



FURTHER MATHEMATICS FOR

**ECONOMIC
ANALYSIS**



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

Under a range of well-known imprints, including
Financial Times/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

FURTHER MATHEMATICS FOR ECONOMIC ANALYSIS

SECOND EDITION

KNUT SYDSÆTER
PETER HAMMOND
ATLE SEIERSTAD
ARNE STRØM

FT Prentice Hall
FINANCIAL TIMES

An imprint of Pearson Education

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

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies throughout the world

Visit us on the World Wide Web at:
www.pearsoned.co.uk

First edition published 2005
Second edition published 2008

© Knut Sydsæter, Peter Hammond, Atle Seierstad and Arne Strøm 2005, 2008

The rights of Knut Sydsæter, Peter Hammond, Atle Seierstad and Arne Strøm to be identified as authors of this work have been asserted by them in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without either the prior written permission of the publisher or a licence permitting restricted copying in the United Kingdom issued by the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

ISBN 978-0-273-71328-9

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

10 9 8 7 6 5 4
12 11 10 09

Typeset in TeX by the authors
Printed and bound by Ashford Colour Press Ltd., Gosport

The publisher's policy is to use paper manufactured from sustainable forests.

CONTENTS

Preface	ix	3 Static Optimization	103
1 Topics in Linear Algebra	1	3.1 Extreme Points	104
1.1 Review of Basic Linear Algebra	2	3.2 Local Extreme Points	110
1.2 Linear Independence	7	3.3 Equality Constraints: The Lagrange Problem	115
1.3 The Rank of a Matrix	11	3.4 Local Second-Order Conditions	125
1.4 Main Results on Linear Systems	14	3.5 Inequality Constraints: Nonlinear Programming	129
1.5 Eigenvalues	19	3.6 Sufficient Conditions	135
1.6 Diagonalization	25	3.7 Comparative Statics	139
1.7 Quadratic Forms	28	3.8 Nonnegativity Constraints	143
1.8 Quadratic Forms with Linear Constraints	35	3.9 Concave Programming	148
1.9 Partitioned Matrices and Their Inverses	37	3.10 Precise Comparative Statics Results	150
2 Multivariable Calculus	43	3.11 Existence of Lagrange Multipliers	153
2.1 Gradients and Directional Derivatives	44	4 Topics in Integration	157
2.2 Convex Sets	50	4.1 Review of One-Variable Integration	157
2.3 Concave and Convex Functions I	53	4.2 Leibniz's Formula	159
2.4 Concave and Convex Functions II	63	4.3 The Gamma Function	164
2.5 Quasiconcave and Quasiconvex Functions	68	4.4 Multiple Integrals over Product Domains	166
2.6 Taylor's Formula	77	4.5 Double Integrals over General Domains	171
2.7 Implicit and Inverse Function Theorems	80	4.6 The Multiple Riemann Integral	175
2.8 Degrees of Freedom and Functional Dependence	89	4.7 Change of Variables	178
2.9 Differentiability	93	4.8 Generalized Double Integrals	186
2.10 Existence and Uniqueness of Solutions of Systems of Equations	98		

5 Differential Equations I: First-Order Equations in One Variable	189	9 Control Theory: Basic Techniques	305	12.4 The Maximum Principle	441	Appendix A	
5.1 Introduction	190	9.1 The Basic Problem	306	12.5 More Variables	444	Sets, Completeness, and Convergence	525
5.2 The Direction is Given: Find the Path!	193	9.2 A Simple Case	308	12.6 Stochastic Optimization	448	A.1 Sets and Functions	525
5.3 Separable Equations	194	9.3 Regularity Conditions	312	12.7 Infinite Horizon Stationary Problems	458	A.2 Least Upper Bound Principle	530
5.4 First-Order Linear Equations	200	9.4 The Standard Problem	314			A.3 Sequences of Real Numbers	533
5.5 Exact Equations and Integrating Factors	206	9.5 The Maximum Principle and the Calculus of Variations	322	13 Topology and Separation	465	A.4 Infimum and Supremum of Functions	541
5.6 Transformation of Variables	208	9.6 Adjoint Variables as Shadow Prices	324	13.1 Point Set Topology in $\mathbb{R}^n$	465	Appendix B	
5.7 Qualitative Theory and Stability	211	9.7 Sufficient Conditions	330	13.2 Topology and Convergence	471	Trigonometric Functions	545
5.8 Existence and Uniqueness	217	9.8 Variable Final Time	336	13.3 Continuous Functions	475	B.1 Basic Definitions and Results	545
		9.9 Current Value Formulations	338	13.4 Maximum Theorems	481	B.2 Differentiating Trigonometric Functions	551
6 Differential Equations II: Second-Order Equations and Systems in the Plane	223	9.10 Scrap Values	341	13.5 Convex Sets	486	B.3 Complex Numbers	555
6.1 Introduction	223	9.11 Infinite Horizon	348	13.6 Separation Theorems	491		
6.2 Linear Differential Equations	226	9.12 Phase Diagrams	353	13.7 Productive Economies and Frobenius's Theorem	495	Answers	559
6.3 Constant Coefficients	228	10 Control Theory with Many Variables	359			References	605
6.4 Stability for Linear Equations	235	10.1 Several Control and State Variables	360	14 Correspondences and Fixed Points	499	Index	609
6.5 Simultaneous Equations in the Plane	237	10.2 Some Examples	366	14.1 Correspondences	500		
6.6 Equilibrium Points for Linear Systems	243	10.3 Infinite Horizon	370	14.2 A General Maximum Theorem	509		
6.7 Phase Plane Analysis	246	10.4 Existence Theorems and Sensitivity	373	14.3 Fixed Points for Contraction Mappings	513		
6.8 Stability for Nonlinear Systems	251	10.5 A Heuristic Proof of the Maximum Principle	377	14.4 Brouwer's and Kakutani's Fixed Point Theorems	516		
6.9 Saddle Points	255	10.6 Mixed Constraints	380	14.5 Equilibrium in a Pure Exchange Economy	521		
7 Differential Equations III: Higher-Order Equations	259	10.7 Pure State Constraints	383				
7.1 Linear Differential Equations	259	10.8 Generalizations	386				
7.2 The Constant Coefficients Case	263	11 Difference Equations	389				
7.3 Stability of Linear Differential Equations	266	11.1 First-Order Difference Equations	390				
7.4 Systems of Differential Equations	269	11.2 Economic Applications	396				
7.5 Stability for Nonlinear Systems	273	11.3 Second-Order Difference Equations	401				
7.6 Qualitative Theory	278	11.4 Linear Equations with Constant Coefficients	404				
7.7 A Glimpse at Partial Differential Equations	280	11.5 Higher-Order Equations	410				
8 Calculus of Variations	287	11.6 Systems of Difference Equations	415				
8.1 The Simplest Problem	288	11.7 Stability of Nonlinear Difference Equations	419				
8.2 The Euler Equation	290	12 Discrete Time Optimization	423				
8.3 Why the Euler Equation is Necessary	293	12.1 Dynamic Programming	424				
8.4 Optimal Savings	298	12.2 The Euler Equation	433				
8.5 More General Terminal Conditions	300	12.3 Infinite Horizon	435				

Supporting resources

Visit www.pearsoned.co.uk/sydsaeter to find valuable online resources

For students

- Downloadable Student's Manual with more extensive answers to selected problems

For instructors

- Downloadable Instructor's Manual, including additional exam style problems (with answers)

For more information please contact your local Pearson Education sales representative or visit www.pearsoned.co.uk/sydsaeter

Custom Publishing

Custom publishing allows academics to pick and choose content from one or more textbooks for their course and combine it into a definitive course text.

Here are some common examples of custom solutions which have helped over 500 courses across Europe:

- Different chapters from across our publishing imprints combined into one book
- Lecturer's own material combined together with textbook chapters or published in a separate booklet
- Third-party cases and articles that you are keen for your students to read as part of the course

The Pearson Education custom text published for your course is professionally produced and bound – just as you would expect from a normal Pearson Education text. Since many of our titles have online resources accompanying them we can even build a custom website that matches your course text.

Many adopters of *Further Mathematics for Economic Analysis* have found that they require just one or two extra chapters from the companion volume *Essential Mathematics for Economic Analysis* or would like to select a range of chapters from both texts.

Custom publishing has allowed these adopters to provide access to additional chapters for their students both online and in print.

If, once you have had time to review this title, you feel Custom publishing might benefit you and your course, please do get in contact. However minor, or major, the change – we can help you out.

For more details on how to make your chapter selection for your course please go to www.pearsoned.co.uk/sydsaeter

You can contact us at: www.pearsoncustom.co.uk or via your local representative at: www.pearsoned.co.uk/replocator

PREFACE

I came to the position that mathematical analysis is not one of many ways of doing economic theory: it is the only way. Economic theory is mathematical analysis. Everything else is just pictures and talk.

—R. E. Lucas, Jr. (2001)

This book is intended for advanced undergraduate and graduate students of economics whose mathematical requirements go beyond the material usually taught in undergraduate courses. In particular, it presents most of the mathematical tools required for typical graduate courses in economic theory—both micro and macro. There are many references to Sydsaeter and Hammond's *Essential Mathematics for Economic Analysis, 3rd Edition*, FT/Prentice Hall, 2008 (generally referred to as EMEA throughout), but that book is by no means a prerequisite. Indeed, this successor volume is designed to be accessible to anybody who has had a basic training in multivariable calculus and linear algebra at the level often encountered in courses taught to economics undergraduates. Like EMFA, the treatment here is deliberately quite rigorous, but rigour is not emphasized for its own sake.

An important aspect of the book is its systematic treatment of the calculus of variations, optimal control theory, and dynamic programming. Recent years may have seen control theory lose some of its prominence in economics, but it is still used in several areas, notably resource economics and industrial organization. Furthermore, in our view the existing economics literature has paid too little attention to some of the subtler issues that frequently arise, especially when the time horizon is infinite.

Some early chapters review and extend elementary matrix algebra, multivariable calculus, and static optimization. Other chapters present multiple integration, as well as ordinary difference and differential equations, including systems of equations. There is a chapter on elementary topology in $\mathbb{R}^n$ and separation theorems. In the final chapter we discuss correspondences and the fixed point theorems that economists most often use.

As the title suggests, this is a mathematics book with the material arranged to allow progressive learning of mathematical topics. If the student acquires some economic insight and intuition at the same time, so much the better. At times, we do emphasize economics not only to motivate a mathematical topic, but also to help acquire mathematical intuition. Obviously, our economic discussions will be more easily understood by students who have some basic familiarity with elementary economics.

In particular, this is not a book about economics or even about mathematical economics. As one reviewer of EMEA put it: "Mathematics is the star of the show". We expect students to learn economic theory systematically in other courses, based on other books or articles. We will have succeeded if they can concentrate on the economics in these courses, having mastered beforehand the relevant mathematical tools we present.

Almost every section includes worked examples and problems for students to solve. Many of the problems are quite easy in order to build the students' confidence in absorbing the material, but there are also a number of more challenging problems. Concise solutions to almost all the problems are suggested in the answers section. More extensive answers to selected problems are included in a Student's Manual that can be downloaded from the book's website, which will also include some proofs and comments that we did not include in the book.

The book is not intended to be studied in a steady progression from beginning to end. Some of the more challenging chapters start with a simple treatment where some technical aspects are played down, while the more complete theory is discussed later. Some of the material, including more challenging proofs, is in small print. Quite often those proofs rely on technical ideas that are only expounded in the last two chapters. So there could be a lot to gain from delaying the more challenging material in earlier chapters till the basic concepts in later chapters have been mastered.

The author team consists of the two co-authors of EMEA, together with two other mathematicians in the Department of Economics at the University of Oslo.

Changes in the Second Edition

We have been gratified by the number of students and their instructors in many countries who appear to have found the first edition useful. One result is that we have been encouraged to produce a revised edition, so we have gone through the text thoroughly. The result is fairly extensive rewriting of some chapters, and numerous other less significant improvements. Some of the more notable changes from the first edition include:

- (1) Answers are now given to (almost) all the problems. (Previously only odd-numbered problems had answers supplied in the text.)
- (2) More extensive answers to selected problems are presented in a new Student's Manual. (Those problems are marked in the text with )
- (3) The Instructor's Manual now has a number of supplementary problems, with answers. As before it has comments on the content. A new feature is many simpler problems that could be suitable for 2-3 hour exams.

System of Cross References

Some information about our system of references might be in order. Theorem 13.3.2 is the second theorem in the third section of Chapter 13. When an example, note, or formula is referred to within a section, we use just a single number, as in Example 4 or formula (12). But if we refer to an example, a note, or a formula from another section we use a three part reference. For instance, (3.6.2) refers to formula 2 in Section 6 of Chapter 3.

Acknowledgements

Over the years we have received help from so many colleagues, lecturers at other institutions, and students, that it is impractical to mention them all. Still, some should be explicitly mentioned.

For arranging production and publication, apart from our very helpful editors, with Ellen Morgan and Tim Parker at Pearson Education in England in charge, we should particularly like to thank Arve Michaelsen at Matematisk Sats in Norway. He created the layout and macros for the book and many of the figures.

We would also like to thank the proofreader David Hemsley for his detailed suggestions and corrections.

To these and all the many unnamed persons and institutions who have helped us make this text possible, including some whose comments on our earlier book were forwarded to us by the publisher, we would like to express our deep appreciation and gratitude, combined with the hope that they may find the resulting product reasonably satisfactory, and even of benefit to their students. That, we can all agree, is what really matters in the end.

Knut Sydsæter, Peter Hammond, Atle Seierstad, Arne Strøm
Oslo and Coventry, May 2008

TOPICS IN LINEAR ALGEBRA

*The economic world is a misty region.
The first explorers used unaided vision.
Mathematics is the lantern by which what was before
dimly visible now looms up in firm, bold outlines.
The old phantasmagoria¹ disappear.
We see better. We also see further.
—Irving Fisher (1892)*

This chapter covers a few topics in linear algebra that are not always treated in standard mathematics courses for economics students. We assume that the reader is familiar with some basic concepts and results, which are nevertheless briefly reviewed in Section 1.1. A fuller treatment, including many practice problems, can be found in EMEA or in many alternative textbooks.

In an economic model described by a linear system of equations, it is important to know when that system has a solution, and when the solution is unique. General conditions for existence and uniqueness are most easily stated using the concept of linear independence, along with the related concept of the rank of a matrix. These topics are treated in Sections 1.2 and 1.3. They are applied to linear equation systems in Section 1.4. Two important theorems give crucial information about the solutions. In particular, Theorem 1.4.2(b) introduces the important concept of degrees of freedom for linear systems.

Section 1.5 discusses eigenvalues. They are indispensable in several areas of mathematics of interest to economists—in particular, stability theory for difference and differential equations. Eigenvalues and the associated eigenvectors are also important in determining when a matrix can be “diagonalized”, which greatly simplifies some calculations involving the matrix. This is discussed in Section 1.6.

Sections 1.7 and 1.8 look at quadratic forms—first without linear constraints, then with them. Such quadratic forms are especially useful in deriving and checking second-order conditions for multivariable optimization.

Lastly, in Section 1.9 we briefly consider partitioned matrices. These are useful for computations involving large matrices, especially when they have a special structure. One area of application is in econometrics.

¹ “Phantasmagoria” is a term invented in 1802 to describe an exhibition of optical illusions produced by means of a magic lantern.

1.1 Review of Basic Linear Algebra

An $m \times n$ matrix is a rectangular array with m rows and n columns:

$$A = (a_{ij})_{m \times n} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \quad (1)$$

Here a_{ij} denotes the element in the i th row and the j th column.

If $A = (a_{ij})_{m \times n}$, $B = (b_{ij})_{n \times n}$, and α is a scalar (a number), we define

$$A+B = (a_{ij}+b_{ij})_{m \times n}, \quad \alpha A = (\alpha a_{ij})_{m \times n}, \quad A-B = A+(-1)B = (a_{ij}-b_{ij})_{m \times n} \quad (2)$$

Suppose that $A = (a_{ij})_{m \times n}$ and that $B = (b_{ij})_{n \times p}$. Then the product $C = AB$ is the $m \times p$ matrix $C = (c_{ij})_{m \times p}$, whose element in the i th row and the j th column is the inner product (or dot product) of the i th row of A and the j th column of B . That is,

$$c_{ij} = \sum_{r=1}^n a_{ir}b_{rj} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj} \quad (3)$$

It is important to note that the product AB is defined only if the number of columns in A is equal to the number of rows in B .

If A , B , and C are matrices whose dimensions are such that the given operations are defined, then the basic properties of matrix multiplication are:

$$(AB)C = A(BC) \quad (\text{associative law}) \quad (4)$$

$$A(B+C) = AB+AC \quad (\text{left distributive law}) \quad (5)$$

$$(A+B)C = AC+BC \quad (\text{right distributive law}) \quad (6)$$

If A and B are matrices, it is possible for AB to be defined even if BA is not. Moreover, even if AB and BA are both defined, AB is not necessarily equal to BA . Matrix multiplication is *not* commutative. In fact,

$$AB \neq BA, \quad \text{except in special cases} \quad (7)$$

$$AB = 0 \quad \text{does not imply that } A \text{ or } B \text{ is } 0 \quad (8)$$

$$AB = AC \quad \text{and } A \neq 0 \quad \text{do not imply that } B = C \quad (9)$$

By using matrix multiplication, one can write a general system of linear equations in a very concise way. Specifically, the system

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

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

$$\vdots$$

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

can be written as $Ax = b$

$$\text{if we define } A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

A matrix is **square** if it has an equal number of rows and columns. If A is a square matrix and n is a positive integer, we define the n th power of A in the obvious way:

$$A^n = \underbrace{AA \cdots A}_{n \text{ factors}} \quad (10)$$

For **diagonal matrices** it is particularly easy to compute powers:

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_m \end{pmatrix} \implies D^n = \begin{pmatrix} d_1^n & 0 & \cdots & 0 \\ 0 & d_2^n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_m^n \end{pmatrix} \quad (11)$$

The **identity matrix** of order n , denoted by I_n (or often just by I), is the $n \times n$ matrix having ones along the main diagonal and zeros elsewhere:

$$I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}_{n \times n} \quad (\text{identity matrix}) \quad (12)$$

If A is any $m \times n$ matrix, then $AI_n = A = I_m A$. In particular,

$$AI_n = I_n A = A \quad \text{for every } n \times n \text{ matrix } A \quad (13)$$

If $A = (a_{ij})_{m \times n}$ is any matrix, the **transpose** of A is defined as $A' = (a_{ji})_{n \times m}$. The subscripts i and j are interchanged because every row of A becomes a column of A' , and every column of A becomes a row of A' .

The following rules apply to matrix transposition:

$$(i) (A')' = A \quad (ii) (A+B)' = A' + B' \quad (iii) (\alpha A)' = \alpha A' \quad (iv) (AB)' = B'A' \quad (14)$$

A square matrix is called **symmetric** if $A = A'$.

Determinants and Matrix Inverses

Recall that the determinants $|A|$ of 2×2 and 3×3 matrices are defined by

$$|A| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$$

$$|A| = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} \\ -a_{12}a_{22}a_{33} + a_{13}a_{23}a_{32} - a_{13}a_{22}a_{31} \end{vmatrix}$$

Determinants of order 2 and 3 have a geometric interpretation which is shown and explained in Fig. 1 for the case $n = 3$.

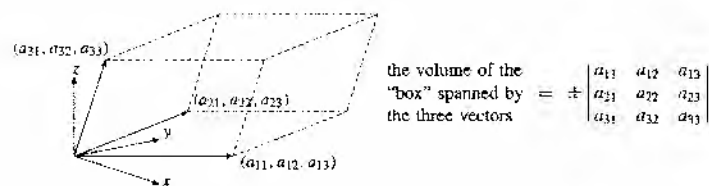


Figure 1

For a general $n \times n$ matrix $A = \{a_{ij}\}$, the determinant $|A|$ can be defined recursively. In fact, for any $i = 1, 2, \dots, n$,

$$|A| = a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{ij}A_{ij} + \cdots + a_{in}A_{in} \quad (15)$$

where each cofactor A_{ij} is the determinant of an $(n-1) \times (n-1)$ matrix given by

$$A_{ij} = (-1)^{i+j} \begin{vmatrix} a_{11} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2,j-1} & a_{2,j+1} & \cdots & a_{2n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i1} & \cdots & a_{i,j-1} & a_{i,j+1} & \cdots & a_{in} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n,j-1} & a_{n,j+1} & \cdots & a_{nn} \end{vmatrix} \quad (16)$$

Lines have been drawn through row i and column j , which are to be deleted from the matrix A to produce A_{ij} . Formula (15) gives the cofactor expansion of $|A|$ along the i th row.

In general,

$$a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in} = |A| \quad (17)$$

$$a_{i1}A_{k1} + a_{i2}A_{k2} + \cdots + a_{in}A_{kn} = 0 \quad (k \neq i)$$

$$a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj} = |A| \quad (18)$$

$$a_{1j}A_{1k} + a_{2j}A_{2k} + \cdots + a_{nj}A_{nk} = 0 \quad (k \neq j)$$

This result says that an expansion of a determinant along row i in terms of the cofactors of row k vanishes when $k \neq i$, and is equal to $|A|$ if $k = i$. Likewise, an expansion along column j in terms of the cofactors of column k vanishes when $k \neq j$, and is equal to $|A|$ if $k = j$.

The following rules for manipulating determinants are often useful:

$$\text{If two rows (or two columns) of } A \text{ are interchanged, the determinant changes sign but its absolute value remains unchanged.} \quad (19)$$

$$\text{If all the elements in a single row (or column) of } A \text{ are multiplied by a number } c, \text{ the determinant is multiplied by } c. \quad (20)$$

$$\text{If two of the rows (or columns) of } A \text{ are proportional, then } |A| = 0. \quad (21)$$

$$\text{The value of } |A| \text{ remains unchanged if a multiple of one row (or one column) is added to another row (or column).} \quad (22)$$

Furthermore,

$$|A'| = |A|, \quad \text{where } A' \text{ is the transpose of } A \quad (23)$$

$$|AB| = |A| \cdot |B| \quad (24)$$

$$|A + B| \neq |A| + |B| \quad (\text{usually}) \quad (25)$$

The inverse A^{-1} of an $n \times n$ matrix A is defined so that it satisfies

$$B = A^{-1} \iff AB = I_n \iff BA = I_n \quad (26)$$

It follows that

$$A^{-1} \text{ exists} \iff |A| \neq 0 \quad (27)$$

If $A = (a_{ij})_{n \times n}$ and $|A| \neq 0$, the unique inverse of A is given by

$$A^{-1} = \frac{1}{|A|} \text{adj}(A), \quad \text{where} \quad \text{adj}(A) = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix} \quad (28)$$

with A_{ij} , the cofactor of the element a_{ij} , given by (16). Note carefully the order of the subscripts in the **adjoint matrix** $\text{adj}(A)$, with the column number preceding the row number. The matrix $(A_{ij})_{n \times n}$ is called the **cofactor matrix**, whose transpose is the adjoint matrix.

In particular, for 2×2 matrices,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad \text{if} \quad \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad-bc \neq 0 \quad (29)$$

The following are important rules for inverses (when the relevant inverses exist):

$$(A^{-1})^{-1} = A, \quad (AB)^{-1} = B^{-1}A^{-1}, \quad (A')^{-1} = (A^{-1})', \quad (cA)^{-1} = c^{-1}A^{-1} \quad (30)$$

Cramer's Rule

Consider a linear system of n equations and n unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \quad (31)$$

This has a unique solution if and only if $|A| = |(a_{ij})_{n \times n}| \neq 0$. The solution is then

$$x_j = |A_j|/|A|, \quad j = 1, 2, \dots, n \quad (32)$$

where $|A_j|$ denotes the determinant of A with its j th column replaced by the column with components $b_1, b_2, \dots, b_n$; that is,

$$|A_j| = \begin{vmatrix} a_{11} & \cdots & a_{1,j-1} & b_1 & a_{1,j+1} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2,j-1} & b_2 & a_{2,j+1} & \cdots & a_{2n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n,j-1} & b_n & a_{n,j+1} & \cdots & a_{nn} \end{vmatrix} \quad (33)$$

If the right-hand side of (31) consists of only zeros, so that it can be written in matrix form as $A\mathbf{x} = \mathbf{0}$, the system is called **homogeneous**. A homogeneous system will always have the **trivial solution** $x_1 = x_2 = \cdots = x_n = 0$. The following result is useful:

$$A\mathbf{x} = \mathbf{0} \text{ (where } A \text{ is square)} \text{ has nontrivial solutions} \iff |A| = 0 \quad (34)$$

Vectors

Recall that an n -vector is an ordered n -tuple of numbers. It is often convenient to regard the rows and columns of a matrix as vectors, and an n -vector can be understood either as a $1 \times n$ matrix $\mathbf{a} = (a_1, a_2, \dots, a_n)$ (a *row vector*) or as an $n \times 1$ matrix $\mathbf{a}' = (a_1, a_2, \dots, a_n)'$ (a *column vector*). The operations of addition and subtraction of vectors, as well as multiplication by scalars, are defined in the obvious way. The **inner product** (**dot product** or **scalar product**) of the n -vectors $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_n)$ is defined as

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^n a_i b_i \quad (35)$$

If $\mathbf{a}$ and $\mathbf{b}$ are regarded as $n \times 1$ matrices, then the inner product of $\mathbf{a}$ and $\mathbf{b}$ is $\mathbf{a}'\mathbf{b}$.

If $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are n -vectors and α is a scalar, then

$$(i) \mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a}, \quad (ii) \mathbf{a} \cdot (\mathbf{b} + \mathbf{c}) = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \cdot \mathbf{c}, \quad (iii) (\alpha \mathbf{a}) \cdot \mathbf{b} = \mathbf{a} \cdot (\alpha \mathbf{b}) = \alpha(\mathbf{a} \cdot \mathbf{b}) \quad (36)$$

The **Euclidean norm** or **length** of the vector $\mathbf{a} = (a_1, a_2, \dots, a_n)$ is

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} \quad (37)$$

Note that $\|\alpha \mathbf{a}\| = |\alpha| \|\mathbf{a}\|$ for all scalars and vectors. The following inequalities also hold:

$$|\mathbf{a} \cdot \mathbf{b}| \leq \|\mathbf{a}\| \cdot \|\mathbf{b}\| \quad (\text{Cauchy-Schwarz inequality}) \quad (38)$$

$$\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\| \quad (\text{triangle inequality for vector norms}) \quad (39)$$

The **angle** θ between nonzero vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^n$ is defined by

$$\cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}, \quad \theta \in [0, \pi] \quad (40)$$

The Cauchy-Schwarz inequality implies that the right-hand side has absolute value ≤ 1 . According to (40), $\cos \theta = 0$ if and only if $\mathbf{a} \cdot \mathbf{b} = 0$. Then $\theta = \pi/2 = 90^\circ$.

By definition, $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^n$ are **orthogonal** if their inner product is 0. In symbols:

$$\mathbf{a} \perp \mathbf{b} \iff \mathbf{a} \cdot \mathbf{b} = 0 \quad (41)$$

The **straight line** through two distinct points $\mathbf{a} = (a_1, \dots, a_n)$ and $\mathbf{b} = (b_1, \dots, b_n)$ in $\mathbb{R}^n$ is the set of all $\mathbf{x} = (x_1, \dots, x_n)$ in $\mathbb{R}^n$ such that, for some real number t ,

$$\mathbf{x} = t\mathbf{a} + (1-t)\mathbf{b} \quad (42)$$

The **hyperplane** in $\mathbb{R}^n$ that passes through the point $\mathbf{a} = (a_1, \dots, a_n)$, and is orthogonal to the nonzero vector $\mathbf{p} = (p_1, \dots, p_n)$, is the set of all points $\mathbf{x} = (x_1, \dots, x_n)$ such that

$$\mathbf{p} \cdot (\mathbf{x} - \mathbf{a}) = 0 \quad (43)$$

1.2 Linear Independence

Any system of linear equations can be written as a vector equation. For instance, the system

$$\begin{aligned} 2x_1 + 2x_2 - x_3 &= -3 \\ 4x_1 &+ 2x_3 = 8 \\ 6x_2 - 3x_3 &= -12 \end{aligned}$$

can be written as the vector equation

$$x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + x_3 \mathbf{a}_3 = \mathbf{b} \quad (*)$$

in the column vectors $\mathbf{a}_1 = \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}$, $\mathbf{a}_2 = \begin{pmatrix} 2 \\ 0 \\ 6 \end{pmatrix}$, $\mathbf{a}_3 = \begin{pmatrix} -1 \\ 2 \\ -3 \end{pmatrix}$, and $\mathbf{b} = \begin{pmatrix} -3 \\ 8 \\ -12 \end{pmatrix}$.

We say that $(*)$ expresses $\mathbf{b}$ as a **linear combination** of the three column vectors of the coefficient matrix A . Solving system $(*)$ we get $x_1 = 1/2$, $x_2 = -1/2$, and $x_3 = 3$. Thus $\mathbf{b} = (1/2)\mathbf{a}_1 + (-1/2)\mathbf{a}_2 + 3\mathbf{a}_3$. In this case, we say that $\mathbf{b}$ is **linearly dependent** on the vectors $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$.

More generally, a set of vectors in $\mathbb{R}^n$ is said to be **linearly dependent** if it has the property that at least one of the vectors can be expressed as a linear combination of the others. Otherwise, if no vector in the set can be expressed as a linear combination of the others, then the set of vectors is **linearly independent**.

It is convenient to have an equivalent but more symmetric definition of linearly dependent and independent vectors:

DEFINITION

The n vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ in $\mathbb{R}^n$ are **linearly dependent** if there exist numbers $c_1, c_2, \dots, c_n$, not all zero, such that

$$c_1 \mathbf{a}_1 + c_2 \mathbf{a}_2 + \cdots + c_n \mathbf{a}_n = \mathbf{0} \quad (1)$$

If this equation holds only in the "trivial" case when $c_1 = c_2 = \cdots = c_n = 0$, then the vectors are **linearly independent**.

