.

- 23** $X_{0.1}(2.2) = 2.923\ 350\ 36$, $X_{0.05}(2.2) = 2.925\ 417\ 56$
Richardson extrapolation estimates the error as
0.000 295 so, to obtain an error less than 5×10^{-7} ,
a step less than 0.0060 should be used.

2.7 Review exercises

1 $X(0.5) = 1.548\ 860$

2 $X(1.2) = 0.524\ 465$

- 3** $X_{0.1}(0.4) = 1.125\ 583$, $X_{0.05}(0.4) = 1.142\ 763$
Richardson extrapolation estimates the error as
0.017 180 so, to obtain an error less than 5×10^{-3} ,
a step less than 0.0146 should be used.

- 4** $X_{0.05}(0.25) = 2.003\ 749$, $X_{0.025}(0.25) = 2.004\ 452$
Richardson extrapolation estimates the error as
0.000 703 so, to obtain an error less than 5×10^{-4} ,
a step less than 0.0178 should be used.

5 $X_1(1.2) = 2.374\ 037$, $X_2(1.2) = 2.374\ 148$,
 $X_3(1.2) = 2.374\ 176$

6 $X(1) = 5.194\ 323$ accurate to 6dp.

- 8** $X_{0.025}(2) = 0.847\ 035$, $X_{0.0125}(2) = 0.844\ 066$
Richardson extrapolation estimates the error as
0.002 969 so we have $X(2) = 0.84$.

9 $X(4) = 0.1458$ (using step size 0.002)

10 $X(2.5) = -0.6532$ (using step size 0.025)

CHAPTER 3

Exercises

- 1** (a) Circles centre $(0, 0)$, $x^2 + y^2 = 1 + e^C$
(b) Straight lines through $(-1, 0)$, $y = (x + 1) \tan C$
- 2** (a) Family of curves $y^2 = 4x^2(x - 1) + C$
(b) Family of curves $y^2 = \frac{1}{12}x^2(x^2 - 12) + C$
- 3** (a) $z - xy = C$
(b) $xy = \ln(C + z)$
- 4** (a) $(A \sec(t + B), \tan(t + B), C e^t)$, curves on
hyperbolic cylinders $(x/A)^2 - y^2 = 1$
(b) Curves defined by the intersections of
mutually orthogonal hyperbolic cylinders,
 $x^2 - y^2 = c$, $x^2 - z^2 = k$
- 5** (a) $f_x = yz - 2x$, $f_y = xz + 1$, $f_z = xy - 1$,
 $f_{xx} = -2$, $f_{xy} = z$, $f_{xz} = y$, $f_{yy} = 0$, $f_{yz} = x$, $f_{zz} = 0$
(b) $f_x = 2xyz^3$, $f_y = x^2z^3$, $f_z = 3x^2yz^2$, $f_{xx} = 2yz^3$,
 $f_{xy} = 2xz^3$, $f_{xz} = 6xyz^2$, $f_{yy} = 0$, $f_{yz} = 3x^2z^2$, $f_{zz} = 6x^2yz$

(c) $f_x = -yz/(x^2 + y^2)$, $f_y = zx/(x^2 + y^2)$,
 $f_z = \tan^{-1}(y/x)$, $f_{xx} = 2xyz/(x^2 + y^2)^2$,
 $f_{xy} = z(y^2 - x^2)/(x^2 + y^2)^2$, $f_{xz} = -y/(x^2 + y^2)$, $f_{zz} = 0$,
 $f_{yy} = -2xyz/(x^2 + y^2)^2$, $f_{yz} = x/(x^2 + y^2)$

6 (a) $6t^2(t^3 - 1) + 8t + \frac{1}{(t - 1)^2}$

(b) $t e^{-2t}(\cos 2t - \sin 2t) + \frac{1}{2} e^{-2t} \sin 2t$

7 $\frac{\partial f}{\partial y} = \frac{\partial f}{\partial r} \sin \theta \sin \phi + \frac{\partial f}{\partial \theta} \frac{\sin \phi \cos \theta}{r} + \frac{\partial f}{\partial \phi} \frac{\cos \phi}{r \sin \theta}$

$\frac{\partial f}{\partial z} = \frac{\partial f}{\partial r} \cos \theta - \frac{\partial f}{\partial \theta} \frac{\sin \theta}{r}$

8 $A/r + B$

13 $e^{2u}, \frac{1}{x^2 + y^2}$

14 ± 1

15 $9, v^2 = u + 2w$

17 $\frac{-2[1 + \sqrt{(1 - 4u^2v^2)}]}{u\sqrt{(1 - 4u^2v^2)}} , \frac{2[1 + \sqrt{(1 - 4u^2v^2)}]}{v\sqrt{(1 - 4u^2v^2)}} , \frac{v}{\sqrt{(1 - 4u^2v^2)}} , \frac{u}{\sqrt{(1 - 4u^2v^2)}}$

18 (a) $xy^2 + x^2y + x + c$ (b) $x^2y^2 + y \sin 3x + c$
(c) Not exact (d) $z^3x - 3xy + 4y^3 + c$

19 $-1, y \sin x - x \cos y + \frac{1}{2}(y^2 - 1)$

20 $m = 2$

$8x^5 + 36x^4y + 62x^3y^2 + 63x^2y^3 + 54xy^4 + 27y^5 + c$

21 (36, 9, 12) (a) $-\frac{117}{7}$ (b) 39, $\frac{1}{13}$ (12, 3, 4)

22 (a) $(2x, 2y, -1)$

(b) $(-yz/(x^2 + y^2), xz/(x^2 + y^2), \tan^{-1}(y/x))$

(c) $\frac{e^{-x-y+z}}{(x^3 + y^2)^{3/2}} \times$

$(-x^3 - y^2 - \frac{3}{2}x^2, -x^3 - y^2 - y, x^3 + y^2)$

(d) $(yz \sin \pi(x + y + z) + \pi xyz \cos \pi(x + y + z),$
 $xz \sin \pi(x + y + z) + \pi xyz \cos \pi(x + y + z),$
 $xy \sin \pi(x + y + z) + \pi xyz \cos \pi(x + y + z))$

23 3

24 $(5i + 4j + 3k)/\sqrt{50}$

25 (a) r/r (b) $-r/r^3$

26 $\phi = x^2y + z^2x + zy$

27 $-9j + 3k, \frac{36}{7}$

28 $54^\circ 25'$

- 29** (a) $x + 2y + 3z = 6$, $x - 1 = \frac{1}{2}(y - 1) = \frac{1}{3}(z - 1)$
 (b) $2x + 2y - 3z = -3$,
 $\frac{1}{2}(x - 1) = \frac{1}{2}(y - 2) = \frac{1}{3}(3 - z)$
 (c) $2x + 4y - z = 6$, $\frac{1}{2}(x - 1) = \frac{1}{4}(y - 2) = 4 - z$

31 (a) $6xy$ (b) 4 **32** -61 **33** $a, a, 3a$ **35** -13 **38** $(y, 6xz - 1, 0)$ **40** $x^2 + y^2 + z^2 + xyz$ **42** $a = 2, b = 2, c = 3$; $\phi = 2x^2y + 2z^3x + 3zy + \text{const}$ **43** $\sqrt{11} \text{ rad s}^{-1}$ **44** $d = -a, c = b$

- 47** (a) $2y^2z^3 + 2x^2z^3 + 6x^2y^2z$
 (b) $2y(1+z)\mathbf{i} + 2(x+xz-z)\mathbf{j} + 2y(x-1)\mathbf{k}$
 (c) $2yz\mathbf{i} + 2(x-z)\mathbf{j} + 2yx\mathbf{k}$

56 156 **57** $-\frac{1}{3}$ **58** $\frac{16}{3}$ **59** (a) $\frac{288}{35}$ (b) 10 (c) 8 **60** $10.5 + 4\pi$ **61** (a) 16 (b) 16

(c) Not necessarily. The value has to be the same for all possible paths.

62 35 **63** $-\frac{9}{10}\mathbf{i} - \frac{2}{3}\mathbf{j} + \frac{7}{5}\mathbf{k}$ **64** $4\pi(7\mathbf{i} + 3\mathbf{j})$ **65** (a) 24 (b) 76 (c) 16 **66** $\frac{8}{3} \ln 2$ **67** $\frac{1}{6}$ **68** (a) $(\ln 2) \tan^{-1}(\frac{1}{3})$ (b) $\frac{1}{3}$ (c) 1 **69** $-8/(3\pi^2)$ **70** (a) $\frac{1}{4}(\sqrt{2} - 1)$ (b) $[(1 - k^2)^{3/2} - 1]/3k^2$ **71** $2(\sqrt{2} - 1)$ **73** 1 **74** $2a(1 - \frac{1}{4}\pi)$ **75** $\frac{1}{3}(6\pi - 20)$ **77** $\frac{1}{4}\pi + 2/\pi - 1$ **78** 0 **79** $\frac{11}{60}$ **80** 0 **81** $\frac{1}{2}a(1 - \frac{1}{4}\pi)$ **83** $\frac{13}{3}\pi$ **84** (a) $\frac{183}{4}$ (b) $\frac{1}{4}\pi$ **85** (a) $\frac{27}{4}$ (b) 0 **87** (a) $\frac{13}{3}\pi$ (b) $\frac{149}{30}\pi$ (c) $\frac{37}{10}\pi$ **88** 24π **90** 90 **91** 0 **92** (a) $\frac{8}{3}$ (b) $\frac{448}{3}$ **94** $\frac{1}{2}\pi^2 - 2$ **95** $\frac{1}{720}$ **96** $\frac{11}{30}$ **97** $\frac{1}{24}\pi$ **98** $\frac{1}{50400}$ **99** $(1 - e^{-1})/6$ **100** $\frac{2}{15}; (\frac{5}{16}, \frac{5}{16}, \frac{11}{16})$ **101** $\frac{1}{8}\pi$ **102** $\frac{1}{16}\pi a^4$ **103** $\frac{3}{2}$ **104** 16π **105** 84π **109** $2\pi ab$ **110** 16π

3.7 Review exercises

2 $\sin(x + 3y)$ **7** $\frac{1}{3}x^3 - y^2x + \frac{1}{2}x^2 - \frac{1}{2}y^2 + c$ **8** (a) $\frac{8}{3}$ (b) 4 **9** $\frac{5}{6}$

10 $I \frac{4}{3}j$

11 0

12 $\frac{13}{80}kc^6$

13 $\frac{2}{3}ha^2 \left[\frac{1}{2}\pi - \sin^{-1}\left(\frac{c}{a}\right) \right] - \frac{1}{3}hcl - \frac{hc^3}{3a} \tanh^{-1}\left(\frac{l}{a}\right)$

$$l = \sqrt{(a^2 - c^2)}$$

14 $\frac{16}{3}a^3$

15 $\pi q_0^2 r^2 l / 4EI$

16 $\frac{1}{3}$

17 0

19 0

20 $\frac{13}{240}$

CHAPTER 4

Exercises

1 (a) $y = \frac{5}{2}x + \frac{5}{4}$ (b) $y = \frac{1}{4}x - \frac{3}{4}$

2 $z = 2, \frac{1}{2}\pi$

3 $u = 6v$

6 Semi-infinite strip $v > 0, |u| < 1$

7 (a) $u = v\sqrt{3} - 4$

(b) $v = -u\sqrt{3}$

(c) $(u+1)^2 + (v-\sqrt{3})^2 = 4$

(d) $u^2 + v^2 = 8$

8 (a) $\alpha = \frac{1}{5}(-2+j), \beta = \frac{3}{5}(1+2j)$

(b) $u + 2v < 3$

(c) $(5u-3)^2 + (5v-6)^2 < 20$

(d) $\frac{3}{10}(1+3j)$

9 Interior of circle, centre $(0, -1/2c)$, radius $1/2c$; half-plane $v < 0$; region outside the circle, centre $(0, -1/2c)$, radius $1/2c$

10 Circle, centre $(\frac{1}{2}, -\frac{2}{3})$, radius $\frac{7}{6}$

11 $\operatorname{Re}(w) = 1/2a$, half-plane $\operatorname{Re}(w) > 1/2a$

12 $w = \frac{z+1}{jz-j}$,

$\operatorname{Re}(z) = \operatorname{const}(k)$ to circles

$$u^2 + \left(v - \frac{k}{1-k}\right)^2 = \frac{1}{(1-k)^2} \text{ plus } v = -1 \text{ (} k = 1 \text{)}$$

$\operatorname{Im}(z) = \operatorname{const}(l)$ to circles $\left(u + \frac{1}{l}\right)^2 + (v+1)^2 = \frac{1}{l^2}$
plus $u = 0$ ($l = 0$)

13 (a) $1+j, j, \infty$

(b) $|w| > \sqrt{2}$

(c) $v = 0, (u-1)^2 + v^2 = 1$

(d) $\pm 2^{1/4}e^{j\pi/8}$

14 Segment of the imaginary axis $|v| \geq 1$

15 (a) Upper segment of the circle, centre $(\frac{2}{3}, -\frac{2}{3})$, radius $\frac{1}{3}\sqrt{5}$, cut off by the line $u - 3v = 1$

16 Circle, centre $(\frac{5}{3}, 0)$, radius $\frac{4}{3}$

17 $z_0 = j, \theta_0 = \pi$

18 $|w-1| < 1; |w - \frac{4}{3}| > \frac{2}{3}$

19 $w = e^{j\theta_0} \frac{z - z_0}{z_0^* z - 1}$, where θ_0 is any real number

20 Region enclosed between the inverted parabola $v = 2 - (u^2/8)$ and the real axis

21 $u = 0, 2mu = (1-m^2)v$

23 $u = x + \frac{x}{x^2+y^2}, v = y - \frac{y}{x^2+y^2}; v = 0$; ellipses,
 $u^2 + v^2 = r^2$ and $x^2 + y^2 = r^2$, r large

24 (a) $e^z(z+1)$ (b) $4 \cos 4z$ (c) Not analytic
(d) $-2 \sin 2z$

25 $a = -1, b = 1$

$w = z^2 + jz^2, dw/dz = 2(1+j)z$

26 $v = 2y + x^2 - y^2$

27 $e^x(x \sin y + y \cos y), ze^z$

28 $\cos x \sinh y, \sin z$

29 (a) $x^4 - 6x^2y^2 + y^4 = \beta$
(b) $2e^{-x} \sin y + x^2 - y^2 = \beta$

30 (a) $(x^2 - y^2) \cos 2x - 2xy \sin 2y$
 $+ j[2xy \cos 2x + (x^2 - y^2) \sin 2y]$
(b) $\sin 2x \cosh 2y + j \cos 2x \sinh 2y$

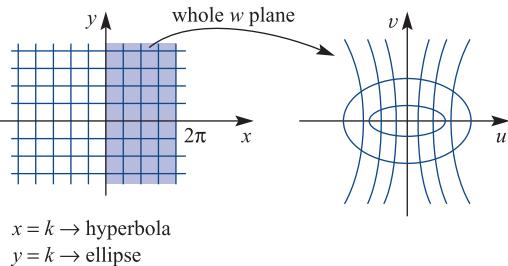
31 $u = \cos^{-1}\{2y^2[x^2 + y^2 - 1 + \sqrt{[(x^2 + y^2 - 1)^2 + 4y^2]}\}$
 $v = \sinh^{-1}\sqrt{\{\frac{1}{2}(x^2 + y^2 - 1) + \frac{1}{2}\sqrt{[(x^2 + y^2 - 1)^2 + 4y^2]}\}}$

33 (a) 0
(b) 3, 4
(c) $\frac{1}{2}, \frac{1}{4}(-1+j\sqrt{3}), \frac{1}{4}(-1-j\sqrt{3})$

34 $z = \pm j$

35 (a) Region outside unit circle
(b) $1 \leq u^2 + v^2 < e^2, 0 \leq v \leq u \tan 1$
(c) Outside unit circle, u and v of opposite sign

36



- 37 $4a$, ellipse centred at origin, semi axes are $\frac{a^2+b^2}{b}$
 and $\frac{|b^2-a^2|}{b}$

38 (a) $j + z - jz^2 - z^3 + jz^4 + \dots$

(b) $\frac{1}{z} + \frac{j}{z^2} - \frac{1}{z^3} - \frac{j}{z^4} + \frac{1}{z^5} + \dots$

(c) $1 - (z - 1 - j) + (z - 1 - j)^2 - (z - 1 - j)^3 + \dots$

39 (a) $1 - 2z^2 + 3z^4 - 4z^6 + \dots$

(b) $1 - 3z^2 + 6z^4 - 10z^6 + \dots$

40 (a) $\frac{1}{2} - \frac{1}{4}(z - 1) + \frac{1}{8}(z - 1)^2 - \frac{1}{16}(z - 1)^3; 2$

(b) $\frac{1}{4} - \frac{1}{16}(z - 2j)^2 + \frac{1}{64}(z - 2j)^4 - \frac{1}{256}(z - 2j)^6; 2$

(c) $-\frac{1}{2}j + \frac{1}{2}(1+j)(z - 1 - j) + \frac{3}{4}(z - 1 - j)^2 + \frac{1}{2}(j - 1)(z - 1 - j)^3; \sqrt{2}$

41 $1 - z + z^3 + \dots$

42 $1, 1, \sqrt{5}$; f is singular at $z = j$

43 $z + \frac{1}{3}z^2 + \frac{2}{15}z^5 + \dots; \frac{1}{2}\pi$

44 (a) $\frac{1}{z} + 2 + 3z + 4z^2 + \dots (0 < |z| < 1)$

(b) $\frac{1}{(z-1)^2} - \frac{1}{z-1} + 1 - (z-1) + (z-1)^2 - \dots (0 < |z-1| < 1)$

45 (a) $\dots + \frac{1}{5!z^3} - \frac{1}{3!z} + z$

(b) $z - \frac{1}{3!z} + \frac{1}{5!z^3} - \dots$

(c) $a^2 \sin \frac{1}{a} + zf'(a) + z^2 f''(a) + \dots$

46 (a) $\frac{1}{2}z + \frac{3}{4}z^2 + \frac{7}{8}z^3 + \frac{15}{16}z^4 + \dots$

(b) $\dots - \frac{1}{z^2} - \frac{1}{z} - 1 - \frac{1}{2}z - \frac{1}{4}z^2 - \frac{1}{8}z^3 - \dots$