So a linear combination of linearly independent vectors can be the zero vector only in the trivial case. To see that the two definitions of linear dependence are equivalent, suppose first that $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ are linearly dependent according to definition (1). Then the equation $c_1 \mathbf{a}_1 + c_2 \mathbf{a}_2 + \cdots + c_n \mathbf{a}_n = \mathbf{0}$ holds with at least one of the coefficients c_i different from 0. After reordering the vectors $\mathbf{a}_i$ and the corresponding scalars c_i , if necessary, we can assume

that $c_1 \neq 0$. Solving equation (1) for $\mathbf{a}_1$ yields $\mathbf{a}_1 = -(c_2/c_1)\mathbf{a}_2 - \cdots - (c_n/c_1)\mathbf{a}_n$. Thus, $\mathbf{a}_1$ is a linear combination of the other vectors.

Suppose on the other hand that $\mathbf{a}_1$, say, can be written as a linear combination of the others, with $\mathbf{a}_1 = d_2\mathbf{a}_2 + d_3\mathbf{a}_3 + \cdots + d_n\mathbf{a}_n$. Then $(-1)\mathbf{a}_1 + d_2\mathbf{a}_2 + d_3\mathbf{a}_3 + \cdots + d_n\mathbf{a}_n = \mathbf{0}$. The first coefficient in this equation is $\neq 0$, so the set $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ is linearly dependent as defined in (1).

EXAMPLE 1

- (a) Prove that $\mathbf{a}_1 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ and $\mathbf{a}_2 = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$ are linearly dependent. Illustrate.
- (b) Prove that $\mathbf{a}_1 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ and $\mathbf{a}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ are linearly independent. Illustrate.

Solution:

- (a) Here $\mathbf{a}_2 = 2\mathbf{a}_1$, so $2\mathbf{a}_1 - \mathbf{a}_2 = \mathbf{0}$. Choosing $c_1 = 2$ and $c_2 = -1$ yields $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 = \mathbf{0}$, so according to definition (1), $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly dependent. The vector $\mathbf{a}_2$ points in the same direction as $\mathbf{a}_1$, and is twice as long. See Fig. 1.

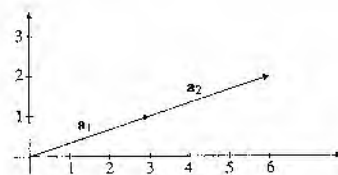


Figure 1 $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly dependent.

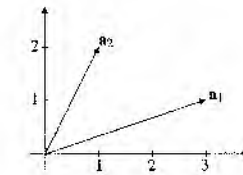


Figure 2 $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly independent.

- (b) In this case the equation $c_1\mathbf{a}_1 + c_2\mathbf{a}_2 = \mathbf{0}$ reduces to

$$\begin{aligned} 3c_1 + c_2 &= 0 \\ c_1 + 2c_2 &= 0 \end{aligned}$$

The only solution is $c_1 = c_2 = 0$, so $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly independent. See Fig. 2. †

It is very helpful to have a geometric feeling for the meaning of linear dependence and independence. For the case of $\mathbb{R}^2$, Example 1 illustrated the possibilities. In $\mathbb{R}^3$, let $\mathbf{a}_1$ and $\mathbf{a}_2$ be two non-parallel 3-vectors starting at the origin. If t_1 and t_2 are real numbers, then the vector $\mathbf{x} = t_1\mathbf{a}_1 + t_2\mathbf{a}_2$ is a linear combination of $\mathbf{a}_1$ and $\mathbf{a}_2$. Geometrically, the set of all linear combinations of $\mathbf{a}_1$ and $\mathbf{a}_2$ is called the plane **spanned** by $\mathbf{a}_1$ and $\mathbf{a}_2$. Any vector in the plane spanned by $\mathbf{a}_1$ and $\mathbf{a}_2$ is linearly dependent on $\mathbf{a}_1$ and $\mathbf{a}_2$.

Suppose we take another 3-vector $\mathbf{a}_3$ that is *not* in the plane spanned by $\mathbf{a}_1$ and $\mathbf{a}_2$. Then the three vectors $\mathbf{a}_1, \mathbf{a}_2$, and $\mathbf{a}_3$ are linearly independent, because no vector in the set can be written as a linear combination of the others. In general, three vectors in $\mathbb{R}^3$ are linearly dependent if and only if they all lie in the same plane. Three vectors in $\mathbb{R}^3$ are linearly

independent if and only if there is no plane that contains all of them. Figures 3 and 4 give geometric illustrations of these statements.

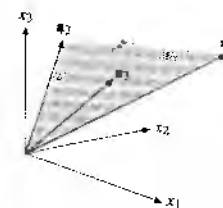


Figure 3 Vectors $\mathbf{a}_1, \mathbf{a}_2$, and $\mathbf{a}_3$ are linearly dependent.



Figure 4 Vectors $\mathbf{a}_1, \mathbf{a}_2$, and $\mathbf{a}_3$ are linearly independent.

In $\mathbb{R}^m$, the two m -vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly dependent if and only if one of the vectors, say $\mathbf{a}_1$, is proportional to the other, so that $\mathbf{a}_1 = c\mathbf{a}_2$. If $c \neq 0$, the two vectors are called **parallel**.

EXAMPLE 2

Suppose that $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are three linearly independent vectors in $\mathbb{R}^n$. Are $\mathbf{a} - \mathbf{b}, \mathbf{b} - \mathbf{c}$, and $\mathbf{a} - \mathbf{c}$ linearly independent?

Solution: Suppose $c_1(\mathbf{a} - \mathbf{b}) + c_2(\mathbf{b} - \mathbf{c}) + c_3(\mathbf{a} - \mathbf{c}) = \mathbf{0}$. Rearranging, we get $(c_1 + c_3)\mathbf{a} + (-c_1 + c_2)\mathbf{b} + (-c_2 - c_3)\mathbf{c} = \mathbf{0}$. Since $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are linearly independent, $c_1 + c_3 = 0$, $-c_1 + c_2 = 0$, and $-c_2 - c_3 = 0$. These equations are satisfied (for example) when $c_1 = c_2 = 1$, and $c_3 = -1$, so $\mathbf{a} - \mathbf{b}, \mathbf{b} - \mathbf{c}$, and $\mathbf{a} - \mathbf{c}$ are linearly dependent. †

Linear Dependence and Systems of Linear Equations

Consider the general system of m equations in n unknowns, written both in its usual form and also as a vector equation:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ \vdots & \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned} \quad \Longleftrightarrow \quad x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b} \quad (2)$$

Here $\mathbf{a}_1, \dots, \mathbf{a}_n$ are the column vectors of coefficients, and $\mathbf{b}$ is the column vector with components $b_1, \dots, b_m$.

Suppose that (2) has two solutions $\mathbf{u}' = (u_1, \dots, u_n)$ and $\mathbf{v}' = (v_1, \dots, v_n)$. Then $u_1\mathbf{a}_1 + \cdots + u_n\mathbf{a}_n = \mathbf{b}$ and $v_1\mathbf{a}_1 + \cdots + v_n\mathbf{a}_n = \mathbf{b}$. Subtracting the second equation from the first yields

$$(u_1 - v_1)\mathbf{a}_1 + \cdots + (u_n - v_n)\mathbf{a}_n = \mathbf{0} \quad (*)$$

Let $c_1 = u_1 - v_1, \dots, c_n = u_n - v_n$. The two solutions $\mathbf{u}'$ and $\mathbf{v}'$ are different if and only if $c_1, \dots, c_n$ are not all equal to 0. We conclude that if system (2) has more than one solution, then the column vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ are linearly dependent. Equivalently: *If the column vectors $\mathbf{a}_1, \dots, \mathbf{a}_n$ are linearly independent, then system (2) has at most one solution.* Without saying more about the right-hand side vector $\mathbf{b}$, however, we cannot know if there are any solutions at all, in general.

Consider, in particular, the case $m = n$.

THEOREM 1.2.1

The n column vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ of the $n \times n$ matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}, \quad \text{where } \mathbf{a}_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix}$$

are linearly independent if and only if $|\mathbf{A}| \neq 0$.

Proof: The vectors $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ are linearly independent if and only if the vector equation $x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = \mathbf{0}$ has only the trivial solution $x_1 = x_2 = \cdots = x_n = 0$. This vector equation is equivalent to a homogeneous system of equations, and according to (1.1.34), the trivial solution is the only one if and only if $|\mathbf{A}| \neq 0$. ■

According to this theorem, three vectors $\mathbf{a}_1, \mathbf{a}_2$, and $\mathbf{a}_3$ in $\mathbb{R}^3$ are linearly dependent if and only if the determinant $|\mathbf{a}_1 \mathbf{a}_2 \mathbf{a}_3|$ of the matrix with columns $\mathbf{a}_1, \mathbf{a}_2$, and $\mathbf{a}_3$ is zero, which is true if and only if the volume shown in Fig. 1.1.1 collapses to zero.

PROBLEMS FOR SECTION 1.2

- Express $\begin{pmatrix} 8 \\ 9 \end{pmatrix}$ as a linear combination of $\begin{pmatrix} 2 \\ 5 \end{pmatrix}$ and $\begin{pmatrix} -1 \\ 3 \end{pmatrix}$.
- Determine which of the following pairs of vectors are linearly independent:
 - $\begin{pmatrix} -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ -6 \end{pmatrix}$
 - $\begin{pmatrix} 2 \\ -1 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \end{pmatrix}$
 - $\begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
- Prove that $\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}$, and $\begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$ are linearly independent. (Use Theorem 1.2.1.)
- Prove that $(1, 1, 1)$, $(2, 1, 0)$, $(3, 1, 4)$, and $(1, 2, -2)$ are linearly dependent.
- If $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are linearly independent vectors in $\mathbb{R}^n$, prove that $\mathbf{a} + \mathbf{b}$, $\mathbf{b} + \mathbf{c}$, and $\mathbf{a} + \mathbf{c}$ are also linearly independent. Is the same true of $\mathbf{a} - \mathbf{b}$, $\mathbf{b} + \mathbf{c}$, and $\mathbf{a} + \mathbf{c}$?
- Suppose that $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$ are all different from $\mathbf{0}$, and that $\mathbf{a} \perp \mathbf{b}$, $\mathbf{b} \perp \mathbf{c}$, and $\mathbf{a} \perp \mathbf{c}$. Prove that $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are linearly independent.
 - Suppose that $\mathbf{a}_1, \dots, \mathbf{a}_n$ are vectors in $\mathbb{R}^n$, all different from $\mathbf{0}$. Suppose that $\mathbf{a}_i \perp \mathbf{a}_j$ for all $i \neq j$. Prove that $\mathbf{a}_1, \dots, \mathbf{a}_n$ are linearly independent.
- A book in mathematics for economists suggests the following definition of linear dependence: A set of vectors $\mathbf{v}_1, \dots, \mathbf{v}_n$ is said to be *linearly dependent* if (and only if) any one of them can be expressed as a linear combination of the remaining vectors. Test this definition on the following three vectors which certainly are linearly dependent: $\mathbf{v}_1 = (1, 0)$, $\mathbf{v}_2 = (1, 1)$, $\mathbf{v}_3 = (2, 2)$.

- Prove that if a set of vectors is linearly dependent, then any superset (that is, any set containing the original set) is also linearly dependent.
- Prove that if a set of vectors is linearly independent, then any subset (that is, any set contained in the original set) is also linearly independent.

1.3 The Rank of a Matrix

Associated with any matrix is an important integer called its *rank*. An $m \times n$ matrix $\mathbf{A}$ has n column vectors, each with m components. The largest number of column vectors in $\mathbf{A}$ that form a linearly independent set is called the *rank* of $\mathbf{A}$, denoted by $r(\mathbf{A})$.

DEFINITION

The **rank** of a matrix $\mathbf{A}$, written $r(\mathbf{A})$, is the maximum number of linearly independent column vectors in $\mathbf{A}$. If $\mathbf{A}$ is the $\mathbf{0}$ matrix, we put $r(\mathbf{A}) = 0$. (1)

This concept is vitally important in stating the main results in the next section concerning the existence and multiplicity of solutions to linear systems of equations.

EXAMPLE 1 The rank of an $n \times n$ matrix $\mathbf{A}$ cannot exceed n , since it has only n columns. In fact, according to Theorem 1.2.1, the n column vectors of $\mathbf{A}$ are linearly independent if and only if $|\mathbf{A}| \neq 0$. We conclude that a square matrix $\mathbf{A}$ of order n has rank n if and only if $|\mathbf{A}| \neq 0$. ■

The rank of a matrix can be characterized in terms of its nonvanishing minors. In general, a **minor** of order k in $\mathbf{A}$ is obtained by deleting all but k rows and k columns, and then taking the determinant of the resulting $k \times k$ matrix.

EXAMPLE 2 Describe all the minors of the matrix $\mathbf{A} = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 2 & 4 & 2 \\ 0 & 2 & 2 & 1 \end{pmatrix}$.

Solution: Because there are only 3 rows, there are minors of order 1, 2, and 3. There are: (a) 4 minors of order 3. These are obtained by deleting any one of the 4 columns:

$$\begin{vmatrix} 1 & 0 & 2 \\ 0 & 2 & 4 \\ 0 & 2 & 2 \end{vmatrix}, \quad \begin{vmatrix} 1 & 0 & 1 \\ 0 & 2 & 2 \\ 0 & 2 & 1 \end{vmatrix}, \quad \begin{vmatrix} 1 & 2 & 1 \\ 0 & 4 & 2 \\ 0 & 2 & 1 \end{vmatrix}, \quad \begin{vmatrix} 0 & 2 & 1 \\ 2 & 4 & 2 \\ 2 & 2 & 1 \end{vmatrix}$$

(b) 18 minors of order 2. These are obtained by deleting one row and two columns, in all possible ways. One of them is:

$$\begin{vmatrix} 0 & 1 \\ 2 & 1 \end{vmatrix} \quad (\text{deleting the second row and the first and third column})$$

(c) 12 minors of order 1. These are all the 12 individual elements of $\mathbf{A}$. ■

NOTE 1 An $m \times n$ matrix has $\binom{n}{k} \binom{m}{k}$ minors of order k . For instance, in Example 2 the 3×4 matrix A has $\binom{3}{2} \binom{4}{2} = \frac{3 \cdot 2}{1 \cdot 2} \cdot \frac{4 \cdot 3}{1 \cdot 2} = 3 \cdot 6 = 18$ minors of order 2. (Recall that the binomial coefficient is defined as $\binom{m}{k} = \frac{m(m-1) \cdots (m-k+1)}{k!} = \frac{m!}{k!(m-k)!}$, with $0! = 1$.)

The relation between the rank and the minors is expressed in the following theorem:²

THEOREM 1.3.1

The rank $r(A)$ of a matrix A is equal to the order of the largest minor of A that is different from 0.

If A is a square matrix of order n , then the largest minor of A is $|A|$ itself. So $r(A) = n$ if and only if $|A| \neq 0$. This agrees with Theorem 1.2.1.

EXAMPLE 3 Find the ranks of the following matrices:

$$(a) \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 2 & 4 & 2 \\ 0 & 2 & 2 & 1 \end{pmatrix} \quad (b) \begin{pmatrix} -1 & 0 & 2 & 1 \\ -2 & 2 & 4 & 2 \\ -3 & 1 & 6 & 3 \end{pmatrix} \quad (c) \begin{pmatrix} -1 & 0 & 2 & 1 \\ -2 & 0 & 4 & 2 \\ -3 & 0 & 6 & 3 \end{pmatrix}$$

Solution:

(a) The rank is 3 because $\begin{vmatrix} 1 & 0 & 2 \\ 0 & 2 & 4 \\ 0 & 2 & 2 \end{vmatrix} = -4$ is a nonzero minor of order 3.

(b) Because columns 1, 3, and 4 are proportional, all four minors of order 3 are 0, whereas

$$\begin{vmatrix} -1 & 0 \\ -2 & 2 \end{vmatrix}, \text{ say, equals } -2. \text{ Hence the rank is 2.}$$

(c) All minors of order 3 and 2 are 0. Because not all the elements are 0, the rank is 1. ■

EXAMPLE 4 Determine the rank of $A = \begin{pmatrix} 5-\lambda & 2 & 1 \\ 2 & 1-\lambda & 0 \\ 1 & 0 & 1-\lambda \end{pmatrix}$ for all values of λ .

Solution: Expanding $|A|$ by the third column, we see that

$$|A| = -(1-\lambda) + (1-\lambda)[(5-\lambda)(1-\lambda) - 4] = \lambda(1-\lambda)(\lambda-6)$$

If $\lambda \neq 0$, $\lambda \neq 1$, and $\lambda \neq 6$, then the rank is 3. Because the minor $\begin{vmatrix} 5-\lambda & 2 \\ 1 & 0 \end{vmatrix} = -2 \neq 0$, whatever the value of λ , we see that the rank of the matrix is 2 when λ is 0, 1, or 6. ■

² For a proof of this theorem, see e.g. Fraleigh and Beauregard (1995).

Recall that according to (1.1.23), the determinant of a matrix is equal to the determinant of its transpose. The following result is therefore not surprising:

THEOREM 1.3.2

The rank of a matrix A is equal to the rank of its transpose: $r(A) = r(A')$.

Proof: Suppose $|D|$ is a minor of A . Then $|D'|$ is a minor of A' , and vice versa. Because $|D'| = |D|$, the result follows from Theorem 1.3.1. ■

It follows from (1) and Theorem 1.3.2 that the rank of a matrix can also be characterized as the maximal number of linearly independent rows of A . So we have three ways of showing that $r(A) = k$:

- Find one set of k columns that is linearly independent, and then show that no set of more than k columns is linearly independent.
- Find one set of k rows that is linearly independent, and then show that no set of more than k rows is linearly independent.
- Find one minor of order k that is not 0, and then show that all minors of order higher than k are 0.

An Efficient Way to Find the Rank of a Matrix

None of the methods (a), (b), and (c) for finding the rank of a matrix is very efficient. A better approach uses the fact that *the rank of a matrix is not affected by elementary operations*.³ In the following, if a matrix A is transformed into a matrix B by means of elementary operations, then we write $A \sim B$.

EXAMPLE 5 Find the rank of $\begin{pmatrix} 1 & 2 & 3 & 2 \\ 2 & 3 & 5 & 1 \\ 1 & 3 & 4 & 5 \end{pmatrix}$.

Solution: We use the elementary operations indicated. That is, we multiply the first row by -2 and add it to the second row, and also multiply the first row by -1 and add it to the third row, etc.

$$\begin{pmatrix} 1 & 2 & 3 & 2 \\ 2 & 3 & 5 & 1 \\ 1 & 3 & 4 & 5 \end{pmatrix} \xrightarrow{\substack{-2 \cdot R_1 \rightarrow R_2 \\ -1 \cdot R_1 \rightarrow R_3}} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & -1 & -1 & -3 \\ 0 & 1 & 1 & 3 \end{pmatrix} \xrightarrow{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 3 \\ 0 & -1 & -1 & -3 \end{pmatrix} \xrightarrow{R_3 \leftarrow R_3 + R_2} \begin{pmatrix} 1 & 2 & 3 & 2 \\ 0 & 1 & 1 & 3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The rank of the last matrix is obviously 2, because there are precisely two linearly independent rows. So the original matrix has rank 2. ■

³ Elementary row (column) operations are: (a) interchanging two rows (columns); (b) multiplying a row (column) by a scalar $\alpha \neq 0$; (c) adding α times a row (column) to a different row (column). See EMFA or Fraleigh and Beauregard (1995).

PROBLEMS FOR SECTION 1.3

501 1. Determine the ranks of the following matrices:

$$(a) \begin{pmatrix} 1 & 2 \\ 8 & 16 \end{pmatrix} \quad (b) \begin{pmatrix} 1 & 3 & 4 \\ 2 & 0 & 1 \end{pmatrix} \quad (c) \begin{pmatrix} 1 & 2 & -1 & 3 \\ 2 & 4 & -4 & 7 \\ -1 & -2 & -1 & -2 \end{pmatrix}$$

$$(d) \begin{pmatrix} 1 & 3 & 0 & 0 \\ 2 & 4 & 0 & -1 \\ 1 & -1 & 2 & 2 \end{pmatrix} \quad (e) \begin{pmatrix} 2 & 1 & 3 & 7 \\ -1 & 4 & 3 & 1 \\ 3 & 2 & 5 & 11 \end{pmatrix} \quad (f) \begin{pmatrix} 1 & -2 & -1 & 1 \\ 2 & 1 & 1 & 2 \\ -1 & 1 & -1 & -3 \\ -2 & -5 & -2 & 0 \end{pmatrix}$$

502 2. Determine the ranks of the following matrices for all values of the parameters:

$$(a) \begin{pmatrix} x & 0 & x^2 - 2 \\ 0 & 1 & 1 \\ -1 & x & x - 1 \end{pmatrix} \quad (b) \begin{pmatrix} t+3 & 5 & 6 \\ -1 & t-3 & -6 \\ 1 & 1 & t+4 \end{pmatrix} \quad (c) \begin{pmatrix} 1 & x & y & 0 \\ 0 & z & w & 1 \\ 1 & x & y & 0 \\ 0 & z & w & 1 \end{pmatrix}$$

3. Give an example where $r(\mathbf{AB}) \neq r(\mathbf{BA})$. (Hint: Try some 2×2 matrices.)

1.4 Main Results on Linear Systems

Consider the general linear system of m simultaneous equations in n unknowns:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \quad \text{or} \quad \mathbf{Ax} = \mathbf{b} \quad (1)$$

where $\mathbf{A}$ is the $m \times n$ coefficient matrix. Define a new $m \times (n+1)$ matrix $\mathbf{A}_b$ that contains $\mathbf{A}$ in the first n columns and $\mathbf{b}$ in column $n+1$, so:

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

Then $\mathbf{A}_b$ is called the **augmented matrix** of the system (1). It turns out that the relationship between the ranks of $\mathbf{A}$ and $\mathbf{A}_b$ is crucial in determining whether system (1) has a solution. Because all the columns in $\mathbf{A}$ occur in $\mathbf{A}_b$, the rank of $\mathbf{A}_b$ is certainly greater than or equal to the rank of $\mathbf{A}$. Moreover, because $\mathbf{A}_b$ contains only one more column than $\mathbf{A}$, the number $r(\mathbf{A}_b)$ cannot be greater than $r(\mathbf{A}) + 1$.

THEOREM 1.4.1

A necessary and sufficient condition for a linear system of equations to be consistent (that is, to have at least one solution) is that the rank of the coefficient matrix is equal to the rank of the augmented matrix. Briefly:

$$\mathbf{Ax} = \mathbf{b} \quad \text{has a solution} \quad \iff \quad r(\mathbf{A}) = r(\mathbf{A}_b)$$

Proof: Let the column vectors in $\mathbf{A}_b$ be $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n, \mathbf{b}$, and suppose that (1) has a solution $(x_1, \dots, x_n)$, so that $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$. Multiply the first n columns in $\mathbf{A}_b$ by $-x_1, \dots, -x_n$, respectively, and add each of the resulting column vectors to the last column in $\mathbf{A}_b$. These elementary column operations make the last column 0. It follows that $\mathbf{A}_b \sim [\mathbf{a}_1, \dots, \mathbf{a}_n, \mathbf{0}]$. Because elementary column operations preserve the rank, this matrix has the same rank as $\mathbf{A}$, so $r(\mathbf{A}_b) = r(\mathbf{A})$.

Suppose, on the other hand, that $r(\mathbf{A}) = r(\mathbf{A}_b) = k$. Then k of the columns of $\mathbf{A}$ are linearly independent. To simplify notation, suppose that the first k columns, $\mathbf{a}_1, \dots, \mathbf{a}_k$, are linearly independent. Because $r(\mathbf{A}_b) = k$, the vectors $\mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{b}$ are linearly dependent. Hence there exist numbers $c_1, \dots, c_k$ and β , not all equal to 0, such that $c_1\mathbf{a}_1 + \cdots + c_k\mathbf{a}_k + \beta\mathbf{b} = \mathbf{0}$. If $\beta = 0$, then $\mathbf{a}_1, \dots, \mathbf{a}_k$ would not be linearly independent. Hence $\beta \neq 0$. Then $\mathbf{b} = x_1^0\mathbf{a}_1 + \cdots + x_k^0\mathbf{a}_k$ where $x_1^0 = -c_1/\beta, \dots, x_k^0 = -c_k/\beta$. It follows that $(x_1^0, \dots, x_k^0, 0, \dots, 0)$ is a solution of $\mathbf{Ax} = \mathbf{b}$. ■

NOTE 1 If $\mathbf{A}$ is $n \times n$ and $r(\mathbf{A}) = n$, then according to Theorem 1.4.1 the system $\mathbf{Ax} = \mathbf{b}$ does have a solution. It is unique according to a remark preceding Theorem 1.2.1.

What happens to system (1) when $r(\mathbf{A}) = r(\mathbf{A}_b) = k$ and either (i) $k < m$ or (ii) $k < n$?

THEOREM 1.4.2

Suppose that system (1) has solutions with $r(\mathbf{A}) = r(\mathbf{A}_b) = k$.

- If $k < m$, i.e. the common rank k is less than the number of equations m , then $m - k$ equations are **superfluous** in the sense that if we choose any subsystem of equations corresponding to k linearly independent rows, then any solution of these k equations also satisfies the remaining $m - k$ equations.
- If $k < n$, i.e. the common rank k is less than the number of unknowns n , then there exist $n - k$ variables that can be chosen freely, whereas the remaining k variables are uniquely determined by the choice of these $n - k$ free variables. The system then has $n - k$ **degrees of freedom**.

Proof: (a) By the definition of rank, there exist k row vectors in $\mathbf{A}_b$ that are linearly independent, and any other row vector in $\mathbf{A}_b$ is a linear combination of those k vectors. We prove that if the vector $(x_1^0, x_2^0, \dots, x_n^0)$ satisfies the k equations corresponding to the k linearly independent row vectors in $\mathbf{A}_b$, then it also satisfies all the other equations corresponding to the remaining rows in $\mathbf{A}_b$. These remaining equations are thus superfluous.

To simplify notation, reorder the equations so that the first k row vectors in $\mathbf{A}_b$ are linearly independent. The other rows are dependent on these first k rows, so for $s = k + 1, \dots, m$,

$$(a_{s1}, a_{s2}, \dots, a_{sn}, b_s) = \sum_{i=1}^k \lambda_{si} (a_{i1}, a_{i2}, \dots, a_{in}, b_i) \quad (*)$$

for suitable constants $\lambda_{s1}, \lambda_{s2}, \dots, \lambda_{sk}$. From (*), we see in particular that $a_{sj} = \sum_{i=1}^k \lambda_{si} a_{ij}$ and $b_s = \sum_{i=1}^k \lambda_{si} b_i$. Suppose that $\sum_{j=1}^n a_{ij} x_j^0 = b_i$ for $i = 1, \dots, k$, so that $(x_1^0, \dots, x_n^0)$ satisfies the first k equations in (1). For $s = k+1, \dots, m$, we then get

$$\sum_{j=1}^n a_{sj} x_j^0 = \sum_{j=1}^n \left(\sum_{i=1}^k \lambda_{si} a_{ij} \right) x_j^0 = \sum_{i=1}^k \lambda_{si} \left(\sum_{j=1}^n a_{ij} x_j^0 \right) = \sum_{i=1}^k \lambda_{si} b_i = b_s$$

This confirms that if the vector $(x_1^0, \dots, x_n^0)$ satisfies the first k equations in (1), then it automatically satisfies the last $m - k$ equations in (1).

(b) Because $r(A) = k$, A has at least one nonzero minor of order k (Theorem 1.3.1). After rearranging the equations and the variables (if necessary), we can assume that the $k \times k$ matrix

$$C = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{pmatrix}$$

in the upper left-hand corner of A has a nonzero determinant. If $k < m$, then we have just proved that the last $m - k$ equations in (1) are superfluous. So the whole system (1) has exactly the same solutions as the first k equations on their own:

$$\sum_{j=1}^n a_{ij} x_j = b_i, \quad i = 1, 2, \dots, k$$

Now move all terms involving $x_{k+1}, x_{k+2}, \dots, x_n$ to the right-hand side:

$$\sum_{j=1}^k a_{ij} x_j = b_i - \sum_{j=k+1}^n a_{ij} x_j, \quad i = 1, 2, \dots, k \quad (**)$$

The $k \times k$ coefficient matrix on the left-hand side of (**) is C , which has rank k . According to Note 1, system (**) has a unique solution for $x_1, x_2, \dots, x_k$ for each choice of $x_{k+1}, x_{k+2}, \dots, x_n$. So the system has $n - k$ degrees of freedom. ■

EXAMPLE 1 Determine whether the following system of equations has any solutions and, if it has, find the number of degrees of freedom and solve the system.

$$\begin{aligned} x_1 + x_2 &= 2x_3 + x_4 + 3x_5 = 1 \\ 2x_1 - x_2 + 2x_3 + 2x_4 + 6x_5 &= 2 \\ 3x_1 + 5x_2 - 10x_3 - 3x_4 - 9x_5 &= 3 \\ 3x_1 + 2x_2 - 4x_3 - 3x_4 - 9x_5 &= 3 \end{aligned}$$

Solution: Here

$$A = \begin{pmatrix} 1 & 1 & -2 & 1 & 3 \\ 2 & -1 & 2 & 2 & 6 \\ 3 & 5 & -10 & -3 & -9 \\ 3 & 2 & -4 & -3 & -9 \end{pmatrix} \quad \text{and} \quad A_b = \begin{pmatrix} 1 & 1 & -2 & 1 & 3 & 1 \\ 2 & -1 & 2 & 2 & 6 & 2 \\ 3 & 5 & -10 & -3 & -9 & 3 \\ 3 & 2 & -4 & -3 & -9 & 3 \end{pmatrix}$$

We know that $r(A_b) \geq r(A)$. All minors of order 4 in A_b are equal to 0 (note that several pairs of columns are proportional), so $r(A_b) \leq 3$. Now, there are minors of order 3 in A that are different from 0. For example, the minor formed by the first, third, and fourth columns, and by the first, second, and fourth rows, is different from 0 because

$$\begin{vmatrix} 1 & -2 & 1 \\ 2 & 2 & 2 \\ 3 & -4 & -3 \end{vmatrix} = -36 \quad (*)$$

Hence, $r(A) = 3$. Because $3 \geq r(A_b) \geq r(A)$, we have $r(A) = r(A_b) = 3$, so the system has solutions. There is one superfluous equation. Because the first, second, and fourth rows in A_b are linearly independent, the third equation can be dropped. The number of variables is 5, and because $r(A) = r(A_b) = 3$, there are 2 degrees of freedom.