(c) $\frac{1}{z} + \frac{3}{z^2} + \frac{7}{z^3} + \frac{15}{z^4} + \dots$

(d) $\frac{1}{z-1} + \frac{2}{(z-1)^2} + \frac{2}{(z-1)^3} + \dots$

(e) $-1 + \frac{2}{z-2} + (z-2) - (z-2)^2 + (z-2)^3 + (z-2)^4 - \dots$

- 47 (a) $z = 0$, double pole

- (b) $z = j$, simple pole; $z = -j$, double pole

- (c) $z = \pm 1, \pm j$, simple poles

- (d) $z = jn\pi$ (n an integer), simple poles

- (e) $z = \pm j\pi$, simple poles

- (f) $z = 1$, essential singularity

- (g) Simple zero at $z = 1$ and simple poles at $z = \pm j$

- (h) Simple zero at $z = -j$, simple pole at $z = 3$ and a pole of order 3 at $z = -2$

- (i) Simple poles at $z = 2 + j, 2 - j$ and a pole of order 2 at $z = 0$

48 (a) $\frac{z}{2!} - \frac{z^3}{4!} + \frac{z^5}{5!} - \dots$ (removable singularity)

(b) $\frac{1}{z^3} + \frac{1}{z} + \frac{z}{2!} + \frac{z^3}{3!} + \frac{z^5}{4!} + \frac{z^7}{5!} + \dots$ (pole of order 3)

(c) $\frac{1}{z} + \frac{1}{2!z^3} + \frac{1}{4!z^5} - \dots$ (essential singularity)

(d) $\tan^{-1} 2 + \frac{2}{5}z - \frac{6}{25}z^2 + \dots$ (analytic point)

49 $2axV_0/(x^2 + y^2)$

- 50 (a) $(0, 0), (0, 1), (0, 7), (7, 0)$

- (b) $v = 0$ (c) $u = 0$

51 $H(x, y) = 2y - y^2 + x^2;$

$W = 2z - jz^2$

- 52 (a) $(0, 0), (1, 0), (-1, 0)$

- (b) $u = 0$ (c) $v = 0$

4.8 Review exercises

1 (a) $3j$ (b) $7 + j4$ (c) 1 (d) $j2$

2 (a) $y = 2x$ gives $3u + v = 3$, $u + 2v = 3$ and $3v - u = 1$ respectively

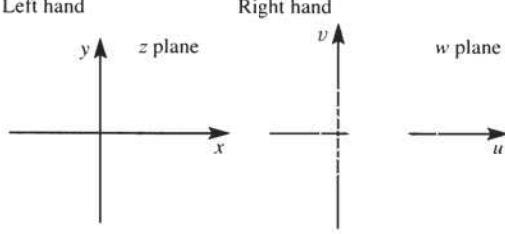
(b) $x + y = 1$ gives $v = 1$, $v - u = 3$ and $u = 1$ respectively

3 (a) $\alpha = -\frac{1}{5}(3 + j4)$, $\beta = 3 + j$

(b) $13 \leqslant 3u + 4v$

(c) $|w - 3 - j| \leqslant 1$ (d) $\frac{1}{4}(7 - j)$

4 (a) $u^2 + v^2 + u - v = 0$ (b) $u = 3v$
 (c) $u^2 + v^2 + u - 2v = 0$ (d) $4(u^2 + v^2) = u$

5 Left hand

$$x = k \rightarrow \left(u - \frac{k}{k-1}\right)^2 + v^2 = \frac{1}{(k-1)^2}$$

$$y = l \rightarrow (u-1)^2 + \left(v + \frac{1}{l}\right)^2 = \frac{1}{l^2}$$

Fixed points: $1 \pm \sqrt{2}$ **6** Fixed points $z = \pm\sqrt{2}/2$

$$r = 1 \Rightarrow u = 0$$

$$7 \quad u = x^3 - 3xy^2, v = 3x^2y - y^3$$

$$8 \quad (z \sin z) v = y \sin x \cosh y + x \cos x \sinh y$$

$$9 \quad w = 1/z$$

$$10 \quad \text{Ellipse is given by } \frac{x^2}{(R+a^2/4k)^2} + \frac{y^2}{(R-a^2/4k)^2} = 1$$

$$11 \quad \begin{aligned} &1 - z^3 + z^6 - z^9 + z^{12} - \dots; \\ &1 - 2z^3 + 3z^6 - 4z^9 + \dots \end{aligned}$$

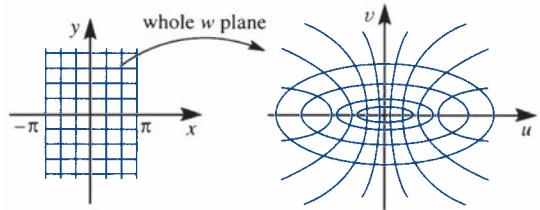
$$12 \quad \begin{aligned} &\text{(a)} \quad 1 - 2z + 2z^2 - 2z^3; \quad 1 \\ &\text{(b)} \quad \frac{1}{2} - \frac{1}{2}(z-1) + \frac{1}{4}(z-1)^2 - \frac{1}{6}(z-1)^4; \quad \sqrt{2} \\ &\text{(c)} \quad \frac{1}{2}(1+j) + \frac{1}{2}j(z-j) - \frac{1}{4}(1+j)(z-j)^2 - \frac{1}{8}(z-j)^3; \quad \sqrt{2} \end{aligned}$$

$$13 \quad 1, 1, 1, \frac{1}{2}\sqrt{5}, 2\sqrt{2} \text{ respectively}$$

$$14 \quad \begin{aligned} &\text{(a)} \quad \frac{1}{z} - z + z^3 - z^5 + \dots \quad 0 < |z| < 1 \\ &\text{(b)} \quad \frac{1}{2} - (z-1) + \frac{5}{4}(z-1)^2 + \dots \quad (|z-1| \sim 1) \end{aligned}$$

$$15 \quad \begin{aligned} &\text{(a)} \quad \text{Taylor series} \\ &\text{(b) and (c) are essential singularities, the principal parts are infinite} \end{aligned}$$

$$\begin{aligned} &\text{(a)} \quad \frac{1}{2}(e^{2x} \cos 2y - 1) + j\frac{1}{2}e^{2x} \sin 2y \\ &\text{(b)} \quad \cos 2x \cosh 2y - j \sin 2x \sinh 2y \\ &\text{(c)} \quad \underline{x \sin x \cosh y + y \cos x \sinh y + j(x \cos x \sinh y - y \sin x \cosh y)} \\ &\text{(d)} \quad \frac{\tan x(1 - \tanh^2 y) + j \tanh y(1 + \tan^2 x)}{1 + \tan^2 x \tanh^2 y} \end{aligned}$$

17 (a) Conformal (b) $j, -1-j$ (c) $\pm 0.465, \pm j0.465$ **18**

$$x = k \rightarrow \text{hyperbolas, } \frac{u^2}{\cos^2 k} - \frac{v^2}{\sin^2 k} = 1$$

$$y = l \rightarrow \text{ellipses, } \frac{u^2}{\cosh^2 l} + \frac{v^2}{\sinh^2 l} = 1$$

- 19** (a) Simple pole at $z = 0$
 (b) Double poles at $z = 2, 2e^{2\pi j/3}, 2e^{4\pi j/3}$
 (c) Simple poles at $z = +1, \pm j$, removable singularity at $z = -1$
 (d) Simple poles at $z = \frac{1}{2}(2n+1)\pi j$ ($n = 0, \pm 1, \pm 2, \dots$)
 (e) No singularities in finite plane (entire)
 (f) Essential singularity at $z = 0$
 (g) Essential (non-isolated) singularity at $z = 0$

CHAPTER 5

Exercises

$$1 \quad f(t) = tH(t) - tH(t-1)$$

$$2 \quad \begin{aligned} &\text{(a)} \quad f(t) = 3t^2 - [3(t-4)^2 + 22(t-4) + 43]H(t-4) \\ &\quad - [2(t-6) + 4]H(t-6) \end{aligned}$$

$$\mathcal{F}(s) = \frac{6}{s^3} - \left(\frac{6}{s^3} + \frac{22}{s^2} + \frac{43}{s} \right) e^{-4s} - \left(\frac{2}{s^3} + \frac{4}{s} \right) e^{-6s}$$

$$\begin{aligned} &\text{(b)} \quad f(t) = t - 2(t-1)H(t-1) + (t-2)H(t-2) \\ &\quad F(s) = \frac{1}{s^2} - \frac{2}{s^2} e^{-s} + \frac{1}{s^2} e^{-2s} \end{aligned}$$

- $$\begin{aligned} &\text{(a)} \quad \frac{1}{6}(t-5)^3 e^{2(t-5)} H(t-5) \\ &\text{(b)} \quad \frac{3}{2}[e^{-(t-2)} - e^{-3(t-2)}]H(t-2) \\ &\text{(c)} \quad [t - \cos(t-1) - \sin(t-1)]H(t-1) \\ &\text{(d)} \quad \frac{\sqrt{3}}{\sqrt[3]{3}} e^{-(t-\pi)/2} \{ \sqrt{3} \cos[\frac{1}{2}\sqrt{3}(t-\pi)] \\ &\quad + \sin[\frac{1}{2}\sqrt{3}(t-\pi)] \} H(t-\pi) \\ &\text{(e)} \quad H(t - \frac{4}{5}\pi) \cos 5t \\ &\text{(f)} \quad [t - \cos(t-1) - \sin(t-1)]H(t-1) \end{aligned}$$

4 $x(t) = e^{-t} + (t - 1)[1 - H(t - 1)]$

5 $x(t) = 2e^{-t/2} \cos(\frac{1}{2}\sqrt{3}t) + t - 1 - 2H(t - 1)$
 $\quad \{t - 2 + e^{-(t-1)/2} \{\cos[\frac{1}{2}\sqrt{3}(t-1)]\}$
 $\quad - \sqrt{\frac{1}{3}} \sin[\frac{1}{2}\sqrt{3}(t-1)]\}$
 $\quad + H(t-2)\{t-3 + e^{-(t-2)/2}$
 $\quad \{\cos[\frac{1}{2}\sqrt{3}(t-2)] - \sqrt{\frac{1}{3}} \sin[\frac{1}{2}\sqrt{3}(t-2)]\}\}$

6 $x(t) = e^{-t} + \frac{1}{10}(\sin t - 3\cos t + 4e^{\pi}e^{-2t}$
 $\quad - 5e^{\pi/2}e^{-t})H(t - \frac{1}{2}\pi)$

7 $f(t) = 3 + 2(t - 4)H(t - 4)$

$$F(s) = \frac{3}{s} + \frac{2}{s^2}e^{-4s}$$

$$x(t) = 3 - 2 \cos t + 2[t - 4 - \sin(t - 4)]H(t - 4)$$

8 $\theta_0(t) = \frac{3}{10}(1 - e^{-3t} \cos t - 3e^{-3t} \sin t)$
 $\quad - \frac{3}{10}[1 - e^{3a}e^{-3t} \cos(t-a) - 3e^{3a}e^{-3t} \sin(t-a)]H(t-a)$

9 $\theta_0(t) = \frac{1}{32}(3 - 2t - 3e^{-4t} - 10t e^{-4t})$
 $\quad + \frac{1}{32}[2t - 3 + (2t-1)e^{-4(t-1)}]H(t-1)$

11 $\frac{3 - 3e^{-2s} - 6se^{-4s}}{s^2(1 - e^{-4s})}$

12 $\frac{K}{T} \frac{1}{s^2} - \frac{K}{s} \frac{e^{-sT}}{1 - e^{-sT}}$

13 (a) $2\delta(t) + 9e^{-2t} - 19e^{-3t}$

(b) $\delta(t) - \frac{5}{2}\sin 2t$

(c) $\delta(t) - e^{-t}(2\cos 2t + \frac{1}{2}\sin 2t)$

14 (a) $x(t) = (\frac{1}{6} - \frac{2}{3}e^{-3t} + \frac{1}{2}e^{-4t})$
 $\quad + (e^{-3(t-2)} - e^{-4(t-2)})H(t-2)$

(b) $x(t) = \frac{1}{2}e^{6\pi}e^{-3t}H(t-2\pi)\sin 2t$

(c) $x(t) = 5e^{-3t} - 4e^{-4t} + (e^{-3(t-3)} - e^{-4(t-3)})H(t-3)$

15 (a) $f'(t) = g'(t) - 43\delta(t-4) - 4\delta(t-6)$

$$g'(t) = \begin{cases} 6t & (0 \leq t < 4) \\ 2 & (4 \leq t < 6) \\ 0 & (t \geq 6) \end{cases}$$

(b) $g'(t) = \begin{cases} 1 & (0 \leq t < 1) \\ -1 & (1 \leq t < 2) \\ 0 & (t \geq 2) \end{cases}$

(c) $f'(t) = g'(t) + 5\delta(t) - 6\delta(t-2) + 15\delta(t-4)$

$$g'(t) = \begin{cases} 2 & (0 \leq t < 2) \\ -3 & (0 \leq t < 4) \\ 2t-1 & (t \geq 4) \end{cases}$$

16 $x(t) = -\frac{19}{9}e^{-5t} + \frac{19}{9}e^{-2t} - \frac{4}{3}t e^{-2t}$

18 $q(t) = \frac{E}{Ln} e^{-\mu t} \sin nt, n^2 = \frac{1}{LC} - \frac{R^2}{4L^2}, \mu = \frac{R}{2L}$
 $i(t) = \frac{E}{Ln} e^{-\mu t} (n \cos nt - \mu \sin nt)$

19 $y(x) = \frac{1}{48EI} [2Mx^4/l + 8W(x - \frac{1}{2}l)^3 H(x - \frac{1}{2}l)$
 $\quad - 4(M+W)x^3 + (2M+3W)l^2 x]$

20 $y(x) = \frac{w(x_2^2 - x_1^2)x^2}{4EI} - \frac{w(x_2 - x_1)x^3}{6EI}$
 $\quad + \frac{w}{24EI} [(x - x_1)^4 H(x - x_1)$
 $\quad - (x - x_2)^4 H(x - x_2)]$
 $y_{\max} = wl^4/8EI$

21 $y(x) = \frac{W}{EI} [\frac{1}{6}x^3 - \frac{1}{6}(x-b)^3 H(x-b) - \frac{1}{2}bx^2]$
 $= \begin{cases} -\frac{Wx^2}{6EI}(3b-x) & (0 < x \leq b) \\ -\frac{Wb^2}{6EI}(3x-b) & (b < x \leq l) \end{cases}$

22 (a) $\frac{3s+2}{s^2+2s+5}$

(b) $s^2 + 2s + 5 = 0$, order 2

(c) Poles $-1 \pm j2$; zero $-\frac{2}{3}$

23 $\frac{s^2 + 5s + 6}{s^3 + 5s^2 + 17s + 13}, s^3 + 5s^2 + 17s + 13 = 0$
 order 3, zeros $-3, -2$, poles $-1, -2 \pm j3$

- 24 (a) Marginally stable (b) Unstable
 (c) Stable (d) Stable (e) Unstable

- 25 (a) Unstable
 (b) Stable
 (c) Marginally stable
 (d) Stable
 (e) Stable

28 $K > \frac{2}{3}$

29 (a) $3e^{-7t} - 3e^{-8t}$ (b) $\frac{1}{3}e^{-4t} \sin 3t$
 (c) $\frac{2}{3}(e^{4t} - e^{-2t})$ (d) $\frac{1}{3}e^{2t} \sin 3t$

30 $\frac{s+8}{(s+1)(s+2)(s+4)}$

33 $\frac{2}{7}, \frac{4}{5}$

35 (a) $\frac{1}{54}[2 - e^{-3t}(9t^2 + 6t + 2)]$
 (b) $\frac{1}{125}[e^{-3t}(5t + 2) + e^{2t}(5t - 2)]$
 (c) $\frac{1}{16}(4t - 1 + e^{-4t})$

37 $e^{-3t} - e^{-4t}$
 $x(t) = \frac{1}{12}A[1 - 4e^{-3t} + 3e^{-4t} - (1 - 4e^{-3(t-T)} + 3e^{-4(t-T)})H(t-T)]$

38 $e^{-2t}\sin t, \frac{1}{5}[1 - e^{-2t}(\cos t + 2\sin t)]$

39 $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -5 & -1 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 \\ 5 \end{bmatrix} u, y = [1 \quad 2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
 $G(s) = \frac{12s+59}{(s+2)(s+4)}$

40 $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -7 & 1 \\ -6 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u, y = [1 \quad 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

41 (a) $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -7 & -5 & -6 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u,$
 $y = [5 \quad 3 \quad 1]\mathbf{x}$

(b) $\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -3 & -4 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u, \quad y = [2 \quad 3 \quad 1]\mathbf{x}$

43 $\mathbf{x}(t) = \begin{bmatrix} -5 + \frac{8}{3}e^{-t} + \frac{10}{3}e^{5t} \\ 3 - \frac{8}{3}e^{-t} + \frac{5}{3}e^{5t} \end{bmatrix}$

44 $x_1 = x_2 = 2e^{-2t} - e^{-3t}$

45 $\mathbf{x}(t) = \begin{bmatrix} 4te^{-t} + e^{-2t} \\ -4te^{-t} - 2e^{-2t} + 2e^{-t} \end{bmatrix}$

46 $\dot{\mathbf{z}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \mathbf{z} + \begin{bmatrix} \frac{1}{2} \\ -1 \\ \frac{1}{2} \end{bmatrix} u, \quad y = [2 \quad 9 \quad 22]\mathbf{z}$

The system is stable, controllable and observable.

47 $\dot{\mathbf{z}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -5 \end{bmatrix} \mathbf{z} + \begin{bmatrix} \frac{1}{5} \\ -\frac{1}{4} \\ \frac{1}{20} \end{bmatrix} u, \quad y = [5 \quad 3 \quad 15]\mathbf{z}$

The system is marginally stable, controllable and observable.

48 $\frac{15}{4} - \frac{5}{2}t + \frac{9}{4}e^{-2t} - 6e^{-t}$

49 $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$

$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$

$G(s) = \frac{1}{(s+1)^2(s^2+1)} \begin{bmatrix} s^2+s+1 & s \\ s & s^2+s+1 \end{bmatrix}$

50 (a) $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 \\ -1 & -3 & -3 \\ -1 & -3 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} -1 & 1 \\ -1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$

$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$

(b) $G(s) = \frac{1}{\Delta} \begin{bmatrix} -2s(2s+3) & 2s(2s+3) \\ -s^2 & s^2 \end{bmatrix},$
 $\Delta = s^3 + 10s^2 + 16s + 6$

(c) $y_1(t) = 1 + 0.578e^{-8.12t} - 1.824e^{-0.56t} + 0.246e^{-1.32t}$
 $y_2(t) = 0.177e^{-8.12t} + 0.272e^{-0.56t} - 0.449e^{-1.32t}$

51 $u(t) = [-\frac{33}{2} \quad -\frac{17}{2}] \mathbf{x}(t) + u_{\text{ext}}$

52 $u(t) = [-\frac{99}{4} \quad -\frac{35}{4}] \mathbf{x}(t) + u_{\text{ext}}$

53 $u(t) = [-\frac{35}{6} \quad -\frac{31}{6}] \mathbf{x}(t) + u_{\text{ext}}$
 $u(t) = [-31 \quad -11] \mathbf{x}(t) + u_{\text{ext}}$

54 $\mathbf{M} = \begin{bmatrix} 2 & -2 \\ 1 & -1 \end{bmatrix}, \text{ rank } 1, \quad \mathbf{M} = \begin{bmatrix} 0 & 1 \\ 1 & \frac{1}{2} \end{bmatrix}, \text{ rank } 2$

5.7 Review exercises

1 (a) (ii) $e^{-(t-\alpha)}[\cos 2(t-\alpha) - \frac{1}{2}\sin 2(t-\alpha)]H(t-\alpha)$
 (b) $y(t) = \frac{1}{10}[e^{-t}(\cos 2t - \frac{1}{2}\sin 2t) + 2\sin t - \cos t] + \frac{1}{10}[e^{-(t-\pi)}(\cos 2t - \frac{1}{2}\sin 2t) + \cos t - 2\sin t]H(t-\pi)$

2 $i(t) = \frac{1}{250}[e^{-40t} - 2H(1 - \frac{1}{2}T)e^{-40(t-T/2)} + 2H(t-T)e^{-40(t-T)} - 2H(t - \frac{3}{2}T)e^{-40(t-3T/2)} + \dots]$

Yes, since time constant is large compared with T

3 $e^{-t} \sin t, \frac{1}{2}[1 - e^{-t}(\cos t + \sin t)]$

4 $EI \frac{d^4y}{dx^4} = 12 + 12H(x-4) - R\delta(x-4),$

$y(0) = y'(0) = y(4) = y^{(2)}(5) = y^{(3)}(0) = 0$

$$y(x) = \begin{cases} \frac{1}{2}x^4 - 4.25x^3 + 9x^2 & (0 \leq x \leq 4) \\ \frac{1}{2}x^4 - 4.25x^3 + 9x^2 + \frac{1}{2}(x-4)^4 - 7.75(x-4)^3 & (4 \leq x \leq 5) \end{cases}$$

25.5 kN, 18 kN m

5 (a) $f(t) = H(t-1) - H(t-2)$

$x(t) = H(t-1)(1 - e^{-(t-1)}) - H(t-2)(1 - e^{-(t-2)})$

(b) $0, E/R$

7 (a) $t - 2 + (t+2)e^{-t}$

(b) $y = t + 2 - 2e^t + 2te^t, y(t) = \frac{1}{2}t^2 + y_1$

8 $EIy = -\frac{2}{9}Wlx^2 + \frac{10}{81}Wx^3 - \frac{W(x-l)^3}{6}H(x-l)$

$EI \frac{d^4y}{dx^4} = -W\delta(x-l) - w[H(x) - H(x-l)]$

9 (a) $x(t) =$

$\frac{1}{6}\{1 + e^{3(t-a)/2}[\sqrt{3}\sin(\frac{1}{2}\sqrt{3}t) - \cos(\frac{1}{2}\sqrt{3}t)]H(t-a)\}$

10 (a) No (b) $\frac{1}{s^2 + 2s + (K-3)}$ (d) $K > 3$

11 (a) 4 (b) $\frac{1}{10}$

12 (c) $4e^{-2t} - 3e^{-3t}, y(t) = 1 (t \geq 0)$

13 $x(t) = \begin{bmatrix} e^{-t} \sin t \\ 1 - e^{-t}(\cos t + \sin t) \end{bmatrix}$

$H(s) = \frac{s+2}{(s+1)^2+1} \quad e^{-t}(\cos t + \sin t)$

14 1, -1, -2; $[1 \ 0 \ -1]^T, [1 \ -1 \ 0]^T, [0 \ 0 \ 1]^T$

$u(t) = -6\{x_1(t) + x_2(t)\}$

15 (a) $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -2 & 4 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u$

$y = [1 \ 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(b) $G(s) = \frac{s+3}{(s+2)(s-1)}$

16 (a) $\frac{K}{s^2 + (1+KK_1)s + K}$

(c) $K = 12.5, K_1 = 0.178$ (d) $0.65 \text{ s}, 2.48 \text{ s}, 1.86 \text{ s}$

17 (a) $K_2 = M_2\omega^2$

18 (b) Unstable (c) $\beta = 2.5 \times 10^{-5}, 92 \text{ dB}$

(d) $-8 \text{ dB}, 24^\circ$

(e) $K = 10^6, \tau_1 = 10^{-6}, \tau_2 = 10^{-7}, \tau_3 = 4 \times 10^{-8}$

(f) $s^3 + 36 \times 10^6 s^2 + 285 \times 10^{12} s + 25 \times 10^{18} (1 + 10^7 \beta) = 0$

CHAPTER 6

Exercises

1 (a) $\frac{4z}{4z-1}, |z| > \frac{1}{4}$ (b) $\frac{z}{z-3}, |z| > 3$

(c) $\frac{z}{z+2}, |z| > 2$ (d) $\frac{-z}{z-2}, |z| > 2$

(e) $3\frac{z}{(z-1)^2}, |z| > 1$

2 $e^{-2\omega kT} \leftrightarrow \frac{z}{z - e^{-2\omega T}}$

4 $\frac{1}{z^3} \frac{2z}{2z-1} = \frac{2}{z^2(2z-1)}$

5 (a) $\frac{5z}{5z+1}$ (b) $\frac{z}{z+1}$

6 $\frac{2z}{2z-1}, \frac{2z}{(2z-1)^2}$

8 (a) $\{e^{-4kT}\} \leftrightarrow \frac{z}{z - e^{-4T}}$

(b) $\{\sin kT\} \leftrightarrow \frac{z \sin T}{z^2 - 2z \cos T + 1}$

(c) $\{\cos 2kT\} \leftrightarrow \frac{z(z - \cos 2T)}{z^2 - 2z \cos 2T + 1}$

11 (a) 1 (b) $(-1)^k$ (c) $(\frac{1}{2})^k$ (d) $\frac{1}{3}(-\frac{1}{3})^k$

(e) j^k (f) $(-\sqrt{2})^k$

(g) 0 ($k=0$), 1 ($k > 0$)

(h) 1 ($k=0$), $(-1)^{k+1}$ ($k > 0$)

12 (a) $\frac{1}{3}[1 - (-2)^k]$ (b) $\frac{1}{7}[3^k - (-\frac{1}{2})^k]$

(c) $\frac{1}{3} + \frac{1}{6}(-\frac{1}{2})^k$ (d) $\frac{2}{3}(\frac{1}{2})^k + \frac{2}{3}(-1)^{k+1}$

(e) $\sin \frac{1}{2}k\pi$ (f) $2^k \sin \frac{1}{6}k\pi$

(g) $\frac{5}{2}k + \frac{1}{4}(1 - 3^k)$ (h) $k + 2\sqrt[3]{3} \cos(\frac{1}{3}k - \frac{3}{2}\pi)$

13 (a) $\{0, 1, 0, 0, 0, 0, 0, 2\}$

(b) $\{1, 0, 3, 0, 0, 0, 0, 0, -2\}$

(c) $\{5, 0, 0, 1, 3\}$ (d) $\{0, 0, 1, 1\} + \{(-\frac{1}{3})^k\}$

(e) $1(k=0), \frac{5}{2}(k=1), \frac{5}{4}(k=2), -\frac{1}{8}(-\frac{1}{2})^{k-3} (k \geq 3)$

(f) $\begin{cases} 0 & (k=0) \\ 3 - 2k + 2^{k-1} & (k \geq 1) \end{cases}$

(g) $\begin{cases} 0 & (k=0) \\ 2 - 2^{k-1} & (k \geq 1) \end{cases}$

14 $y_{k+2} + \frac{1}{2}y_{k+1} = x_k, y_{k+2} + \frac{1}{4}y_{k+1} - \frac{1}{5}y_k = x_k$

15 (a) $y_k = k$ (b) $y_k = \frac{3}{10}(9^k) + \frac{17}{10}(-1)^k$
 (c) $2^{k-1} \sin \frac{1}{2}k\pi$ (d) $2(-\frac{1}{2})^k + 3^k$

16 (a) $y_k = \frac{2}{5}(-\frac{1}{2})^k - \frac{9}{10}(\frac{1}{3})^k + \frac{1}{2}$
 (b) $y_k = \frac{7}{2}(3^k) - 6(2^k) + \frac{5}{2}$
 (c) $y_n = \frac{2}{5}(3^n) - \frac{2}{3}(2^n) + \frac{4}{15}(\frac{1}{2})^n$
 (d) $y_n = -2(\sqrt{3})^{n-1} \sin \frac{1}{6}n\pi + 1$
 (e) $y_n = -\frac{2}{5}(-\frac{1}{2})^n + \frac{12}{5}(2)^n - 2n - 1$
 (f) $y_n = -\frac{1}{2}[2^n + (-2)^n] + 1 - n$

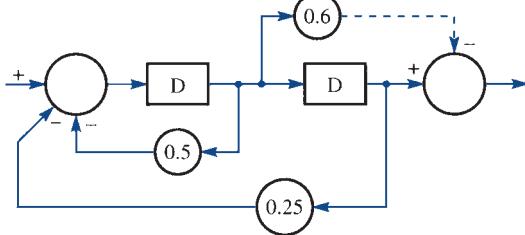
17 (b) 7, £4841

18 $y_k = 2^k - \frac{1}{2}(3^k) + \frac{1}{2}$

19 As $k \rightarrow \infty$, $I_k \rightarrow 2G$ as a damped oscillation

21 (a) $\frac{1}{z^2 - 3z + 2}$
 (b) $\frac{z - 1}{z^2 - 3z + 1}$
 (c) $\frac{z + 1}{z^3 - z^2 + 2z + 1}$

22



23 (a) $\frac{1}{2}\{(-\frac{1}{4})^k - (-\frac{1}{2})^k\}$ (b) $2(3^k) \sin \frac{1}{6}(k+1)\pi$
 (c) $\frac{2}{3}(0.4)^k + \frac{1}{3}(-0.2)^k$ (d) $4^{k+1} + 2^k$

24 $\begin{cases} 0 & (k=0) \\ 2^{k-1} - 1 & (k \geq 1) \end{cases}$
 $\begin{cases} 0 & (k=0) \\ 2^{k-1} & (k \geq 1) \end{cases}$

25 (a), (b) and (c) are stable; (d) is unstable;
 (e) is marginally stable

26 $2 - (\frac{1}{2})^k$

28 $y_n = -4(\frac{1}{2})^n + 2(\frac{1}{3})^n + 2(\frac{2}{3})^n$

30 (a) $\frac{2^k}{4} \begin{bmatrix} 2 & 1 \\ 4 & 2 \end{bmatrix} + \frac{(-2)^k}{4} \begin{bmatrix} 2 & -1 \\ -4 & 2 \end{bmatrix}$

(b) $\frac{(-4)^k}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + 2^{k-1} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$
 (c) $(-1)^k \begin{bmatrix} 1 & -k \\ 0 & 1 \end{bmatrix}$

31 $x(k) = 5^k(\cos k\theta + \sin k\theta), y(k) = 5^k(2 \cos k\theta),$
 $\cos \theta = -\frac{3}{5}$

32 $x(k) = \begin{bmatrix} \frac{25}{18} - \frac{17}{6}(-0.2)^k + \frac{22}{9}(-0.8)^k \\ \frac{7}{18} - (3.4/6)(-0.2)^k - (17.6/9)(-0.8)^k \end{bmatrix}$

33 $y(k) = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2} \right)^k - \left(\frac{1-\sqrt{5}}{2} \right)^k \right]$

34 $\begin{bmatrix} x_1[(k+1)T] \\ x_2[(k+1)T] \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{2}(1-e^{-2T}) \\ 0 & e^{-2T} \end{bmatrix} \begin{bmatrix} x_1(kT) \\ x_2(kT) \end{bmatrix} + \begin{bmatrix} \frac{1}{2}T + \frac{1}{4}(e^{-2T} - 1) \\ \frac{1}{2}(1 - e^{-2T}) \end{bmatrix} u(kT)$

35 (a) $\begin{bmatrix} x_1[(k+1)T] \\ x_2[(k+1)T] \end{bmatrix} = \begin{bmatrix} 1 & T \\ -T & 1-T \end{bmatrix} \begin{bmatrix} x_1(kT) \\ x_2(kT) \end{bmatrix} + \begin{bmatrix} 0 \\ T \end{bmatrix} u(kT)$
 $y(kT) = [1 \ 0] \mathbf{x}(kT)$
 (b) $\mathbf{x}[(k+1)T] = \mathbf{G}\mathbf{x}(kT) + \mathbf{H}u(kT)$
 $y(kT) = [1 \ 0] \mathbf{x}(kT)$

$$\mathbf{G} = e^{-T/2} \begin{bmatrix} \cos(\frac{\sqrt{3}}{2}T) + \frac{1}{\sqrt{3}}\sin(\frac{\sqrt{3}}{2}T) \\ -\frac{2}{\sqrt{3}}\sin(\frac{\sqrt{3}}{2}T) \\ \frac{2}{\sqrt{3}}\sin(\frac{\sqrt{3}}{2}T) \\ \cos(\frac{\sqrt{3}}{2}T) - \frac{1}{\sqrt{3}}\sin(\frac{\sqrt{3}}{2}T) \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 1 - e^{-T/2} \cos(\frac{\sqrt{3}}{2}T) - \frac{1}{\sqrt{3}}e^{-T/2} \sin(\frac{\sqrt{3}}{2}T) \\ \frac{2}{\sqrt{3}}e^{-T/2} \sin(\frac{\sqrt{3}}{2}T) \end{bmatrix}$$

37 (a) $\mathbf{x}(k+1) = \begin{bmatrix} 0.368 & 0 \\ 0.632 & 1 \end{bmatrix} \mathbf{x}(k) + \begin{bmatrix} 0.632k_1 & 0 \\ 0.368k_1 & -1 \end{bmatrix} \mathbf{u}(k)$
 (b) $\mathbf{x}(k+1) = \begin{bmatrix} 0.368 & -0.1185 \\ 0.632 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} 0.1185 & 0 \\ 0.069 & -1 \end{bmatrix} \begin{bmatrix} k_c \\ 1.1x_1(0) \end{bmatrix}$

(c) $x_1(t) = x_1(0)[1.1 - 2.15e^{-1/4t} + 2.05e^{-3/4t}]$
 $x_2(t) = k_c + x_1(0)[-5.867 + 8.6e^{-1/4t} - 2.71e^{-3/4t}]$

38 q form:

$$(Aq^2 + Bq + C)y_k = \Delta^2(q^2 + 2q + 1)u_k$$

 δ form:

$$[A\Delta^2\delta^2 + (2\Delta A + \Delta B)\delta + (A + B + C)]y_k$$

$$= \Delta^2(4 + 4\Delta\delta + \Delta^2\delta^2)u_k$$

$$A = 2\Delta^2 + 6\Delta + 4$$

$$B = 4\Delta^2 - 8$$

$$C = 2\Delta^2 - 6\Delta + 4$$

$$\frac{1}{s^3 + 2s^2 + 2s + 1}$$

$$[(\Delta^3 + 4\Delta^2 + 8\Delta + 8)\delta^3 + (6\Delta^2 + 16\Delta + 16)\delta^2 + (12\Delta + 16)\delta + 8]y_k = (2 + T\delta)^3 u_k$$

$$\frac{12(z^2 - z)}{(12 + 5\Delta)z^2 + (8\Delta - 12)z - \Delta}$$

$$\frac{12\gamma(1 + \Delta\gamma)}{\Delta(12 + 5\Delta)\gamma^2 + (8\Delta - 12)\gamma + 12}$$

6.12 Review exercises

4 $3 + 2k$

5 $\frac{1}{6} + \frac{1}{3}(-2)^k - \frac{1}{2}(-1)^k$

7 $\frac{2z}{(z - e)^{3T}} - \frac{z}{z - e^{-2T}}$

8 (a) $\left\{ \frac{1}{a-b}(a^n - b^n) \right\}$

(b) (i) $3^{k-1}k$ (ii) $2\sqrt[3]{3} \sin \frac{1}{3}k\pi$

9 $\frac{3}{2} - \frac{1}{2}(-1)^k - 2^k$

10 $(-1)^k$

13 $\frac{1}{2}A[2 - 2(\frac{1}{2})^n - n(\frac{1}{2})^{n-1}]$

17 $[1 \quad 3 \quad 1]^T, [3 \quad 2 \quad 1]^T, [1 \quad 0 \quad 1]^T$

$$\mathbf{x}(k) = \begin{bmatrix} -\frac{1}{6}(-1)^k - \frac{1}{3}(2^k) + \frac{3}{2} \\ 1 - 2^k \\ -\frac{1}{6}(-1)^k - \frac{1}{3}(2^k) + \frac{1}{2} \end{bmatrix}$$

18 $D(z) = \frac{z+3}{z^2 + 4z - 5}$ (i) $\mathbf{M}_c = \begin{bmatrix} 1 & -3 \\ 0 & -2 \end{bmatrix}$

(ii) $\mathbf{M}_c^{-1} = \begin{bmatrix} 1 & -\frac{3}{2} \\ 0 & -\frac{1}{2} \end{bmatrix}$ (iii) $\mathbf{v}^T = [0 \quad -\frac{1}{2}]$

(iv) $\mathbf{T} = \begin{bmatrix} 0 & -\frac{1}{2} \\ 1 & \frac{1}{2} \end{bmatrix}$ (v) $\mathbf{T}^{-1} = \begin{bmatrix} 1 & 1 \\ -2 & 0 \end{bmatrix}$

$\alpha = -5, \beta = 4$

CHAPTER 7

Exercises

2 $f(t) = -\frac{8}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos(2n-1)\pi t$

3 (a) $f(t) = \frac{2}{3} - \frac{1}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{n^2} \cos 2n\pi t$
 $+ \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin 2n\pi t$

(b) $f(t) = \frac{1}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin 2n\pi t$
 $+ \frac{2}{\pi} \sum_{n=1}^{\infty} \left[\frac{1}{2n-1} + \frac{4}{\pi^2(2n-1)^3} \right]$
 $\times \sin(2n-1)\pi t$

(c) $f(t) = \frac{2}{3} + \frac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^2} \cos n\pi t$

4 $f(t) = \frac{1}{6}\pi^2 - \sum_{n=1}^{\infty} \frac{1}{n^2} \cos 2nt$

$f(t) = \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^3} \sin(2n-1)t$

5 $f(x) = \frac{8a}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)^2} \sin \frac{(2n-1)\pi x}{l}$

6 $f(x) = \frac{2l}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)^2} \sin \frac{2(2n-1)\pi x}{l}$