Next we find all the solutions to the system of equations. The determinant in (*) is different from 0, so we rewrite the subsystem of three independent equations as

$$\begin{aligned} x_1 - 2x_3 + x_4 + x_2 + 3x_5 &= 1 \\ 2x_1 + 2x_3 + 2x_4 - x_2 + 6x_5 &= 2 \\ 3x_1 - 4x_3 - 3x_4 + 2x_2 - 9x_5 &= 3 \end{aligned} \quad (**)$$

or, in matrix form, as

$$\begin{pmatrix} 1 & -2 & 1 \\ 2 & 2 & 2 \\ 3 & -4 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 1 & 3 \\ -1 & 6 \\ 2 & -9 \end{pmatrix} \begin{pmatrix} x_2 \\ x_5 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

The 3×3 coefficient matrix corresponding to x_1, x_3 , and x_4 in (**) has a determinant different from 0, so it has an inverse. Therefore,

$$\begin{pmatrix} x_1 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 \\ 2 & 2 & 2 \\ 3 & -4 & -3 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} - \begin{pmatrix} 1 & -2 & 1 \\ 2 & 2 & 2 \\ 3 & -4 & -3 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 3 \\ -1 & 6 \\ 2 & -9 \end{pmatrix} \begin{pmatrix} x_2 \\ x_5 \end{pmatrix}$$

It is easy to verify that

$$\begin{pmatrix} 1 & -2 & 1 \\ 2 & 2 & 2 \\ 3 & -4 & -3 \end{pmatrix}^{-1} = \frac{1}{18} \begin{pmatrix} -1 & 5 & 3 \\ -6 & 3 & 0 \\ 7 & 1 & -3 \end{pmatrix}$$

Then, after some routine algebra, we have

$$\begin{pmatrix} x_1 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ -\frac{1}{2}x_2 \\ 3x_5 \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{1}{2}x_2 \\ -3x_5 \end{pmatrix}$$

So if $x_2 = a$ and $x_5 = b$ are arbitrary real numbers, then there is a solution $x_1 = 1, x_2 = a, x_3 = \frac{1}{2}a, x_4 = -3b, x_5 = b$. This confirms that there are two degrees of freedom. (You should verify that the values found for $x_1, \dots, x_5$ do satisfy the original system of equations for all values of a and b .)

The concept of degrees of freedom is very important. Note that if a linear system of equations has k degrees of freedom, then there exist k variables that can be chosen freely. These may not be the first k variables. For instance, in Example 1 there are two degrees of freedom, but x_1 cannot be chosen freely because $x_1 = 1$.

PROBLEMS FOR SECTION 1.4

1. Use Theorem 1.4.1 to examine whether the following systems of equations have solutions. If they do, determine the number of degrees of freedom. Find all the solutions. Check the results.

$$(a) \begin{cases} -2x_1 - 3x_2 + x_3 = 3 \\ 4x_1 + 6x_2 - 2x_3 = 1 \end{cases}$$

$$(b) \begin{cases} x_1 + x_2 - x_3 + x_4 = 2 \\ 2x_1 - x_2 + x_3 - 3x_4 = 1 \end{cases}$$

$$(c) \begin{cases} x_1 - x_2 + 2x_3 + x_4 = 1 \\ 2x_1 + x_2 - x_3 + 3x_4 = 3 \\ x_1 + 5x_2 - 8x_3 + x_4 = 1 \\ 4x_1 + 5x_2 - 7x_3 + 7x_4 = 7 \end{cases}$$

$$(d) \begin{cases} x_1 + x_2 + 2x_3 + x_4 = 5 \\ 2x_1 + 3x_2 - x_3 - 2x_4 = 2 \\ 4x_1 + 5x_2 + 3x_3 = 7 \end{cases}$$

2. Solve the following systems and determine the number of degrees of freedom:

$$(a) \begin{cases} x_1 - x_2 + x_3 = 0 \\ x_1 + 2x_2 - x_3 = 0 \\ 2x_1 + x_2 + 3x_3 = 0 \end{cases}$$

$$(b) \begin{cases} x_1 + x_2 + x_3 + x_4 = 0 \\ x_1 + 3x_2 + 2x_3 + 4x_4 = 0 \\ 2x_1 + x_2 - x_4 = 0 \end{cases}$$

3. Discuss the number of solutions of the following system for all values of a and b .

$$\begin{cases} x + 2y + 3z = 1 \\ -x + ay - 21z = 2 \\ 3x + 7y + az = b \end{cases}$$

4. Let $Ax = b$ be a linear system of equations in matrix form. Prove that if x_1 and x_2 are both solutions of the system, then so is $\lambda x_1 + (1 - \lambda)x_2$ for every real number λ . Use this fact to prove that a linear system of equations that is consistent has either one solution or infinitely many solutions. (For instance, it cannot have exactly three solutions.)

5. Let $a_1, \dots, a_n$ be n linearly independent vectors in $\mathbb{R}^n$. Prove that if a vector b in $\mathbb{R}^n$ is orthogonal to all the vectors $a_1, \dots, a_n$, then $b = 0$.

6. (a) Find the rank of $A_t = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 5 & t \\ 4 & 7-t & -6 \end{pmatrix}$ for all real numbers t .

- (b) When $t = -3$, find all vectors x that satisfy the vector equation $A_{-3}x = \begin{pmatrix} 11 \\ 3 \\ 6 \end{pmatrix}$.

7. In an economic model the endogenous variables $x_1, x_2, \dots, x_n$ are related to the exogenous variables $b_1, b_2, \dots, b_n$ by the linear system (1.1.31), or in matrix form $Ax = b$. Assume that the $n \times n$ -matrix A is nonsingular. For each choice of b the vector x is uniquely determined by $x = A^{-1}b$. Suppose b_j changes to $b_j + \Delta b_j$, but that all the other b_i 's are unchanged. The corresponding values of the endogenous variables will then (in general) all be changed. Let the change in x_i be denoted by Δx_i . Prove that $\Delta x_i = a_{ij}^{(-1)} \Delta b_j$, where $a_{ij}^{(-1)}$ is the (i, j) th element of the matrix A^{-1} .

1.5 Eigenvalues

Many applied problems, especially in dynamic economics, involve the powers A^n , $n = 1, 2, \dots$, of a square matrix A . If the dimension of A is very large and x is a given nonzero vector, then computing A^5x or, even worse, $A^{100}x$, is usually a major problem. But suppose there happens to be a scalar λ with the special property that

$$Ax = \lambda x \quad (*)$$

In this case, we would have $A^2x = A(Ax) = A(\lambda x) = \lambda Ax = \lambda \lambda x = \lambda^2 x$ and, in general, $A^n x = \lambda^n x$. Many of the properties of A and A^n can be deduced by finding the pairs (λ, x) , $x \neq 0$, that satisfy $(*)$.

A nonzero vector x that solves $(*)$ is called an **eigenvector**, and the associated λ is called an **eigenvalue**. Zero solutions are not very interesting, of course, because $A0 = \lambda 0$ for every scalar λ .

In optimization theory, in the theory of difference and differential equations, in statistics, in population dynamics, and in many other applications of mathematics, there are important arguments and results based on eigenvalues. One contemporary example is how search engines like Google use eigenvalue methods to order web pages so quickly and efficiently.⁴

Eigenvalues for Matrices of Order 2

In the case when $n = 2$, we have $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Then $(*)$ reduces to

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{or} \quad \begin{cases} a_{11}x_1 + a_{12}x_2 = \lambda x_1 \\ a_{21}x_1 + a_{22}x_2 = \lambda x_2 \end{cases}$$

This system can be written as

$$\begin{cases} (a_{11} - \lambda)x_1 + a_{12}x_2 = 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 = 0 \end{cases} \quad \text{or in matrix form} \quad (A - \lambda I)x = 0 \quad (1)$$

where I is the identity matrix of order 2. According to (1.1.34), this homogeneous system has a solution $x \neq 0$ if and only if the coefficient matrix has determinant equal to 0—that is, if and only if $|A - \lambda I| = 0$. Evaluating this 2×2 determinant, we get

$$|A - \lambda I| = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = \lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0 \quad (2)$$

So the eigenvalues are the real or complex solutions of this quadratic equation, and the eigenvectors are the nonzero vectors $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ that satisfy system (1).

EXAMPLE 3 Find the eigenvalues and the associated eigenvectors of the matrices

$$(a) \quad A = \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix} \quad (b) \quad B = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

⁴ See P. Fernandez Gallardo: "Google's secret and linear algebra", *Newsletter of the European Mathematical Society*, March 2007.

Solution: (a) The eigenvalue equation is $|\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} 1-\lambda & 2 \\ 3 & -\lambda \end{vmatrix} = \lambda^2 - \lambda - 6 = 0$.

This equation has the solutions $\lambda_1 = -2$ and $\lambda_2 = 3$, which are the eigenvalues of $\mathbf{A}$.

For $\lambda = \lambda_1 = -2$, the two equations of system (1) both reduce to $3x_1 + 2x_2 = 0$. Choosing $x_2 = t$, we have $x_1 = -\frac{2}{3}t$. The eigenvectors associated with $\lambda_1 = -2$ are therefore $\mathbf{x} = t \begin{pmatrix} -2/3 \\ 1 \end{pmatrix}$, $t \neq 0$. If we put $t = -3s$, we can equivalently represent the eigenvectors as $\mathbf{x} = s \begin{pmatrix} 2 \\ -3 \end{pmatrix}$, $s \neq 0$.

For $\lambda = 3$, system (1) implies that $x_1 = x_2$, so the eigenvectors are $s \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $s \neq 0$.

(b) The eigenvalue equation is $|\mathbf{B} - \lambda\mathbf{I}| = \begin{vmatrix} -\lambda & 1 \\ -1 & -\lambda \end{vmatrix} = \lambda^2 + 1 = 0$, which has the complex roots $\lambda = \pm i$. In this case the eigenvectors are also complex, and they are $s \begin{pmatrix} 1 \\ i \end{pmatrix}$ and $t \begin{pmatrix} 1 \\ -i \end{pmatrix}$, with $s \neq 0$ and $t \neq 0$.⁵

The eigenvalues λ_1 and λ_2 of the matrix $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ are the roots of equation (2), whose left-hand side can be written as

$$\lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2 \quad (3)$$

We see from (3) that the sum $\lambda_1 + \lambda_2$ of the eigenvalues is equal to $a_{11} + a_{22}$, the sum of the diagonal elements (also called the **trace** of the matrix and denoted by $\text{tr}(\mathbf{A})$). The product $\lambda_1\lambda_2$ of the eigenvalues is equal to $a_{11}a_{22} - a_{12}a_{21} = |\mathbf{A}|$. In symbols:

$$(i) \lambda_1 + \lambda_2 = \text{tr}(\mathbf{A}) \quad (ii) \lambda_1\lambda_2 = |\mathbf{A}| \quad (4)$$

Many dynamic economic models involve square matrices whose eigenvalues determine their stability properties. In the 2×2 case, important questions are when the two eigenvalues are real and what are their signs. The roots of the quadratic equation (2) are

$$\lambda = \frac{1}{2}(a_{11} + a_{22}) \pm \sqrt{\frac{1}{4}(a_{11} + a_{22})^2 - (a_{11}a_{22} - a_{12}a_{21})} \quad (5)$$

These roots are real if and only if $(a_{11} + a_{22})^2 \geq 4(a_{11}a_{22} - a_{12}a_{21})$, which is equivalent to $(a_{11} - a_{22})^2 + 4a_{12}a_{21} \geq 0$. In particular, both eigenvalues are real if the matrix is symmetric, because then $a_{12} = a_{21}$ and so we have the sum of two squares. (But a matrix may well have real eigenvalues even if it is not symmetric, as in Example 1(a).)

It follows from (4) that for a 2×2 matrix $\mathbf{A}$ with *real eigenvalues*,

(A) both eigenvalues are positive $\iff |\mathbf{A}| > 0$ and $\text{tr}(\mathbf{A}) > 0$

(B) both eigenvalues are negative $\iff |\mathbf{A}| > 0$ and $\text{tr}(\mathbf{A}) < 0$

(C) the two eigenvalues have opposite signs $\iff |\mathbf{A}| < 0$

Moreover, 0 is an eigenvalue if and only if $|\mathbf{A}| = 0$. The other eigenvalue is then equal to $a_{11} + a_{22}$.

⁵ For complex numbers, see Appendix B.3. One can do matrix algebra with complex numbers in the same way as with real matrices.

The General Case

Let us turn to the general case in which $\mathbf{A}$ is an $n \times n$ matrix:

EIGENVALUES AND EIGENVECTORS

If $\mathbf{A}$ is an $n \times n$ matrix, then a scalar λ is an **eigenvalue** of $\mathbf{A}$ if there is a nonzero vector $\mathbf{x}$ in $\mathbb{R}^n$ such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad (6)$$

Then $\mathbf{x}$ is an **eigenvector** of $\mathbf{A}$ (associated with λ).

It should be noted that if $\mathbf{x}$ is an eigenvector associated with the eigenvalue λ , then so is $\alpha\mathbf{x}$ for every scalar $\alpha \neq 0$. Eigenvalues and eigenvectors are also called **characteristic roots (values)** and **characteristic vectors**, respectively.

How to Find Eigenvalues

The eigenvalue equation (6) can be written as

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \quad (7)$$

where $\mathbf{I}$ denotes the identity matrix of order n . According to (1.1.34), this homogeneous linear system of equations has a solution $\mathbf{x} \neq \mathbf{0}$ if and only if the coefficient matrix has determinant *equal to 0*—that is, if and only if $|\mathbf{A} - \lambda\mathbf{I}| = 0$. Letting $p(\lambda) = |\mathbf{A} - \lambda\mathbf{I}|$, where $\mathbf{A} = (a_{ij})_{n \times n}$, we have the equation

$$p(\lambda) = |\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0 \quad (8)$$

This is called the **characteristic equation** (or **eigenvalue equation**) of $\mathbf{A}$. The polynomial $p(\lambda)$ is called the **characteristic polynomial** of $\mathbf{A}$. It follows from (8) that $p(\lambda)$ is a polynomial of degree n in λ . According to the fundamental theorem of algebra, equation (8) has exactly n roots (real or complex), provided that any multiple roots are counted appropriately.

If the components of the vector $\mathbf{x}$ are $x_1, \dots, x_n$, then (7) can be written as

$$\begin{aligned} (a_{11} - \lambda)x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \cdots + a_{2n}x_n &= 0 \\ \vdots &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + (a_{nn} - \lambda)x_n &= 0 \end{aligned} \quad (9)$$

An eigenvector associated with λ is a nontrivial solution $(x_1, \dots, x_n)$ of (9).

EXAMPLE 2 Find all the eigenvalues of the matrices and also the eigenvectors associated with the real eigenvalues.

$$(a) \quad A = \begin{pmatrix} 0 & 0 & 6 \\ 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \end{pmatrix} \quad (b) \quad B = \begin{pmatrix} 5 & -6 & -6 \\ -1 & 4 & 2 \\ 3 & -6 & -4 \end{pmatrix}$$

Solution: (a) The characteristic equation is

$$|A - \lambda I| = \begin{vmatrix} -\lambda & 0 & 6 \\ 1/2 & -\lambda & 0 \\ 0 & 1/3 & -\lambda \end{vmatrix} = -\lambda^3 + 1 = 0$$

which has $\lambda = 1$ as its only real root. (Because $-\lambda^3 + 1 = (1 - \lambda)(\lambda^2 + \lambda + 1)$, there are two complex eigenvalues, $\lambda = -\frac{1}{2} \pm \frac{\sqrt{3}}{2}i$.) The eigenvectors associated with $\lambda = 1$ satisfy (9), which becomes

$$\begin{aligned} -x_1 + 6x_3 &= 0 \\ \frac{1}{2}x_1 - x_2 &= 0 \\ \frac{1}{3}x_2 - x_3 &= 0 \end{aligned}$$

with eigenvectors $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = t \begin{pmatrix} 6 \\ 3 \\ 1 \end{pmatrix}$, where $t \neq 0$.

(b) The characteristic equation is

$$|B - \lambda I| = \begin{vmatrix} 5-\lambda & -6 & -6 \\ -1 & 4-\lambda & 2 \\ 3 & -6 & -4-\lambda \end{vmatrix} = -(\lambda - 2)^2(\lambda - 1) = 0$$

Thus, $\lambda_1 = 1$ and $\lambda_2 = 2$ are the eigenvalues.

For $\lambda_1 = 1$, the eigenvectors are $\mathbf{x} = t \begin{pmatrix} 3 \\ -1 \\ 3 \end{pmatrix}$ with $t \neq 0$.

For $\lambda_2 = 2$, the eigenvectors are the nonzero solutions of the equation system

$$\begin{aligned} 3x_1 - 6x_2 - 6x_3 &= 0 \\ -x_1 + 2x_2 + 2x_3 &= 0 \\ 3x_1 - 6x_2 - 6x_3 &= 0 \end{aligned}$$

The three equations are all proportional, so the system has solutions with two degrees of freedom. They can be written as $\mathbf{x} = \begin{pmatrix} 2s + 2t \\ s \\ t \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}s + \begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix}t$, with s and t in $\mathbb{R}$, not both equal to 0.

EXAMPLE 3 Let $D = \text{diag}(a_1, \dots, a_n)$ be an $n \times n$ diagonal matrix with diagonal elements $a_1, \dots, a_n$. The characteristic polynomial is $|D - \lambda I| = (a_1 - \lambda)(a_2 - \lambda) \cdots (a_n - \lambda)$. Hence, the eigenvalues of D are the diagonal elements. Let $\mathbf{e}_j$ denote the j th unit vector in $\mathbb{R}^n$, having all components 0, except for the j th component which is 1. Because $D\mathbf{e}_j = a_j\mathbf{e}_j$, it follows that any nonzero multiple of $\mathbf{e}_j$ is an eigenvector associated with the eigenvalue a_j of D .

Suppose we rewrite $p(\lambda)$ in (8) as a polynomial in $-\lambda$:

$$p(\lambda) = (-\lambda)^n + b_{n-1}(-\lambda)^{n-1} + \cdots + b_1(-\lambda) + b_0 \quad (10)$$

The zeros of this polynomial are precisely the eigenvalues of A . Denoting the eigenvalues by $\lambda_1, \lambda_2, \dots, \lambda_n$, we have

$$p(\lambda) = (-1)^n(\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) \quad (11)$$

Consider the particular coefficients b_0 and b_{n-1} in the characteristic polynomial (10). Putting $\lambda = 0$ in (8) and in (10), we see that $p(0) = b_0 = |A|$. But $\lambda = 0$ in (11) gives $p(0) = (-1)^n(-1)^n\lambda_1\lambda_2 \cdots \lambda_n = \lambda_1\lambda_2 \cdots \lambda_n$. Hence, $b_0 = |A| = \lambda_1\lambda_2 \cdots \lambda_n$.

As for b_{n-1} , the product of the elements on the main diagonal of the determinant in (8) is $(a_{11} - \lambda)(a_{22} - \lambda) \cdots (a_{nn} - \lambda)$. If we choose a_{jj} from the j th factor and $-\lambda$ from the remaining $n - 1$, and then add over $j = 1, 2, \dots, n$, we get

$$(a_{11} + a_{22} + \cdots + a_{nn})(-\lambda)^{n-1} \quad (*)$$

Now each term in the expansion of the determinant in (8) contains one element from each row and one from each column. Hence, except for the product of all the elements on the main diagonal, at most $n - 2$ factors in each term come from the main diagonal, so there are no other terms with $(-\lambda)^{n-1}$. Hence, $b_{n-1} = a_{11} + a_{22} + \cdots + a_{nn}$, the trace of A . By expanding (11) we see that the coefficient of $(-\lambda)^{n-1}$ in (10) is $b_{n-1} = \lambda_1 + \lambda_2 + \cdots + \lambda_n$. Thus we have:

THEOREM 1.5.1

If A is an $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$, then

$$(a) \quad |A| = \lambda_1\lambda_2 \cdots \lambda_n$$

$$(b) \quad \text{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn} = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

In other words: If A is an $n \times n$ matrix, the product of all the eigenvalues is equal to the determinant of A , while the sum of all the eigenvalues is equal to the trace of A . This confirms the results we found for $n = 2$.

NOTE 1 Theorem 1.5.1 gives us the coefficients b_{n-1} and b_0 in the characteristic polynomial (10). One can prove in general that each coefficient $b_{n-1}, \dots, b_1, b_0$ in (10) can be characterized as follows:

$$b_k = \text{the sum of all principal minors of } A \text{ of order } n - k \quad (12)$$

Thus b_0 equals the determinant of A , since it is the only principal minor of order n , and b_{n-1} is the sum of all the principal minors of order 1, i.e. the sum $a_{11} + a_{22} + \cdots + a_{nn}$, the trace of A . (For the definition of principal minors, see Section 1.7.)

PROBLEMS FOR SECTION 1.5

- SM 1. For the following matrices, find the eigenvalues and also those eigenvectors that correspond to the real eigenvalues:

$$\begin{array}{lll} \text{(a)} \begin{pmatrix} 2 & -7 \\ 3 & -8 \end{pmatrix} & \text{(b)} \begin{pmatrix} 2 & 4 \\ -2 & 6 \end{pmatrix} & \text{(c)} \begin{pmatrix} 1 & 4 \\ 6 & -1 \end{pmatrix} \\ \text{(d)} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{pmatrix} & \text{(e)} \begin{pmatrix} 2 & 1 & -1 \\ 0 & 1 & 1 \\ 2 & 0 & -2 \end{pmatrix} & \text{(f)} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \end{array}$$

- SM 2. (a) Compute $\mathbf{X}^T \mathbf{A} \mathbf{X}$, $\mathbf{A}^2$, and $\mathbf{A}^3$ when $\mathbf{A} = \begin{pmatrix} a & a & 0 \\ a & a & 0 \\ 0 & 0 & b \end{pmatrix}$ and $\mathbf{X} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$.

(b) Find all the eigenvalues of $\mathbf{A}$.

(c) The characteristic polynomial $p(\lambda)$ of $\mathbf{A}$ is a cubic function of λ . Show that if we replace λ by $\mathbf{A}$, then $p(\mathbf{A})$ is the zero matrix. (This is a special case of the Cayley–Hamilton theorem. See (1.6.6).)

3. $\mathbf{A} = \begin{pmatrix} a & b & c \\ b & d & e \\ c & e & f \end{pmatrix}$ has the eigenvectors $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$, $\mathbf{v}_2 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$, $\mathbf{v}_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$, with associated eigenvalues $\lambda_1 = 3$, $\lambda_2 = 1$, and $\lambda_3 = 4$. Determine the matrix $\mathbf{A}$.

- SM 4. (a) Find the eigenvalues of $\mathbf{A} = \begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 4 & 1 & 1 \\ 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 4 \end{pmatrix}$. (Hint: Problem 1.9.7(b) might be useful.)

(b) One of the eigenvalues has multiplicity 3. Find three linearly independent eigenvectors associated with this eigenvalue.

- SM 5. Let $\mathbf{A} = \begin{pmatrix} -2 & -1 & 4 \\ 2 & 1 & -2 \\ -1 & -1 & 3 \end{pmatrix}$, $\mathbf{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$, $\mathbf{x}_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$, $\mathbf{x}_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

(a) Verify that $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ are eigenvectors of $\mathbf{A}$, and find the associated eigenvalues.

(b) Let $\mathbf{B} = \mathbf{A} \mathbf{A}$. Show that $\mathbf{B} \mathbf{x}_2 = \mathbf{x}_2$ and $\mathbf{B} \mathbf{x}_3 = \mathbf{x}_3$. Is $\mathbf{B} \mathbf{x}_1 = \mathbf{x}_1$?

(c) Let $\mathbf{C}$ be an arbitrary $n \times n$ matrix such that $\mathbf{C}^3 = \mathbf{C}^2 + \mathbf{C}$. Prove that if λ is an eigenvalue for $\mathbf{C}$, then $\lambda^3 = \lambda^2 + \lambda$. Show that $\mathbf{C} + \mathbf{I}_n$ has an inverse.

6. Prove that λ is an eigenvalue of the matrix $\mathbf{A}$ if and only if λ is an eigenvalue of $\mathbf{A}'$.

7. Suppose $\mathbf{A}$ is a square matrix and let λ be an eigenvalue of $\mathbf{A}$. Prove that if $|\mathbf{A}| \neq 0$, then $\lambda \neq 0$. In this case show that $1/\lambda$ is an eigenvalue of the inverse $\mathbf{A}^{-1}$.

8. Let $\mathbf{A} = (a_{ij})_{n \times n}$ be a matrix where all column sums are 1—that is, $\sum_{i=1}^n a_{ij} = 1$ for $j = 1, 2, \dots, n$. Prove that $\lambda = 1$ is an eigenvalue of $\mathbf{A}$.

1.6 Diagonalization

We begin by noting a simple and useful result. Let $\mathbf{A}$ and $\mathbf{P}$ be $n \times n$ matrices with $\mathbf{P}$ invertible. Then

$$\mathbf{A} \text{ and } \mathbf{P}^{-1} \mathbf{A} \mathbf{P} \text{ have the same eigenvalues} \quad (1)$$

This is true because the two matrices have the same characteristic polynomial:

$$|\mathbf{P}^{-1} \mathbf{A} \mathbf{P} - \lambda \mathbf{I}| = |\mathbf{P}^{-1} (\mathbf{A} - \lambda \mathbf{I}) \mathbf{P}| = |\mathbf{P}^{-1}| |\mathbf{A} - \lambda \mathbf{I}| |\mathbf{P}| = |\mathbf{A} - \lambda \mathbf{I}|$$

where we made use of rule (1.1.24) for determinants, and the fact that $|\mathbf{P}^{-1}| = 1/|\mathbf{P}|$.

An $n \times n$ matrix $\mathbf{A}$ is **diagonalizable** if there exist an invertible $n \times n$ matrix $\mathbf{P}$ and a diagonal matrix $\mathbf{D}$ such that

$$\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \mathbf{D} \quad (2)$$

By Example 1.5.3, the eigenvalues of a diagonal matrix are the diagonal elements. Hence, if $\mathbf{A}$ is diagonalizable, so that (2) holds, then $\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \text{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $\mathbf{A}$. Two questions arise:

(A) Which square matrices are diagonalizable?

(B) If $\mathbf{A}$ is diagonalizable, how do we find the matrix $\mathbf{P}$ in (2)?

The answers to both of these questions are given in the next theorem:

THEOREM 1.6.1 (DIAGONALIZABLE MATRICES)

An $n \times n$ matrix $\mathbf{A}$ is diagonalizable if and only if it has a set of n linearly independent eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$. In that case,

$$\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \text{diag}(\lambda_1, \dots, \lambda_n) \quad (3)$$

where $\mathbf{P}$ is the matrix with $\mathbf{x}_1, \dots, \mathbf{x}_n$ as its columns, and $\lambda_1, \dots, \lambda_n$ are the corresponding eigenvalues.

Proof: Suppose $\mathbf{A}$ has n linearly independent eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$, with corresponding eigenvalues $\lambda_1, \dots, \lambda_n$. Let $\mathbf{P}$ denote the matrix whose columns are $\mathbf{x}_1, \dots, \mathbf{x}_n$. Then $\mathbf{A} \mathbf{P} = \mathbf{P} \mathbf{D}$, where $\mathbf{D} = \text{diag}(\lambda_1, \dots, \lambda_n)$. Because the eigenvectors are linearly independent, $\mathbf{P}$ is invertible, so $\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \mathbf{D}$.

Conversely, if $\mathbf{A}$ is diagonalizable, (2) must hold. Then $\mathbf{A} \mathbf{P} = \mathbf{P} \mathbf{D}$. The columns of $\mathbf{P}$ must be eigenvectors of $\mathbf{A}$, and the diagonal elements of $\mathbf{D}$ the corresponding eigenvalues. ■

EXAMPLE 1 Verify Theorem 1.6.1 for $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix}$. (See Example 1.5.1(a).)

Solution: From Example 1.5.1(a), the eigenvalues are $\lambda_1 = -2$ and $\lambda_2 = 3$. For the matrix $\mathbf{P}$ we can choose $\mathbf{P} = \begin{pmatrix} 2 & 1 \\ -3 & 1 \end{pmatrix}$, whose inverse is $\mathbf{P}^{-1} = \begin{pmatrix} 1/5 & -1/5 \\ 3/5 & 2/5 \end{pmatrix}$. Now direct multiplication shows that $\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \text{diag}(-2, 3)$. Theorem 1.6.1 is confirmed. ■

EXAMPLE 2 It follows from (3) that $A = P \operatorname{diag}(\lambda_1, \dots, \lambda_n) P^{-1}$. If m is a natural number, then

$$A^m = P \operatorname{diag}(\lambda_1^m, \dots, \lambda_n^m) P^{-1} \quad (4)$$

(See Problem 3.) This provides a simple formula for computing A^m when A is diagonalizable.

NOTE 1 A matrix P is called **orthogonal** if $P' = P^{-1}$, i.e. $P'P = I$. If $x_1, \dots, x_n$ are the n column vectors of P , then $x'_1, \dots, x'_n$ are the row vectors of the transposed matrix, P' . The condition $P'P = I$ then reduces to the n^2 equations $x'_i x_j = 1$ if $i = j$ and $x'_i x_j = 0$ if $i \neq j$. Thus P is orthogonal if and only if $x_1, \dots, x_n$ all have length 1 and are mutually orthogonal.

Many of the matrices encountered in economics are symmetric. For symmetric matrices we have the following important result:

THEOREM 1.6.2 (THE SPECTRAL THEOREM FOR SYMMETRIC MATRICES)

If the matrix $A = (a_{ij})_{n \times n}$ is symmetric, then:

- All the n eigenvalues $\lambda_1, \dots, \lambda_n$ are real.
- Eigenvectors that correspond to different eigenvalues are orthogonal.
- There exists an **orthogonal** matrix P (i.e. with $P' = P^{-1}$) such that

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \quad (5)$$

The columns $v_1, v_2, \dots, v_n$ of the matrix P are eigenvectors of unit length corresponding to the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$.