7 $f(t) = \frac{1}{2} \sin t + \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{n(-1)^{n+1}}{4n^2 - 1} \sin 2nt$

8 $f(x) = -\frac{1}{2}A - \frac{4A}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos \frac{(2n-1)\pi x}{l}$

9 $T(x) = \frac{8KL^2}{\pi^3} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^3} \sin \frac{(2n-1)\pi x}{L}$

10 $f(t) = \frac{1}{2} + \frac{1}{2} \cos \pi t + \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{4n^2 - 1} \sin 2n\pi t$
 $- \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin(2n-1)\pi t$

12 (a) $\frac{1}{6}\pi^2 + \sum_{n=1}^{\infty} \frac{2}{n^2}(-1)^n \cos nt$
 $+ \sum_{n=1}^{\infty} \frac{1}{\pi} \left[-\frac{\pi^2}{n}(-1)^n + \frac{2}{n^3}(-1)^n - \frac{2}{n^3} \right] \sin nt$

(b) $a_n = 0$

$$\begin{aligned} b_n &= \frac{4}{n\pi} \left(\cos n\pi - \cos \frac{1}{2}n\pi \right) \\ &+ 2 \left(\frac{3\pi}{4n^2} \sin \frac{1}{2}n\pi - \frac{\pi^2}{8n} \cos \frac{1}{2}n\pi \right. \\ &\quad \left. + \frac{3}{n^3} \cos \frac{1}{2}n\pi - \frac{6}{\pi n^4} \sin \frac{1}{2}n\pi \right), \\ &\frac{1}{\pi} \left[\left(\frac{3}{2}\pi^2 - 16 \right) \sin t + \frac{1}{8}(32 + \pi^3 - 6\pi) \sin 2t \right. \\ &\quad \left. - \frac{1}{3} \left(\frac{32}{9} + \frac{1}{2}\pi^2 \right) \sin 3t + \dots \right] \end{aligned}$$

(c) $-\frac{4}{\pi^2} \sum_{n=1}^{\infty} \frac{\cos(2n-1)\pi t}{(2n-1)^2} + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{\sin(2n-1)t}{(2n-1)}$

(d) $\frac{1}{4} + \frac{2}{\pi^2} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos 2(2n-1)\pi t$

13 $e(t) = 5 + \frac{20}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin(2n-1)100\pi t$

$$i_{ss}(t) \approx 0.008 \cos(100\pi t - 1.96) + 0.005 \cos(300\pi t - 0.33)$$

14 $f(t) = \frac{400}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin(2n-1)t$

$$x_{ss}(t) \approx 0.14 \sin(\pi t - 0.1) + 0.379 \sin(3\pi t - 2.415) + 0.017 \sin(5\pi t - 2.83)$$

15 $f(t) = \frac{100}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} \sin 2\pi nt$

$$x_{ss}(t) \approx 0.044 \sin(2\pi t - 3.13) - 0.0052 \sin(4\pi t - 3.14)$$

16 $e(t) = \frac{100}{\pi} + 50 \sin 50\pi t - \frac{200}{\pi} \sum_{n=1}^{\infty} \frac{\cos 100\pi nt}{4n^2 - 1}$

$$i_{ss}(t) \approx 0.78 \cos(50\pi t + (-0.17)) - 0.01 \sin(100\pi t + (-0.48))$$

18 $f(t) = \frac{1}{2} + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{j}{2n\pi} [(-1)^n - 1] e^{jn\pi t/2}$

19 (a) $\frac{3}{4}\pi + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{1}{2\pi} \left\{ \frac{j\pi}{n} - \frac{1}{n^2} [1 + (-1)^n] \right\} e^{jnt}$

(b) $\frac{a}{2} \sin \omega t - \sum_{n=-\infty}^{\infty} \frac{a}{2\pi(n^2 - 1)} [(-1)^n + 1] e^{jn\omega t}$

(c) $\frac{3}{2} + \sum_{\substack{n=-\infty \\ n \neq 0}}^{\infty} \frac{j}{2\pi n} [1 - (-1)^n] e^{jnt}$

(d) $\frac{2}{\pi} \sum_{n=-\infty}^{\infty} \frac{1}{1 - 4n^2} e^{2jn\omega t}$

21 (b) (i) 17.74, (ii) 17.95

(c) 18.14; (i) 2.20%, (ii) 1.05%

22 (a) $c_0 = 15, c_n = \frac{30}{jn\pi} (1 - e^{-jn\pi/2})$

$$15, \frac{30}{\pi}(1-j), -\frac{30}{\pi}j, -\frac{10}{\pi}(1+j), 0, \frac{6}{\pi}(1-j)$$

(b) 15 W, 24.30 W, 12.16 W, 2.70 W, 0.97 W

(c) 60 W

(d) 91.9%

23 0.19, 0.10, 0.0675

24 (c) $c_0 = 0, c_1 = \frac{3}{2}, c_2 = 0, c_3 = -\frac{7}{8}$

25 (c) $c_0 = \frac{1}{4}, c_1 = \frac{1}{2}, c_2 = \frac{5}{16}, c_3 = 0$

26 (b) $c_0 = 0, c_1 = \sqrt{(2\pi)}, c_2 = 0, \text{MSE} = 0$

7.7 Review exercises

$$\begin{aligned} \mathbf{1} \quad f(t) &= \frac{1}{6}\pi^2 + \sum_{n=1}^{\infty} \frac{2}{n^2} (-1)^n \cos nt \\ &+ \sum_{n=1}^{\infty} \left[\frac{\pi}{2n-1} - \frac{4}{\pi(2n-1)^3} \right] \sin(2n-1)t \\ &- \sum_{n=1}^{\infty} \frac{\pi}{2n} \sin 2nt \end{aligned}$$

Taking $T = \pi$ gives the required sum.

2 $f(t) = \frac{1}{9}\pi + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1}{n^2} \left\{ \cos \frac{1}{3}n\pi - \frac{1}{3}[2 + (-1)^n] \right\} \cos nt, \quad \frac{2}{9}\pi$

3 (a) $f(t) = \frac{2T}{\pi^2} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)^2} \sin \frac{2(2n-1)\pi t}{T}$

(b) $-\frac{1}{4}T$

(c) Taking $t = \frac{1}{4}T$ gives $S = \frac{1}{8}\pi^2$

5 $f(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^n \sin(2n-1)t}{(2n-1)^2}$

8 $f(x) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{(2n-1)^2} \sin(2n-1)x$

$$f(x) = \frac{1}{4}\pi - \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{\cos 2(2n-1)x}{(2n-1)^2}$$

10 (a) $f(t) = \sum_{\substack{n=1 \\ n \neq 0}}^{\infty} \frac{2}{n} \sin nt$

(b) $f(t) = \frac{1}{2}\pi + \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos(2n-1)t$

13 (a) $f(t) = \frac{1}{2}\pi - \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^2} \cos((2n-1)t)$

(b) $g(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin((2n-1)t)$

15 (a) $v(t) = \frac{10}{\pi} + 5 \sin \frac{2\pi t}{T} - \frac{20}{\pi} \sum_{n=1}^{\infty} \frac{1}{4n^2-1} \cos \frac{4n\pi t}{T}$
 (b) 2.5 W, 9.01%

16 (b) $g(t) = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin((2n-1)t)$
 $f(t) = 1 + g(t)$

18 (b) $\frac{\sin \omega t - \omega \cos \omega t}{1 + \omega^2} = \frac{4}{\pi} \sum_{n=1}^{\infty} \frac{\sin \alpha t - \alpha \cos \alpha t}{(2n-1)(1 + \alpha^2)}$
 $\alpha = (4n-2)\pi/T$

19 (c) $T_0 = 1, T_1 = t_1, T_2 = 2t^2 - 1, T_3 = 4t^3 - 3t$

(d) $\frac{1}{16}T_5 - \frac{5}{8}T_4 + \frac{33}{16}T_3 - \frac{5}{2}T_2 + \frac{95}{5}T_1 - \frac{79}{8}T_0$

(e) $\frac{33}{4}t^3 - 5t^2 + \frac{91}{16}t - \frac{59}{8}, t = -1$

CHAPTER 8

Exercises

1 $\frac{2a}{a^2 + \omega^2}$

2 $AT^2 j\omega \operatorname{sinc}^2 \frac{\omega T}{2}$

3 $AT \operatorname{sinc}^2 \frac{\omega t}{2}$

4 $8K \operatorname{sinc} 2\omega, 2K \operatorname{sinc} \omega, 2K(4 \operatorname{sinc} 2\omega - \operatorname{sinc} \omega)$

5 $4 \operatorname{sinc} \omega - 4 \operatorname{sinc} 2\omega$

7 $\frac{\omega_0}{(a + j\omega)^2 + \omega_0^2}$

10 $F_s = \frac{x}{x^2 + a^2}, F_c = \frac{x}{x^2 + a^2}$

12 $\frac{1}{(1 - \omega^2) + 3j\omega}$

13 $4 \operatorname{sinc} 2\omega - 2 \operatorname{sinc} \omega$

14 $\frac{1}{2}T [\operatorname{sinc} \frac{1}{2}(\omega_0 - \omega)T + \operatorname{sinc} \frac{1}{2}(\omega_0 + \omega)T]$

15 $\frac{1}{2}Te^{-j\omega_0 T/2} [e^{j\omega_0 T/2} \operatorname{sinc} \frac{1}{2}(\omega - \omega_0)T + e^{-j\omega_0 T/2} \operatorname{sinc} \frac{1}{2}(\omega + \omega_0)T]$

16 $j[\operatorname{sinc}(\omega + 2) - \operatorname{sinc}(\omega - 2)]$

18 $4AT \cos \omega \tau \operatorname{sinc} \omega T$

19 High-pass filter

20 $\pi e^{-a|\omega|}$

21 $T[\operatorname{sinc}(\omega - \omega_0)T + \operatorname{sinc}(\omega + \omega_0)T]$

26 $\frac{1}{2}\pi j[\delta(\omega + \omega_0) - \delta(\omega - \omega_0)] - \frac{\omega_0}{\omega_0^2 - \omega^2}$

28 {2, 0, 2, 0}

29 {2, 0, 2, 0}

32 $D(z) = 0.06366 - 0.10660z^{-2} + 0.31831z^{-4} + 0.5z^{-5} + 0.31831z^{-6} - 0.010660z^{-8} + 0.06366z^{-10}$

33 $D(z) = 0.00509(1 + z^{-10}) - 0.04221(z^{-2} + z^{-8}) + 0.29035(z^{-4} + z^{-6}) + 0.5z^{-5}$

8.10 Review exercises

1 $\frac{\sin \omega}{\omega^2} - \frac{\cos 2\omega}{\omega}$

2 $-\frac{\pi j}{\omega} \operatorname{sinc} 2\omega$

7 (a) $\frac{1}{a-b}(e^{at} - e^{bt})H(t)$
 (b) (i) $te^{2t}H(t)$ (ii) $(t-1+e^{-t})H(t)$

8 (a) $-\sin \omega_0(t + \frac{1}{4}\pi)$ (b) $\cos \omega_0 t$
 (c) $je^{j\omega_0 t}$ (d) $-je^{-j\omega_0 t}$

17 (a) $\frac{a + 2\pi s}{a^2 + 4\pi^2 s^2}$
 (b) $\frac{1}{2\pi s}(\sin 2\pi sT - \cos 2\pi sT + 1)$

CHAPTER 9

Exercises

1 $a^2 = b^2 c^2$

2 $\alpha = \pm c$

5 For $\alpha = 0$: $V = A + Bx$

For $\alpha > 0$:

$V = A \sinh at + B \cosh at$, where $a^2 = \alpha/\kappa$
 or $C e^{at} + D e^{-at}$

For $\alpha < 0$:

$V = A \cos bt + B \sin bt$, where $b^2 = -\alpha/\kappa$

6 $n = -3, 2$

8 $a = -3$

10 (a) $I_{xx} = (Lc)I_n$

(b) $v_{xx} = (rg)v + (rc)v_t$, and $(rc)W_t = W_{xx}$ (c) $w_{xx} = (Lc)w_n$

12 $g(z) = (1 + 2z)/(1 + z)^4$

15 $u = \sin x \cos ct$

16 $u = \frac{1}{c}(\sin x \sin ct + \frac{1}{4} \sin 2x \sin 2ct)$

19 $u = \frac{2l}{\pi^2} \left\{ \left[\sin \frac{\pi(x-ct)}{l} - \frac{1}{9} \sin \frac{3\pi(x-ct)}{l} + \frac{1}{25} \sin \frac{5\pi(x-ct)}{l} + \dots \right] + \left[\sin \frac{\pi(x+ct)}{l} - \frac{1}{9} \sin \frac{3\pi(x+ct)}{l} + \frac{1}{25} \sin \frac{5\pi(x+ct)}{l} + \dots \right] \right\}$

20 $u = \frac{1}{4c} [\exp\{-(x-ct)^2\} - \exp\{-(x+ct)^2\}]$

21 $u = \frac{1}{2} F(x-ct) + \frac{1}{2} F(x+ct)$, where

$$F(z) = \begin{cases} 1-z & (0 \leq z \leq 1) \\ 1+z & (-1 \leq z \leq 0) \\ 0 & (|z| \geq 1) \end{cases}$$

22 $x + (-3 - \sqrt{6})y = \text{constant}$

and

$x + (-3 + \sqrt{6})y = \text{constant}$

23 $u = \frac{1}{5} [4(x+2t)^2 + (x-3t)^2 - 5]$

25

x	$f(x)$	$u(x, 0)$	$u(x, 0.5)$	$u(x, 1)$	$u(x, 1.5)$	$u(x, 2)$
-3.0	0.024893	0	0.025943	0.058509	0.106010	0.180570
-2.5	0.041042	0	0.042774	0.096466	0.174781	0.297710
-2.0	0.067667	0	0.070522	0.159046	0.288166	0.490842
-1.5	0.111565	0	0.116272	0.262222	0.475106	0.681635
-1.0	0.183939	0	0.191700	0.432332	0.655692	0.791166
-0.5	0.303265	0	0.316060	0.585169	0.748392	0.847392
0	0.5	0	0.393469	0.632120	0.776869	0.864664
0.5	0.696734	0	0.316060	0.585169	0.748392	0.847392
1.0	0.816060	0	0.191700	0.432332	0.655692	0.791166
1.5	0.888434	0	0.116272	0.262222	0.475106	0.681635
2.0	0.932332	0	0.070522	0.159046	0.288166	0.490842
2.5	0.958957	0	0.042774	0.096466	0.174781	0.297710
3.0	0.975106	0	0.025943	0.058509	0.106010	0.180570

26 $u = \frac{8}{\pi} \sum_{n=1}^{\infty} \frac{1}{(2n-1)^3} \sin(2n-1)x \cos(2n-1)ct$

28 Explicit with $\lambda = 0.5$

x	$t = 0$	$t = 0.25$	$t = 0.5$	$t = 1$	$t = 1.5$
0	0	0	0	0	0
0.25	0	0.0625	0.125	0.179687	0.210937
0.50	0	0.125	0.21875	0.265625	0.269531
0.75	0	0.0625	0.125	0.179687	0.210937
1.00	0	0	0	0	0

Implicit with $\lambda = 0.5$

x	$t = 0$	$t = 0.25$	$t = 0.5$	$t = 1$
0	0	0	0	0
0.25	0	0.0625	0.12245	0.17407
0.50	0	0.125	0.22449	0.2815
0.75	0	0.0625	0.12245	0.17407
1.00	0	0	0	0

29 Explicit

x	$t = 0$	$t = 0.02$	$t = 0.04$	$t = 0.06$	$t = 0.08$
0	0	0.031410	0.062790	0.094108	0.125333
0.2	0	0	0.000314	0.001249	0.003101
0.4	0	0	0	0.000003	0.000018
0.6	0	0	0	0	0.000000
0.8	0	0	0	0	0
1.0	0	0	0	0	0

30 Explicit

x	$t = 0$	$t = 0.2$	$t = 0.4$	$t = 0.6$
0	0	0	0	0
0.2	0.16	0.19	0.2725	0.38875
0.4	0.24	0.27	0.36	0.508125
0.6	0.24	0.27	0.36	0.508125
0.8	0.16	0.19	0.2725	0.38875
1.0	0	0	0	0

Implicit (symmetric as in the explicit case)

x	$t = 0$	$t = 0.2$	$t = 0.4$	$t = 0.6$
0	0	0	0	0
0.2	0.16	0.19	0.2319	0.2785
0.4	0.24	0.27	0.3191	0.3849

31 Explicit

x	$t = 0$	$t = 0.2$	$t = 0.4$	$t = 0.6$
0	0	0.03	0.12	0.27
0.2	0.16	0.19	0.28	0.43
0.4	0.24	0.27	0.36	0.51
0.6	0.24	0.27	0.36	0.51
0.8	0.16	0.19	0.28	0.43
1.0	0	0.03	0.12	0.27

Implicit (symmetric as in the explicit case)

x	$t = 0$	$t = 0.2$	$t = 0.4$	$t = 0.6$
0	1	0.03	0.08	0.1495
0.2	0.16	0.19	0.24	0.3099
0.4	0.24	0.27	0.32	0.39

32 $u = \frac{1}{2}a[\exp(-\frac{9}{4}\kappa\pi^2t)\cos\frac{3}{2}\pi x + \exp(-\frac{1}{4}\kappa\pi^2t)\cos\frac{1}{2}\pi x]$

33 $A_N = 2/\pi N$

34 $\alpha = -\frac{1}{2}$, $\kappa = -\frac{1}{4}$

35 $\beta = 2$, $u = -u_0 e^{-x} \sin(x - 2t)$

36 The term represents heat loss at a rate proportional to the excess temperature over θ_0 .

37 $u = \sum_{n=0}^{\infty} a_n \exp\left[\frac{-\kappa(n + \frac{1}{2})^2\pi^2 t}{l^2}\right] \cos\left[\left(n + \frac{1}{2}\right)\frac{\pi x}{l}\right]$

where

$$a_n = u_0 \left[\frac{8}{(2n+1)^2\pi^2} - \frac{2(-1)^n}{(2n+1)\pi} \right]$$

39 $u(0, t) = u(l, t) = 0$ for all t

$u(x, 0) = 10$ for $0 < x < l$

x	$t = 0$	$t = 0.02$	$t = 0.04$	$t = 0.06$	$t = 0.08$	$t = 0.1$
0	0	0	0	0	0	0
0.2	0.04	0.08	0.1	0.12	0.135	0.1475
0.4	0.16	0.2	0.24	0.27	0.295	0.315
0.6	0.36	0.4	0.44	0.47	0.495	0.515
0.8	0.64	0.68	0.7	0.72	0.735	0.7475
1.0	1	1	1	1	1	1