Proof: (a) Suppose λ is an eigenvalue for A , possibly complex, so that $Ax = \lambda x$ for some vector $x \neq 0$ that may have complex components. According to (B.3.9), $(\overline{\lambda x})'(Ax)$ is a real number ≥ 0 . Because A is a symmetric matrix with real entries, one has $\overline{A} = A$ and $(\overline{Ax})' = (\overline{A \overline{x}})' = (\overline{A} \overline{x})' = \overline{x}' A' = \overline{x}' A$. Therefore,

$$0 \leq (\overline{\lambda x})'(Ax) = (\overline{x}' A)(\lambda x) = (\overline{x}' A)(\lambda x) = \lambda \overline{x}'(Ax) = \lambda \overline{x}'(\lambda x) = \lambda^2 \overline{x}' x$$

Since $\overline{x}' x$ is a positive real number, λ^2 is real and ≥ 0 . We conclude that λ is real. (See Example B.3.2.)

(b) Suppose $Ax_i = \lambda_i x_i$ and $Ax_j = \lambda_j x_j$ with $\lambda_i \neq \lambda_j$. Multiplying these equalities from the left by x'_j and x'_i , respectively, we get $x'_j Ax_i = \lambda_i x'_j x_i$ and $x'_i Ax_j = \lambda_j x'_i x_j$. Since A is symmetric, transposing the first equality yields $x'_i Ax_j = \lambda_i x'_i x_j$. But then $\lambda_i x'_i x_j = \lambda_j x'_i x_j$, or $(\lambda_i - \lambda_j)x'_i x_j = 0$. Since $\lambda_i \neq \lambda_j$, it follows that $x'_i x_j = 0$, and thus x_i and x_j are orthogonal.

(c) Suppose that all the (real) eigenvalues are different. Then according to (b), the eigenvectors are mutually orthogonal. Problem 1.2.6(b) tells us that the eigenvectors must be linearly independent. By Theorem 1.6.1, it follows that A is diagonalizable and (5) is valid with P as the matrix where the columns are the eigenvectors $x_1, \dots, x_n$. We can choose the eigenvectors so that they all have length 1, by replacing each x_j with $x_j/\|x_j\|$. According to Note 1, the matrix P is then orthogonal.

For a proof of the general case where some of the eigenvalues are equal, see e.g. Fraleigh and Bearegard (1995).

NOTE 2 A famous and striking result is Cayley–Hamilton's theorem which says that any square matrix satisfies its own characteristic equation. Thus, with reference to the characteristic polynomial (1.5.10) of A ,

$$(-A)^n + b_{n-1}(-A)^{n-1} + \cdots + b_1(-A) + b_0 I = 0 \quad (\text{Cayley–Hamilton}) \quad (6)$$

For a proof, see Faddeeva (1959) or Lewis (1991).

EXAMPLE 3 Test the Cayley–Hamilton theorem on the matrix $A = \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix}$.

Solution: Example 1.5.1(a) showed that the characteristic polynomial is $\lambda^2 - \lambda - 6 = 0$. According to (6), the matrix A satisfies the matrix equation $A^2 - A - 6I = 0$. In fact,

$$A^2 = \begin{pmatrix} 7 & 2 \\ 3 & 6 \end{pmatrix}, \text{ so we find}$$

$$A^2 - A - 6I = \begin{pmatrix} 7 & 2 \\ 3 & 6 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix} - 6 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

PROBLEMS FOR SECTION 1.6

1. Verify (5) for the following matrices by finding the matrix P explicitly:

$$(a) \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad (b) \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \quad (c) \begin{pmatrix} 1 & 3 & 4 \\ 3 & 1 & 0 \\ 4 & 0 & 1 \end{pmatrix}$$

2. (a) Let the matrices A_k and P be given by

$$A_k = \begin{pmatrix} 1 & k & 0 \\ 3 & -2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 1/\sqrt{10} & -3/\sqrt{35} & 3/\sqrt{14} \\ 0 & 5/\sqrt{35} & 2/\sqrt{14} \\ 3/\sqrt{10} & 1/\sqrt{35} & -1/\sqrt{14} \end{pmatrix}$$

Find the characteristic equation of A_k and determine the values of k that make all the eigenvalues real. What are the eigenvalues if $k = 3$?

(b) Show that the columns of P are eigenvectors of A_3 , and compute the matrix product $P'A_3P$. What do you see?

3. (a) Prove that if $A = PDP^{-1}$, where P and D are $n \times n$ matrices, then $A^2 = PD^2P^{-1}$.

(b) Show by induction that $A^m = PD^mP^{-1}$ for every positive integer m .

4. Use (1) to prove that if A and B are both invertible $n \times n$ matrices, then AB and BA have the same eigenvalues.

5. Cayley–Hamilton's theorem can be used to compute powers of matrices. In particular, if $A = (a_{ij})$ is 2×2 , then

$$A^2 = \operatorname{tr}(A)A - |A|I \quad (*)$$

Multiplying this equation by A and using $(*)$ again yields A^3 expressed in terms of A and I , etc. Use this method to find A^4 when $A = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$.

1.7 Quadratic Forms

Many applications of mathematics make use of a special kind of function called a **quadratic form**. A general quadratic form in two variables is

$$Q(x_1, x_2) = a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_2x_1 + a_{22}x_2^2 \quad (1)$$

It follows from the definition of matrix multiplication that

$$Q(x_1, x_2) = (x_1, x_2) \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (2)$$

Of course, $x_1x_2 = x_2x_1$, so we can write $a_{12}x_1x_2 + a_{21}x_2x_1 = (a_{12} + a_{21})x_1x_2$. If we replace each of a_{12} and a_{21} by $\frac{1}{2}(a_{12} + a_{21})$, then the new numbers a_{12} and a_{21} become equal without changing $Q(x_1, x_2)$. Thus, we can assume in (1) that $a_{12} = a_{21}$. Then the matrix $(a_{ij})_{2 \times 2}$ in (2) becomes symmetric, and

$$Q(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \quad (3)$$

We are often interested in conditions on the coefficients a_{11} , a_{12} , and a_{22} ensuring that $Q(x_1, x_2)$ in (3) has the same sign for all (x_1, x_2) . Both $Q(x_1, x_2)$ and its associated symmetric matrix in (2) are called **positive definite**, **positive semidefinite**, **negative definite**, or **negative semidefinite** according as $Q(x_1, x_2) > 0$, $Q(x_1, x_2) \geq 0$, $Q(x_1, x_2) < 0$, or $Q(x_1, x_2) \leq 0$ for all $(x_1, x_2) \neq (0, 0)$. The quadratic form $Q(x_1, x_2)$ is **indefinite** if there exist vectors (x_1^*, x_2^*) and (y_1^*, y_2^*) such that $Q(x_1^*, x_2^*) < 0$ and $Q(y_1^*, y_2^*) > 0$. Thus an indefinite quadratic form assumes both negative and positive values.

Sometimes we can see the sign of a quadratic form immediately, as in the next example.

EXAMPLE 1 Determine the definiteness of the following quadratic forms:

$$Q_1 = x_1^2 + x_2^2, \quad Q_2 = -x_1^2 - x_2^2, \quad Q_3 = (x_1 - x_2)^2 = x_1^2 - 2x_1x_2 + x_2^2 \\ Q_4 = -(x_1 - x_2)^2 = -x_1^2 + 2x_1x_2 - x_2^2, \quad Q_5 = x_1^2 - x_2^2$$

Solution: Q_1 is positive definite because it is always ≥ 0 and it is 0 only if both x_1 and x_2 are 0. Q_3 is positive semidefinite because it is always ≥ 0 , but it is not positive definite because it is 0 if, say, $x_1 = x_2 = 1$. Q_5 is indefinite, because it is 1 for $x_1 = 1, x_2 = 0$, but it is -1 for $x_1 = 0, x_2 = 1$. Evidently, $Q_2 = -Q_1$ is negative definite and $Q_4 = -Q_3$ is negative semidefinite.

In Example 1 it was very easy to determine the sign of the quadratic forms. In general, it is harder, but the old trick of completing the square gives us necessary and sufficient conditions quite easily. We claim that:

$$\begin{aligned} \text{(a) } Q(x_1, x_2) \text{ is pos. semidef.} &\iff a_{11} \geq 0, a_{22} \geq 0, \text{ and } a_{11}a_{22} - a_{12}^2 \geq 0 \\ \text{(b) } Q(x_1, x_2) \text{ is neg. semidef.} &\iff a_{11} \leq 0, a_{22} \leq 0, \text{ and } a_{11}a_{22} - a_{12}^2 \geq 0 \end{aligned} \quad (4)$$

Proof: To prove (a), suppose first that $a_{11} \geq 0$, $a_{22} \geq 0$, and $a_{11}a_{22} - a_{12}^2 \geq 0$. If $a_{11} = 0$, then $a_{11}a_{22} - a_{12}^2 \geq 0$ implies $a_{12} = 0$, and so $Q(x_1, x_2) = a_{22}x_2^2 \geq 0$ for all (x_1, x_2) . If $a_{11} > 0$, then by completing the square, we can write

$$Q(x_1, x_2) = a_{11} \left(x_1 + \frac{a_{12}}{a_{11}}x_2 \right)^2 + \left(a_{22} - \frac{a_{12}^2}{a_{11}} \right) x_2^2 \quad (*)$$

Because $a_{11} > 0$ and $a_{11}a_{22} - a_{12}^2 \geq 0$, we see that $Q(x_1, x_2) \geq 0$ for all (x_1, x_2) .

To prove the reverse implication, suppose $Q(x_1, x_2) \geq 0$ for all (x_1, x_2) . Then, in particular, $Q(1, 0) = a_{11} \geq 0$ and $Q(0, 1) = a_{22} \geq 0$. If $a_{11} = 0$, then $Q(x_1, 1) = 2a_{12}x_1 + a_{22}$, which is ≥ 0 for all x_1 if and only if $a_{12} = 0$. (If $a_{12} > 0$, then choosing x_1 as a large negative number makes $Q(x_1, 1)$ negative. If $a_{12} < 0$, then choosing x_1 as a large positive number makes $Q(x_1, 1)$ negative.) Thus, $a_{11}a_{22} - a_{12}^2 = 0$. If $a_{11} > 0$, then (*) is valid and $Q(-a_{12}/a_{11}, 1) = [a_{11}a_{22} - a_{12}^2]/a_{11} \geq 0$, so $a_{11}a_{22} - a_{12}^2 \geq 0$.

The proof of (b) is easy when you observe that $Q(x_1, x_2) = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$ is ≥ 0 if and only if $-Q(x_1, x_2) = -a_{11}x_1^2 - 2a_{12}x_1x_2 - a_{22}x_2^2 \leq 0$.

The following equivalences are often useful:

$$\begin{aligned} \text{(a) } Q(x_1, x_2) \text{ is positive definite} &\iff a_{11} > 0 \text{ and } a_{11}a_{22} - a_{12}^2 > 0 \\ \text{(b) } Q(x_1, x_2) \text{ is negative definite} &\iff a_{11} < 0 \text{ and } a_{11}a_{22} - a_{12}^2 > 0 \end{aligned} \quad (5)$$

Proof: To prove (a), suppose first that $a_{11} > 0$ and $a_{11}a_{22} - a_{12}^2 > 0$. By (*), $Q(x_1, x_2) \geq 0$ for all (x_1, x_2) . If $Q(x_1, x_2) = 0$, then $x_1 + a_{12}x_2/a_{11} = 0$ and $x_2^2 = 0$, so $x_1 = x_2 = 0$. This proves $\Leftarrow$.

To prove the reverse implication, suppose $Q(x_1, x_2) > 0$ for all $(x_1, x_2) \neq (0, 0)$. In particular, $Q(1, 0) = a_{11} > 0$, so (*) is valid. Also $Q(-a_{12}/a_{11}, 1) = [a_{11}a_{22} - a_{12}^2]/a_{11} > 0$. The conclusion follows.

EXAMPLE 2 Use (4) and (5) to investigate the definiteness of

$$\text{(a) } Q(x_1, x_2) = 5x_1^2 - 2x_1x_2 + x_2^2 \quad \text{(b) } Q(x_1, x_2) = -4x_1^2 + 12x_1x_2 - 9x_2^2$$

Solution: (a) Note that $a_{11} = 5$, $a_{12} = -1$ (not -2 !), and $a_{22} = 1$. Thus $a_{11} > 0$ and $a_{11}a_{22} - a_{12}^2 = 5 - 1 = 4 > 0$, so according to (5), $Q(x_1, x_2)$ is positive definite.

(b) $a_{11} = -4$, $a_{12} = 6$ (not 12), and $a_{22} = -9$. Thus $a_{11} \leq 0$, $a_{22} \leq 0$, and $a_{11}a_{22} - a_{12}^2 = 36 - 36 = 0 \geq 0$, so according to (4), $Q(x_1, x_2)$ is negative semidefinite.

NOTE 1 In (5) we say nothing about the sign of a_{22} . But if $a_{11}a_{22} - a_{12}^2 > 0$, then $a_{11}a_{22} > a_{12}^2 \geq 0$, and so $a_{11}a_{22} > 0$. Since a_{11} is positive, so is a_{22} . So we could have added the condition $a_{22} > 0$ to the right-hand side in (5)(a), but it is superfluous. Note, however, that in (4)(a) one cannot drop the condition $a_{22} \geq 0$. For instance, $Q(x_1, x_2) = 0 \cdot x_1^2 - x_2^2$ has $a_{11} = 0$, $a_{12} = 0$, and $a_{22} = -1$, so $a_{11} \geq 0$ and $a_{11}a_{22} - a_{12}^2 \geq 0$, yet $Q(0, 1) = -1 < 0$.

The General Case

A quadratic form in n variables is a function Q of the form

$$Q(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j = a_{11} x_1^2 + a_{12} x_1 x_2 + \dots + a_{ij} x_i x_j + \dots + a_{nn} x_n^2 \quad (6)$$

where the a_{ij} are constants. Each term in the double sum either contains the square of a variable or a product of exactly two of the variables. In fact, Q is homogeneous of degree 2. We can see the structure of the quadratic form better if we write it this way:

$$Q(x_1, \dots, x_n) = a_{11}x_1^2 + a_{12}x_1x_2 + \dots + a_{1n}x_1x_n + a_{21}x_2x_1 + a_{22}x_2^2 + \dots + a_{2n}x_2x_n + \dots + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \dots + a_{nn}x_n^2 \quad (7)$$

Suppose we put $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ and $\mathbf{A} = (a_{ij})_{n \times n}$. Then it follows from the definition of matrix multiplication that

$$Q(x_1, \dots, x_n) = Q(\mathbf{x}) = \mathbf{x}' \mathbf{A} \mathbf{x} \quad (8)$$

By the same argument as in the case $n = 2$, in (7) we can assume that $a_{ij} = a_{ji}$ for all i and j , which means that the matrix A in (8) is symmetric. Then A is called the *symmetric matrix associated with Q* , and Q is called a *symmetric quadratic form*.

EXAMPLE 3 Write $Q(x_1, x_2, x_3) = 3x_1^2 + 6x_1x_3 + x_2^2 - 4x_2x_3 + 8x_3^2$ in matrix form with $\mathbf{A}$ symmetric.

Solution: We first write Q as follows:

$$Q = 3x_1^2 + 0 \cdot x_1x_2 + 3x_1x_3 + 0 \cdot x_2x_1 + x_2^2 - 2x_2x_3 + 3x_3x_1 - 2x_3x_2 + 8x_3^2$$

Then $Q = x'Ax$, where $A = \begin{pmatrix} 3 & 0 & 3 \\ 0 & 1 & -2 \\ 3 & -2 & 8 \end{pmatrix}$ and $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

Next, we want to generalize the definitions after (3) and the associated results in (4) and (5) to general quadratic forms.

DEFINITENESS OF A QUADRATIC FORM

A quadratic form $Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$, as well as its associated symmetric matrix $\mathbf{A}$, are said to be **positive definite**, **positive semidefinite**, **negative definite**, or **negative semidefinite** according as

$$Q(\mathbf{x}) > 0, \quad Q(\mathbf{x}) \geq 0, \quad Q(\mathbf{x}) < 0, \quad Q(\mathbf{x}) \leq 0 \quad (9)$$

for all $\mathbf{x} \neq \mathbf{0}$. The quadratic form $Q(\mathbf{x})$ is **indefinite** if there exist vectors $\mathbf{x}^*$ and $\mathbf{y}^*$ such that $Q(\mathbf{x}^*) < 0$ and $Q(\mathbf{y}^*) > 0$. Thus an indefinite quadratic form assumes both negative and positive values.

In order to generalize (4) and (5) we need a few new concepts. In Section 1.3 we defined the minors of a matrix. To study the sign of quadratic forms we need some particular minors.

An arbitrary **principal minor** of order r of an $n \times n$ matrix $\mathbf{A} = (a_{ij})$, is the determinant of a matrix obtained by deleting $n - r$ rows and $n - r$ columns such that if the i th row (column) is selected, then so is the i th column (row). In particular, a principal minor of order r always includes exactly r elements of the main (principal) diagonal. Also, if the matrix $\mathbf{A}$ is symmetric, then so is each matrix whose determinant is a principal minor. The determinant $|\mathbf{A}|$ itself is also a principal minor. (No rows or columns are deleted.)

A principal minor is called a **leading principal minor** of order r ($1 \leq r \leq n$), if it consists of the first ("leading") r rows and columns of $[A]$.

EXAMPLE 4 Write down the principal and the leading principal minors of

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

Solution: By deleting row 2 and column 2 in $\mathbf{A}$ we got (a_{11}) , which has determinant a_{11} . (Remember that the determinant of a 1×1 matrix (a) is the number a itself.) Deleting row 1 and column 1 in $\mathbf{A}$ we got (a_{22}) , which has determinant a_{22} . The principal minors of $\mathbf{A}$ are therefore a_{11} , a_{22} , and $|\mathbf{A}|$. The *leading* principal minors are a_{11} and $|\mathbf{A}|$.

The principal minors of $\mathbf{B}$ are $|\mathbf{B}|$ itself, and

$$b_{11}, \quad b_{22}, \quad b_{33}, \quad \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix}, \quad \begin{vmatrix} b_{11} & b_{13} \\ b_{31} & b_{33} \end{vmatrix}, \quad \text{and} \quad \begin{vmatrix} b_{22} & b_{23} \\ b_{32} & b_{33} \end{vmatrix}$$

while the leading principal minors are b_{11} , $\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix}$, and $|\mathbf{B}|$ itself.

Suppose $\mathbf{A}$ is an arbitrary $n \times n$ matrix. The leading principal minors of $\mathbf{A}$ are

$$D_k = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{vmatrix}, \quad k = 1, 2, \dots, n \quad (10)$$

These determinants are obtained from the elements in $|A|$ according to the pattern suggested by the following arrangement:

$$\begin{array}{ccccccc} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{array} \quad (11)$$

Note that there are many more principal minors than there are leading principal minors.⁶

⁶ There are $\binom{n}{k}$ principal minors of order k , but only one leading principal minor of order k .

It is notationally cumbersome to represent a specific principal minor, but we use Δ_k as a generic symbol to denote an arbitrary principal minor of order k .

The above concepts make it possible to formulate the following theorem:

THEOREM 1.7.1

Consider the symmetric matrix A and the associated quadratic form

$$Q(\mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j \quad (a_{ij} = a_{ji})$$

Let D_k be the leading principal minor defined by (10) and let Δ_k denote an arbitrary principal minor of A of order k . Then we have:

- (a) Q is positive definite $\iff D_k > 0$ for $k = 1, \dots, n$
- (b) Q is positive semidefinite $\iff \begin{cases} \Delta_k \geq 0 \text{ for all principal minors} \\ \text{of order } k = 1, \dots, n \end{cases}$
- (c) Q is negative definite $\iff (-1)^k D_k > 0$ for $k = 1, \dots, n$
- (d) Q is negative semidefinite $\iff \begin{cases} (-1)^k \Delta_k \geq 0 \text{ for all principal} \\ \text{minors of order } k = 1, \dots, n \end{cases}$

If we change the sign of each element in a $k \times k$ matrix, then the determinant of the new matrix is $(-1)^k$ times the determinant of the original matrix. Since $Q = \sum \sum a_{ij}x_i x_j$ is negative (semi)definite if and only if $-Q = \sum \sum (-a_{ij})x_i x_j$ is positive (semi)definite, we see that (c) and (d) in Theorem 1.7.1 follow from (a) and (b). For $n = 2$ we proved (a) and (b) in (5) and (4) above. For a general proof, we refer to e.g. Fraleigh and Bearegard (1995).

WARNING: A rather common misconception is that weakening each inequality $D_k > 0$ in (a) to $D_k \geq 0$, gives a necessary and sufficient condition for A to be positive semidefinite. Theorem 1.7.1 tells us that to check for positive semidefiniteness requires considering the signs of *all* principal minors, not only the leading ones. For a counterexample see Note 1 above.

EXAMPLE 5 Use Theorem 1.7.1 to determine the definiteness of

$$(a) Q = 3x_1^2 + 6x_1x_3 + x_2^2 - 4x_2x_3 + 8x_3^2 \quad (b) Q = -x_1^2 + 6x_1x_2 - 9x_2^2 - 2x_3^2$$

Solution: It makes sense to check the leading principal minors first, in case the matrix turns out to be definite rather than merely semidefinite.

(a) The associated symmetric matrix A is given in Example 3, and its leading principal minors are

$$D_1 = 3, \quad D_2 = \begin{vmatrix} 3 & 0 \\ 0 & 1 \end{vmatrix} = 3, \quad \text{and} \quad D_3 = \begin{vmatrix} 3 & 0 & 3 \\ 0 & 1 & -2 \\ 3 & -2 & 8 \end{vmatrix} = 3$$

We conclude that Q is positive definite.

(b) The symmetric matrix A associated with Q has leading principal minors $D_1 = -1$, $D_2 = 0$, $D_3 = 0$. It follows that the conditions in part (a) of Theorem 1.7.1 are not satisfied, nor are those in (b) or (c). In order to check the conditions in (d), we must examine all the principal minors of A . As described in Example 4, the 3×3 matrix A has three principal minors of order 1, whose values $\Delta_1^{(1)}$, $\Delta_1^{(2)}$, and $\Delta_1^{(3)}$ are the diagonal elements. These satisfy

$$(-1)^1 \Delta_1^{(1)} = (-1)(-1) = 1, \quad (-1)^1 \Delta_1^{(2)} = (-1)(-9) = 9, \quad (-1)^1 \Delta_1^{(3)} = (-1)(-2) = 2$$

There are also three second-order principal minors, which satisfy

$$\Delta_2^{(1)} = \begin{vmatrix} -1 & 3 \\ 3 & -9 \end{vmatrix} = 0, \quad \Delta_2^{(2)} = \begin{vmatrix} -1 & 0 \\ 0 & -2 \end{vmatrix} = 2, \quad \Delta_2^{(3)} = \begin{vmatrix} -9 & 0 \\ 0 & -2 \end{vmatrix} = 18$$

Hence $(-1)^2 \Delta_2^{(1)} = 0$, $(-1)^2 \Delta_2^{(2)} = 2$, and $(-1)^2 \Delta_2^{(3)} = 18$. Finally, $(-1)^3 \Delta_3 = (-1)^3 D_3 = 0$. Thus $(-1)^k \Delta_k \geq 0$ for all principal minors Δ_k of order k ($k = 1, 2, 3$) in the matrix A . It follows from (d) that Q is negative semidefinite.

The definiteness of a quadratic form can often be determined more easily from the signs of the eigenvalues of the associated matrix. (By Theorem 1.6.2 these eigenvalues are all real.)

THEOREM 1.7.2

Let $Q = \mathbf{x}'A\mathbf{x}$ be a quadratic form, where the matrix A is symmetric, and let $\lambda_1, \dots, \lambda_n$ be the (real) eigenvalues of A . Then:

- (a) Q is positive definite $\iff \lambda_1 > 0, \dots, \lambda_n > 0$
- (b) Q is positive semidefinite $\iff \lambda_1 \geq 0, \dots, \lambda_n \geq 0$
- (c) Q is negative definite $\iff \lambda_1 < 0, \dots, \lambda_n < 0$
- (d) Q is negative semidefinite $\iff \lambda_1 \leq 0, \dots, \lambda_n \leq 0$
- (e) Q is indefinite $\iff A$ has both pos. and neg. eigenvalues

Proof: By Theorem 1.6.2, there exists an orthogonal matrix P such that $P'AP = \text{diag}(\lambda_1, \dots, \lambda_n)$. Let $\mathbf{y} = (y_1, \dots, y_n)'$ be the $n \times 1$ matrix defined by $\mathbf{y} = P'\mathbf{x}$. Then $\mathbf{x} = P\mathbf{y}$, so

$$\mathbf{x}'A\mathbf{x} = \mathbf{y}'P'AP\mathbf{y} = \mathbf{y}'\text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \dots + \lambda_n y_n^2 \quad (12)$$

It follows that $\mathbf{x}'A\mathbf{x} \geq 0$ (resp. ≤ 0) for all $\mathbf{x}$ if and only if $\mathbf{y}'P'AP\mathbf{y} \geq 0$ (resp. ≤ 0) for all $\mathbf{y} \neq \mathbf{0}$. Also, $\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{y} = \mathbf{0}$, and so $\mathbf{x}'A\mathbf{x} > 0$ (resp. < 0) for all $\mathbf{x} \neq \mathbf{0}$ if and only if $\mathbf{y}'P'AP\mathbf{y} > 0$ (resp. < 0) for all $\mathbf{y} \neq \mathbf{0}$. The conclusion follows immediately. ■

EXAMPLE 6 Use Theorem 1.7.2 to determine the definiteness of the quadratic form in Example 5(b).

Solution: The associated matrix A has characteristic equation $-\lambda(\lambda + 2)(\lambda + 10) = 0$, so the eigenvalues are 0, -2, and -10. Theorem 1.7.2 tells us that the quadratic form is negative semidefinite. ■

PROBLEMS FOR SECTION 1.7

- Use (4) and (5) to investigate the definiteness of
 - $Q(x_1, x_2) = -x_1^2 + 2x_1x_2 - 6x_2^2$
 - $Q(x_1, x_2) = 4x_1^2 + 2x_1x_2 + 25x_2^2$
- Write out the double sum in (6) when $n = 3$ and $a_{ij} = a_{ji}$ for $i, j = 1, 2, 3$.
- What is the symmetric matrix $\mathbf{A}$ associated with each of the following quadratic forms?
 - $x^2 + 2xy + y^2$
 - $ax^2 + bxy + cy^2$
 - $3x_1^2 - 2x_1x_2 + 3x_1x_3 + x_2^2 + 3x_3^2$
- What is the symmetric matrix $\mathbf{A}$ associated with the following quadratic form?

$$3x_1^2 - 2x_1x_2 + 4x_1x_3 + 8x_1x_4 + x_2^2 + 3x_2x_3 + x_3^2 - 2x_3x_4 + x_4^2$$

- 5.** Using Theorem 1.7.1 or Theorem 1.7.2, or otherwise, determine the definiteness of

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

6. Let $A = (a_{ij})_{n \times n}$ be symmetric and positive semidefinite. Prove that

$$A \text{ is positive definite} \iff |A| \neq 0$$

- 7.** (a) For what values of c is the quadratic form

$$Q(x, y) = 3x^2 - (5 + c)xy + 2cy^2$$

- (i) positive definite, (ii) positive semidefinite, (iii) indefinite?

- (b) Let $\mathbf{B}$ be an $n \times n$ matrix. Show that the matrix $\mathbf{A} = \mathbf{B}^T \mathbf{B}$ is positive semidefinite. Can you find a necessary and sufficient condition on $\mathbf{B}$ for $\mathbf{A}$ to be positive definite, not just semidefinite?

8. Show that if $Q = \mathbf{x}'\mathbf{A}\mathbf{x}$ in (8) is positive definite, then

$$(a) \quad a_{ii} > 0, \quad i = 1, \dots, n \qquad (b) \quad \begin{vmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{vmatrix} > 0, \quad i, j = 1, \dots, n$$

9. Let A be a symmetric matrix. Write its characteristic polynomial (1.5.10) as

$$\varphi(\lambda) = (-1)^n (\lambda^n + a_{n-1} \lambda^{n-1} + \dots + a_1 \lambda + a_0)$$

Prove that $\mathbf{A}$ is negative definite if and only if $a_i > 0$ for $i = 0, 1, \dots, n-1$.

- SM 10.** (a) Consider the problem

$$\max(\min) Q = 2x_1^2 + 14x_1x_2 + 2x_2^2 \quad \text{subject to} \quad x_1^2 + x_2^2 = 1$$

The quadratic form Q can be written as $Q = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $\mathbf{A} = \begin{pmatrix} 2 & 7 \\ 7 & 2 \end{pmatrix}$.

Use Lagrange's method (see Section 3.3) to show that the first-order condition for $\mathbf{x}$ to solve either problem is that $\mathbf{x}$ is an eigenvector for $\mathbf{A}$. (The eigenvalue is the Lagrange multiplier.)

- (b) Prove that the largest (smallest) eigenvalue of A is the maximum (minimum) value of Q subject to the constraint. (Hint: Multiply $Ax = \lambda x$ from the left by x' .)

1.8 Quadratic Forms with Linear Constraints

In constrained optimization theory the second-order conditions involve the signs of quadratic forms subject to homogeneous linear constraints.

Consider first the quadratic form $Q = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2$ and assume that the variables are subject to the homogeneous linear constraint $b_1x_1 + b_2x_2 = 0$, where $b_1 \neq 0$. Solving the constraint for x_1 , we have $x_1 = -b_2x_2/b_1$. Substituting this value for x_1 into the expression for Q yields

$$Q = a_{11} \left(-\frac{b_2 x_2}{b_1} \right)^2 + 2a_{12} \left(-\frac{b_2 x_2}{b_1} \right) x_2 + a_{22} x_2^2 = \frac{1}{b_1^2} (a_{11} b_2^2 - 2a_{12} b_1 b_2 + a_{22} b_1^2) x_2^2 \quad (*)$$

We say that $Q(x_1, x_2)$ is **positive (negative) definite subject to the constraint** $b_1x_1 + b_2x_2 = 0$ if Q is positive (negative) for all $(x_1, x_2) \neq (0, 0)$ that satisfy $b_1x_1 + b_2x_2 = 0$. By expanding the determinant below, it is easy to verify that

$$a_{11}b_2^2 - 2a_{12}b_1b_2 + a_{22}b_1^2 = - \begin{vmatrix} 0 & b_1 & b_2 \\ b_1 & a_{11} & a_{12} \\ b_2 & a_{12} & a_{22} \end{vmatrix} \quad (1)$$

Combining this with (*) gives the following equivalence:

$$\left. \begin{array}{l} Q = a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \text{ is positive definite} \\ \text{subject to the constraint } b_1x_1 + b_2x_2 = 0 \end{array} \right\} \iff \begin{vmatrix} 0 & b_1 & b_2 \\ b_1 & a_{11} & a_{12} \\ b_2 & a_{12} & a_{22} \end{vmatrix} < 0 \quad (2)$$

This is also valid when $b_1 = 0$ but $b_2 \neq 0$. The condition for negative definiteness is that the determinant on the right-hand side of (2) is > 0 .

EXAMPLE 1 Use (2) to verify the definiteness of the quadratic form $x_1^2 - 8x_1x_2 + 16x_2^2$ subject to $2x_1 - x_2 = 0$.

Solution: The determinant in (2) is $D = \begin{vmatrix} 0 & 2 & -1 \\ 2 & 1 & -4 \\ -1 & -4 & 16 \end{vmatrix}$. We find that $D = -49$,

so the quadratic form is positive definite subject to the given constraint.

Consider next the general quadratic form

$$Q(\mathbf{x}) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j \quad (a_{ij} = a_{ji}) \quad (3)$$

in n variables subject to m linear, homogeneous constraints

[illegible]

or $\mathbf{B}\mathbf{x} = \mathbf{0}$, where $\mathbf{B} = (b_{ij})$ is an $m \times n$ matrix.

We say that Q is **positive (negative) definite** subject to the linear constraints (4) if $Q(\mathbf{x}) > 0$ (< 0) for all $\mathbf{x} = (x_1, \dots, x_n) \neq (0, \dots, 0)$ that satisfy (4).

Define the symmetric determinants

$$B_r = \begin{vmatrix} 0 & \cdots & 0 & b_{11} & \cdots & b_{1r} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & b_{m1} & \cdots & b_{mr} \\ b_{11} & \cdots & b_{m1} & a_{11} & \cdots & a_{1r} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ b_{1r} & \cdots & b_{mr} & a_{r1} & \cdots & a_{rr} \end{vmatrix}, \quad r = 1, \dots, n$$

Notice that the determinant B_r is the $(m+r)$ th leading principal minor of the $(m+n) \times (m+n)$ bordered matrix $\begin{pmatrix} 0_{m \times m} & \mathbf{B} \\ \mathbf{B}' & \mathbf{A} \end{pmatrix}$. Then we have the following result:

THEOREM 1.8.1

Assume that the first m columns in the matrix $(b_{ij})_{m \times n}$ in (4) are linearly independent. Then a necessary and sufficient condition for the quadratic form

$$Q = \sum_{i=1}^m \sum_{j=1}^n a_{ij} x_i x_j$$

to be positive definite subject to the linear constraints (4), is that

$$(-1)^m B_r > 0, \quad r = m+1, \dots, n \quad (5)$$

The corresponding necessary and sufficient condition for negative definiteness subject to the constraints (4) is

$$(-1)^r B_r > 0, \quad r = m+1, \dots, n \quad (6)$$

Note that the number of determinants to check is $n - m$. The more degrees of freedom (i.e. the smaller m is), the more determinants must be checked. If there is only one variable more than there are constraints, we need only examine B_n . If $n = 2$ and $m = 1$, then (5) reduces to $(-1)B_r > 0$ for $r = 2$, that is, $B_2 < 0$. (This confirms the result in (2).)

EXAMPLE 2 Examine the definiteness of $Q = 3x_1^2 - x_2^2 + 4x_3^2$ subject to $x_1 + x_2 + x_3 = 0$.

Solution: Here $n = 3$, $m = 1$. According to (5) and (6), we must examine the determinant B_r for $r = 2$ and 3. We find that

$$B_2 = \begin{vmatrix} 0 & 1 & 1 \\ 1 & 3 & 0 \\ 1 & 0 & -1 \end{vmatrix} = -2, \quad B_3 = \begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 3 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & 4 \end{vmatrix} = -5$$

Hence $(-1)^1 B_2 = 2$ and $(-1)^1 B_3 = 5$, so (5) shows that Q is positive definite subject to the given condition. (This is easy to check directly by substituting $-x_1 - x_2$ for x_3 in Q .)

PROBLEMS FOR SECTION 1.8

1. Determine the definiteness of $x_1^2 - 2x_1x_2 + x_2^2$ subject to $x_1 + x_2 = 0$ both directly and using Theorem 1.8.1.
2. Examine the definiteness of the following quadratic forms subject to the given linear constraint using Theorem 1.8.1:
 - (a) $2x_1^2 - 4x_1x_2 + x_2^2$ subject to $3x_1 + 4x_2 = 0$
 - (b) $-x_1^2 + x_1x_2 - x_2^2$ subject to $5x_1 - 2x_2 = 0$

SM 3. Examine the definiteness of $-5x^2 + 2xy + 4xz - y^2 - 2z^2$ s.t. $\begin{cases} x + y + z = 0 \\ 4x - 2y + z = 0 \end{cases}$

SM 4. Examine the definiteness of $x^2 + 2xy + y^2 + z^2$ s.t. $\begin{cases} x + 2y + z = 0 \\ 2x - y - 3z = 0 \end{cases}$

5. Find a necessary and sufficient condition for the quadratic form

$$\mathcal{L}_{11}''(x_1^*, x_2^*)h_1^2 + 2\mathcal{L}_{12}''(x_1^*, x_2^*)h_1h_2 + \mathcal{L}_{22}''(x_1^*, x_2^*)h_2^2$$

to be positive definite subject to the constraint $g_1'(x_1^*, x_2^*)h_1 + g_2'(x_1^*, x_2^*)h_2 = 0$.

1.9 Partitioned Matrices and Their Inverses

Some applications of linear algebra deal with matrices of high order. Sometimes, in order to see the structure of such matrices and to ease the computational burden in dealing with them, especially if a matrix has many zero elements, it helps to consider the matrix as an array of submatrices. operation of subdividing a matrix into submatrices is called **partitioning**.

EXAMPLE 1

Consider the 3×5 matrix $\mathbf{A} = \begin{pmatrix} 2 & 0 & 1 & 0 & 4 \\ 1 & 2 & 1 & 3 & 4 \\ 0 & 0 & 2 & 1 & 4 \end{pmatrix}$. The matrix $\mathbf{A}$ can be partitioned in a number of ways. For example,

$$\mathbf{A} = \left(\begin{array}{cc|cc|c} 2 & 0 & 1 & 0 & 4 \\ 1 & 2 & 1 & 3 & 4 \\ 0 & 0 & 2 & 1 & 4 \end{array} \right) = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \quad (*)$$

where $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are submatrices of dimensions 2×2 , 2×3 , 1×2 , and 1×3 , respectively. This is useful because $\mathbf{A}_{21}$ is a zero matrix. It is less useful to partition $\mathbf{A}$ into three row vectors, or into five column vectors.

Though Example 1 raises the possibility of partitioning a matrix into arbitrarily many submatrices, the rest of this section considers only partitionings into 2×2 arrays of submatrices, as in (*).

Operations on Partitioned Matrices

One can perform standard matrix operations on partitioned matrices, treating the submatrices as if they were ordinary matrix elements. This requires obeying the rules for sums, differences, and products.

Adding or subtracting partitioned matrices is simple. For example,

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} + B_{21} & A_{22} + B_{22} \end{pmatrix} \quad (1)$$

as long as the dimensions of A_{11} are those of B_{11} , the dimensions of A_{12} are those of B_{12} , and so on. The result follows directly from the definition of matrix addition. The rule for subtracting partitioned matrices is similar.

The rule for multiplying a partitioned matrix by a number is obvious. For example,

$$\alpha \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} \alpha A_{11} & \alpha A_{12} \\ \alpha A_{21} & \alpha A_{22} \end{pmatrix} \quad (2)$$

The following example shows how to multiply partitioned matrices.

EXAMPLE 2 Let A be the partitioned 3×5 matrix (*) in Example 1, and let B be the 5×4 matrix

$$B = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

with the indicated partitioning. The product AB is defined, and the ordinary rules of matrix multiplication applied to the entire matrices yield

$$AB = \begin{pmatrix} 2 & 0 & 5 & 6 \\ 1 & 2 & 6 & 15 \\ 0 & 0 & 3 & 4 \end{pmatrix}$$

Consider next how to take advantage of the partitioning of the two matrices to compute the product AB . The dimensions of the partitioned matrices allow us to multiply them as if the submatrices were ordinary matrix elements to obtain

$$\begin{aligned} AB &= \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix} \\ &= \begin{pmatrix} \begin{pmatrix} 2 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \end{pmatrix} & \begin{pmatrix} 4 & 2 \end{pmatrix} + \begin{pmatrix} 1 & 4 \end{pmatrix} \\ \begin{pmatrix} 1 & 2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \end{pmatrix} & \begin{pmatrix} 2 & 11 \end{pmatrix} + \begin{pmatrix} 4 & 4 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 \end{pmatrix} + \begin{pmatrix} 3 & 4 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} 2 & 0 & 5 & 6 \\ 1 & 2 & 6 & 15 \\ 0 & 0 & 3 & 4 \end{pmatrix} \quad \mathbf{I} \end{aligned}$$

The method suggested by Example 2 is valid in general. It is not difficult to formulate and prove the general result, though the notation becomes cumbersome. If you work through Problem 1 in detail, the general idea should become clear enough.

Multiplying matrices using partitioning is particularly convenient if the matrices have a special structure and involve simple submatrices (like identity or zero matrices).

EXAMPLE 3 Use the indicated partitioning to find M^n for all $n = 2, 3, \dots$

$$M = \begin{pmatrix} 1/4 & 1/2 & 1/6 & 0 \\ 1/2 & 1/3 & 0 & 1/6 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} P & Q \\ 0 & I \end{pmatrix}$$

Solution: Multiplication gives

$$M^2 = \begin{pmatrix} P & Q \\ 0 & I \end{pmatrix} \begin{pmatrix} P & Q \\ 0 & I \end{pmatrix} = \begin{pmatrix} P^2 & (P+I)Q \\ 0 & I \end{pmatrix} \quad \text{and} \quad M^3 = \begin{pmatrix} P^3 & (P^2+P+I)Q \\ 0 & I \end{pmatrix}.$$

In general, for all natural numbers n , it can be shown by induction that

$$M^n = \begin{pmatrix} P^n & (P^{n-1} + \dots + P^2 + P + I)Q \\ 0 & I \end{pmatrix} \quad \mathbf{I}$$

Inverses by Partitioning

Inverting large square matrices is often made much easier using partitioning. Consider an $n \times n$ matrix A which has an inverse. Assume that A is partitioned as follows:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad \text{where } A_{11} \text{ is a } k \times k \text{ matrix with an inverse} \quad (3)$$

Hence A_{12} is a $k \times (n-k)$ matrix, A_{21} is $(n-k) \times k$, while A_{22} is an $(n-k) \times (n-k)$ matrix. Since A has an inverse, there exists an $n \times n$ matrix B such that $AB = I_n$. Partitioning B in the same way as A yields

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

The equality $AB = I_n$ implies the following four matrix equations for determining B_{11} , B_{12} , B_{21} , and B_{22} :

$$\begin{aligned} \text{(i)} \quad A_{11}B_{11} + A_{12}B_{21} &= I_k & \text{(ii)} \quad A_{11}B_{12} + A_{12}B_{22} &= 0_{k \times (n-k)} \\ \text{(iii)} \quad A_{21}B_{11} + A_{22}B_{21} &= 0_{(n-k) \times k} & \text{(iv)} \quad A_{21}B_{12} + A_{22}B_{22} &= I_{n-k} \end{aligned}$$

where the subscripts attached to I and 0 indicate the dimensions of these matrices. Because A_{11} has an inverse, (ii) gives $B_{12} = -A_{11}^{-1}A_{12}B_{22}$. Inserting this into (iv) gives $\Delta B_{22} = I_{n-k}$, where $\Delta = A_{22} - A_{21}A_{11}^{-1}A_{12}$. So Δ has an inverse and $B_{22} = \Delta^{-1}$. Next, solve (i) for B_{11} to

obtain $B_{11} = A_{11}^{-1} - A_{11}^{-1}A_{12}B_{21}$. Inserting this into (iii) yields $A_{21}A_{11}^{-1} - A_{21}A_{11}^{-1}A_{12}B_{21} + A_{22}B_{21} = 0$, or $\Delta B_{21} = -A_{21}A_{11}^{-1}$. It follows that $B_{21} = -\Delta^{-1}A_{21}A_{11}^{-1}$ and so $B_{11} = A_{11}^{-1} + A_{11}^{-1}A_{12}\Delta^{-1}A_{21}A_{11}^{-1}$. The conclusion is:

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}\Delta^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}\Delta^{-1} \\ -\Delta^{-1}A_{21}A_{11}^{-1} & \Delta^{-1} \end{pmatrix} \quad (4)$$

where $\Delta = A_{22} - A_{21}A_{11}^{-1}A_{12}$.

Formula (4) is valid if both A_{11}^{-1} and A^{-1} exist. Similarly, if both A_{22}^{-1} and A^{-1} exist, then

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \bar{\Delta}^{-1} & -\bar{\Delta}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}\bar{\Delta}^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}\bar{\Delta}^{-1}A_{12}A_{22}^{-1} \end{pmatrix} \quad (5)$$

where $\bar{\Delta} = A_{11} - A_{12}A_{22}^{-1}A_{21}$.

EXAMPLE 4

Compute A^{-1} when $A = \begin{pmatrix} 2 & -3 & 0 & 0 & 0 \\ 3 & -4 & 0 & 0 & 0 \\ \hline 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ -5 & 7 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$.

Solution: Because $A_{22} = I$, it is easier to use formula (5) than (4). Note that $\bar{\Delta} = A_{11}$, so by formula (1.1.29), $\bar{\Delta}^{-1} = \begin{pmatrix} -4 & 3 \\ -3 & 2 \end{pmatrix}$. Then

$$-A_{22}^{-1}A_{21}\bar{\Delta}^{-1} = -I_3 \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ -5 & 7 \end{pmatrix} \begin{pmatrix} -4 & 3 \\ -3 & 2 \end{pmatrix} = - \begin{pmatrix} -1 & 1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix}$$

and so (you should check the result by computing AA^{-1} using partitioning):

$$A^{-1} = \begin{pmatrix} \begin{pmatrix} -4 & 3 \\ -3 & 2 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{pmatrix} = \begin{pmatrix} -4 & 3 & 0 & 0 & 0 \\ -3 & 2 & 0 & 0 & 0 \\ \hline 1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

To find this inverse without partitioning would be very time consuming.

We conclude this section by giving two useful formulas for the determinant of an $n \times n$ matrix A partitioned as in (3) (see Problem 6):

$$\text{If } A_{11}^{-1} \text{ exists, then } \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{11}| \cdot |A_{22} - A_{21}A_{11}^{-1}A_{12}| \quad (6)$$

$$\text{If } A_{22}^{-1} \text{ exists, then } \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{22}| \cdot |A_{11} - A_{12}A_{22}^{-1}A_{21}| \quad (7)$$

PROBLEMS FOR SECTION 1.9

1. Compute the following matrix product using: (i) ordinary matrix multiplication; (ii) the suggested partitioning.

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ \hline a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ \hline b_{31} & b_{32} \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} \\ B_{21} \end{pmatrix}$$

2. Compute the following matrix product using the suggested partitioning. Check the result by ordinary matrix multiplication.

$$\begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & -1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$$

3. Use partitioning to compute the inverses of the following matrices:

$$(a) \begin{pmatrix} 2 & 3 & 0 & 0 \\ 5 & 2 & 0 & 0 \\ 0 & 0 & 4 & 3 \\ 0 & 0 & 3 & 2 \end{pmatrix} \quad (b) \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad (c) \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}$$

4. Let $A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$ and $X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, where $|A| \neq 0$. Show that

$$|A + XX'| = \begin{vmatrix} 1 & -x_1 & \cdots & -x_n \\ x_1 & a_{11} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_n & a_{n1} & \cdots & a_{nn} \end{vmatrix} = |A| \cdot (1 + X'A^{-1}X) \quad (*)$$

where the 1×1 matrix $X'A^{-1}X$ is treated as a number. (This formula is useful in econometrics.)

5. If P and Q are invertible square matrices, prove that $\begin{pmatrix} P & R \\ 0 & Q \end{pmatrix}^{-1} = \begin{pmatrix} P^{-1} & -P^{-1}RQ^{-1} \\ 0 & Q^{-1} \end{pmatrix}$.

HARDER PROBLEMS

6. (a) Show that if A_{12} or A_{21} in (3) is the zero matrix, then $|A| = |A_{11}| \cdot |A_{22}|$. (Hint: Use the definition of a determinant. See e.g. EMFA, Chapter 16.)

(b) Show that $\begin{pmatrix} I_k & -A_{12}A_{22}^{-1} \\ 0 & I_{n-k} \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ 0 & A_{22} \end{pmatrix}$, and use this result to prove (7).

7. (a) Suppose A is $n \times m$ and B is $m \times n$. Prove that $|I_n + AB| = |I_m + BA|$. (Hint: Define $D = \begin{pmatrix} I_n & A \\ 0 & I_m \end{pmatrix}$, $E = \begin{pmatrix} I_n & -A \\ B & I_m \end{pmatrix}$. Then $|I_n + AB| = |DE| = |D||E| = |E||D| = |ED| = |I_m + BA|$.)

(b) Use the result in (a) to prove that if $a_1, \dots, a_n$ are all different from 1, then

$$\begin{vmatrix} a_1 & 1 & 1 & \dots & 1 \\ 1 & a_2 & 1 & \dots & 1 \\ 1 & 1 & a_3 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & a_n \end{vmatrix} = (a_1 - 1)(a_2 - 1) \cdots (a_n - 1) \left(1 + \sum_{i=1}^n \frac{1}{a_i - 1} \right)$$

(Hint: Let $F = I_n + A_{n \times n}$ where $A_{n \times n} = \left(\frac{1}{a_1 - 1}, \frac{1}{a_2 - 1}, \dots, \frac{1}{a_n - 1} \right)$ and $B_{1 \times n} = (1, 1, \dots, 1)$.)

2

MULTIVARIABLE
CALCULUS

Wisdom and maturity are the last settlers in pioneering communities.

—P. A. Samuelson (1985)

In this chapter we discuss a number of topics in multivariable calculus.

First, Section 2.1 deals with gradients and directional derivatives, which are useful concepts in optimization theory. The important properties of the gradient vector are set out in Theorem 2.1.1.

The elementary theory of convex sets in $\mathbb{R}^n$ is discussed in Section 2.2. (Some additional results are presented in Sections 13.5 and 13.6.)

Sections 2.3 and 2.4 give a rather detailed introduction to concave (and convex) functions. It is important that you learn to appreciate arguments based directly on the definition of concavity. Many results in economic theory become so much simpler when one does not have to "differentiate everything in sight". Jensen's inequality, which finds many uses in economic analysis, is essentially an easy extension of the definition.

Although important in theoretical arguments, the definition of concavity is essentially useless when we want to check if a particular function is concave. In establishing concavity it is useful to know that a nonnegative linear combination of concave functions is concave (Theorem 2.3.4), and that an increasing concave function of a concave function is concave (Theorem 2.3.5). However, we often have to rely on the standard second-order conditions in Theorems 2.3.2 and 2.3.3 based on the signs of the minors of the Hessian matrix.

Quasiconcave (and quasiconvex) functions are important in economics, mostly in utility theory. It is worth noting that a sum of quasiconcave functions need not be quasiconcave.

In utility theory it is useful to know that an increasing transformation of a quasiconcave function is still quasiconcave (Theorem 2.5.2). Properties of the Cobb–Douglas and generalized CES functions are described in some detail.

Arguments based on the basic characterizations of quasiconcavity are important in economic theory, but are less useful in deciding whether a specified function is quasiconcave. The standard second-order conditions based on the signs of the minors of the bordered Hessian matrix are set out at the end of the section.

Taylor's formula is a major result in mathematical analysis, and the brief Section 2.6 presents the formula for functions of several variables.

Even elementary books in mathematics for economists need to discuss implicit function theorems. In Section 2.7 we formulate precise results.

Section 2.8 begins with a discussion on degrees of freedom in systems of equations. The tricky concept of functional dependence is dealt with next. Many texts state erroneously that functional dependence of n functions is equivalent to the vanishing of the associated Jacobian determinant.

Section 2.9 is rather demanding and discusses linear approximations and differentiability.

Finally, Section 2.10 gives some results on the existence and uniqueness of solutions to systems of equations. In particular, some results on global univalence are discussed.

2.1 Gradients and Directional Derivatives

If F is a function of two variables and C is any number, the graph of the equation $F(x, y) = C$ is called a **level curve** for F . Recall that the slope of the level curve $F(x, y) = C$ at a point (x, y) is given by the formula

$$F(x, y) = C \Rightarrow y' = -\frac{F'_1(x, y)}{F'_2(x, y)} \quad (1)$$

According to (1), if (x_0, y_0) is a particular point on the level curve $F(x, y) = C$, the slope at (x_0, y_0) is $-F'_1(x_0, y_0)/F'_2(x_0, y_0)$. The equation for the tangent line T shown in Fig. 1 is $y - y_0 = -[F'_1(x_0, y_0)/F'_2(x_0, y_0)](x - x_0)$ or, rearranging,

$$F'_1(x_0, y_0)(x - x_0) + F'_2(x_0, y_0)(y - y_0) = 0 \quad (2)$$

The inner product notation (see (1.1.35)) allows us to write equation (2) as

$$(F'_1(x_0, y_0), F'_2(x_0, y_0)) \cdot (x - x_0, y - y_0) = 0 \quad (3)$$

The vector $(F'_1(x_0, y_0), F'_2(x_0, y_0))$ is called the **gradient** of F at (x_0, y_0) , and is often denoted by $\nabla F(x_0, y_0)$. The vector $(x - x_0, y - y_0)$ is a vector on the tangent line T in Fig. 1, and (3) means that $\nabla F(x_0, y_0)$ is **orthogonal** to the tangent line T at (x_0, y_0) .

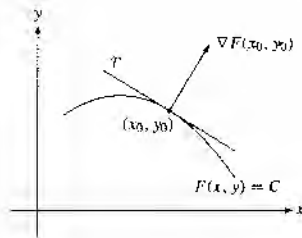


Figure 1 $\nabla F(x_0, y_0)$ is orthogonal to the tangent line T at (x_0, y_0) .

EXAMPLE 1 Compute $\nabla F(\frac{1}{2}, 2)$ for $F(x, y) = xy$, and find the equation for the tangent to the curve $xy = 1$ at $(\frac{1}{2}, 2)$.

Solution: $F'_1(x, y) = y$ and $F'_2(x, y) = x$, so $\nabla F(\frac{1}{2}, 2) = (2, \frac{1}{2})$. Hence, by (3), the equation of the tangent is

$$(2, \frac{1}{2}) \cdot (x - \frac{1}{2}, y - 2) = 0, \quad \text{i.e. } 2x - 1 + \frac{1}{2}y - 1 = 0, \quad \text{or } y = -4x + 4 \quad \blacksquare$$

Suppose more generally that $F(\mathbf{x}) = F(x_1, \dots, x_n)$ is a function of n variables defined on an open set A in $\mathbb{R}^n$, and let $\mathbf{x}^0 = (x_1^0, \dots, x_n^0)$ be a point in A . The **gradient** of F at $\mathbf{x}^0$ is the vector

$$\nabla F(\mathbf{x}^0) = \left(\frac{\partial F(\mathbf{x}^0)}{\partial x_1}, \dots, \frac{\partial F(\mathbf{x}^0)}{\partial x_n} \right)$$

of first-order partial derivatives. Alternative notations for $\nabla F(\mathbf{x}^0)$ are $F'(\mathbf{x}^0)$ and $DF(\mathbf{x}^0)$.

Consider the level surface of $F(x_1, \dots, x_n)$ corresponding to the level C , i.e. the set of points that satisfy

$$F(x_1, \dots, x_n) = C$$

If $\mathbf{x}^0 = (x_1^0, \dots, x_n^0)$ lies on this level surface, i.e. if $F(\mathbf{x}^0) = C$, then the **tangent hyperplane** to the level surface at $\mathbf{x}^0$ is the set of all $\mathbf{x} = (x_1, \dots, x_n)$ such that

$$F'_1(\mathbf{x}^0)(x_1 - x_1^0) + \dots + F'_n(\mathbf{x}^0)(x_n - x_n^0) = 0$$

Using the scalar product we can write this as

$$\nabla F(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0) = 0 \quad (4)$$

Since any point $\mathbf{x}$ in the tangent hyperplane satisfies (4), the gradient $\nabla F(\mathbf{x}^0)$ is orthogonal to the tangent hyperplane at $\mathbf{x}^0$. (See (1.1.43).) An illustration is given in Fig. 2 for $n = 3$. (We assume that $\nabla F(\mathbf{x}^0) \neq \mathbf{0}$.)

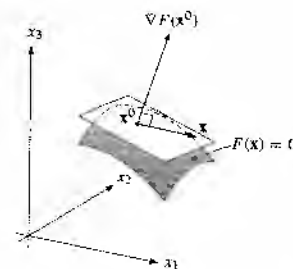


Figure 2 The gradient $\nabla F(\mathbf{x}^0)$ is orthogonal to the tangent plane of $F(\mathbf{x}) = C$ at $\mathbf{x}^0$.

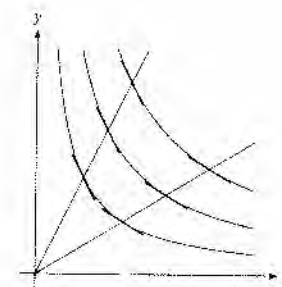


Figure 3 f is homogeneous of degree k . The level curves are parallel along each ray from the origin.

NOTE 1 If the graph in Fig. 2 is defined by $z = f(x_1, x_2)$, then the tangent plane to this graph at the point (x_1^0, x_2^0, z^0) is (see e.g. EMEA, Section 12.8),

$$f'_1(x_1^0, x_2^0)(x_1 - x_1^0) + f'_2(x_1^0, x_2^0)(x_2 - x_2^0) - 1 \cdot (z - z^0) = 0$$

Thus the vector $(f'_1(x_1^0, x_2^0), f'_2(x_1^0, x_2^0), -1)$ is orthogonal to the tangent plane of $z = f(x_1, x_2)$ at (x_1^0, x_2^0, z^0) . This tangent plane is precisely the plane obtained from (4) by putting $F(x_1, x_2, z) = f(x_1, x_2) - z$.

EXAMPLE 2 Let $f(x_1, \dots, x_n) = f(\mathbf{x})$ be homogeneous of degree k . Then $f'_i(\mathbf{x})$ is homogeneous of degree $k-1$ for $i = 1, \dots, n$ (see e.g. EMEA, Section 12.7). Hence, for $\lambda > 0$ we have $\nabla f(\lambda \mathbf{x}) = (f'_1(\lambda \mathbf{x}), \dots, f'_n(\lambda \mathbf{x})) = (\lambda^{k-1} f'_1(\mathbf{x}), \dots, \lambda^{k-1} f'_n(\mathbf{x})) = \lambda^{k-1} \nabla f(\mathbf{x})$. So, at each point on a given ray through the origin the gradients are proportional, which implies that the tangent planes at each point on the ray are parallel. See Fig. 3. To summarize:

$$f(\mathbf{x}) \text{ is homogeneous of degree } k \implies \begin{cases} \text{the level surfaces are parallel} \\ \text{along each ray from the origin.} \end{cases} \quad (5)$$

The Directional Derivative

Let $z = f(\mathbf{x})$ be a function of n variables. The partial derivative $\partial f / \partial x_i$ measures the rate of change of $f(\mathbf{x})$ in the direction parallel to the i th coordinate axis. Each partial derivative says nothing about the behaviour of f in other directions. We introduce the concept of *directional derivative* in order to measure the rate of change of f in an arbitrary direction.

Consider the vector $\mathbf{x} = (x_1, \dots, x_n)$ and let $\mathbf{a} = (a_1, \dots, a_n) \neq \mathbf{0}$ be a given vector. If we move a distance $h\|\mathbf{a}\| > 0$ from $\mathbf{x}$ in the direction given by $\mathbf{a}$, we arrive at $\mathbf{x} + h\mathbf{a}$. The average rate of change of f from $\mathbf{x}$ to $\mathbf{x} + h\mathbf{a}$ is then $(f(\mathbf{x} + h\mathbf{a}) - f(\mathbf{x})) / h$. We define the *derivative of f along the vector $\mathbf{a}$* , denoted $f'_\mathbf{a}(\mathbf{x})$, by

$$f'_\mathbf{a}(\mathbf{x}) = \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{a}) - f(\mathbf{x})}{h} \quad (6)$$

or, with its components written out,

$$f'_\mathbf{a}(x_1, \dots, x_n) = \lim_{h \rightarrow 0} \frac{f(x_1 + ha_1, \dots, x_n + ha_n) - f(x_1, \dots, x_n)}{h}$$

(We assume that $\mathbf{x} + h\mathbf{a}$ lies in the domain of f for all sufficiently small h .) In particular, with $a_i = 1$ and $a_j = 0$ for $j \neq i$, the derivative in (6) is the partial derivative of f w.r.t. x_i .