42 At $t = 1$ with $\lambda = 0.4$ and $\Delta t = 0.05$

Explicit

x	0	0.2	0.4	0.6	0.8	1.0
u	0	0.1094	0.2104	0.2939	0.3497	0.3679

Implicit

x	0	0.2	0.4	0.6	0.8	1.0
u	0	0.1082	0.2095	0.2954	0.3551	0.3679

43

x	$t = 0$	$t = 0.02$	$t = 0.04$	t large
0	0	-0.04	-0.0799	$\rightarrow -1$
0.2	0.16	0.12	0.0803	$\rightarrow -0.8$
0.4	0.24	0.2002	0.1613	$\rightarrow -0.6$
0.6	0.24	0.2012	0.1657	$\rightarrow -0.4$
0.8	0.16	0.1269	0.1034	$\rightarrow -0.2$
1	0	0	0	$\rightarrow 0$

44 $u = \frac{5}{8}e^{-\pi y} \sin \pi x - \frac{5}{16}e^{-3\pi y} \sin 3\pi x + \frac{1}{16}e^{-5\pi y} \sin 5\pi x$

46 $\phi = x^2y + \sin \pi x \frac{\sinh \pi y}{\sinh \pi}$

47 $u(r, \theta) = \frac{3}{4}\frac{r}{a} \sin \theta - \frac{1}{4}\left(\frac{r}{a}\right)^3 \sin(3\theta)$

48 $v = \text{const}$ gives circles with centre $\left(\frac{-v}{v-1}, 0\right)$ and radius $1/|v-1|$

$u = \text{const}$ gives circles with centre $\left(1, \frac{-1}{u}\right)$ and radius $1/|u|$

50 $u = x + \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{\sin n\pi x}{n \sinh n\pi} \times \{\sinh n\pi y + (-1)^n \sinh n\pi(1-y)\}$

51 Boundary conditions are $u(0, y) = u(a, y) = 0$, $0 \leq y < a$; and $u(x, 0) = 0$, $0 \leq x \leq a$, $u(x, a) = u_0$, $0 < x < a$

54 $V = \frac{2\alpha}{3} \frac{\left(1 - \frac{a}{r}\right)}{\left(1 - \frac{a}{b}\right)} - \frac{\alpha}{3} (2 - 3 \sin^2 \theta) \frac{\left(r^2 - \frac{a^5}{r^3}\right)}{\left(b^2 - \frac{a^5}{b^3}\right)}$

55 For $\Delta x = \Delta y = 0.5$ $u(0.5, 0.5) = 0.3125$
For $\Delta x = \Delta y = 0.25$ $u(0.5, 0.5) = 0.3047$

56 At two sample points

For $\Delta x = \Delta y = \frac{1}{2}$, $u(0.5, 0.5) = 0.6429$ and $u(0.5, 1) = 0.5714$

For $\Delta x = \Delta y = \frac{1}{4}$, $u(0.5, 0.5) = 0.6379$ and $u(0.5, 1) = 0.5602$

57 $u(1, 1) = 10.674$, $u(2, 1) = 12.360$, $u(3, 1) = 8.090$, $u(1, 2) = 10.337$, $u(2, 2) = 10.674$

58 $h = 1/2$ gives $\phi(0.5, 0.5) = 1.8611$ and

$$\phi(0.5, 1) = 1.3194$$

For $h = 1/4$ ϕ is given in the table

y	2	1.6015666	1.2867647	1.0565216	1
x	0	0.25	0.5	0.75	1
1	2.4375	1.9679551	1.5818015	1.2572287	1
0.75	2.75	2.2665772	1.8465073	1.4374669	1
0.5	2.9375	2.5174715	2.1314338	1.6930064	1
0.25	3	2.75	2.5	2.25	1

59 $\phi(0, 0) = 1.5909$, $\phi(0, \frac{1}{3}) = 2.0909$, $\phi(0, \frac{2}{3}) = 4.7727$, $\phi(\frac{1}{3}, 0) = 1.0909$, $\phi(\frac{2}{3}, 0) = 0.7727$ and other values can be obtained by symmetry.

60 (a) $u_1 = 1/35$, $u_2 = 6/35$

(b) $u_1 = 0.1024$, $u_2 = 0.0208$, $u_3 = 0.2920$, $u_4 = 0.2920$, $u_5 = 0.0208$

61 Has the same solution as Exercise 57.

62 $u(0, 0) = 1.6818$, $u(0, \frac{1}{3}) = 2.2485$, $u(0, \frac{2}{3}) = 5.3121$, $u(\frac{1}{3}, 0) = 1.1152$, $u(\frac{2}{3}, 0) = 0.7727$ and other values by symmetry. Compare with Exercise 59.

63 $T(r, \theta) = \frac{T_0}{\pi} \left[\tan^{-1} \left(\frac{\alpha+r}{\alpha-r} \tan \frac{\theta}{2} \right) + \tan^{-1} \left(\frac{\alpha+r}{\alpha-r} \cot \frac{\theta}{2} \right) \right]$

64 $\frac{2}{\pi} \tan^{-1} \left(\frac{x_0}{y_0} \right)$

66 $T(x, y, z) = \frac{q_L}{4\rho c \kappa \pi} \left[\sinh^{-1} \left(\frac{z+L}{x-a^2+y^2} \right) - \sinh^{-1} \left(\frac{z-L}{x-a^2+y^2} \right) - \frac{q_L}{4\rho c \kappa \pi} \left[\sinh^{-1} \left(\frac{z+L}{x+a^2+y^2} \right) - \sinh^{-1} \left(\frac{z-L}{x+a^2+y^2} \right) \right] \right]$

68 Parabolic; $r = x - y$ and $s = x + y$ gives $u_{ss} = 0$

Elliptic; $r = -3x + y$ and $s = x + y$ gives

$$8(u_{ss} + u_{rr}) - 9u_r + 3u_s + u = 0$$

Hyperbolic; $r = 9x + y$ and $s = x + y$ gives $49u_{rr} - u_{ss} = 0$

69 $u = f(2x + y) + g(x - 3y)$

71 (a) Elliptic

(b) Parabolic

(c) Hyperbolic

For $y < 0$ characteristics are $(-y)^{3/2} \pm \frac{3}{2}x = \text{constant}$

72 Elliptic if $|y| < 1$;
parabolic if $x = 0$ or $y = \pm 1$;
hyperbolic if $|y| > 1$

73 $p > q$ or $p < -q$ then hyperbolic; $p = q$ then parabolic;
 $-q < p < q$ then elliptic

9.11 Review exercises

3 $y = 4 e^{-t/2\tau} \sin \left(3\pi \frac{x}{l} \right) \left(\cos \omega_3 t + \frac{1}{2\omega_3 \tau} \sin \omega_3 t \right)$

where $\omega_3 = 3\pi \frac{c}{l} \sqrt{1 - \frac{l^2}{36\pi^2 c^2 \tau^2}}$

5 $A_{2n+1} = 8\theta_0 l / \pi^2 (2n+1)^2$

6 $T = T_0 + \phi_0 [1 - \text{erf}(x/2\sqrt{\kappa t})]$

7 Explicit

x	$t = 0$	$t = 0.004$	$t = 0.008$
0	1.0000	0.9600	0.9296
0.2	1.0000	1.0000	0.9960
0.4	1.0000	1.0000	1.0000
0.6	1.0000	1.0000	1.0000
0.8	1.0000	1.0000	0.9960
1.0	1.0000	0.9600	0.9296

Implicit

x	$t = 0$	$t = 0.004$	$t = 0.008$
0	1.0000	0.9641	0.9354
0.2	1.0000	0.9984	0.9941
0.4	1.0000	0.9999	0.9996
0.6	1.0000	0.9999	0.9996
0.8	1.0000	0.9984	0.9941
1.0	1.0000	0.9641	0.9354

$y = 1$	$y = 0.5$	$y = 1$	$x = 0$	$x = 0.5$	$x = 1$	$x = 1.5$
			1	0.9875743	0.9569621	0.9379995
		1	0.9849808	0.9647746	0.9601934	

11 $k = -\frac{3}{2}$

12 $z = x - y$, valid in the region $x \geq y$

14 $A_{2n+1} = \frac{32a^2(-1)^{n+1}}{\pi^3 (2n+1)^3 \cosh \left[\frac{(2n+1)\pi b}{2a} \right]}$

15 $u(x, t) = \frac{2}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin n\pi x \cos n\pi t$

17 $\phi = A \cos(px)e^{-Kt/2} \cos \omega t$, where $\omega^2 = c^2 p^2 - \frac{1}{4} K^2$

18 On $r = a$, $v_r = 0$, so there is no flow through the cylinder $r = a$. As $r \rightarrow \infty$, $v_r \rightarrow U \cos \theta$ and $v_\theta \rightarrow -U \sin \theta$, so the flow is steady at infinity and parallel to the x axis.

CHAPTER 10

Exercises

1 $x = 1, y = 1, f = 9$

3 Original problem:

20 of type 1, 50 of type 2, profit = £1080, 70 m chipboard remain

Revised problem:

5 of type 1, 75 of type 2, profit = £1020, 5 m chipboard remain

4 4 kg nails, 2 kg screws, profit 14 p

5 9 of CYL1, 6 of CYL2 and profit £54

6 LP solution gives $x_1 = 66.67, x_2 = 50, f = £3166.67$. Profit is improved if more cloth is bought, up to a maximum when the amount of cloth is increased to 600 m then $x_1 = 0, x_2 = 150$ and $f = £4500$.

7 For $k \geq 60$: $x_1 = \frac{25}{3}, x_2 = 0, z = \frac{25}{3}k$

For $60 \geq k \geq 10$: $x_1 = 6, x_2 = 7, z = 140 + 6k$

For $k \leq 10$: $x_1 = 0, x_2 = 10, z = 200$

8 $x_1 = 1, x_2 = 0.5, x_3 = 1, x_4 = 0, f = 6.5$

9 B1, 0; B2, 15 000; B3, 30 000; profit £21 000

10 Long range 15, medium range 0, short range 0, estimated profit £6 million

11 Many solutions of the form $x_1 = 1.5 - 1.5t, x_2 = 0, x_3 = 2.5 - 1.5t, x_4 = 3t$ where $0 \leq t \leq 1$ giving $f = 14$

12 $x = 1, y = 4, f = 9$

13 $x_1 = 1, x_2 = 10, f = 20$

14 Boots 50, shoes 150, profit £1150

15 B1, 0; B2, 10 000; B3, 40 000; profit is down to £20 000

16 $x = 3, y = 0, z = \frac{4}{3}, f = \frac{13}{3}$

17 $x_1 = 2, x_2 = 0, x_3 = 2, x_4 = 0, f = 12$

18 36.63% of A, 44.55% of B, 18.81% of C, profit per 100 litres £134.26

19 6 of style 1, 11 of style 2, 6 of style 3, total profit £37 500

20 $x_1 = 2500 \text{ m}^2, x_2 = 1500 \text{ m}^2, x_3 = 1000 \text{ m}^2$, profit £9500

21 $x = \frac{1}{2}, y = \frac{1}{2}$

22 $x = \pm a$ and $y = 0$

23 $x = a/\sqrt{2}, y = b/\sqrt{2}$, area = $2ab$

24 Several possible optima: $(0, 3, 0); (\frac{3}{2}, \frac{3}{2}, \frac{1}{2}); (6 - 3t, 0, t)$ for any t

25 $(0, 1, 1); (0, -1, -1); (2, -1, 1)/\sqrt{7}; -(2, -1, 1)/\sqrt{7}$

26 For given surface area S , $b = c = 2a$, where $a^2 = \frac{1}{12}S$ and $V = 4a^3$

27 $A = -1.83, B = 0.609, I = 81.4$

28 For $\alpha \geq 0$ minimum at $(0, 0)$; for $\alpha < 0$ minimum at $(2\alpha, -3\alpha)/5$

29 (a) Bracket (without using derivatives) $0.7 < x < 3.1$
 (b) Iteration 1:

a	0.7	$f(a)$	2.7408
b	1.9	$f(b)$	2.177
c	3.1	$f(c)$	3.2041

Iteration 2:

a	0.7	$f(a)$	2.7408
b	1.7253	$f(b)$	2.0612
c	1.9	$f(c)$	2.177

gives $b = 1.5127$ and $f(b) = 1.9497$

(c)	Iteration 1		Iteration 2	
	x	$f(x)$	x	$f(x)$
	0.7	2.7408	0.7	1.5129
	3.1	3.2041	2.7408	1.9498
	-4.8309	0.9329	-4.8309	0.4224

gives $x = 1.1684$ and $f = 1.9009$

30 (a) Iteration 1:

a	0	$f(a)$	0
b	1	$f(b)$	0.42074
c	3	$f(c)$	0.01411

Iteration 2:

a	0	$f(a)$	0
b	1	$f(b)$	0.42074
c	1.5113	$f(c)$	0.30396

gives $x = 0.98979$ and $f = 0.42224$

(b)

x	Iteration 1		Iteration 2	
	0	1	0	0.8667
$f(x)$	0	0.4207	0	0.4352
$f'(x)$	1	-0.1506	1	-0.0612

gives $x = 0.8242$, $f = 0.4371$, $f' = -0.0247$

31 (a) Iteration 1:

a	1	$f(a)$	0.23254
b	1.6667	$f(b)$	0.25533
c	3	$f(c)$	0.14193

Iteration 2:

a	1	$f(a)$	0.23254
b	1.6200	$f(b)$	0.25715
c	1.6667	$f(c)$	0.25533

gives $x = 1.4784$ and $f = 0.26022$

(b) Iteration 1:

x	1	3
$f(x)$	0.23254	0.14193
$f'(x)$	0.13534	-0.08718

Iteration 2:

x	1	1.5077
$f(x)$	0.23254	0.25990
$f'(x)$	0.13534	-0.01368

gives $x = 1.4462$, $f = 0.26035$, $f' = -0.00014$ (c) Convergence in 6 and 3 iterations to $x = 1.446$, $f = 0.2603$ 32 $x = 1$, $\lambda_{\max} = 2$; $x = 0$, $\lambda_{\max} = 1.41421$; $x = -1$, $\lambda_{\max} = 1.73205$. One application of the quadratic algorithm gives $x = -0.14826$ and $\lambda_{\max} = 1.3854$.34 (a) After five iterations $x = 2.0492$ and $f = 1.8191$.
(b) After five iterations $x = 2.1738$, $f = 0.0267059$.

35 Iteration 1:

$$\mathbf{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad f = \frac{3}{2}, \quad \nabla f = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Iteration 2:

$$\mathbf{a} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \quad f = -\frac{1}{2}, \quad \nabla f = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Iteration 3:

$$\mathbf{a} = \begin{bmatrix} -\frac{7}{5} \\ -\frac{3}{5} \end{bmatrix}, \quad f = -0.9, \quad \nabla f = \begin{bmatrix} \frac{1}{5} \\ \frac{1}{5} \end{bmatrix}$$

36 Steepest descent gives the point $(-0.382, -0.255)$ and $f = -0.828$

37

f	-29.0000	-1.5023	-0.4523	-0.0764	-0.0248	-0.0165	$\rightarrow 0$
x	2.0000	1.1523	0.5022	0.6214	0.4948	0.5185	$\rightarrow 0.5$
y	2.0000	2.1695	1.8214	1.7539	1.6654	1.6394	$\rightarrow 1.5$
z	2.0000	0.4741	0.7943	0.9630	1.0170	1.0301	$\rightarrow 1$

38

f	29.0000	1.2448	0.1056	0.0026	0.0000	$\rightarrow 0$
x	2.0000	0.2727	0.4245	0.4873	0.4995	$\rightarrow 0.5$
y	2.0000	1.7273	1.5755	1.5127	1.5005	$\rightarrow 1.5$
z	2.0000	1.4545	1.1510	1.0253	1.0009	$\rightarrow 1$

39 $y = 0.2294x$ and $y = 0.5x - 0.2706$, cost = 5.974

40 (a) After step 1

$$\mathbf{a} = \begin{bmatrix} 1.2 \\ 2 \end{bmatrix}, \quad f = 0.8, \quad \mathbf{g} = \begin{bmatrix} 0 \\ 1.6 \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 0.1385 & 0.1923 \\ 0.1923 & 0.9615 \end{bmatrix}$$

After step 2 the exact solution $x = 1, y = 1$ is obtained

(b) After cycle 1

$$\mathbf{a}_1 = \begin{bmatrix} 0.5852 \\ 0 \\ 0.2926 \end{bmatrix}, \quad f = 1.0662, \quad \mathbf{g}_1 = \begin{bmatrix} -0.3918 \\ -1.7557 \\ 0.7822 \end{bmatrix}$$

$$\mathbf{H}_1 = \begin{bmatrix} 0.3681 & 0.1593 & -0.4047 \\ 0.1593 & 0.9632 & 0.1002 \\ -0.4047 & 0.1002 & 0.7418 \end{bmatrix}$$

After cycle 2

$$\mathbf{a}_2 = \begin{bmatrix} 1.0190 \\ 0.9813 \\ -0.0372 \end{bmatrix}, \quad f = 2.999 \times 10^{-6},$$

$$\mathbf{g}_2 = \begin{bmatrix} 0.0046 \\ -0.0012 \\ 0.0027 \end{bmatrix}$$

41 (a) After cycle 1

$$\mathbf{a} = \begin{bmatrix} 0.485 \\ -0.061 \end{bmatrix}, \quad f = 0.2424, \quad \mathbf{g} = \begin{bmatrix} 0.970 \\ -0.242 \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 0.995 & -0.062 \\ -0.062 & 0.258 \end{bmatrix}$$

After cycle 2 the minimum at $x = 0, y = 0$ is obtained

(b) After cycle 1

$$\mathbf{a}_1 = \begin{bmatrix} -0.0732 \\ 0.8344 \\ 0.4522 \end{bmatrix}, \quad f = 0.1563, \quad \mathbf{g}_1 = \begin{bmatrix} 0.0386 \\ 0.1564 \\ 0.6296 \end{bmatrix}$$

$$\mathbf{H}_1 = \begin{bmatrix} 0.4425 & 0.3669 & 0.0998 \\ 0.3669 & 0.7585 & -0.0657 \\ 0.0198 & -0.0657 & 0.9821 \end{bmatrix}$$

After cycle 2

$$\mathbf{a}_2 = \begin{bmatrix} -0.1628 \\ 0.7747 \\ 0.0525 \end{bmatrix}, \quad f = 0.0321, \quad \mathbf{g}_2 = \begin{bmatrix} -0.2006 \\ -0.1207 \\ 0.0630 \end{bmatrix}$$

$$\mathbf{H}_2 = \begin{bmatrix} 0.2820 & 0.1819 & -0.0498 \\ 0.1819 & 0.5452 & -0.2380 \\ -0.0498 & -0.2380 & 0.8429 \end{bmatrix}$$

The method converges to $x = 0, y = 1, z = 0$.