Suppose f is C^1 in an open set A , and let $\mathbf{x}$ be a point in A .¹ For an arbitrary vector $\mathbf{a}$, define the function g by $g(h) = f(\mathbf{x} + h\mathbf{a}) = f(x_1 + ha_1, \dots, x_n + ha_n)$. Then

¹ Open sets and related topological concepts are reviewed in Section 13.1. A function f is said to be of class C^k ($k \geq 1$) in a set U if f and all its partial derivatives of order $\leq k$ exist and are continuous throughout U .

$(g(h) - g(0))/h = (f(\mathbf{x} + h\mathbf{a}) - f(\mathbf{x}))/h$. Letting h tend to 0, equation (6) implies that $g'(0) = f'_\mathbf{a}(\mathbf{x})$. But according to the chain rule for differentiating composite functions,

$$g(h) = f(\mathbf{x} + h\mathbf{a}) \implies g'(h) = \sum_{i=1}^n f'_i(\mathbf{x} + h\mathbf{a}) a_i \quad (7)$$

so that $g'(0) = \sum_{i=1}^n f'_i(\mathbf{x}) a_i$. Hence we have shown that

$$f'_\mathbf{a}(\mathbf{x}) = \sum_{i=1}^n f'_i(\mathbf{x}) a_i = \nabla f(\mathbf{x}) \cdot \mathbf{a} \quad (8)$$

Formula (8) shows that *the derivative of f along the vector $\mathbf{a}$ is equal to the inner product of the gradient of f and $\mathbf{a}$.*

If $\|\mathbf{a}\| = 1$, the number $f'_\mathbf{a}(\mathbf{x})$ is called the **directional derivative** of f at $\mathbf{x}$, in the direction $\mathbf{a}$. It is precisely when $\mathbf{a}$ is a vector with length 1 that we have the following nice interpretation of $f'_\mathbf{a}(\mathbf{x})$: moving a distance h in the direction given by $\mathbf{a}$ changes the value of f by approximately $f'_\mathbf{a}(\mathbf{x})h$, provided h is small.

EXAMPLE 3 Find the directional derivative of $f(x, y) = x^2 + 2xy + y^3$ at the point $(x, y) = (1, 1)$ in the direction given by the vector $\mathbf{a} = \mathbf{b}/\|\mathbf{b}\|$, where $\mathbf{b} = (1, 3)$.

Solution: Let $\mathbf{b} = (1, 3)$. Then $\|\mathbf{b}\| = (1^2 + 3^2)^{1/2} = \sqrt{10}$. The vector $\mathbf{a} = (a_1, a_2) = (1/\sqrt{10}, 3/\sqrt{10})$ is a vector of length 1 in the same direction as $\mathbf{b}$. Now, $f'_1(x, y) = 2x + 2y$ and $f'_2(x, y) = 2x + 3y^2$, so that $f'_1(1, 1) = 4$, $f'_2(1, 1) = 5$. According to (8),

$$f'_\mathbf{a}(1, 1) = f'_1(1, 1)a_1 + f'_2(1, 1)a_2 = 4 \frac{1}{\sqrt{10}} + 5 \frac{3}{\sqrt{10}} = \frac{19}{10} \sqrt{10}$$

By introducing φ as the angle between the vectors $\nabla f(\mathbf{x})$ and $\mathbf{a}$ (see (1.1.40)), we have

$$f'_\mathbf{a}(\mathbf{x}) = \|\nabla f(\mathbf{x})\| \|\mathbf{a}\| \cos \varphi \quad (9)$$

Remember that $\cos \varphi \leq 1$ for all φ and $\cos 0 = 1$. So when $\|\mathbf{a}\| = 1$, it follows that at points where $\nabla f(\mathbf{x}) \neq \mathbf{0}$, the number $f'_\mathbf{a}(\mathbf{x})$ is largest when $\varphi = 0$, i.e. when $\mathbf{a}$ points in the same direction as $\nabla f(\mathbf{x})$, while $f'_\mathbf{a}(\mathbf{x})$ is smallest when $\varphi = \pi$ (and hence $\cos \varphi = -1$), i.e. when $\mathbf{a}$ points in the opposite direction of $\nabla f(\mathbf{x})$. Moreover, it follows that the length of $\nabla f(\mathbf{x})$ equals the magnitude of the maximum directional derivative.

The most important observations about the gradient are gathered in this theorem:

THEOREM 2.1.1 (PROPERTIES OF THE GRADIENT)

Suppose that $f(\mathbf{x}) = f(x_1, \dots, x_n)$ is C^1 in an open set A . Then, at points $\mathbf{x}$ where $\nabla f(\mathbf{x}) \neq \mathbf{0}$, the gradient $\nabla f(\mathbf{x}) = (f'_1(\mathbf{x}), f'_2(\mathbf{x}), \dots, f'_n(\mathbf{x}))$ satisfies:

- $\nabla f(\mathbf{x})$ is orthogonal to the level surface through $\mathbf{x}$.
- $\nabla f(\mathbf{x})$ points in the direction of maximal increase of f .
- $\|\nabla f(\mathbf{x})\|$ measures how fast the function increases in the direction of maximal increase.

We assumed above that $\nabla f(\mathbf{x}) \neq 0$. Points $\mathbf{x}$ where $\nabla f(\mathbf{x}) = 0$ are called *stationary points* for f . If f is C^1 , it follows from (8) that all directional derivatives are equal to 0 at a stationary point.

The Mean Value Theorem

The mean value theorem for functions of one variable (see e.g. EMEA, Section 8.4) can easily be generalized to functions of several variables. Suppose that $\mathbf{x}$ and $\mathbf{y}$ are points in $\mathbb{R}^n$. Then define the **closed** and **open line segments** between $\mathbf{x}$ and $\mathbf{y}$ as the sets

$$[\mathbf{x}, \mathbf{y}] = \{\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} : \lambda \in [0, 1]\} \quad \text{and} \quad (\mathbf{x}, \mathbf{y}) = \{\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} : \lambda \in (0, 1)\}$$

respectively.

THEOREM 2.1.2 (THE MEAN VALUE THEOREM)

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is C^1 in an open set containing $[\mathbf{x}, \mathbf{y}]$. Then there exists a point $\mathbf{w}$ in $(\mathbf{x}, \mathbf{y})$ such that

$$f(\mathbf{x}) - f(\mathbf{y}) = \nabla f(\mathbf{w}) \cdot (\mathbf{x} - \mathbf{y}) \quad (10)$$

Proof: Define $\varphi(\lambda) = f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y})$. Then $\varphi'(\lambda) = \nabla f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \cdot (\mathbf{x} - \mathbf{y})$. According to the mean value theorem for functions of one variable, there exists a number λ_0 in $(0, 1)$ such that $\varphi(1) - \varphi(0) = \varphi'(\lambda_0)$. Putting $\mathbf{w} = \lambda_0 \mathbf{x} + (1 - \lambda_0)\mathbf{y}$, we get (10). ■

PROBLEMS FOR SECTION 2.1

1. Compute the gradients of the following functions at the given points.

(a) $f(x, y) = y^2 + xy$ at $(2, 1)$ (b) $g(x, y, z) = xe^{xy} - z^2$ at $(0, 0, 1)$
 (c) $h(x, y, z) = e^x + e^{2y} + e^{3z}$ at $(0, 0, 0)$ (d) $k(x, y, z) = e^{x+2y+3z}$ at $(0, 0, 0)$

2. Let $f(t)$ be a C^1 function of t with $f'(t) \neq 0$.

(a) Put $F(x, y) = f(x^2 + y^2)$. Find the gradient ∇F at an arbitrary point and show that it is parallel to the straight line segment joining the point and the origin.
 (b) Put $G(x, y) = f(y/x)$. Find ∇G at an arbitrary point where $x \neq 0$, and show that it is orthogonal to the straight line segment joining the point and the origin.

3. Compute the directional derivatives of the following functions at the given points and in the given directions.

(a) $f(x, y) = 2x + y - 1$ at $(2, 1)$, in the direction given by $(1, 1)$.
 (b) $g(x, y, z) = xe^{xy} - xy - z^2$ at $(0, 1, 1)$, in the direction given by $(1, 1, 1)$.

4. Let

$$f(x_1, \dots, x_n) = x_1^2 + x_2^2 + \dots + x_n^2$$

Find the directional derivative of f in the direction given by the vector $\mathbf{a} = (a_1, a_2, \dots, a_n)$ by using definition (6). Check the result by using (8). (Suppose that $\|\mathbf{a}\| = 1$.)

5. (a) Find the directional derivative of

$$f(x, y, z) = xy \ln(x^2 + y^2 + z^2)$$

at $(1, 1, 1)$ in the direction given by the vector from the point $(3, 2, 1)$ to the point $(-1, 1, 2)$.

- (b) Determine also the direction of maximal increase from the point $(1, 1, 1)$.

6. Suppose that $f(x, y)$ has continuous partial derivatives. Suppose too that the maximum directional derivative of f at $(0, 0)$ is equal to 4, and that it is attained in the direction given by the vector from the origin to the point $(1, 3)$. Find $\nabla f(0, 0)$.

7. Let $\mathbf{b} = (b_1, \dots, b_n)$ be a given vector and define the function $f(\mathbf{x}) = f(x_1, \dots, x_n) = \mathbf{b} \cdot \mathbf{x}$. Show that the derivative of f along the vector $\mathbf{a} = (a_1, \dots, a_n)$ is $\mathbf{b} \cdot \mathbf{a}$.

8. Let $f(\mathbf{v}) = f(v_1, \dots, v_n)$ denote a positive valued differentiable function of n variables defined whenever $v_i > 0$, $i = 1, 2, \dots, n$. The **directional elasticity** of f at the point $\mathbf{v}$ along the vector $\mathbf{v}/\|\mathbf{v}\| = \mathbf{a}$, hence in the direction from the origin to $\mathbf{v}$, is denoted by $\text{El}_{\mathbf{a}} f(\mathbf{v})$ and is, by definition,

$$\text{El}_{\mathbf{a}} f(\mathbf{v}) = \frac{\|\mathbf{v}\|}{f(\mathbf{v})} f'_{\mathbf{a}}(\mathbf{v}), \quad \mathbf{a} = \frac{\mathbf{v}}{\|\mathbf{v}\|}$$

where $f'_{\mathbf{a}}(\mathbf{v})$ is the directional derivative of f in the direction given by $\mathbf{a}$. Use (8) to show that

$$\text{El}_{\mathbf{a}} f(\mathbf{v}) = \sum_{i=1}^n \text{El}_i f(\mathbf{v})$$

where $\text{El}_i f(\mathbf{v})$ denotes the partial elasticity of f w.r.t. v_i . (When $f(\mathbf{v})$ is a production function, $\text{El}_{\mathbf{a}} f(\mathbf{v})$ is called the **scale elasticity**.)

HARDER PROBLEMS

9. (a) Prove that if F is C^2 and $F(x, y) = C$ defines y as a twice differentiable function of x , then

$$y'' = -\frac{1}{(F_2')^3} [F_{11}''(F_2')^2 - 2F_{12}''F_1'F_2' + F_{22}''(F_1')^2] = \frac{1}{(F_2')^3} \begin{vmatrix} 0 & F_1' & F_2' \\ F_1' & F_{11}'' & F_{12}'' \\ F_2' & F_{21}'' & F_{22}'' \end{vmatrix}$$

(Differentiate the expression for y' in (1) w.r.t. x .)

- (b) Let $F(x, y) = x^2y = 8$. Use the formula in (a) to compute y'' at $(x, y) = (2, 2)$. Check the result by differentiating $y = 8/x^2$ twice.

2.2 Convex Sets

Convexity plays an important role in theoretical economics. In this book we shall see many examples of its importance.

A set S in the plane is called **convex** if each pair of points in S can be joined by a line segment lying entirely within S . Examples are given in Fig. 1. In (d) the non-convex set is the union $S_1 \cup S_2$ of two ovals, each of which is convex on its own.

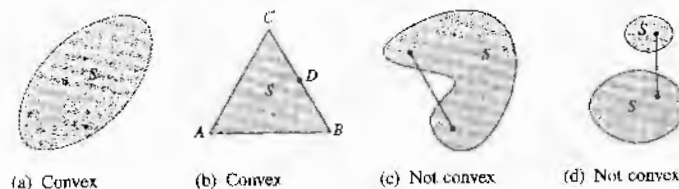


Figure 1 Convex and non-convex sets.

This definition of a convex set can be extended to sets in $\mathbb{R}^n$. Let x and y be any two points in $\mathbb{R}^n$. The closed line segment between x and y is the set

$$[x, y] = \{z : \text{there exists } \lambda \in [0, 1] \text{ such that } z = \lambda x + (1 - \lambda)y\} \quad (1)$$

whose members are the **convex combinations** $z = \lambda x + (1 - \lambda)y$, with $0 \leq \lambda \leq 1$, of the two points x and y . See Fig. 2. If $\lambda = 0$, then $z = y$. Moreover, $\lambda = 1$ gives $z = x$, and $\lambda = 1/2$ gives $z = \frac{1}{2}x + \frac{1}{2}y$, the midpoint between x and y . Note that if we let λ run through *all* real values, then z describes the whole of the straight line L through x and y . This line passes through y and has the direction determined by $x - y$. (See Fig. 3 and (1.1.42).)

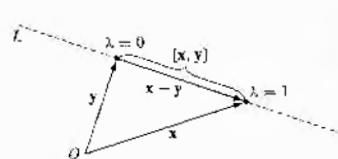


Figure 2 The closed segment $[x, y]$.

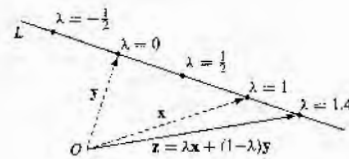


Figure 3 The straight line through x and y .

The definition of a convex set in $\mathbb{R}^n$ is now easy to formulate:

DEFINITION OF A CONVEX SET

A set S in $\mathbb{R}^n$ is called **convex** if $[x, y] \subseteq S$ for all x, y in S , or, equivalently, if

$$\lambda x + (1 - \lambda)y \in S \text{ for all } x, y \text{ in } S \text{ and all } \lambda \text{ in } [0, 1] \quad (2)$$

Note in particular that the empty set and also any set consisting of one single point are convex. Intuitively speaking, a convex set must be "connected" without any "holes", and its boundary must not "bend inwards" at any point.

EXAMPLE 1 The set H of all points $x = (x_1, x_2, \dots, x_n)$ in $\mathbb{R}^n$ that satisfy

$$p \cdot x = p_1 x_1 + p_2 x_2 + \dots + p_n x_n = m \quad (*)$$

where $p \neq 0$, is a hyperplane in $\mathbb{R}^n$. (See (1.1.43).) The hyperplane H divides $\mathbb{R}^n$ into two convex sets,

$$H_+ = \{x \in \mathbb{R}^n : p \cdot x \geq m\} \quad \text{and} \quad H_- = \{x \in \mathbb{R}^n : p \cdot x \leq m\}$$

These two sets are called *half spaces*. To show that H_+ is convex, take two arbitrary points x^1 and x^2 in H_+ . Then $p \cdot x^1 \geq m$ and $p \cdot x^2 \geq m$. For each λ in $[0, 1]$, we have to show that $\lambda x^1 + (1 - \lambda)x^2 \in H_+$, i.e. that $p \cdot (\lambda x^1 + (1 - \lambda)x^2) \geq m$. It follows easily from the rules for the inner product, (1.1.36), that

$$\begin{aligned} p \cdot (\lambda x^1 + (1 - \lambda)x^2) &= p \cdot \lambda x^1 + p \cdot (1 - \lambda)x^2 \\ &= \lambda p \cdot x^1 + (1 - \lambda)p \cdot x^2 \geq \lambda m + (1 - \lambda)m = m \end{aligned}$$

(Where did we use the assumption that $\lambda \in [0, 1]$?) Convexity of H_- is shown in the same way, and it is equally easy to show that the hyperplane H itself is convex. (Convexity of H also follows from (3) below, since $H = H_+ \cap H_-$.)

If S and T are two convex sets in $\mathbb{R}^n$, then their intersection $S \cap T$ is also convex (see Fig. 4). More generally:

$$S_1, \dots, S_m \text{ are convex sets in } \mathbb{R}^n \implies S_1 \cap \dots \cap S_m \text{ is convex} \quad (3)$$

Proof: (One of the world's simplest.) Suppose that x and y both lie in $S = S_1 \cap \dots \cap S_m$. Then x and y both lie in S_i for each $i = 1, \dots, m$. Because S_i is convex, the line segment $[x, y]$ must lie in S_i for each $i = 1, \dots, m$ and hence in the intersection $S_1 \cap \dots \cap S_m = S$. This means that S is convex. ■

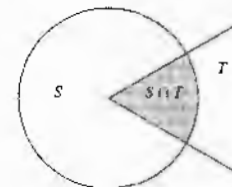


Figure 4 $S \cap T$ is convex, but $S \cup T$ is not.

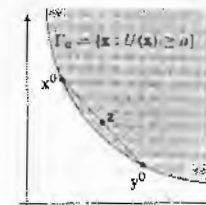


Figure 5 T_x is a convex set.

The union of convex sets is usually not convex. See Fig. 4 again, as well as Fig. 1(d).

EXAMPLE 2 Let $U(\mathbf{x}) = U(x_1, \dots, x_n)$ denote a consumer's utility function. If $U(\mathbf{x}^0) = a$, then the **upper level set** or **upper contour set** $\Gamma_a = \{\mathbf{x} : U(\mathbf{x}) \geq a\}$ consists of all commodity vectors $\mathbf{x}$ that the consumer weakly prefers to $\mathbf{x}^0$. In consumer theory, Γ_a is often assumed to be a convex set for every a . (The function U is then called *quasiconcave*.) Figure 5 shows a typical upper level set for the case of two goods.

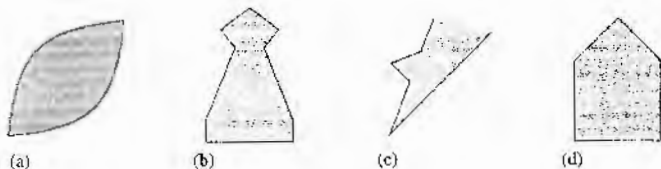
Let $\mathbf{x} = (x_1, \dots, x_n)$ represent a commodity vector and $\mathbf{p} = (p_1, \dots, p_n)$ the corresponding price vector. Then $\mathbf{p} \cdot \mathbf{x} = p_1x_1 + \dots + p_nx_n$ is the cost of buying $\mathbf{x}$. A consumer with m dollars to spend on the commodities has a **budget set** $\mathcal{B}(\mathbf{p}, m)$ defined by the inequalities

$$\mathbf{p} \cdot \mathbf{x} = p_1x_1 + \dots + p_nx_n \leq m \quad \text{and} \quad x_1 \geq 0, \dots, x_n \geq 0 \quad (4)$$

The budget set $\mathcal{B}(\mathbf{p}, m)$ consists of all commodity vectors that the consumer can afford. Let $\mathbb{R}_+^n$ denote the set of all $\mathbf{x}$ for which $x_1 \geq 0, \dots, x_n \geq 0$. Then $\mathcal{B}(\mathbf{p}, m) = H_- \cap \mathbb{R}_+^n$, where H_- is the convex half space introduced in Example 1. It is easy to see that $\mathbb{R}_+^n$ is a convex set. (If $\mathbf{x} \geq \mathbf{0}$ and $\mathbf{y} \geq \mathbf{0}$ and $\lambda \in [0, 1]$, then evidently $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \geq \mathbf{0}$.) Hence $\mathcal{B}(\mathbf{p}, m)$ is convex according to (3). Note that this means that if the consumer can afford either of the commodity vectors $\mathbf{x}$ and $\mathbf{y}$, she can also afford any convex combination of these two vectors. $\blacksquare$

PROBLEMS FOR SECTION 2.2

1. Determine which of the following four sets are convex:



2. Determine which of the following sets are convex by drawing each in the xy -plane.

- (a) $\{(x, y) : x^2 + y^2 < 2\}$ (b) $\{(x, y) : x \geq 0, y \geq 0\}$
 (c) $\{(x, y) : x^2 + y^2 > 8\}$ (d) $\{(x, y) : x \geq 0, y \geq 0, xy \geq 1\}$
 (e) $\{(x, y) : xy \leq 1\}$ (f) $\{(x, y) : \sqrt{x} + \sqrt{y} \leq 2\}$

3. Let S be the set of all points $(x_1, \dots, x_n)$ in $\mathbb{R}^n$ that satisfy all the m inequalities

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &\leq b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &\leq b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &\leq b_m \end{aligned}$$

and moreover are such that $x_1 \geq 0, \dots, x_n \geq 0$. Show that S is a convex set.

4. If S and T are two sets in $\mathbb{R}^n$ and a and b are scalars, let $W = aS + bT$ denote the set of all points of the form $a\mathbf{x} + b\mathbf{y}$, where $\mathbf{x} \in S$ and $\mathbf{y} \in T$. (Then W is called a **linear combination** of the two sets.) Prove that if S and T are both convex, then so is $W = aS + bT$.
5. If S and T are any two sets, the **Cartesian product** $S \times T$ of S and T is defined by $S \times T = \{(\mathbf{x}, \mathbf{t}) : \mathbf{x} \in S, \mathbf{t} \in T\}$, as illustrated in the figure for the case when S and T are intervals of the real line. Prove that if S and T are convex sets in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, then $S \times T$ is also convex (in $\mathbb{R}^{n+m}$).



Figure for Problem 5

6. (a) Let $S = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| \leq r\}$ be the closed n -dimensional ball centred at the origin and with radius $r > 0$. Prove that S is convex.
 (b) If we replace $\leq$ with $<$, $=$, or $\geq$ in the definition of S , we get three new sets S_1 , S_2 , and S_3 . Which of them are convex?

HARDER PROBLEMS

7. (a) Let S be a set of real numbers with the property that if $x_1, x_2 \in S$, then the midpoint $\frac{1}{2}(x_1 + x_2)$ also belongs to S . Show by an example that S is not necessarily convex.
 (b) Does it make any difference if S is closed?
8. Show that if a convex subset S of $\mathbb{R}$ contains more than one point, it must be an interval. (Hint: Show first that if S is bounded, then S must be an interval with endpoints $\inf S$ and $\sup S$. See Section A.2.)

2.3 Concave and Convex Functions I

Recall that a C^2 function of one variable $y = f(x)$ is called **concave** on the interval I if $f''(x) \leq 0$ for all x in I —the graph then turns its hollow side downwards: $\cap$.

We need a definition that is valid more generally, preferably for functions of n variables that may not even be differentiable. Here is our first geometric attempt.

The function f is called **concave (convex)** if it is defined on a convex set and the line segment joining any two points on the graph is never above (below) the graph.

This definition is difficult to use directly. After all, for a function that is specified by a complicated formula, it is far from evident whether the condition is satisfied or not. For concave functions of two variables the definition is illustrated in Fig. 1.

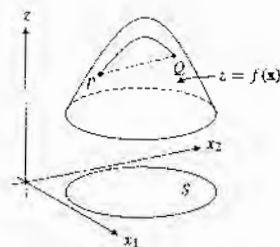


Figure 1 f is concave; for all points P and Q on the graph of f , the line segment PQ lies below the graph.

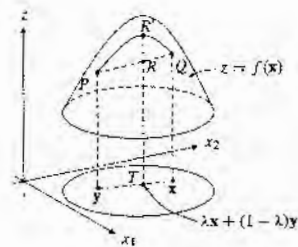


Figure 2 $TR' = f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$.

The two points P and Q in Fig. 2 correspond to points y and x in the domain of f such that $P = (y, f(y))$ and $Q = (x, f(x))$. An arbitrary point R on the line segment PQ has the coordinates $\lambda Q + (1 - \lambda)P = (\lambda x + (1 - \lambda)y, \lambda f(x) + (1 - \lambda)f(y))$ for a suitable λ in $[0, 1]$. This point lies directly above the point $\lambda x + (1 - \lambda)y$ on the line segment between x and y in the " x -plane". The corresponding point on the graph of f can be expressed as $R' = (\lambda x + (1 - \lambda)y, f(\lambda x + (1 - \lambda)y))$. The fact that R does not lie above R' can be expressed by the following inequality:

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y)$$

This motivates the following algebraic definition:

DEFINITION OF A CONCAVE/CONVEX FUNCTION

A function $f(x) = f(x_1, \dots, x_n)$ defined on a convex set S is **concave** on S if

$$f(\lambda x + (1 - \lambda)y) \geq \lambda f(x) + (1 - \lambda)f(y) \quad (1)$$

for all x and y in S and all λ in $[0, 1]$.

A function $f(x)$ is **convex** if (1) holds with $\geq$ replaced by $\leq$.

Note that (1) holds with equality for $\lambda = 0$ and $\lambda = 1$. If we have *strict* inequality in (1) whenever $x \neq y$ and $\lambda \in (0, 1)$, then f is **strictly concave**. The function whose graph is drawn in Fig. 1 is, therefore, strictly concave.

Note that a function f is **convex** on S if and only if $-f$ is concave. Furthermore, f is **strictly convex** if and only if $-f$ is strictly concave.

It is usually impractical to apply the definition directly to show that a function is concave or convex in a certain set. We shall later develop a number of theorems that often help us to decide with ease whether a function is concave or convex. Even so, here is one example where we use the definition directly.

EXAMPLE 1 Use definition (1) to show that the function f defined for all (x_1, x_2) by

$$f(x_1, x_2) = 1 - x_1^2 \quad (\text{so } x_2 \text{ does not appear in the formula for } f)$$

is concave. Is it strictly concave?

Solution: Let (x_1, x_2) and (y_1, y_2) be arbitrary points in the plane. We must show that for all λ in $[0, 1]$,

$$f(\lambda x_1 + (1 - \lambda)y_1, \lambda x_2 + (1 - \lambda)y_2) \geq \lambda f(x_1, x_2) + (1 - \lambda)f(y_1, y_2) \quad (i)$$

Using the definition of f , we see that (i) is equivalent to $1 - [\lambda x_1 + (1 - \lambda)y_1]^2 \geq \lambda(1 - x_1^2) + (1 - \lambda)(1 - y_1^2)$. Expanding and collecting all terms on the left-hand side yields

$$\lambda(1 - \lambda)[x_1^2 - 2x_1y_1 + y_1^2] = \lambda(1 - \lambda)(x_1 - y_1)^2 \geq 0 \quad (ii)$$

This inequality is obviously satisfied for all λ in $[0, 1]$. Thus $f(x, y)$ is concave.

When $x_1 = y_1$, we have equality in (ii), and thus equality in (i) for all values of x_2 and y_2 , even when $x_2 \neq y_2$. So f cannot be strictly concave.

NOTE 1 The one-variable function $g(x_1) = 1 - x_1^2$ is concave. Example 1 showed that it is also concave considered as a function of two variables, x_1 and x_2 . In general, it follows directly from the definitions that if $g(x_1, \dots, x_p)$ is concave (convex) in $(x_1, \dots, x_p)$, then for $n > p$, $f(x_1, \dots, x_p, x_{p+1}, \dots, x_n) = g(x_1, \dots, x_p)$ is concave (convex) in $(x_1, \dots, x_n)$.

Figure 3 shows a portion of the graph of a function of the form $f(x_1, x_2) = h(x_1)$. Here h is concave, and therefore so is f . Through each point on the graph there is a straight line parallel to the x_2 -axis that lies in the graph. This shows that f cannot be strictly concave, even though h is.

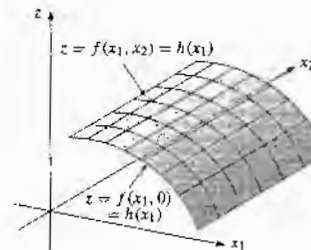


Figure 3 h is strictly concave; f is concave, but not strictly concave.

EXAMPLE 2 The linear (affine) function $f(x) = a \cdot x + b = a_1x_1 + \dots + a_nx_n + b$, where $u = (a_1, \dots, a_n)$ and b are constants, is both concave and convex. This follows immediately from the definition. The graph is a hyperplane in $\mathbb{R}^n$. All points on the line segment between two points in the hyperplane lie in the hyperplane.

Concavity/Convexity for C^2 Functions

Checking the sign of the second derivative is often a quick way to decide where a C^2 function of one variable is concave or convex. For functions of two variables there is a corresponding test which is often used (it is a special case of Theorems 2.3.2 and 2.3.3 below).

THEOREM 2.3.1

Let $z = f(x, y)$ be a C^2 function defined on an open convex set S in the plane. Then (all inequalities must hold throughout S):

- (a) f is convex $\iff f''_{11} \geq 0$, $f''_{22} \geq 0$, and $f''_{11}f''_{22} - (f''_{12})^2 \geq 0$.
- (b) f is concave $\iff f''_{11} \leq 0$, $f''_{22} \leq 0$, and $f''_{11}f''_{22} - (f''_{12})^2 \geq 0$.
- (c) $f''_{11} > 0$ and $f''_{11}f''_{22} - (f''_{12})^2 > 0 \implies f$ is strictly convex.
- (d) $f''_{11} < 0$ and $f''_{11}f''_{22} - (f''_{12})^2 > 0 \implies f$ is strictly concave.

NOTE 2 The implications in parts (c) and (d) cannot be reversed. For example, $f(x, y) = x^4 + y^4$ is strictly convex in the whole plane, even though $f''_{11}(0, 0) = 0$. (See Problem 2.4.6.)

NOTE 3 From the two inequalities specified in part (c), it follows that $f''_{22}(x, y) > 0$ as well. (In fact, the second inequality implies $f''_{11}f''_{22} > (f''_{12})^2 \geq 0$. Thus if $f''_{11} > 0$, then $f''_{22} > 0$ as well.) In a similar way, the two inequalities in part (d) imply that $f''_{22} < 0$.

EXAMPLE 3 Let $f(x, y) = 2x - y - x^2 + 2xy - y^2$ for all (x, y) . Is f concave/convex?

Solution: $f''_{11} = -2$, $f''_{12} = f''_{21} = 2$, and $f''_{22} = -2$. Hence $f''_{11}f''_{22} - (f''_{12})^2 = 0$. Thus the conditions in Theorem 2.3.1(b) are satisfied, so f is concave for all (x, y) . $\blacksquare$

EXAMPLE 4 Find the largest domain S on which $f(x, y) = x^2 - y^2 - xy - x^3$ is concave.