43 (a) $(0, 0)$

(b)

f	1.3125	0.0764	0.0072	0.0007	0.0004	0.0000	$\rightarrow 0$
x	0.5000	-0.0950	0.0057	-0.0251	-0.0079	-0.0032	$\rightarrow 0$
y	0.5000	0.9165	0.9276	0.9633	0.9742	0.9978	$\rightarrow 1$
z	0.5000	0.7380	0.9674	1.0009	1.0044	1.0014	$\rightarrow 1$

10.7 Review exercises

1 $x_1 = 250, x_2 = 100, F = 3800$

2 $x_1 = 22, x_2 = 0, x_3 = 6$, profit £370

3 Standard 20, super 10, deluxe 40, profit £21 000

4 2 kg bread and 0.5 kg cheese, cost 210 p

5 Maximum at $(1, 1)$ and $(-1, -1)$, with distance $= \sqrt{2}$

Minimum at $(\frac{1}{\sqrt{3}}, -\frac{1}{\sqrt{3}})$ and $(-\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}})$, with distance $= \frac{\sqrt{2}}{\sqrt{3}}$

6 Sides are $3\sqrt{\frac{1}{10}}$ and $2\sqrt{\frac{1}{10}}$

7 $(\frac{17}{5}, 0, \frac{6}{5})$, with distance 2.683

8 $(1, 2, 3)$ with $F = 14$, and $(-1, -2, -3)$ with $F = -14$. $(1, 2, 3)$ gives the global maximum and $(0, 0, 0)$ gives the global minimum.

9 (i) $b = c$ (ii) $a = b = c$

10 $h^2 = 3\pi^2/b, r^2 = 3a^2/2b$

11 $k = 2.19$

12 Bracket:

R	3.5	5.5	9.5
Cost	1124	704	1418

Quadratic algorithm gives $R = 6.121$ and cost = 802, so $R = 5.5$ still gives the best result. After many iterations $R = 4.4$ and cost = 579

13 Quadratic algorithm always gives $x = 0.5$ for any intermediate value. However,

f	0.7729	0.7584	0.7524	0.7508	$\rightarrow 0.7500$
a	0.3147	0.5000	0.5629	0.6051	
b	0.5000	0.5629	0.6051	0.6243	$\rightarrow 0.6514$
c	1.0000	1.0000	1.0000	1.0000	

14 Maximum at $\theta = 5.01$ rad, minimum at $\theta = 1.28$ rad

15 44 mph

16 At iteration 2

(a) $x = -0.0916, y = -0.1375, f = 0.0326$

(b) $x = -0.1023, y = -0.1534, f = 0.0323$

(c) $x = -0.1007, y = 0.1519, f = 0.0323$. The exact minimum is at $x = -0.1026, y = -0.1540, f = 0.0323$.

17 Maximum of 1.056 at $X = 0, Y = 0.4736$, minimum of 0.5278 at $X = \pm 0.25, Y = 2$

18 Partan

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad f = 1 \qquad \mathbf{x}_2 = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}, \quad f = 0.5$$

$$\mathbf{z}_2 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad f = 0.25 \qquad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad f = 0$$

Steepest descent

$$\mathbf{x}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad f = 1 \qquad \mathbf{x}_2 = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}, \quad f = 0.5$$

$$\mathbf{x}_3 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad f = 0.25$$

$$\mathbf{x}_4 = \begin{bmatrix} 0.75 \\ 0.5 \end{bmatrix}, \quad f = 0.125$$

19 Start values:

$$\mathbf{a}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad f = 1$$

$\lambda = 0$:

$$\mathbf{a}_1 = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}, \quad f = 1 \quad (\text{no improvement})$$

$\lambda = 1$:

$$\mathbf{a}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \quad f = -5 \quad (\text{ready for next iteration})$$

20 (a) $\mathbf{a}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad F = 0.0625$

$$\mathbf{a}_1 = \begin{bmatrix} -0.25 \\ -0.25 \end{bmatrix}, \quad F = 0.0039$$

$$\mathbf{a}_2 = \begin{bmatrix} -0.375 \\ -0.375 \end{bmatrix}, \quad F = 0.0002$$

(b) $\mathbf{a}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad F = 0.3125$

$$\mathbf{a}_1 = \begin{bmatrix} 0.333 \\ 3.667 \end{bmatrix}, \quad F = 0.0664$$

21 $\mathbf{a}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad F = 1.29$

$$\mathbf{a}_1 = \begin{bmatrix} 1.07 \\ 0.27 \end{bmatrix}, \quad F = 0.239$$

$$\mathbf{a}_{\min} = \begin{bmatrix} 0.987 \\ 0.956 \end{bmatrix}, \quad F_{\min} = 0.032$$

23 $\mathbf{a}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad f = 1$

$$\mathbf{a}_1 = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}, \quad f = 0.5$$

$$\mathbf{a}_2 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad F = 0.25$$

24 Bracket gives

α	$[F(\alpha) - 3]^2$
1.4	0.0776
1.5	0.0029
1.6	0.0369

Quadratic algorithm gives $\alpha^* = 1.5218$ and $f = 9 \times 10^{-5}$

CHAPTER 11

Exercises

1 (a) (762, 798) (b) 97

2 76.1, (65.7, 86.5)

3 (8.05, 9.40)

4 (71.2, 75.2), accept

5 (2.92, 3.92)

6 (24.9, 27.9)

7 95% confidence interval (53.9, 58.1), criterion satisfactory

8 (-1900, 7500), reject

9 90%: (34, 758), 95%: (-45, 837), reject at 10% but accept at 5%

10 90%: (0.052, 0.089), 95%: (0.049, 0.092), reject at 10% but accept at 5%. Test statistic leads to rejection at both 10% and 5% levels, and is more accurate

11 203, (0.223, 0.327)

12 90%: (-0.28, -0.08), 95%: (-0.30, -0.06), accept at 10% but reject at 5%

13 (0.003, 0.130), carcinogenic

14 (a) X : (0.34, 0.53, 0.13), Y : (0.25, 0.31, 0.44)
(b) 0.472, (c) $E(X) = 1.79$, $\text{Var}(X) = 0.426$, $E(Y) = 2.19$, $\text{Var}(Y) = 0.654$, $\rho_{X,Y} = -0.246$

17 (a) 0.552 (b) 0.368

18 0.934

19 0.732

20 (0.45, 0.85)

21 (0.67, 0.99)

22 0.444, 90%: (0.08, 0.70), 95%: (0.00, 0.74), just significant at 5%, rank correlation 0.401, significant at 10%

23 (a) 6 (b) 0.484

(c) $f_X(x) = 6\left(\frac{1}{2} - x + \frac{1}{2}x^2\right)$, $f_Y(y) = 6(1-y)y$

24 0.84**25** $a = 1.22$, $b = 2.13$ **26** $a = 6.315$, $b = 14.64$, $y = 226$ **28** (a) $a = 343.7$, $b = 3.221$, $y = 537$

(b) (0.46, 5.98), reject (c) (459, 615)

29 $a = 0.107$, $b = 1.143$, (14.4, 17.8)**31** 120Ω **32** $\lambda = 2.66$, $C = 2.69 \times 10^6$, $P = 22.9$ **33** $a = 7533$, $b = -1.059$, $y = 17.9$ **34** $\chi^2 = 2.15$, accept**35** $\chi^2 = 12.3$, significant at 5%**36** $\chi^2 = 1.35$, accept Poisson**37** $\chi^2 = 12.97$, accept Poisson**39** $\chi^2 = 1.30$, not significant**40** $\chi^2 = 20.56$, significant at 5%**41** $\chi^2 = 20.7$, significant at 0.5%**42** $\chi^2 = 11.30$, significant at 5% but not at 1%, for proportion 95%: (0.111, 0.159), 99%: (0.104, 0.166), significant at 1%**43** Warning 9.5, action 13.5, sample 12, UCL = 11.4, sample 9**44** UK sample 28, US sample 25**45** Action 2.93, sample 12**46** Action 14.9, sample 19 but repeated warnings**47** (a) Sample 9 (b) Sample 9**48** (a) Sample 10 (b) Sample 12**49** Sample 10**50** Sample 16**51** (a) Repeated warnings (b) Sample 15

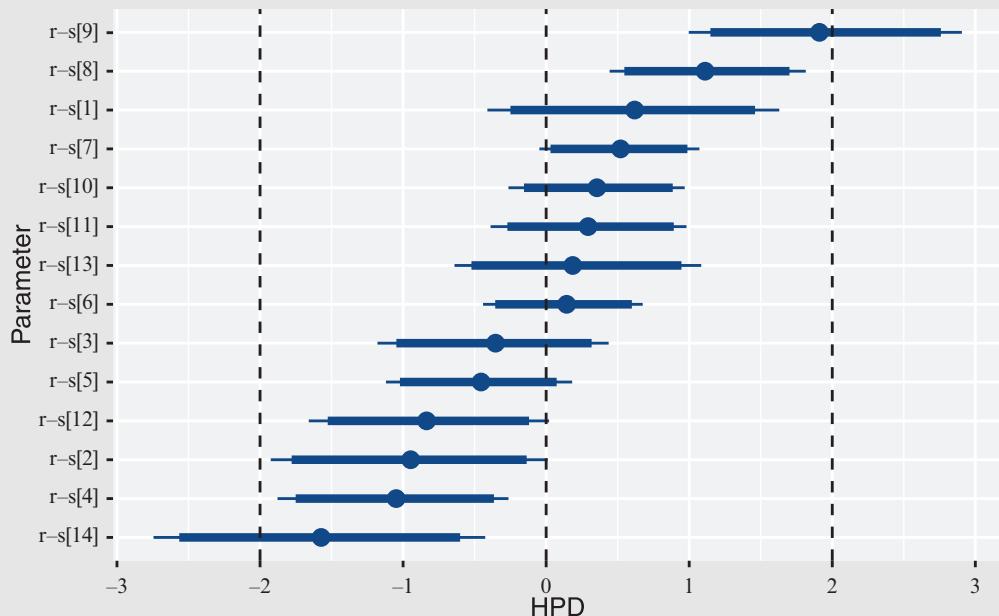
(c) Sample 14

53 Sample 11**54** Shewhart, sample 26; cusum, sample 13; moving-average, sample 11**55** 0.132**56** 0.223, 0.042**58** (a) $\frac{1}{4}$ (b) $2\frac{1}{4}$ (c) 0.237
(d) 45 min (e) 0.276**60** Mean costs per hour: A, £200; B, £130**61** 6**62** Second cash desk**63** 29.4%**64** Sabotage**65** $P(C|\text{two hits}) = 0.526$ **66** $\frac{1}{2}$ **67** (a) 0.0944 (b) 0.81**68** (a) $\frac{3}{4}$ (b) $[1 + (\frac{1}{3})^k 2^{n-k}]^{-1}$ **69** AAAA**70** 1.28:1 in favour of Poisson**72** 2.8:1 in favour of H_2 **73** 12.8:1 in favour of H_1 **74** R and BUGS code that perform the required Bayesian inference tasks and the resulting caterpillar plot of the posterior distribution of the residuals are as follows:

```

X <- 1:14 # Integer sequence from 1 to 14
Y <- c(4.4, 4.9, 6.4, 7.3, 8.8, 10.3, 11.7,
      13.2, 14.8, 15.3, 16.5, 17.2, 18.9, 19.3)
# New values of X
X_new <- 8.5
X_new_2 <- 10
# Required value of Y
Y_required <- 13.8
# Regression model
reg_model <- function(){
  # Define the model as stated above
  for(i in 1:n){
    Y[i] ~ dnorm(mu[i], tau)
    # Parametrized by the precision tau = 1 / sigma^2
    mu[i] <- a + b * X[i]
    r_s[i] <- sqrt(tau) * (Y[i] - a - b * X[i]) # Define r_si
  }
  # Priors
  a ~ dnorm(0.0, 1.0E-4)
  b ~ dnorm(0.0, 1.0E-4)
  tau ~ dgamma(1.0E-3, 1.0E-3)
  # We allow tau to take a large range of possible values
  #
  sigma <- 1.0 / sqrt(tau) # Definition of sigma, a transformation of tau
  # Monitor a + b * X_new and a + b * X_new_2
  mu_new <- a + b * X_new
  mu_new_2 <- a + b * X_new_2
  # and (Y_required - a) / b
  tension <- (Y_required - a) / b
  # There would be a problem if b were zero!
  # This could be overcome using an ifelse construction
  # to ensure that the divisor is never actually zero
}
# Regression data
n <- length(X)
reg_data <- list("Y", "X", "n", "X_new", "X_new_2", "Y_required")
# Load package
require(R2jags)
set.seed(14)
# Set the seed of the random number generator for reproducibility
##> DIC is an estimate of expected predictive error
##> (lower deviance is better).
# Plot
reg_posterior.mcmc <- as.mcmc(reg_posterior)
require(ggmcmc)
reg_posterior.gss <- ggs(reg_posterior.mcmc)
ggs_caterpillar(reg_posterior.gss, family = "r_s") +
  geom_vline(xintercept = c(-2, 0, 2), lty = "dashed")

```



We can see from the caterpillar plot that the posterior medians of $\tau_{s,i}$ lie between -2 and 2 .

11.12 Review exercises

1 $Z = 0.27$, accept

2 $(0.202, 0.266)$

3 $(96.1 \times 10^6, 104.9 \times 10^6)$

4 $\chi^2 = 3.35$ (using class intervals of length 5, with a single class for all values greater than 30), accept exponential

5 Outlier 72 significant at 5%, outlier included $(7.36, 11.48)$; excluded $(7.11, 10.53)$

6 $\chi^2 = 20.0$, significant at 2.5%

7 Operate if $p > \frac{4}{13}$

8 Cost per hour: A, £632.5; B, £603.4

9 (a) $P(\text{input } 0 | \text{output } 0) = \frac{p\alpha}{p\alpha + p\bar{\alpha}}$ etc.

(b) $p < \alpha < 1 - p$



Index

A

abscissa of convergence of Laplace transforms 361–2
AC circuits (application) 304–5
action limits in control charts 891
Adams–Bashforth formulae 131
Adams–Morton formulae 138
addition of matrices 4
addition rule 801
adjoint matrix 5
adjusted residual 872
algebraic multiplicity of eigenvalues 22
aliasing error 587
alternative solutions in linear programming 746
amplified gain 392
amplified input 392
amplitude gain 392
amplitude ratio 392
amplitude spectrum 515, 548, 604
analogue filters 471
 application 599–601
analytic function 275
applied probability 800
arbitrary constant 627
arbitrary function 627–32
arbitrary inputs in transfer functions 374–7
Argand diagram 251
artificial variable 754
associative law 4
asymptotically stable system 101
attenuated input 392
attribute 891
augmented matrix 8
average power 521

B

basic variables 741
Bayes’ theorem 930–43, 931
 applications 933–5
 derivation 930–2
 statistical inference 935–43
beams, bending of 352–5
bending of beams 352–5
Bernoulli distribution 804–5
Bessel’s equality 528
Best, N. 937
BFGS method 784
bilateral Laplace transform 558
bilinear mappings of complex functions 265–71
bilinear transform 473
bilinear transform method 475
binding constraints 741
binomial distribution 804–5
Blackman window 609–11
block diagram algebra 357
blood-flow model (application) 726–9
Bode plots 394
Boole, George 408
boundary conditions in partial differential equations 718–22
boundary-value problems in differential equations 159–60
bracket 770
bracketing procedure 770–2
bracketing triple 772
Branseum, A. 943
breakpoint 395
Brigham, E.E. 592
Broyden 784
Burgers’ equation 632
Butterworth approximation 599
Butterworth filter 471

C

canonical form of matrices, reduction to **36–50**
 diagonal form **36–9**
 Jordan canonical form **39–42**
 quadratic forms **44–50**
 canonical representation of equations **96**
 capacitor microphone (application) **104–7**
 Carlin, J.B. **943**
 Carslaw, H.S. **715**
 Cauchy–Riemann equations **275–9, 284**
 Cauchy's conditions in partial differentiation
 equations **718, 719**
 causal sequences **409**
 Cayley–Hamilton theorem **53**
 central difference scheme **138**
 central limit theorem **806**
 chain rule **181**
 Chapman, M.J. **471, 475, 601, 602**
 characteristic curves **628**
 characteristic equation **357, 436**
 characteristic polynomial **14**
 characteristics in partial differential
 equations **637**
 chemical processing plant
 (application) **790–2**
 Chi-square distribution and test **863–7**
 Christensen, R. **943**
 circular frequency **486**
 closed boundary **718**
 column rank matrix **63, 64**
 column vector **3**
 columns **653**
 commutative law **4**
 not satisfied **4**
 companion form **80**
 complement of event **801**
 complementary function of matrices **87**
 complex differentiation **274–86**
 Cauchy–Riemann equations **275–9, 284**
 conjugate functions **280–2**
 harmonic functions **280–2**
 mapping **282–6**
 complex form of Fourier series **509**
 complex frequency domain **548**
 complex functions **251–74**
 bilinear mappings **265–71**
 inversion **260–5**
 linear mappings **253–9**
 polynomial mappings **272–74**
 complex series **287–99**
 Laurent series **295–9**
 power series **287–91**
 Taylor series **291–4**
 composite-function rule **181**
 conditional distribution **829**

conditional probability **802**
 confidence interval for mean **812–15, 813**
 conformal mapping **282**
 conjugate functions **280–2**
 conjugate-gradient methods **782**
 conservative force **210, 238**
 contingency tables **867–72**
 continuity correction **806**
 continuity equation **240**
 continuous Fourier spectra **548–50**
 continuous Fourier transform **583–91**
 continuous random variables **802**
 continuous source **713**
 continuous variables **802**
 control charts **891**
 controllable modes in matrix **97**
 convergence rate of eigenvalues **31**
 convolution
 in discrete-linear systems **450–4**
 in Fourier transforms **572–4**
 for Laplace transforms **372–4**
 convolution integral **371**
 convolution sum **451**
 Cooley, J.W. **538, 592**
 corner frequency **395**
 correlation **829–33, 831**
 partial **832**
 rank **838–40**
 and regression **856**
 sample **833–5**
 coupled first order equations **149–54**
 Courant, Fredricks and Levy (CFL)
 condition **657**
 covariance **829–33**
 Cowell, R.G. **935**
 Crank–Nicolson method for solution of
 heat-conduction/diffusion
 equation **674**
 cumulative distribution function **802**
 curl of a vector field **199–201**
 curl-free motion **201**
 current in field-effect transistor
 (application) **307–9**
 customers **909**
 Cusum control charts **898–901**