Solution: $f''_{11} = 2 - 6x$, $f''_{12} = f''_{21} = -1$, and $f''_{22} = -2$. Hence $f''_{11} \leq 0$ iff $x \geq 1/3$. Moreover, $f''_{11}f''_{22} - (f''_{12})^2 = 12x - 5 \geq 0$ iff $x \geq 5/12$. Since $5/12 > 1/3$, we conclude that the set S consists of all (x, y) where $x \geq 5/12$. $\blacksquare$

EXAMPLE 5 Check the concavity/convexity of the Cobb–Douglas function

$$f(x, y) = x^a y^b$$

defined on the set $S = \{(x, y) : x > 0, y > 0\}$, assuming that $a + b \leq 1$, $a \geq 0$, and $b \geq 0$.

Solution: $f''_{11} = a(a-1)x^{a-2}y^b$, $f''_{12} = abx^{a-1}y^{b-1}$, and $f''_{22} = b(b-1)x^a y^{b-2}$. Since a and b belong to $[0, 1]$, one has $f''_{11} \leq 0$ and $f''_{22} \leq 0$. Moreover, $f''_{11}f''_{22} - (f''_{12})^2 = abx^{2a-2}y^{2b-2}(1-a-b) \geq 0$ in S . Thus the conditions in Theorem 2.3.1(b) are satisfied and $f(x, y)$ is concave in S . If a and b are positive and $a + b < 1$, then $f''_{11} < 0$ and $f''_{11}f''_{22} - (f''_{12})^2 > 0$, so f is strictly concave according to Theorem 2.3.1(d). $\blacksquare$

The results in Theorem 2.3.1 on concavity/convexity of functions of two variables can be generalized to functions of n variables. (The proofs are given at the end of this section.)

Suppose that $z = f(\mathbf{x}) = f(x_1, \dots, x_n)$ is a C^2 function in an open convex set S in $\mathbb{R}^n$. Then the symmetric matrix

$$\mathbf{f}''(\mathbf{x}) = (f''_{ij}(\mathbf{x}))_{n \times n} \quad (2)$$

is called the **Hessian (matrix)** of f at $\mathbf{x}$, and the n determinants

$$D_r(\mathbf{x}) = \begin{vmatrix} f''_{11}(\mathbf{x}) & f''_{12}(\mathbf{x}) & \cdots & f''_{1r}(\mathbf{x}) \\ f''_{21}(\mathbf{x}) & f''_{22}(\mathbf{x}) & \cdots & f''_{2r}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f''_{r1}(\mathbf{x}) & f''_{r2}(\mathbf{x}) & \cdots & f''_{rr}(\mathbf{x}) \end{vmatrix}, \quad r = 1, 2, \dots, n \quad (3)$$

are the **leading principal minors** of $\mathbf{f}''(\mathbf{x})$ —see Section 1.7.

Theorem 2.3.1(c) and (d) can then be generalized as follows:

THEOREM 2.3.2 (STRICT CONVEXITY/CONCAVITY SUFFICIENT CONDITIONS)

Suppose that $f(\mathbf{x}) = f(x_1, \dots, x_n)$ is a C^2 function defined on an open, convex set S in $\mathbb{R}^n$. Let $D_r(\mathbf{x})$ be defined by (3). Then:

- (a) $D_r(\mathbf{x}) > 0$ for all $\mathbf{x}$ in S and all $r = 1, \dots, n \implies f$ is strictly convex on S .
- (b) $(-1)^r D_r(\mathbf{x}) > 0$ for all $\mathbf{x}$ in S and all $r = 1, \dots, n \implies f$ is strictly concave on S .

When $n = 2$ and $\mathbf{x} = (x, y)$, then $D_1(x, y) = f''_{11}(x, y)$. Since $f''_{12}(x, y) = f''_{21}(x, y)$, we have $D_2(x, y) = f''_{11}(x, y)f''_{22}(x, y) - (f''_{12}(x, y))^2$. Hence, the conditions in Theorem 2.3.2(a) and (b) reduce to those in Theorem 2.3.1(c) and (d), respectively.

EXAMPLE 6 Prove that the function $f(x_1, x_2, x_3) = 100 - 2x_1^2 - x_2^2 - 3x_3 - x_1x_2 - e^{x_1+x_2+x_3}$, defined for all x_1, x_2 , and x_3 , is strictly concave.

Solution: The Hessian matrix of f is $\mathbf{f}''(x_1, x_2, x_3) = \begin{pmatrix} -4 - e^u & -1 - e^u & -e^u \\ -1 - e^u & -2 - e^u & -e^u \\ -e^u & -e^u & -e^u \end{pmatrix}$, where $u = x_1 + x_2 + x_3$. The three leading principal minors are $D_1 = f''_{11} = -4 - e^u$ and

$$D_2 = \begin{vmatrix} -4 - e^u & -1 - e^u \\ -1 - e^u & -2 - e^u \end{vmatrix} = 7 + 4e^u, \quad D_3 = \begin{vmatrix} -4 - e^u & -1 - e^u & -e^u \\ -1 - e^u & -2 - e^u & -e^u \\ -e^u & -e^u & -e^u \end{vmatrix} = -7e^u$$

Thus $D_1 < 0$, $D_2 > 0$, and $D_3 < 0$. By Theorem 2.3.2(b), f is strictly concave. $\blacksquare$

Theorem 2.3.1(a) and (b) can also be generalized to n variables. To do so, however, we must consider the signs of *all* the principal minors of the Hessian matrix $\mathbf{f}''(\mathbf{x}) = (f''_{ij}(\mathbf{x}))_{n \times n}$. Recall from Section 1.7 that a principal minor $\Delta_r(\mathbf{x})$ of order r in $\mathbf{f}''(\mathbf{x}) = (f''_{ij}(\mathbf{x}))_{n \times n}$ is obtained by deleting $n - r$ rows and the $n - r$ columns with the same numbers.

THEOREM 2.3.3 (CONVEXITY/CONCAVITY: NECESSARY/SUFFICIENT CONDITIONS)

Suppose that $f(\mathbf{x}) = f(x_1, \dots, x_n)$ is a C^2 function defined on an open, convex set S in $\mathbb{R}^n$. Let $\Delta_r(\mathbf{x})$ denote an arbitrary principal minor of order r in the Hessian matrix. Then:

- (a) f is convex in $S \iff \Delta_r(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in S and all $\Delta_r(\mathbf{x}), r = 1, \dots, n$.
 (b) f is concave in $S \iff (-1)^r \Delta_r(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in S and all $\Delta_r(\mathbf{x}), r = 1, \dots, n$.

NOTE 4 If $f(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$ is a symmetric quadratic form, then by differentiating the double sum (1.7.6) we get $\mathbf{f}''(\mathbf{x}) = 2\mathbf{A}$. For any fixed open convex set S in $\mathbb{R}^n$, if we combine the statements in Theorems 2.3.2 and 2.3.3 with Theorem 1.7.1, we get the following results:

The Hessian matrix $\mathbf{f}''(\mathbf{x})$ is positive definite in $S \implies f(\mathbf{x})$ is strictly convex in S . (4)

The Hessian matrix $\mathbf{f}''(\mathbf{x})$ is negative definite in $S \implies f(\mathbf{x})$ is strictly concave in S . (5)

$f(\mathbf{x})$ is convex in $S \iff$ the Hessian matrix $\mathbf{f}''(\mathbf{x})$ is positive semidefinite in S . (6)

$f(\mathbf{x})$ is concave in $S \iff$ the Hessian matrix $\mathbf{f}''(\mathbf{x})$ is negative semidefinite in S . (7)

EXAMPLE 7 Let $f(x_1, x_2, x_3) = -x_1^2 + 6x_1x_2 - 9x_2^2 - 2x_3^2$. Prove that f is concave.

Solution: The Hessian matrix is

$$\mathbf{f}''(x_1, x_2, x_3) = \begin{pmatrix} -2 & 6 & 0 \\ 6 & -18 & 0 \\ 0 & 0 & -4 \end{pmatrix} = 2 \begin{pmatrix} -1 & 3 & 0 \\ 3 & -9 & 0 \\ 0 & 0 & -2 \end{pmatrix} \quad (*)$$

In Example 1.7.5(b) we saw that the last matrix in (*) is negative semidefinite. But then so is $\mathbf{f}''$. We conclude from (7) that f is concave. ■

Useful Results

Using Theorems 2.3.2 and 2.3.3 to decide convexity/concavity can be quite hard, although easier than relying directly on the definitions. The following two theorems can sometimes ease the task of establishing concavity/convexity.

If f and g are C^2 functions of one variable, they are concave if and only if $f''(x) \leq 0$ and $g''(x) \leq 0$. Suppose that a and b are ≥ 0 . Then $G(x) = af(x) + bg(x)$ has $G''(x) = af''(x) + bg''(x) \leq 0$, so G is also concave. If f and g are C^2 functions of n variables,

the same result holds, but a proof based on Theorem 2.3.3 would be quite messy. You should therefore appreciate the extreme simplicity of the proof of the following much more general result:

THEOREM 2.3.4

If $f_1, \dots, f_m$ are functions defined on a convex set S in $\mathbb{R}^n$, then:

- (a) $f_1, \dots, f_m$ concave and $a_1 \geq 0, \dots, a_m \geq 0 \implies a_1f_1 + \dots + a_mf_m$ concave.
 (b) $f_1, \dots, f_m$ convex and $a_1 \geq 0, \dots, a_m \geq 0 \implies a_1f_1 + \dots + a_mf_m$ convex.

Proof: We prove (a). The proof of (b) is similar. Put $G(\mathbf{x}) = a_1f_1(\mathbf{x}) + \dots + a_mf_m(\mathbf{x})$. For λ in $[0, 1]$ and $\mathbf{x}, \mathbf{y}$ in S , the definition (1) of a concave function implies that

$$\begin{aligned} G(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= a_1f_1(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) + \dots + a_mf_m(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \\ &\geq a_1[\lambda f_1(\mathbf{x}) + (1-\lambda)f_1(\mathbf{y})] + \dots + a_m[\lambda f_m(\mathbf{x}) + (1-\lambda)f_m(\mathbf{y})] \\ &= \lambda[a_1f_1(\mathbf{x}) + \dots + a_mf_m(\mathbf{x})] + (1-\lambda)[a_1f_1(\mathbf{y}) + \dots + a_mf_m(\mathbf{y})] \\ &= \lambda G(\mathbf{x}) + (1-\lambda)G(\mathbf{y}) \end{aligned}$$

Using definition (1) again, this shows that G is concave. ■

The composition of two concave functions is not necessarily concave. If, for example, $f(x) = -x^2$ and $F(u) = -e^u$, then f and F are both (strictly) concave, but the composite function $F(f(x)) = -e^{-x^2}$ is actually convex in an interval about the origin. But if we also require the exterior function to be *increasing*, then the composite function is concave. In general we have the following important result:

THEOREM 2.3.5

Suppose that $f(\mathbf{x})$ is defined for all $\mathbf{x}$ in a convex set S in $\mathbb{R}^n$ and that F is defined over an interval in $\mathbb{R}$ that contains $f(\mathbf{x})$ for all $\mathbf{x}$ in S . Then:

- (a) $f(\mathbf{x})$ concave, $F(u)$ concave and increasing $\implies U(\mathbf{x}) = F(f(\mathbf{x}))$ concave.
 (b) $f(\mathbf{x})$ convex, $F(u)$ convex and increasing $\implies U(\mathbf{x}) = F(f(\mathbf{x}))$ convex.
 (c) $f(\mathbf{x})$ concave, $F(u)$ convex and decreasing $\implies U(\mathbf{x}) = F(f(\mathbf{x}))$ convex.
 (d) $f(\mathbf{x})$ convex, $F(u)$ concave and decreasing $\implies U(\mathbf{x}) = F(f(\mathbf{x}))$ concave.

Proof: (a) Let $\mathbf{x}, \mathbf{y} \in S$ and let $\lambda \in [0, 1]$. Then

$$\begin{aligned} U(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= F(f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y})) \geq F(\lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})) \\ &\geq \lambda F(f(\mathbf{x})) + (1-\lambda)F(f(\mathbf{y})) = \lambda U(\mathbf{x}) + (1-\lambda)U(\mathbf{y}) \end{aligned}$$

The first inequality uses the concavity of f and the fact that F is increasing. The second inequality is due to the concavity of F .

The statement in (b) is shown in the same way. We then obtain (c) and (d) from (a) and (b) by replacing F with $-F$. ■

NOTE 5 Suppose the functions f and F in (a) are C^2 functions of one variable, and that $U(x) = F(f(x))$. Then $U'(x) = F'(f(x))f'(x)$ and $U''(x) = F''(f(x))(f'(x))^2 + F'(f(x))f''(x)$. Because F is concave and increasing and f is concave, $F'' \leq 0$, $F' \geq 0$, and $f'' \leq 0$. It follows immediately that $U''(x) \leq 0$. This “calculus” proof of (a) is valid for C^2 functions of one variable.

If f is a C^2 function of n variables, it is possible to prove (a) partly relying on Theorem 2.3.3(b), but such a proof would be rather complicated. After attempting it, however, you might come to appreciate better the easy proof above, which needs no differentiability assumptions at all.

If we drop the requirement in (a) that F be concave, then $U(x)$ is not necessarily concave. For example, if $f(x) = \sqrt{x}$ and $F(u) = u^3$, then f is concave for $x \geq 0$ and F is increasing, but $U(x) = F(f(x)) = x^{3/2}$ is convex rather than concave.

NOTE 6 The results in Theorem 2.3.5 can easily be generalized to the case $U(x) = F(f_1(x), \dots, f_m(x))$, where the functions $f_1(x), \dots, f_m(x)$ are all concave (convex) and $F(u_1, \dots, u_m)$ is concave (convex) and moreover increasing in each variable.

If $f(x) = a \cdot x + b$, so that f is an affine function of x , then in (a) and (b), the assumption that F is increasing can be dropped. Indeed, in the proof of (a), the first inequality used the concavity of f and the fact that F is increasing. When f is affine, this inequality becomes an equality, and the rest of the argument is as before. Thus:

A concave (convex) function of an affine function is concave (convex). (8)

EXAMPLE 8 Examine the concavity/convexity of the following functions:

(a) $f(x, y, z) = ax^2 + by^2 + cz^2$ (a, b , and c are nonnegative)

(b) $g(x, y, z) = e^{ax^2 + by^2 + cz^2}$ (a, b , and c are nonnegative)

(c) $h(x_1, \dots, x_n) = (a_1x_1 + \dots + a_nx_n)^2$

Solution: The function f is convex as a sum of convex functions. The function g is also convex. In fact, $g(x, y, z) = e^u$, with $u = f(x, y, z) = ax^2 + by^2 + cz^2$. Here the transformation $u \mapsto e^u$ is convex and increasing, and u is convex, so by Theorem 2.3.5(b), g is convex. Finally, h is a convex function ($u \mapsto u^2$) of a linear function, and thus convex according to (8).

EXAMPLE 9 Define the function G on $S = \{(x, y) : x^2 + y^2 < a^2\}$ by

$$G(x, y) = Ax + By + \ln[a^2 - (x^2 + y^2)]$$

Show that G is concave in S . (A, B , and a are constants. The domain of G is S , because $\ln u$ is defined only when $u > 0$.)

Solution: The function $g(x, y) = Ax + By$ is linear and hence concave by Example 2. Define $h(x, y) = \ln[a^2 - (x^2 + y^2)]$. If we put $f(x, y) = a^2 - x^2 - y^2$ and $F(u) = \ln u$, then $h(x, y) = F(f(x, y))$. As a sum of concave functions, $f(x, y)$ is concave. Moreover, $F'(u) = 1/u$ and $F''(u) = -1/u^2$, so F is increasing and concave. According to Theorem 2.3.5(a), the function $h(x, y)$ is concave, so $G(x, y) = g(x, y) + h(x, y)$ is concave as a sum of concave functions.

We end this section by proving Theorems 2.3.3 and 2.3.2.

Proof of Theorem 2.3.3: Let us first show the implication $\Leftarrow$ in part (a). Take two points x, y in S and let $t \in [0, 1]$. Define $g(t) = f(y + t(x - y)) = f(tx + (1 - t)y)$. Then by using formula (2.1.7), $g'(t) = \sum_{i=1}^n f'_i(y + t(x - y))(x_i - y_i)$. Using the chain rule again, we get (for more details see (2.6.6)):

$$g''(t) = \sum_{i=1}^n \sum_{j=1}^n f''_{ij}(y + t(x - y))(x_i - y_i)(x_j - y_j) \quad (i)$$

By the assumption in (a) that $\Delta_i(y) \geq 0$ for all y in S and all $i = 1, \dots, n$, Theorem 1.7.1(b) implies that the quadratic form in (i) is ≥ 0 for t in $[0, 1]$. This shows that g is convex. In particular,

$$g(t) = g(t \cdot 1 + (1 - t) \cdot 0) \leq tg(1) + (1 - t)g(0) = tf(x) + (1 - t)f(y) \quad (ii)$$

But this shows that f is convex, since the inequality in (i) is satisfied with $\leq$.

To prove that $\Rightarrow$ is valid in case (a), suppose f is convex in S . According to Theorem 1.7.1(b), it suffices to show that for all x in S and all $h_1, \dots, h_n$ we have

$$Q = \sum_{i=1}^n \sum_{j=1}^n f''_{ij}(x)h_ih_j \geq 0 \quad (iii)$$

Now S is an open set, so if $x \in S$ and $h = (h_1, \dots, h_n)$ is an arbitrary vector, there exists a positive number a such that $x + th \in S$ for all t with $|t| < a$. Let $I = (-a, a)$. Define the function p on I by $p(t) = f(x + th)$. According to (8), p is convex in I . Hence $p''(t) \geq 0$ for all t in I . But

$$p''(t) = \sum_{i=1}^n \sum_{j=1}^n f''_{ij}(x + th)h_ih_j \quad (iv)$$

Putting $t = 0$, we get inequality (iii).

This proves the equivalence in part (a) of the theorem. The equivalence in (b) follows from (a) if we simply replace f with $-f$.

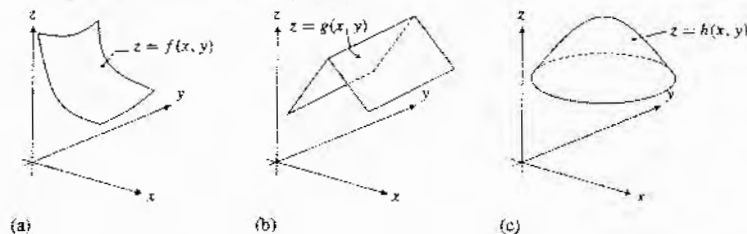
Proof of Theorem 2.3.2: Define g as in the proof of Theorem 2.3.3.

(a) If the specified conditions are satisfied, the Hessian matrix $F''(x)$ is positive definite according to Theorem 1.7.1(a). So for $x \neq y$ the sum in (i) is > 0 for all t in $[0, 1]$. It follows that g is strictly convex. The inequality in (ii) of the proof above is then strict for t in $[0, 1]$, so f is strictly convex.

(b) Follows from (a) by replacing f with $-f$.

PROBLEMS FOR SECTION 2.3

1. Which of the functions whose graphs are shown in the figure below are (presumably) convex/concave, strictly convex/strictly concave?



2. (a) Let f be defined for all x, y by $f(x, y) = x - y - x^2$. Show that f is concave (i) by using Theorem 2.3.1, (ii) by using Theorem 2.3.4.
 (b) Show that $-e^{-f(x, y)}$ is concave.
3. (a) Show that $f(x, y) = ax^2 + 2bxy + cy^2 + px + qy + r$ is strictly concave if $ac - b^2 > 0$ and $a < 0$, whereas it is strictly convex if $ac - b^2 > 0$ and $a > 0$.
 (b) Find necessary and sufficient conditions for $f(x, y)$ to be concave/convex.
4. For what values of the constant a is the following function concave/convex?

$$f(x, y) = -6x^2 + (2a + 4)xy - y^2 + 4ay$$

5. Examine the convexity/concavity of the following functions:

$$(a) z = x + y - e^x - e^{x+y} \quad (b) z = e^{x+y} + e^{x-y} - \frac{1}{2}y \quad (c) w = (x + 2y + 3z)^2$$

6. Suppose $y = f(x)$ is a production function determining output y as a function of the vector x of nonnegative factor inputs, with $f(0) = 0$. Show that:

- (a) If f is concave, then $f''_{ii}(x) \leq 0$ (so each marginal product $f'_i(x)$ is decreasing).
 (b) If f is concave, then $f(\lambda x)/\lambda$ is decreasing as a function of λ .
 (c) If f is homogeneous of degree 1 (constant returns to scale), then f is not strictly concave.

7. Let f be defined for all x in $\mathbb{R}^n$ by $f(x) = \|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$. Prove that f is convex. Is f strictly convex? (Hint: Use (1.1.39).)

HARDER PROBLEMS

8. Use Theorem 2.3.1 to show that the CES function f defined for $v_1 > 0, v_2 > 0$ by

$$f(v_1, v_2) = A(\delta_1 v_1^{-\rho} + \delta_2 v_2^{-\rho})^{-1/\rho} \quad (A > 0, \rho \neq 0, \mu > 0, \delta_1, \delta_2 > 0)$$

is concave for $\rho \geq -1$ and convex for $\rho \leq -1$, and that it is strictly concave if $0 < \mu < 1$ and $\rho > -1$. (See also Example 2.5.6.)

9. (a) The Cobb–Douglas function $z = f(x) = x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$ ($a_1 > 0, \dots, a_n > 0$) is defined for all $x_1 > 0, \dots, x_n > 0$. Prove that the k th leading principal minor of the Hessian $f''(x)$ is

$$D_k = \frac{a_1 \cdots a_k}{(x_1 \cdots x_k)^2} z^k \begin{vmatrix} a_1 - 1 & a_1 & \cdots & a_1 \\ a_2 & a_2 - 1 & \cdots & a_2 \\ \vdots & \vdots & \ddots & \vdots \\ a_k & a_k & \cdots & a_k - 1 \end{vmatrix}$$

- (b) Prove that $D_k = (-1)^{k-1} (\sum_{i=1}^k a_i - 1) z^k \frac{a_1 \cdots a_k}{(x_1 \cdots x_k)^2}$. (Hint: Add all the other rows to the first row, extract the common factor $\sum_{i=1}^k a_i - 1$, and then subtract the first column in the new determinant from all the other columns.)
 (c) Prove that the function is strictly concave if $a_1 + \cdots + a_n < 1$.

2.4 Concave and Convex Functions II

We continue our discussion of concave/convex functions. Our first result has a geometric interpretation. Consider Fig. 1, which concerns the case $n = 1$.



Figure 1 f is concave. The tangent is above the graph.

The tangent at any point on the graph will lie above the graph. The following algebraic version of this geometric statement is extremely important in both static and dynamic optimization theory:

THEOREM 2.4.1 / CONCAVITY FOR DIFFERENTIABLE FUNCTIONS

Suppose that $f(x)$ is a C^1 function defined on an open, convex set S in $\mathbb{R}^n$. Then:

(a) f is concave in S if and only if

$$f(x) - f(x^0) \leq \nabla f(x^0) \cdot (x - x^0) = \sum_{i=1}^n \frac{\partial f(x^0)}{\partial x_i} (x_i - x_i^0) \quad (1)$$

for all x and x^0 in S .

(b) f is strictly concave iff the inequality (1) is always strict when $x \neq x^0$.

(c) The corresponding results for convex (strictly convex) functions are obtained by changing $\leq$ to $\geq$ ($<$ to $>$) in the inequality (1).

Proof: (a) Suppose f is concave, and let $x^0, x \in S$. According to definition (2.3.1),

$$f(x) - f(x^0) \leq \frac{f(x^0 + \lambda(x - x^0)) - f(x^0)}{\lambda} \quad (i)$$

for all λ in $(0, 1]$. Let $\lambda \rightarrow 0^+$. The right-hand side of (i) then approaches the right-hand side in (1) (see the argument for formula (2.1.7)). Because the weak inequality is preserved when passing to the limit, we have shown inequality (1).

To prove the reverse implication let $x, x^0 \in S$ and $\lambda \in (0, 1]$. Put $z = \lambda x + (1 - \lambda)x^0$. Then z belongs to S and, according to (1),

$$f(x) - f(z) \leq \nabla f(z) \cdot (x - z), \quad f(x^0) - f(z) \leq \nabla f(z) \cdot (x^0 - z) \quad (ii)$$

where we used the gradient notation from Section 2.1.

Multiply the first inequality in (ii) by $\lambda > 0$ and the second by $1 - \lambda > 0$, and add the resulting inequalities. This gives

$$\lambda(f(\mathbf{x}) - f(\mathbf{z})) + (1 - \lambda)(f(\mathbf{x}^0) - f(\mathbf{z})) \leq \nabla f(\mathbf{z}) \cdot [\lambda(\mathbf{x} - \mathbf{z}) + (1 - \lambda)(\mathbf{x}^0 - \mathbf{z})] \quad (\text{iii})$$

Here $\lambda(\mathbf{x} - \mathbf{z}) + (1 - \lambda)(\mathbf{x}^0 - \mathbf{z}) = \lambda\mathbf{x} + (1 - \lambda)\mathbf{x}^0 - \mathbf{z} = \mathbf{0}$, so the right-hand side of (iii) is 0. Then rearranging (iii) we see that f is concave.

(b) Suppose that f is strictly concave in S . Then inequality (i) is strict for $\mathbf{x} \neq \mathbf{x}^0$. With $\mathbf{z} = \mathbf{x}^0 + \lambda(\mathbf{x} - \mathbf{x}^0)$, we have

$$f(\mathbf{x}) - f(\mathbf{x}^0) < \frac{f(\mathbf{x}) - f(\mathbf{x}^0)}{\lambda} \leq \frac{\nabla f(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0)}{\lambda} = \nabla f(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0)$$

where we used inequality (i), which we have already proved, and the fact that $\mathbf{z} - \mathbf{x}^0 = \lambda(\mathbf{x} - \mathbf{x}^0)$. This shows that (1) holds with strict inequality.

On the other hand, if (1) holds with strict inequality for $\mathbf{x} \neq \mathbf{x}^0$, then (ii) and (iii) hold with $\leq$ replaced by $<$, and thus f is strictly concave. ■

The next theorem lists several interesting properties of concave/convex functions:

THEOREM 2.4.2

Let $f(\mathbf{x}) = f(x_1, \dots, x_n)$ and $g(\mathbf{x}) = g(x_1, \dots, x_n)$ be defined on a convex set S in $\mathbb{R}^n$. Then:

- If f is concave, the set $P_a = \{\mathbf{x} \in S : f(\mathbf{x}) \geq a\}$ is convex for every number a .
- If f is convex, the set $P^a = \{\mathbf{x} \in S : f(\mathbf{x}) \leq a\}$ is convex for every number a .
- f is concave $\iff M_f = \{(\mathbf{x}, y) : \mathbf{x} \in S \text{ and } y \leq f(\mathbf{x})\}$ is convex.
- f is convex $\iff M^f = \{(\mathbf{x}, y) : \mathbf{x} \in S \text{ and } y \geq f(\mathbf{x})\}$ is convex.
- f and g are concave $\implies h(\mathbf{x}) = \min(f(\mathbf{x}), g(\mathbf{x}))$ is concave.
- f and g are convex $\implies H(\mathbf{x}) = \max(f(\mathbf{x}), g(\mathbf{x}))$ is convex.

Proof: (a) Let $\mathbf{x}$ and $\mathbf{y}$ be points in P_a . Then $\mathbf{x}$ and $\mathbf{y}$ belong to S , while $f(\mathbf{x}) \geq a$ and $f(\mathbf{y}) \geq a$. If $\lambda \in [0, 1]$, then $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$ also belongs to S (since S is convex). Because f is concave, $f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}) \geq \lambda a + (1 - \lambda)a = a$. This shows that $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in P_a$, which confirms that P_a is convex. (If P_a is empty, P_a is convex by definition.)

Part (b) is shown in the same way. Parts (c) and (d) follow easily from the definitions.

(e) The function h maps $\mathbf{x}$ to the smaller of the numbers $f(\mathbf{x})$ and $g(\mathbf{x})$. Using the notation from part (c), $M_h = M_f \cap M_g$. The hypothesis that f and g are concave implies that M_f and M_g are convex, by part (c). Since the intersection of convex sets is convex, it follows from (c) that M_h is convex and so h is concave. Part (f) is shown in the same way. ■

The result in part (c) is illustrated in Fig. 2, while Fig. 3 illustrates part (e).

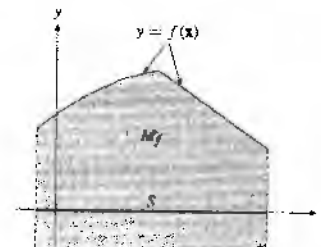


Figure 2 $M_f = \{(\mathbf{x}, y) : \mathbf{x} \in S, y \leq f(\mathbf{x})\}$

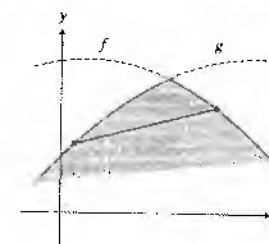


Figure 3 $M_{\min(f,g)} = M_f \cap M_g$

Jensen's Inequality

Suppose we put $\mathbf{x} = \mathbf{x}_1$, $\mathbf{y} = \mathbf{x}_2$, $\lambda_1 = \lambda$ and $\lambda_2 = 1 - \lambda$ in the definition (2.3.1) of a concave function. This leads to the equivalent definition: f is concave on S if and only if

$$f(\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2) \geq \lambda_1 f(\mathbf{x}_1) + \lambda_2 f(\mathbf{x}_2)$$

for all $\mathbf{x}_1$ and $\mathbf{x}_2$ in S and for all $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ with $\lambda_1 + \lambda_2 = 1$.

Jensen's inequality extends to m points this characterization of a concave function:

THEOREM 2.4.3 (JENSEN'S INEQUALITY, DISCRETE VERSION)

A function f is concave on the convex set S in $\mathbb{R}^n$ if and only if

$$f(\lambda_1 \mathbf{x}_1 + \dots + \lambda_m \mathbf{x}_m) \geq \lambda_1 f(\mathbf{x}_1) + \dots + \lambda_m f(\mathbf{x}_m) \quad (2)$$

holds for all $\mathbf{x}_1, \dots, \mathbf{x}_m$ in S and all $\lambda_1 \geq 0, \dots, \lambda_m \geq 0$ with $\lambda_1 + \dots + \lambda_m = 1$.