D

D'Alembert solution in partial differential
 equations **634–43, 637**
 Danzig **739**
 Davidon **782**
 Dawid, P. **935**
 decision variables **740**
 deflation methods in matrices **33**

- degeneracy of a matrix **25**
 degree of belief **933**
 degrees of freedom **817**
 delay theorem **325**
 delta (shift) operator **474–5**
 delta function **341**
 delta operator (application) **473–9**
 q (shift) operator **474–5**
 delta operator **474–5**
 \mathcal{D} transform **479**
 Denis, J.-B. **935**
 dependent variable **251**
 derivatives
 Laplace transforms of **318–19**
 of scalar point function **192–5**
 of vector point function **196–206**
 curl of a vector field **199–201**
 divergence of a vector field **196–8**
 vector operator **202–6**
 determinants
 of mappings **266**
 of a matrix **5**
 DFP method **782, 783–4**
 diagonal matrix **3, 37, 63**
 diagonalization **37**
 difference between means **819–21**
 difference equation **408**
 in discrete-time systems **428–9**
 solutions **430–4**
 differential **189**
 differential equations **114**
 Laplace transform methods on **331–5**
 step and impulse functions **320–56**
 numerical solution of
 boundary-value problems **159–60**
 coupled first order equations **149–54**
 first order **115–49**
 on engineering problems **123–5**
 Euler's method **116–22**
 local and global truncation errors
 132–4
 multi-step methods **126–32**
 predictor–corrector methods **134–9**
 Runge–Kutta methods **139–42**
 software libraries on **147–9**
 stiff equations **145–7**
 higher order systems, state-space
 representation of **154–7**
 method of shooting **160–2**
 see also partial differential equations
 diffusion equation in partial differential
 equations **617, 620–3**
 solution of **660–76**
 Laplace transform method **664–9**
 numerical solution **671–6**
 separation method **660–4**
 sources and sinks for **712–15**
 digital filters (application) **602–7**
 and windows **607–11**
 digital replacement filters **472–3**
 Dirac delta function **341**
 direct form of state equations **86–8**
 directional derivative **193**
 directional field **116**
 Dirichlet's conditions
 for Fourier integral **541**
 in partial differentiation equations
 718, 720, 721
 discrete Fourier transform (DFT) pair **582**
 discrete Fourier transform **579–83**
 discrete frequency spectra **515, 539**
 discrete variables **802**
 discrete-linear systems **435–54**
 convolution **450–4**
 impulse response **441–4**
 stability **444–50**
 discrete-time Fourier transform (DTFT) **603**
 discrete-time signal **408**
 discrete-time system
 constructing **475–7**
 Delta transform **479**
 design of (application) **470–3**
 analogue filters **471**
 digital replacement filters **472–3**
 difference equations in **428–9**
 implementation **477–9**
 discretization of continuous-time state-space
 models **464–9**
 Euler's method **464–6**
 step-invariant method **466–9, 467**
 disjoint events **801**
 dissipative force **210**
 distensibility **728**
 distribution **342, 802**
 of sample average **810–12**
 distributive law **4**
 divergence of vector **197**
 divergence theorem *see* Gauss's divergence
 theorem
 domain of dependence **638**
 domain of function **251**
 domain of influence **638**
 dominant eigenvalue **30**
 double integrals **211–16, 212**
 duality property **556**
 Duhamel integral **371**
 Dunson, D.B. **943**
 Dyke, Phil **538**
 dynamic equations **81**
- E**
- echelon form of a matrix **8**
 Efron, B. **935**

- eigenvalues **2, 13–29, 16**
 characteristic equation **14–16**
 method of Faddeev **15**
 and eigenvectors **16–22**
 pole location **398–9**
 and poles **398**
 repeated **22–6**
 useful properties **26–8**
- eigenvectors **2, 13, 16**
- electrical fuse, heating of (application) **167–71**
- element of a matrix **3**
- elliptic equations **716**
- energy **563**
 energy signals **568**
 energy spectral density **564**
 energy spectrum **564**
- engine performance data (application) **874–91**
 dependence of running time on temperature **880–7**
 mean running times and temperatures **877–80**
 normality test **888–90**
- entire function **294**
- equal matrices **3**
- equality-constrained optimization **736**
- equality constraints in Lagrange multipliers **764–7**
- equivalent linear systems **96**
- error function **246**
- essential singularity **300**
- Euler's formula **487**
- Euler's method **464–6**
- Euler's method on differential equations **116–22, 118**
 analysis **120–2**
- even periodic extension **494**
- events **801**
- Everitt, B.S. **872**
- exact differential **190**
- expected value **802**
- explicit formula for solution of heat-conduction/diffusion equation **671**
- explicit methods in partial differential equations **657**
- exponential distribution with parameter **910**
- exponential form of Fourier series **509**
- F
- Faddeev method on eigenvalues **15**
- faltung integral **371**
- Fannin, D.R. **609**
- feasible basic solution **741**
- feasible region **739**
- Fermat, Pierre de **765**
- Feshbach, H. **722**
- Fick's law **621**
- field-effect transistor (application) **307–9**
- filter length **605**
- filters **471**
- final-value theorem **367–70**
- finite calculus **408**
- finite difference methods **408**
- finite-difference representation **654**
- finite elements in partial differential equations **694–706**
- finite impulse response (FIR) **607**
- first harmonic **487**
- first order method on differential equations **121**
- first shift property of z transforms **416–17**
- first shift theorem in inverse Laplace transform **318**
- fixed point **253**
- Fletcher, R. **782, 784**
- fluid dynamics, streamline in (application) **240–2**
- folding integral **371**
- Forbeniuc, G. **53**
- Forsyth, R. **935**
- Forsythe, W. **474**
- Fourier coefficients **487**
- Fourier cosine integral **542**
- Fourier integral representation **541**
- Fourier law **621**
- Fourier series **486**
 complex forms **508–23**
 complex representation **508–12**
 discrete frequency spectra **515–21**
 multiplication theorem **512–13**
 Parseval's theorem **512, 514–15**
 power spectrum **521–3, 522**
 functions of period 2B **488–92**
 of jumps at discontinuities **499–502**
 orthogonal functions **524–9**
 convergence of generalized series **527–9**
 definitions **524–6**
 generalized series **526–7**
 periodic functions **486**
- Fourier series expansion **487**
- Fourier sine integral **542**
- Fourier transforms **538, 539–50, 544**
 continuous Fourier spectra **548–50**
 in discrete time **575–98**
 continuous transform **583–91**
 fast Fourier transform **592–8**
 sequences **575–9**
- Fourier integral **539–43**
- Fourier transform pair **544–8**
- frequency response **560–2**
 and Laplace transform **558–60**
 properties of **552–7**

frequency-shift 554–5, **555**
 linearity 552
 time-differentiation 552–3, **553**
 time-shift 553–4
 step and impulse functions 563–74
 convolution 572–4
 energy and power 563–72
 Fourier's theorem 487–8
 Fredricks 657
 frequency **486**
 frequency components in Fourier series **515**
 frequency-domain portrait **548**
 frequency response (applications) 390–7, **392**
 frequency response plot **397**
 frequency response
 in Fourier series 502–6
 in Fourier transform 560–2
 frequency-shift property 554–5, **555**
 frequency spectrum **515**
 frequency transfer function **561**, 578
 Fryba, L. 723
 full-rank matrix **63**, 64, **74**
 functions **251**
 describing functions (application) **532**–3
 of period 2B 488–92
 fundamental mode **487**

G

Gabel, R.R. 346–7
 Gauss's divergence theorem **233**–6
 Gelman, A. 943
 generalized calculus **342**
 generalized derivatives **348**
 generalized form of Parseval's theorem **528**
 generalized Fourier coefficients **527**
 generalized Fourier series **527**
 generalized Fourier transforms **566**
 generalized functions **341**
 generating function **410**
 geometric distribution **918**
 geometric moving-average (GMA) charts **901**
 geometric multiplicity of eigenvalue **25**
 Gibbs' phenomenon **606**
 Gill, K.F. 370
 global truncation errors on differential equations **132**–4
 Golden search algorithm 772
 Goldfarb 784
 Goodall, D.P. 471, 475, 601, 602
 Goodall, R.M. 474

goodness-of-fit tests 863–74
 Chi-square **863**–7
 contingency tables **867**–72
 Goodwin, G.C. 474, 479
 gradient of scalar point function 192–5, **193**
 Green's functions 684, **711**
 Green's theorem 217–21, **218**, 721
 Grimmett, G.R. 923

H

Haberman, R. 617, 711
 half-range cosine series expansion **495**
 half-range Fourier series expansion **495**
 half-range sine series expansion **495**
 Hamming window 609–10
 Hanson, T.E. 943
 harmonic components in Fourier series **515**
 harmonic functions (application) 305–9
 harmonic functions **280**–2
 Hastie, T. 935
 heat-conduction in partial differential equations **617**, 620–3
 solution of 660–76
 Laplace transform method 664–9
 numerical solution 671–6
 separation method 660–4
 sources and sinks for 712–15
 heat transfer (application) 242–6
 using harmonic functions 305–7
 heating fin (application) 792–4
 Heaviside step function 320–3
 and impulse function 346–51
 Heaviside theorem **325**
 Helmholtz equation **626**
 Hessian matrix **778**
 higher order systems, state-space representation of 154–7
 hill climbing **769**–89
 advanced multivariable searches 782–5
 least squares 786–9
 single multivariable searches 775–81
 single-variable search 769–74
 holomorphic function **275**
 Householder methods **34**
 Howsion, S. 723
 Hunter, S.C. 704
 Hush, D.R. 609
 hyperbolic equations **717**
 hypothesis tests **810**
 simple, testing 815–16

I

- ideal low-pass filter **471**
- identity matrix **3**
- Ifeachor, E.C. **609**
- image set **251**
- implicit formula for solution of heat-conduction/diffusion equation **674**
- implicit methods in partial differential equations **657**
- impulse forces **341**
- impulse functions **341–2**
 - in Fourier transforms **563–74**
 - Laplace transforms on **343–6**
- impulse invariant technique **473**
- impulse response in transfer functions **364–5**
- impulse sequence **412, 441**
- indefinite quadratic forms **46**
- independent events **802, 828**
- independent variable **251**
- inequality constraints in Lagrange multipliers **768**
- inequality-constrained optimization **736**
- infinite sequence **409**
- initial-value theorem
 - of Laplace transforms **365–7**
 - of z transforms **419**
- inner (scalar) product **28**
- in-phase quadrature components **487**
- input-output block diagram **356**
- instantaneous source **713**
- integral equations **114**
 - integral solutions to partial differential equations **707–15**
 - separated solutions **707–9**
 - singular solutions **709–12**
- integrals, Laplace transforms of **319**
- integration
 - in vector calculus **206–39**
 - double integrals **211–16**
 - Gauss's divergence theorem **233–6**
 - Green's theorem **217–21**
 - line integral **207–10**
 - Stokes' theorem **236–9**
 - surface integrals **222–8**
 - volume integrals **229–32**
- integro-differential equations **114**
- inter-arrival time **909**
- interval and test
 - for correlation **835–7**
 - for proportion **821–4**
- interval estimate **810**
- inverse mapping **253**
 - with respect to the circle **263**
 - of complex functions **260–5**

- inverse matrix **5–6**
- properties **6**
- inverse Nyquist approach **397**
- inverse polar plot **397**
- inverse z transform operator **420**
- inverse z transformation **420**
- inverse z transforms **420–7**
 - techniques **421–7**

- inversion of complex functions **260–5**
- irrotational motion **201, 238**

J

- Jackson, C. **937**
- Jackson, L.B. **609**
- Jacobi methods **34**
- Jacobian **219**
- Jacobian matrix **185, 778**
- Jaeger, J.C. **715**
- Jervis, B.W. **609**
- Johnson, W. **943**
- joint density function **827**
- joint distributions **825–8, 826**
 - independence **828–9**
 - and marginal distributions **825–8, 826**
- Jones, O. **922**
- Jong, M.T. **607**
- Jordan, Marie Ennemond Camille **40**
- Jordan canonical form **39–42**
- Jordan normal form **40**
- jump discontinuities, coefficients of Fourier series at **499–502**
- Jury stability criterion **446–8**
- Jury, E.I. **446**

K

- Kirchhoff's laws **84**
- Kraniauskas, P. **453**
- Kuhn **768**

L

- Lagrange interpolation formula **772**
- Lagrange multipliers **764–8**
 - equality constraints **764–7**

- inequality constraints **768**
 Laplace equation in partial differential equations **617**, 623–5
 solution of 677–693
 numerical solution 686–693
 separated solutions 677–84
 Laplace transform 316–89
 bending of beams 352–5
 definition and notation 316–18
 derivative of 318–19
 differential equations 331–5
 and Fourier transform 558–60
 frequency response (application) 390–7
 Heaviside step function 320–3
 and impulse functions 346–51
 impulse functions **341**–2, 343–6
 and Heaviside step function 346–51
 periodic functions 335–9
 pole placement (application) 398–9
 second shift theorem 325–8
 inversion 328–31
 sifting property **342**–3
 solution to wave equation 648–51
 state-space equations, solution of 378–89
 transfer functions **356**–77
 and arbitrary inputs 374–7
 convolution **371**–4
 definitions 356–9
 final-value theorem **367**–70
 impulse response **364**–5
 initial-value theorem **365**–7
 stability in 359–64
 unit step function **320**, 323–5
 and z transforms 455–6
 Laplace transform method for solution of heat-conduction/diffusion equation 664–9
 Laplacian operator **203**
 Laurent series **295**–9, 300
 Lauritzen, S.L. 935
 leading diagonal **3**
 leading principle minor of matrices **47**
 least squares in hill climbing 786–9
 left inverse matrix **74**
 left singular vector matrix **68**
 Levy **657**
 Lewis, P.E. 694
 likelihood ratio **933**
 limit-cycle behaviour **532**
 Lindley, D.V. 935
 line integral **207**–10
 line spectra **515**
 linear dependence **10**
 linear equations of matrices 7–8
 linear independence of vector spaces **10**–11
 linear mappings of complex functions 253–9
 linear operator
 in Fourier transforms **552**
 on z transforms **415**
 linear programming **739**–62
 equality constraints/variables, unrestricted in sign **761**–2
 simplex algorithm 741–51
 two-phase method **753**–61
 linear regression **843**
 linearity property
 of Fourier transforms 552
 of z transforms 415–16
 local truncation errors on differential equations **132**–4
 LR methods in matrices **35**
 Lunn, D. 937
 Lyapunov function **101**, **102**
 Lyapunov stability analysis (application) 101–3

M

- Maclaurin series expansion **292**
 magnification 255, 256
 Maillardet, R. 922
 main lobe **608**
 main lobe width **608**
 MAPLE
 on Fourier transforms 571
 on Laplace transforms 318, 324, 325, 328, 329, 334, 344–5, 346, 650
 on linear programming 748, 758, 761
 on matrices 9, 20, 24, 36, 77–8, 91
 on ordinary differential equations 115–16, 119, 120, 123, 124, 137, 142, 147, 148, 152, 154, 157, 167, 171
 on partial differentiation equations 625, 630, 631, 666, 668
 on vector calculus 179, 184, 198, 201
 on z transforms 413, 414, 423, 425, 431, 432, 434, 443
 mapping **251**
 in complex differentiation 282–6
 determinants of **266**
 polynomial mapping 272–4
 marginal density function **827**
 marginal distributions 825–8, **826**
 marginally stable linear system **359**–**60**
 marginally stable system **445**
 MATLAB
 on Fourier transforms 547–8, 566, 569, 571–2
 on hill climbing 769–72, 773, 774, 776, 781, 784, 785, 787, 788–9
 on Laplace transforms 318, 323–4, 325, 328, 329, 334, 344–5, 346, 359, 365, 387, 388, 397, 650
 on linear programming 748, 758, 761

- on matrices 2, 6, 7, 9, 20, 21, 23–4, 35, 36, 42, 61, 76–7, 91
 on ordinary differential equations 115–16, 125, 132, 147, 148–9, 171
 on partial differentiation equations 631, 656, 659, 668, 673, 676, 694, 698, 699, 701, 703, 704, 705
 on vector calculus 179, 184, 198, 201, 216, 232
 on z transforms 412, 414, 422–3, 424, 425, 426, 425, 431, 432, 434, 443, 468–9
- matrices 2–111
 eigenvalues 13–29
 characteristic equation 14–16
 method of Faddeev 15
 and eigenvectors 16–22
 functions 51–62
 repeated 22–6
 useful properties 26–8
- matrix 3
- matrix algebra 2–9
 adjoint matrix 5
 basic operations 3–4
 definitions 3
 determinants 5
 inverse matrix 5–6
 properties 6
 linear equations 7–8
 rank 8–9
 numerical methods 29–35
 power method 29–35
 reduction to canonical form 36–50
 diagonal form 36–9
 Jordan canonical form 39–42
 quadratic forms 44–50
 singular value decomposition 63–78
 pseudo inverse 72–8
 SVD 69–72
 singular values 65–9
 solution of state equation 86–99
 canonical representation 95–9
 direct form 86–8
 spectral representation of response 92–5
 transition matrix 88
 evaluating 89–91
 state-space representation 79–85
 multi-input-multi-output (MIMO)
 systems 84–5
 single-input-single-output (SISO)
 systems 79–83
 symmetric 28–9
 vector spaces 9–12
 linear dependence 10–11
 transformation between bases 11–12
 maximum of objective function 745
 McElreath, R. 943
 mean 802
 when variance unknown 817–19
 mean square error in Fourier series 527
 means, difference between 819–21
 memoryless property 910
 meromorphic poles 301
 method of separation of variables 643
 Middleton, R.M. 474, 479
 minimal form 385
 modal form in matrix 93
 modal matrix 36
 modes in matrix 93
 modulation in Fourier transforms 555
 Moore-Penrose pseudo inverse square matrix 73
 Morse, P.M. 722
 motion in a viscous fluid (application) 114–15
 moving-average control charts 901–5
 multi-input-multi-output (MIMO) systems
 in Laplace transforms 383–4
 in matrices 84–5
 multiple service channels queues 921–3
 multiplication by scalar, matrix 3
 multiplication of matrices 4
 multiplication theorem in Fourier series 512–13
 multi-step methods on differential equations 126–32, 129
 Murdoch, J. 898, 901
- N**
- negative-definite quadratic forms 46, 48
 negative-semidefinite quadratic forms 46, 48
 net circulation integral 210
 Neumann conditions in partial differentiation equations 718, 720, 721
 Newton method 776, 778
 Newton-Raphson methods 778
 Nichols diagram 397
 nodes 653
 non-basic variables 741, 746
 non-binding constraints 741
 non-conservative force 210
 nonlinear regression 856–61
 non-negative eigenvalues 65
 non-square matrix 63
 non-trivial solutions of matrices 8
 normal distribution 804–8
 normalizing eigenvectors 19
 n th harmonic 487
 Ntzoufras, I. 935, 943
 null matrix 25
 Nyquist approach 397
 Nyquist interval 587
 Nyquist–Shannon sampling theorem 587, 591

O

- objective function **740**
- observable state of matrix **97**
- odd periodic extension **495**
- offsets **368**
- one-dimensional heat equation **621**
- open boundary **718**
- Oppenheim, A.V. 609
- optimization
 - chemical processing plant (application) **790–2**
 - heating fin (application) **792–4**
 - hill climbing **769**, **769–89**
 - Lagrange multipliers **764–8**
 - linear programming **739–62**
- order of pole **300**
- order of the system **357**, **436**
- orthogonal functions **524–9**
- orthogonal matrix **12**
- orthogonal set **524**
- orthonormal set **525**
- oscillating systems (application) **502–6**
- oscillations of a pendulum (application) **162–7**
- over determined matrix **72**, **75**

P

- Page, E. 923
- parabolic equations **717**
- parameters **804**
 - estimating **810–24**
 - confidence interval for mean **812–15**, **813**
 - difference between means **819–21**
 - distribution of sample average **810–12**
 - hypothesis tests **810**
 - interval and test for proportion **821–4**
 - interval estimate **810**
 - mean when variance unknown **817–19**
 - testing simple hypotheses **815–16**
 - parasitic solutions in differential equations **130**
 - Parseval's theorem **512**, **514**, **564**
 - partial correlation **832**
 - partial derivative **179**
 - partial differential equations **616**
 - arbitrary functions and first-order equations **627–32**
 - boundary conditions **718–22**
 - finite elements **694–706**
 - formal classification of **716–18**
 - heat-conduction or diffusion equation **617**, **620–3**
 - solution of **660–76**
 - Laplace transform method **664–9**
 - numerical solution **671–6**
 - separation method **660–4**
 - sources and sinks for **712–15**
 - Helmholtz equation **626**
 - integral solutions **707–15**
 - separated solutions **707–9**
 - singular solutions **709–12**
 - Laplace equation **617**, **623–733**
 - solution of **677–693**
 - numerical solution **686–693**
 - separated solutions **677–84**
 - Poisson equation **626**
 - Reynolds number **625**
 - Schrödinger equation **626**
 - wave equation **617–20**
 - solution of **634–59**
 - D'Alembert solution **634–43**, **637**
 - Laplace transform solution **648–51**
 - numerical solution **653–9**
 - method of separation of variables **643–8**
 - particular integral **87**
 - Paterson, Colin 474
 - path of line integral **207**
 - pendulum, oscillations of (application) **162–7**
 - period **486**
 - periodic extension **492**
 - periodic functions **335–9**, **486–7**
 - phase angle **487**
 - phase plane **80**
 - phase quadrature components **487**
 - phase shift **392**
 - phase spectrum **515**, **548**, **604**
 - phases in linear programming **753–61**
 - point at infinity **295**
 - Poisson distribution **804–5**
 - Poisson equation **626**
 - Poisson process in queues **909–16**, **911**
 - polar plot **397**
 - pole placement (application) **398–9**, **399**
 - poles **300**, **436**
 - and eigenvalues 399
 - pole-zero plot **357**
 - polynomial approximation **772**
 - polynomial mapping **272–4**
 - population **808**
 - population mean **810–11**
 - positive constant in matrices **101**
 - positive definite function **101**
 - positive-definite quadratic forms **46**, **48**
 - positive-semidefinite quadratic forms **46**, **48**
 - posterior odds **933**
 - posterior probabilities **933**

Powell 782
 power **565**
 power method on matrices 29–35, **31**
 power series **287**–91
 power signals **568**
 power spectrum 521–3, **522**
 practical signal **541**
 predictor–corrector methods on differential equations 134–9, **136**
 principal diagonal **3**
 principal part of Laurent series **296**
 principle minor of matrices **47**
 principle of superposition **374**
 prior odds **933**
 prior probabilities **933**
 probability density function **802**
 probability theory 800–8
 Bernoulli distribution **804**–5
 binomial distribution **804**–5
 central limit theorem **806**
 normal distribution 806–8
 Poisson distribution **804**–5
 random variables **802**–4
 rules 801–2
 sample measures 808–10
 product of eigenvalues 27
 product rule **801**
 proportion, interval and test for 821–4
 pseudo inverse square matrix 72–8, **73**
 punctured disc **297**

Q

QR methods in matrices **35**
 quadratic forms of matrices **44**–50, **102**
 quadratic polynomial **772**
 quasi-Newton method 782
 queues 908–29
 multiple service channels queue 921–3
 Poisson process in 909–16
 problems 909
 simulation 923–9
 single service channel queue 916–21
 quiescent state **356**

R

radius of convergence **288**
 random variables **802**–4
 range (R) charts **905**–7
 range of function **251**

rank correlation **838**–40
 rank of a matrix **8**–9
 rate of arrival **910**
 real vector space **9**
 realization problem **385**
 reciprocal basis vectors of matrix **92**
 rectangular matrix **63**
 rectangular window **605**
 reduction to Jordan normal **40**
 regression 841–61, **842**
 and correlation 856
 least squares method 842–52
 linear **843**
 nonlinear 856–61
 residuals 852–5
 regression coefficients **843**
 regular function **275**
 regular point of $f(z)$ **300**
 removable singularity **301**
 repeated eigenvalues 22–6
 residuals in regression 852–5
 resonance **503**
 Reynolds number in partial differential equations **625**
 Richardson extrapolation **134**
 Riemann sphere **295**
 right inverse matrix **74**
 right singular vector matrix **68**
 Roberts, R.A. 346–7
 Robinson, A. 922
 robust methods **800**
 root mean square (RMS) **514**
 rotation 255, 256
 rotational motion **201**
 Routh–Hurwitz criterion **362**
 row rank matrix **63**–4
 row vector **3**
 rows **653**
 Rubin, D.B. 943
 rule of total probability **931**
 Runge–Kutta methods on differential equations 139–42

S

sample **808**
 sample average **808**
 distribution of 810–12
 sample correlation **833**–5
 sample measures in probability theory 808–10
 sample range **905**
 sample space **801**
 sample variance **808**
 sampling **408**, 413–14

- sampling function **521**
 scalar field **177**, 202
 scalar Lyapunov function **101**
 scalar point function 176
 derivatives of 192–5
 gradient 192–4
 scalar product **28**
 scatter diagrams **833**
 Schafer, R.W. 609
 Schrödinger equation **626**
 Schwarzenbach, J. 370
 second shift property of z transforms
 417–18
 second shift theorem 325–8
 inversion 328–31
 separation method for solution of heat-conduction/diffusion equation
 660–4
 separation of variables method
 in Laplace equation method 677–84
 of partial differential equations 707–9
 to wave equation **643**–8
 service channel **909**
 service discipline **909**
 service time **909**
 set of vectors **10**
 Seutari, M. 935
 Shanno 784
 Shewart attribute control charts 891–4
 Shewart variable control charts 894–7
 shooting method in differential equations
 160–2, **162**
 sifting property **342**–3
 significance levels **816**
 signum function **571**
 similarity transform **36**
 simple pole **301**
 simplex algorithm 741–4, **742**, **745**
 general theory 745–51
 simplex method **740**
 simplification **2**
 simulation, queues 923–9
 simultaneous differential equations, Laplace transform on 318–19
 sine function **520**
 Singer, A. 727
 single-input-single-output (SISO) systems
 79–**83**
 in Laplace transforms 378–82
 in matrices **79**–**83**
 single multivariable searches in hill climbing
 775–81
 single service channel queue 916–21
 distribution of number of customers in the system 917–18
 queue length and waiting time 918–21
 single-variable search in hill climbing
 769–74
 singular points **765**
 singular solutions of partial differential equations 709–12
 singular value decomposition of matrices
 63–78, **69**
 pseudo inverse 72–8
 singular value matrix **65**
 singularities **289**, **295**, **300**–3
 sinks in solution of heat-conduction/diffusion equation 712–15
 skew symmetric matrix **3**
 slack variables **741**
 Snell's law **766**
 solenoidal vectors **198**
 sources in solution of heat-conduction/diffusion equation 712–15
 Spearman rank correlation coefficient **838**
 spectral form in matrices **93**
 spectral leakage **606**
 spectral matrix **36**
 spectral pairs in matrix **92**
 spectral representation of response of state equations 92–5
 Spiegelhalter, D.J. 935, 937
 square matrix **3**, 65
 square non-singular matrix **73**
 stability
 in differential equations **130**
 in discrete-linear systems 444–50
 in transfer functions 359–64
 stable linear system **359**
 standard deviation **802**
 standard form of transfer function **394**
 standard normal distribution **806**
 standard tableau **745**
 state equation **80**
 state equation, solution of 86–99
 canonical representation 95–9
 direct form 86–8
 spectral representation of response 92–5
 transition matrix 88
 evaluating 89–91
 state feedback **398**
 state variables **80**
 state vector **80**
 state-space **2**, **80**
 state-space form **478**
 state-space model **81**
 state-space representation
 of higher order systems 154–7
 in Laplace transforms 378–89
 multi-input-multi-output (MIMO) systems **383**–**9**
 single-input-single-output (SISO) systems 378–82
 in matrices
 multi-input-multi-output (MIMO) systems **84**–**5**
 single-input-single-output (SISO) systems **79**–**83**

- statistical quality control (application) 891–907
 Cusum control charts 898–901
 moving-average control charts 901–5
 range charts 905–7
 Shewart attribute control charts 891–4
 Shewart variable control charts 894–7
statistics 800
 steady-state errors 368
 steady-state gain 368
 Stearns, S.J. 609
 Steele, N.C. 471, 475, 601, 602
 steepest ascent/descent 776
 step functions
 in Fourier transforms 563–74
 Laplace transforms on 320–56
 step size in Euler's method 118
 step-invariant method 466–9, 467
 Stern, H.S. 943
 stiff differential equations 145–7
 stiffness matrix 698
 Stirzaker, D.R. 923
 Stokes' theorem 236–9, 237
 stream function 240
 streamline in fluid dynamics (application) 240–2
 subdominant eigenvalue 30
 successive over-relaxation (SOR) method 689–90
 sum of eigenvalues 27
 superposition integral 371
 superposition principle 374
 surface integrals 222–8
 surplus variable 754
 Sylvester's conditions 47, 48, 102
 symmetric matrix 28–9, 65
 symmetry property 555–7, 556
 system discrete 436
 system frequency response 561
- top hat function 321
 total differential 189
 in vector calculus 188–91
 trace 3
 trajectory 80
 transfer functions 356–77
 and arbitrary inputs 374–7
 convolution 371–4
 definitions 356–9
 final-value theorem 367–70
 impulse response 364–5
 initial-value theorem 365–7
 transfer matrix 384
 transformations 185, 251
 in vector calculus 185–7
 of vector spaces 11–12
 transition matrix 88
 in discrete-time state equations 459
 evaluating 89–91
 transition property 88
 translation 255, 256
 transmission line 728
 transposed matrix 3, 4, 65
 properties 4
 Tranter, W.H. 609
 travelling waves 636
 triangular window 609
 Tucker 768
 Tukey, J.W. 538, 592
 Tustin transform 473, 475
 two phase strategy 756
 two-dimensional heat equation 623
 two-phase method 753–61
 two-sided Laplace transform 317
 type I error 815
 type II error 815

T

- tableau form 742
 Taylor series 291–4
 Taylor series expansion 292
 Taylor theorem 778
 testing simple hypotheses 815–16
 text statistic 815
 thermal diffusivity 243, 621
 thermally isotropic medium 242
 Thomas, A. 937
 Thomas algorithm 658
 time-differentiation property 552–3, 553
 time-shift property 553–4, 554

U

- unbounded region 746
 uncontrollable modes in matrix 97
 under determined matrix 72
 unilateral Laplace transform 558
 unit impulse function 341
 unit matrix 3
 unit pulse 412
 unit step function 320, 323–5
 unitary matrix 65
 unobservable state of matrix 97
 unrestricted in sign in linear programming 761–2
 upper control limit (UCL) in control charts 894
 utilization 921

V

- variable 891
- variance **802**
 - unknown, mean when 817–19
- variational problems **793**
- vector calculus 175–239
 - basic concepts 177–84
 - derivatives of scalar point function 192–5
 - gradient 192–5, **193**
 - derivatives of vector point function 196–206
 - curl of a vector field 199–201
 - divergence of a vector field 196–8
 - vector operator 202–6
- domain **176**
- integration 206–39
 - double integrals 211–16
 - Gauss’s divergence theorem 233–6
 - Green’s theorem 217–21
 - line integral 207–10
 - Stokes’ theorem 236–9
 - surface integrals 222–8
 - volume integrals 229–32
- rule **176**
- total differential 188–91
 - transformations 185–7
- vector field **177**, 202
 - divergence of 196–8
- vector-matrix differential equation **80**
- vector point function **176**
 - derivatives of vector point function 196–206
 - curl of a vector field 199–201
 - divergence of a vector field 196–8
 - vector operator 202–6
- vector spaces in matrices 9–12
 - linear independence 10–11
 - transformation between bases 11–12
- vectors **9**
- Vehtari, A. 943
- viscous fluid, motion in (application) 114–15
- volume integrals 229–32
- vortex **242**

W

- Ward, J.P. 694
- warning in control charts **891**
- wave equations in partial differential equations **617**–20
 - solution of 634–60
 - D’Alembert solution 634–43

- Laplace transform solution 648–51
- numerical solution 653–9
- method of separation of variables **643**–8
- wave propagation under moving load (application) 723–6
- ‘weak’ form **696**
- weighting factor **76**
- weighting function **364**
 - in transfer functions **364**
- window functions **605**, 607–11

Z

- z* transform function **436**
- z* transform method for solving linear constant-coefficient difference equations 431
- z* transform operator **409**
- z* transform pair **409**
- z* transforms **409**
 - definition and notation 409–13
 - discrete-linear systems 435–54
 - convolution 450–4
 - impulse response 441–4
 - stability 444–50
 - discrete-time state equations 459–63
 - discrete-time state-space equations in 456–63
 - discrete-time systems in 428–34
 - design of (application) 470–3
 - state-space model in 456–8
 - discretization of continuous-time state-space models 464–9
 - Euler’s method **464**–6
 - inverse *see* inverse *z* transforms and Laplace transform 455–6
 - properties 414–19
 - final-value theorem 419
 - first shift property **416**–17
 - initial-value theorem 419
 - linearity 415–16
 - multiplication 418–19
 - second shift property 417–18
 - sampling 413–14
 - table of 419
 - zero crossing **522**
 - zero matrix **3**
 - zero of $f(z)$ **300**–3
 - zero-order hold device **408**
 - zeros of discrete transfer function **436**
 - zeros of transfer function **357**
 - Ziemer, R.E. 609

Building on the foundations laid in the companion text *Modern Engineering Mathematics*, this book gives an extensive treatment of key advanced areas of mathematics that have applications in various fields of engineering, particularly as tools for computer-based system modelling, analysis and design.

The philosophy of learning by doing is retained throughout this advanced level text, with continuing emphasis on the development of students' ability to use mathematics with understanding to solve engineering problems.

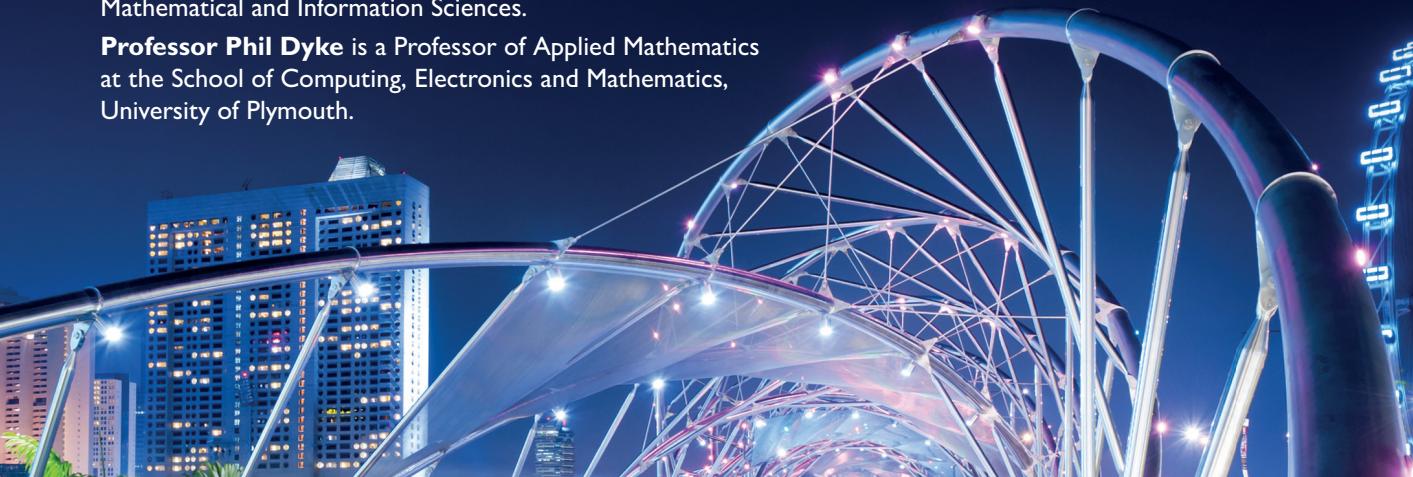
Key features of this new edition:

- A set of new, relevant engineering examples are incorporated to develop an applied understanding of these mathematical techniques
- Exercises with worked solutions, and review exercises at the end of each chapter, help you to test your understanding and master these techniques
- A wide range of MATLAB and MAPLE examples, along with basic commands and illustrations, are included throughout the book
- Examples in R are integrated in the Applied Probability and Statistics section
 - Includes state-of-the-art tools for data manipulation and visualization
 - Emphasises learning by doing through repeated practice
- Lecturer solutions manual and PowerPoint slides are available to download from www.pearsoned.co.uk/james

About the authors:

Professor Glyn James is currently Emeritus Professor in Mathematics at Coventry University, having previously been Dean of the School of Mathematical and Information Sciences.

Professor Phil Dyke is a Professor of Applied Mathematics at the School of Computing, Electronics and Mathematics, University of Plymouth.



Cover image © Tomml/E+/Getty Images

www.pearson-books.com

ISBN 978-1-292-17434-1

A standard linear barcode representing the ISBN 9781292174341. The barcode is oriented vertically and is used for book identification and tracking.

9 781292 174341 >