The result for convex functions is obtained by reversing the inequality sign in (2).

It is obvious that if (2) holds, then f is concave: with $m = 2$ and $\lambda = \lambda_1 = 1 - \lambda_2$, (2) reduces to the inequality in the definition of a concave function. Problem 3 suggests an argument showing that the inequality (2) holds for any concave function.

EXAMPLE 1

(Production smoothing) Consider a manufacturing firm producing a single commodity. The cost of maintaining an output level y per year for a fraction λ of a year is $\lambda C(y)$, where $C'(y) > 0$ and $C''(y) \geq 0$ for all $y \geq 0$. In fact, the firm's output level can fluctuate over the year. Show that, given the total output Y that the firm produces over the whole year, the firm's total cost per year is minimized by choosing a constant flow of output.

Solution: Suppose the firm chooses different output levels $y_1, \dots, y_n$ per year for fractions of the year $\lambda_1, \dots, \lambda_n$, respectively. Then the total output is $\sum_{i=1}^n \lambda_i y_i = Y$, which is produced at total cost $\sum_{i=1}^n \lambda_i C(y_i)$. Applying Jensen's inequality to the convex function C gives the inequality $\sum_{i=1}^n \lambda_i C(y_i) \geq C(\sum_{i=1}^n \lambda_i y_i) = C(Y)$. The right-hand side is the cost of maintaining the constant output level Y over the whole year, and this is the minimum cost. ■

There is also a continuous version of Jensen's inequality that involves integrals. We restrict our attention to functions of one real variable. A proof of the next theorem is indicated in Problem 4.

THEOREM 2.4.4 (JENSEN'S INEQUALITY, CONTINUOUS VERSION)

Let $x(t)$ and $\lambda(t)$ be continuous functions in the interval $[a, b]$, with $\lambda(t) \geq 0$ and $\int_a^b \lambda(t) dt = 1$. If f is a concave function defined on the range of $x(t)$, then

$$f\left(\int_a^b \lambda(t)x(t) dt\right) \geq \int_a^b \lambda(t)f(x(t)) dt \quad (3)$$

NOTE 1 Jensen's inequality is important in statistics. One application is this: If f is concave in an interval I and if X is a random variable with values in I whose expectation $E(X)$ is finite, then $f(E(X)) \geq E(f(X))$.

EXAMPLE 2

(Consumption smoothing in continuous time) Suppose that a consumer expects to live from now (time $t = 0$) until time T . Let $c(t)$ denote consumption expenditure flow at time t , and $y(t)$ the given income flow. Let w_0 be wealth at time 0. Assume that the consumer would like to choose $c(t)$ so as to maximize the lifetime intertemporal utility function

$$\int_0^T e^{-\alpha t} u(c(t)) dt \quad (i)$$

where $\alpha > 0$ is the rate of impatience or of utility discount, and $u(c)$ is a strictly increasing concave utility function (such as $\ln c$ or $-c^{-\gamma}$). Suppose that r is the instantaneous rate of interest on savings, and that the consumer is not allowed to pass time T in debt. The initial wealth together with the present discounted value (PDV) of future income is $w_T = w_0 + \int_0^T e^{-rt} y(t) dt$. The intertemporal budget constraint imposes the requirement that the PDV of consumption cannot exceed w_T :

$$\int_0^T e^{-rt} c(t) dt \leq w_T \quad (\text{for all admissible } c(t)) \quad (ii)$$

Finding an optimal time path of consumption for a problem like this generally involves techniques from optimal control theory. (See Example 9.4.2.) In the special case when $r = \alpha$, however, an optimal time path can easily be found by means of Jensen's inequality. Let $\bar{c}$ be the (constant) level of consumption that satisfies the equation

$$\int_0^T e^{-\alpha t} \bar{c} dt = w_T = w_0 + \int_0^T e^{-\alpha t} y(t) dt \quad (iii)$$

Note how $\bar{c} = \bar{y}$ in the special case when $w_0 = 0$ and $y(t) = \bar{y}$ for all t . Our claim is that an optimal path is to choose $c(t) = \bar{c}$ for all t , which we call "consumption smoothing" because all fluctuations in income are smoothed out through saving and borrowing in a way that leaves consumption constant.

To establish this claim, define the constant $\bar{\alpha} = \int_0^T e^{-\alpha t} dt$. Then (iii) implies $\bar{c} = w_T/\bar{\alpha}$. Now apply Jensen's inequality to the concave function u with weights $\lambda(t) = (1/\bar{\alpha})e^{-\alpha t}$. This yields

$$u\left(\int_0^T (1/\bar{\alpha})e^{-\alpha t} c(t) dt\right) \geq \int_0^T (1/\bar{\alpha})e^{-\alpha t} u(c(t)) dt = (1/\bar{\alpha}) \int_0^T e^{-\alpha t} u(c(t)) dt \quad (iv)$$

Inequalities (iv) and (ii), together with the fact that $\bar{c} = w_T/\bar{\alpha}$ and the definition of $\bar{\alpha}$, imply that

$$\int_0^T e^{-\alpha t} u(c(t)) dt \leq \bar{\alpha} u\left(\frac{1}{\bar{\alpha}} \int_0^T e^{-\alpha t} c(t) dt\right) \leq \bar{\alpha} u\left(\frac{w_T}{\bar{\alpha}}\right) = \bar{\alpha} u(\bar{c}) = \int_0^T e^{-\alpha t} u(\bar{c}) dt \quad (v)$$

This proves that no other consumption plan satisfying budget constraint (ii) can yield a higher value of lifetime utility, given by (i), than does the "smoothed consumption" path with $c(t) = \bar{c}$ for all t . ■

NOTE 2 Consumption smoothing is actually quite an important topic in economics. John R. Hicks defined "income" as the level of consumption that can be sustained without change. Milton Friedman called a similar measure "permanent income", and enunciated the "permanent income hypothesis" according to which a measure of "permanent consumption" equals permanent income.

Supergradients

Even if the concave function f in Theorem 2.4.1 is not necessarily C^1 , we still have:

THEOREM 2.4.5 (EXISTENCE OF A SUPERGRADIENT)

Let f be concave on a convex set $S \subseteq \mathbb{R}^n$, and let $\mathbf{x}^0$ be an interior point in S . Then there exists a vector $\mathbf{p}$ in $\mathbb{R}^n$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^0) \leq \mathbf{p} \cdot (\mathbf{x} - \mathbf{x}^0) \quad \text{for all } \mathbf{x} \text{ in } S \quad (4)$$

A vector $\mathbf{p}$ that satisfies (4) is called a **supergradient** for f at $\mathbf{x}^0$. Symmetrically, if f is convex, there exists a vector $\mathbf{q}$ in $\mathbb{R}^n$ such that (4) is valid with $\leq$ replaced by $\geq$ and $\mathbf{p}$ by $\mathbf{q}$. Such a vector is called a **subgradient** for f .

Proof: Let $M_f = \{(\mathbf{x}, y) : \mathbf{x} \in S \text{ and } y \leq f(\mathbf{x})\}$. According to Theorem 2.4.2(c), M_f is convex. The point $\mathbf{z}^0 = (\mathbf{x}^0, f(\mathbf{x}^0)) \in M_f$ is a boundary point of M_f , since $(\mathbf{x}^0, f(\mathbf{x}^0) + \gamma) \notin M_f$ for all $\gamma > 0$. By the supporting hyperplane theorem (Theorem 13.6.2) there exists a vector $\mathbf{q}^0 = (\mathbf{p}^0, r) \neq (\mathbf{0}, 0)$, with $\mathbf{p}^0 \in \mathbb{R}^n$ and $r \in \mathbb{R}$, such that

$$(\mathbf{p}^0, r) \cdot \mathbf{z} \leq (\mathbf{p}^0, r) \cdot (\mathbf{x}^0, f(\mathbf{x}^0)) \quad \text{for all } \mathbf{z} \text{ in } M_f \quad (*)$$

Given any $v > 0$, one has $\mathbf{z} = (\mathbf{x}^0, f(\mathbf{x}^0) - v) \in M_f$, so $\mathbf{p}^0 \cdot \mathbf{x}^0 + rf(\mathbf{x}^0) - rv \leq \mathbf{p}^0 \cdot \mathbf{x}^0 + rf(\mathbf{x}^0)$ by (*). Hence $-rv \leq 0$ for all $v > 0$. It follows that $r \geq 0$.

We want to prove that $r \neq 0$. To this end, note that for all $\mathbf{x}$ in S , one has $\mathbf{z} = (\mathbf{x}, f(\mathbf{x})) \in M_f$. So (*) implies that $(\mathbf{p}^0, r) \cdot (\mathbf{x}, f(\mathbf{x})) \leq (\mathbf{p}^0, r) \cdot (\mathbf{x}^0, f(\mathbf{x}^0))$. That is,

$$\mathbf{p}^0 \cdot \mathbf{x} + rf(\mathbf{x}) \leq \mathbf{p}^0 \cdot \mathbf{x}^0 + rf(\mathbf{x}^0) \quad \text{for all } \mathbf{x} \text{ in } S \quad (**)$$

Suppose by way of contradiction that $r = 0$. Then $\mathbf{p}^0 \neq \mathbf{0}$. By (**) one has $\mathbf{p}^0 \cdot \mathbf{x} \leq \mathbf{p}^0 \cdot \mathbf{x}^0$ for all $\mathbf{x}$ in S . But $\mathbf{p}^0 \cdot (\mathbf{x}^0 + \varepsilon \mathbf{p}^0) > \mathbf{p}^0 \cdot \mathbf{x}^0$ for all $\varepsilon > 0$, so $\mathbf{x}^0 + \varepsilon \mathbf{p}^0 \notin S$ for all $\varepsilon > 0$. This would imply that $\mathbf{x}^0$ is a boundary point, contradicting the hypothesis that it is an interior point. Hence $r > 0$. If we define $\mathbf{p} = -\mathbf{p}^0/r$, and divide (**) by r , we get inequality (4). ■

NOTE 3 Suppose f is defined on a set $S \subseteq \mathbb{R}^n$, and that $\mathbf{x}^0$ is an interior point in S at which f is differentiable. Then, if $\mathbf{p}$ is a supergradient for f at $\mathbf{x}^0$, i.e. a vector that satisfies (4), one has $\mathbf{p} = \nabla f(\mathbf{x}^0)$. This result follows because $\phi(\mathbf{x}) = f(\mathbf{x}) - f(\mathbf{x}^0) - \mathbf{p} \cdot (\mathbf{x} - \mathbf{x}^0)$ has a maximum at $\mathbf{x}^0$, so $\nabla \phi(\mathbf{x}^0) = \mathbf{0}$. Hence, if a concave function f is differentiable, its gradient is the only supergradient.

PROBLEMS FOR SECTION 2.4

1. Use Theorem 2.4.1 to prove that $f(x, y) = 1 - x^2 - y^2$ defined in $\mathbb{R}^2$ is concave.

2. Apply Jensen's inequality (2) to $f(x) = \ln x$, with $\lambda_1 = \dots = \lambda_n = 1/n$ to prove that

$$\sqrt[n]{x_1 x_2 \dots x_n} \leq \frac{1}{n}(x_1 + x_2 + \dots + x_n) \quad \text{for } x_1 > 0, \dots, x_n > 0$$

(The geometric mean is less than or equal to the arithmetic mean.)

HARDER PROBLEMS

3. Prove Jensen's inequality (2) for $m = 3$. (Hint: If $m = 3$ and $\lambda_3 = 1$, the inequality in (2) is trivial. If $\lambda_3 \neq 1$, then $\lambda_1 + \lambda_2 \neq 0$, and

$$f(\lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 + \lambda_3 \mathbf{x}_3) = f\left((\lambda_1 + \lambda_2)\left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \mathbf{x}_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2} \mathbf{x}_2\right) + \lambda_3 \mathbf{x}_3\right) \quad (*)$$

Show how (2) can now be derived from the result for $m = 2$. The general proof of (2) is based on mathematical induction.)

4. Prove Jensen's inequality (3) for the case in which f is C^1 by using the following idea: by (1), concavity of f implies that $f(x(t)) - f(z) \leq f'(z)(x(t) - z)$. Multiply both sides of this inequality by $\lambda(t)$ and integrate w.r.t. t . Then let $z = \int_a^b \lambda(t)x(t) dt$.
5. Suppose S is a convex subset of $\mathbb{R}^n$ and that the function $f: S \rightarrow \mathbb{R}^n$ has a supergradient at every point of S . Prove that f is concave.
6. With reference to Note 2.3.2 prove that $f(x, y) = x^4 + y^4$ is strictly convex. (Hint: Use Theorem 2.4.1.)

2.5 Quasiconcave and Quasiconvex Functions

Let $f(\mathbf{x})$ be a function defined over a convex set S in $\mathbb{R}^n$. For each real number a , define the subset P_a of S by

$$P_a = \{\mathbf{x} \in S : f(\mathbf{x}) \geq a\} \quad (1)$$

Then P_a is called an **upper level set** for f . It consists of those points in S that give values of f that are greater than or equal to a . Fig. 1 shows the graph of a quasiconcave function of two variables and Fig. 2 shows one of its upper level sets.

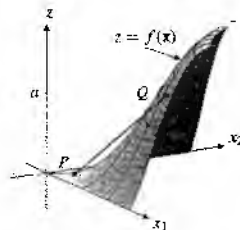


Figure 1 A quasiconcave function of two variables.

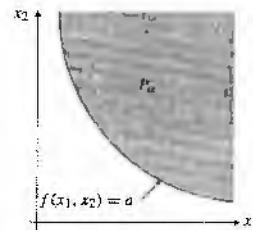


Figure 2 An upper level set for the function in Fig. 1.

DEFINITION OF QUASICONCAVE AND QUASICONVEX FUNCTIONS

A function f , defined over a convex set $S \subseteq \mathbb{R}^n$, is **quasiconcave** if the upper level set $P_a = \{\mathbf{x} \in S : f(\mathbf{x}) \geq a\}$ is convex for each number a .

We say that f is **quasiconvex** if $-f$ is quasiconcave. So f is quasiconvex if the lower level set $P^a = \{\mathbf{x} : f(\mathbf{x}) \leq a\}$ is convex for each number a .

(2)

EXAMPLE 1

Show that $f(x, y) = e^{-x^2-y^2}$ (which is proportional to the bivariate normal distribution function in statistics) is quasiconcave. (Its graph is often called "bell-shaped".)

Solution: Note that $f(x, y) \leq 1$ for all (x, y) and that $f(0, 0) = 1$. For each number a , let $P_a = \{(x, y) : e^{-x^2-y^2} \geq a\}$. If $a > 1$, P_a is empty, and thus convex. If $a = 1$, the set P_a consists only of the point $(0, 0)$, which is convex. If $a \leq 0$, P_a is the whole xy -plane, which again is convex. Suppose $a \in (0, 1)$. Then $e^{-x^2-y^2} \geq a$ iff $-x^2 - y^2 \geq \ln a$ iff $x^2 + y^2 \leq -\ln a$. The points (x, y) satisfying the last inequality are those on or inside the circle centered at the origin with radius $\sqrt{-\ln a}$. (Since $a \in (0, 1)$, $-\ln a$ is positive.) This is a convex set, so all upper level sets of f are convex, and thus f is quasiconcave. (In Example 3 we give a much simpler argument.)

EXAMPLE 2

Let $y = f(x)$ be any function of one variable that is either increasing or decreasing on an interval. Explain why f is quasiconcave as well as quasiconvex. (In particular, it follows that a quasiconcave function is not necessarily "bell-shaped".)

Solution: The level sets are either intervals or empty. Figure 3 shows a typical case.

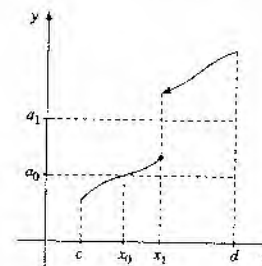


Figure 3 $P_{a_0} = [x_0, d]$, $P_{a_1} = (x_1, d]$.

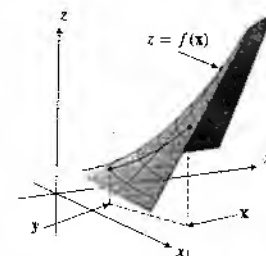


Figure 4 Illustration of (4) and (5).

The function in Fig. 1 is not concave. For example, the line segment joining points P and Q on the graph lies above the graph of f , not below. On the other hand, according to Theorem 2.4.2 (a), a concave function has convex upper level sets. Likewise, a convex function has convex lower level sets. Thus:

If $f(\mathbf{x})$ is concave, then $f(\mathbf{x})$ is quasiconcave.

If $f(\mathbf{x})$ is convex, then $f(\mathbf{x})$ is quasiconvex.

(3)

Theorem 2.3.4 implies that a sum of concave functions is again concave. The corresponding result is not valid for quasiconcave functions (see Problem 6):

WARNING

A sum of quasiconcave functions need not be quasiconcave.

Some useful properties of quasiconcave functions, illustrated in Fig. 4, are these:

THEOREM 2.5.1

Let f be a function of n variables defined on a convex set S in $\mathbb{R}^n$. Then f is quasiconcave if and only if either of the following equivalent conditions is satisfied for all $\mathbf{x}$ and $\mathbf{y}$ in S and all λ in $[0, 1]$:

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq \min\{f(\mathbf{x}), f(\mathbf{y})\} \quad (4)$$

$$f(\mathbf{x}) \geq f(\mathbf{y}) \implies f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq f(\mathbf{y}) \quad (5)$$

Proof: Suppose that f is quasiconcave. Let $\mathbf{x}, \mathbf{y} \in S$, $\lambda \in [0, 1]$, and define $a = \min\{f(\mathbf{x}), f(\mathbf{y})\}$. Then $\mathbf{x}$ and $\mathbf{y}$ both belong to the set $P_a = \{\mathbf{u} \in S : f(\mathbf{u}) \geq a\}$. Since P_a is convex, the vector $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ is also in P_a , meaning that $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq a$. So (4) is satisfied.

Suppose on the other hand that (4) is valid and let a be an arbitrary number. We must show that P_a is convex. If P_a is empty or consists only of one point, P_a is evidently convex. If P_a contains more than one point, take two arbitrary points $\mathbf{x}$ and $\mathbf{y}$ in P_a . Then $f(\mathbf{x}) \geq a$ and $f(\mathbf{y}) \geq a$. Also, for all λ in $[0, 1]$, (4) implies that $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq \min\{f(\mathbf{x}), f(\mathbf{y})\} \geq a$, i.e. $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ lies in P_a . Hence P_a is convex.

Finally, it is easy to prove that (4) holds for all $\mathbf{x}$ and $\mathbf{y}$ in S and all λ in $[0, 1]$ iff (5) holds for all such $\mathbf{x}, \mathbf{y}, \lambda$. ■

The following result is useful in utility theory because it shows that quasiconcavity (or quasiconvexity) is preserved by any increasing transformation F :

THEOREM 2.5.2

Let $f(\mathbf{x})$ be defined on a convex set S in $\mathbb{R}^n$ and let F be a function of one variable whose domain includes $f(S)$.

- (a) If $f(\mathbf{x})$ is quasiconcave (quasiconvex) and F is increasing, then $F(f(\mathbf{x}))$ is quasiconcave (quasiconvex).
- (b) If $f(\mathbf{x})$ is quasiconcave (quasiconvex) and F is decreasing, then $F(f(\mathbf{x}))$ is quasiconvex (quasiconcave).

Proof: (a) Suppose that $f(\mathbf{x})$ is quasiconcave and F is increasing. If $F(f(\mathbf{x}))$ is not quasiconcave, then by (4) in Theorem 2.5.1 there must exist points $\mathbf{u}$ and $\mathbf{v}$ in S and a point

$\mathbf{w} = \lambda\mathbf{u} + (1-\lambda)\mathbf{v}$ on the line segment between them such that $F(f(\mathbf{w})) < F(f(\mathbf{u}))$ and $F(f(\mathbf{w})) < F(f(\mathbf{v}))$. Since F is increasing, this would imply that $f(\mathbf{w}) < f(\mathbf{u})$ and $f(\mathbf{w}) < f(\mathbf{v})$. By Theorem 2.5.1 again, that is impossible if f is quasiconcave. It follows that $F(f(\mathbf{x}))$ is quasiconcave.

The quasiconvex case is treated in similar manner, and finally the proof of (b) is analogous to the proof of (a). ■

EXAMPLE 3

In Example 1 we showed that $f(x, y) = e^{-x^2-y^2}$ is quasiconcave. Because of Theorem 2.5.2 we can give a simpler argument. The function $-x^2 - y^2$ is concave, and therefore quasiconcave. The function $u \mapsto e^u$ is increasing, so $f(x, y)$ is quasiconcave. ■

EXAMPLE 4

Economists usually think of an individual's utility function as representing preferences, rather than as a numerical measurement of "happiness" associated with a certain commodity bundle. So economists are more concerned with the level sets of the utility function than with the numerical values taken by the function. Given a utility function, any increasing transformation of that function represents the same level sets, although the numerical values assigned to the level curves are different. Note that, according to Theorem 2.5.2, the property of quasiconcavity is preserved by an arbitrary increasing transformation. (This is not the case for concavity. The function f in Example 3 is not concave.) ■

A set K in $\mathbb{R}^n$ is called a **cone** if $t\mathbf{x} \in K$ whenever $\mathbf{x} \in K$ and $t > 0$. A function f defined on a cone K is **homogeneous of degree q** if $f(t\mathbf{x}) = t^q f(\mathbf{x})$ for all $\mathbf{x}$ in K and all $t > 0$.

THEOREM 2.5.3

Let $f(\mathbf{x})$ be a function defined on a convex cone K in $\mathbb{R}^n$. Suppose that f is quasiconcave and homogeneous of degree q , where $0 < q \leq 1$, that $f(\mathbf{0}) = 0$, and that $f(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$ in K . Then f is concave.

Proof: Suppose $q = 1$. We need to show that if $\mathbf{x}$ and $\mathbf{y}$ are points in K and $\lambda \in [0, 1]$, then

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) \quad (*)$$

If $\mathbf{x} = \mathbf{0}$, then $f(\mathbf{x}) = 0$, and (*) is satisfied with equality for all $\mathbf{y}$ since f is homogeneous of degree 1. The same is true for all $\mathbf{x}$ if $\mathbf{y} = \mathbf{0}$. Suppose next that $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{y} \neq \mathbf{0}$. Then $f(\mathbf{x}) > 0$ and $f(\mathbf{y}) > 0$, by hypothesis. Given any $\lambda \in [0, 1]$, put $\alpha = f(\mathbf{x})/f(\mathbf{y})$, $\beta = \alpha\lambda + (1-\lambda)$, $\mu = \alpha\lambda/\beta$, and let $\mathbf{x}' = (\beta/\alpha)\mathbf{x}$, $\mathbf{y}' = \beta\mathbf{y}$. Note that $\mu \in [0, 1]$. Also, $\mu\mathbf{x}' + (1-\mu)\mathbf{y}' = \lambda\mathbf{x} + (\beta-\alpha\lambda)\mathbf{y} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$. Moreover, $f(\mathbf{x}') = (\beta/\alpha)f(\mathbf{x}) = \beta f(\mathbf{y}) = f(\mathbf{y}')$. Since f is quasiconcave,

$$\begin{aligned} f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) &= f(\mu\mathbf{x}' + (1-\mu)\mathbf{y}') \geq f(\mathbf{x}') = f(\mathbf{y}') = \mu f(\mathbf{x}') + (1-\mu)f(\mathbf{y}') \\ &= (\mu\beta/\alpha)f(\mathbf{x}) + (\beta-\beta\mu)f(\mathbf{y}) = \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y}) \end{aligned}$$

This proves the theorem in the case when $q = 1$.

Finally, suppose that $q \in (0, 1]$, and define a new function g by $g(\mathbf{x}) = (f(\mathbf{x}))^{1/q}$ for all $\mathbf{x}$ in K . Then g is quasiconcave and homogeneous of degree 1, and $g(\mathbf{x}) > 0$ for $\mathbf{x} \neq \mathbf{0}$. According to the argument above, g is concave. Theorem 2.3.5(a) shows that $f = g^q$ is also concave. ■

EXAMPLE 5* The Cobb–Douglas function is defined for all $x_1 > 0, \dots, x_n > 0$ by

$$z = Ax_1^{a_1} x_2^{a_2} \cdots x_n^{a_n} \quad (a_1, a_2, \dots, a_n, \text{ and } A \text{ are positive constants}) \quad (*)$$

Taking the natural logarithm of each side yields $\ln z = \ln A + a_1 \ln x_1 + \cdots + a_n \ln x_n$. As a sum of concave functions of $(x_1, \dots, x_n)$, $\ln z$ is concave, and hence quasiconcave. Now, $z = e^{\ln z}$, and the function $u \mapsto e^u$ is increasing. By Theorem 2.5.2(a), z is quasiconcave.

Let $a = a_1 + \cdots + a_n$. For $a < 1$, the Cobb–Douglas function is strictly concave. (See Problem 2.3.9.) For $a > 1$, it is not concave. (Along the ray $x_1 = \cdots = x_n = x$, one has $z = Ax^a$, which is strictly convex for $a > 1$.)

If $a \leq 1$, Theorem 2.5.3 shows that the function is concave. The following display sets out some of the most important properties of the Cobb–Douglas function.

PROPERTIES OF THE COBB–DOUGLAS FUNCTION

The Cobb–Douglas function $z = Ax_1^{a_1} \cdots x_n^{a_n}$, defined for $x_1 > 0, \dots, x_n > 0$, with A and $a_1, \dots, a_n$ positive, is homogeneous of degree $a = a_1 + \cdots + a_n$, and:

- (a) quasiconcave for all $a_1, \dots, a_n$;
- (b) concave for $a \leq 1$;
- (c) strictly concave for $a < 1$.

(6)

EXAMPLE 6 The generalized CES function is defined for $x_1 > 0, x_2 > 0, \dots, x_n > 0$ by

$$z = A(\delta_1 x_1^{-\rho} + \delta_2 x_2^{-\rho} + \cdots + \delta_n x_n^{-\rho})^{-1/\rho}, \quad A > 0, \mu > 0, \rho \neq 0, \delta_i > 0, i = 1, \dots, n$$

We have $z = Au^{-\mu/\rho}$ where $u = \delta_1 x_1^{-\rho} + \delta_2 x_2^{-\rho} + \cdots + \delta_n x_n^{-\rho}$. If $\rho \leq -1$, then u is quasiconvex (in fact convex as a sum of convex functions) and $u \mapsto Au^{-\mu/\rho}$ is increasing, so z is quasiconvex, according to Theorem 2.5.2(a). If $\rho \in [-1, 0)$, then u is quasiconcave (in fact concave) and $u \mapsto Au^{-\mu/\rho}$ is increasing, so z is quasiconcave according to Theorem 2.5.2(a). If $\rho > 0$, then u is quasiconvex (in fact convex) and $u \mapsto Au^{-\mu/\rho}$ is decreasing, so z is quasiconcave according to Theorem 2.5.2(b).

It is easy to see that z is homogeneous of degree μ . It follows from Theorem 2.5.3 that if $0 < \mu \leq 1$ and $\rho \geq -1$, then z is concave. Part (c) of the display below follows from Problem 11.

PROPERTIES OF THE GENERALIZED CES FUNCTION

The CES function $z = A(\delta_1 x_1^{-\rho} + \delta_2 x_2^{-\rho} + \cdots + \delta_n x_n^{-\rho})^{-1/\rho}$, $A > 0$, $\mu > 0$, $\rho \neq 0$, $\delta_i > 0$, $i = 1, \dots, n$ is homogeneous of degree μ , and:

- (a) quasiconvex for $\rho \leq -1$, quasiconcave for $\rho \geq -1$;
- (b) concave for $0 < \mu \leq 1$, $\rho \geq -1$;
- (c) strictly concave for $0 < \mu < 1$, $\rho > -1$.

(7)

Quasiconcave C^1 Functions

Theorem 2.4.1 implies that the graph of a concave C^1 function lies below its tangent hyperplane. We now consider a somewhat similar characterization of quasiconcave C^1 functions. The geometric idea is suggested in Fig. 5, where the upper level set $P_a = \{x : f(x) \geq a\}$ is convex. Here $x = (x_1, x_2)$ and $f(x^0) = a$. The gradient $\nabla f(x^0)$ of f at x^0 is orthogonal to the level surface $f(x) = a$ at x^0 (the level curve in the two-dimensional case), and points in the direction of maximal increase of f . (See Theorem 2.1.1.) Points z that satisfy $\nabla f(x^0) \cdot (z - x^0) = 0$ are on the tangent hyperplane to the level curve at x^0 . If we choose an x in P_a , then $f(x) \geq a = f(x^0)$, and it seems that the angle α between the vectors $x - x^0$ and $\nabla f(x^0)$ is always acute ($\leq 90^\circ$), in the sense that $\nabla f(x^0) \cdot (x - x^0) = \|\nabla f(x^0)\| \|x - x^0\| \cos \alpha \geq 0$. (See (1.1.40).) Thus, the points x in P_a all lie “above” the tangent hyperplane to the level surface. In that sense, the tangent hyperplane “supports” the upper level set. On the other hand, in Fig. 6 the level set P_a is not convex, and the tangent hyperplane at x^0 does not support the level set.

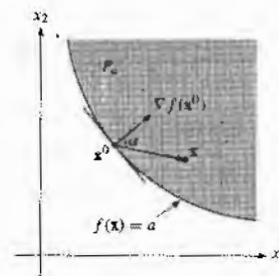


Figure 5 P_a is convex and $\nabla f(x_0) \cdot (x - x_0) \geq 0$.

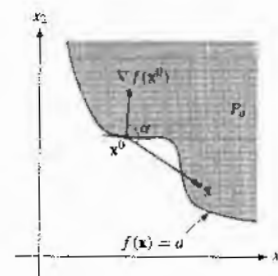


Figure 6 P_a is not convex and $\nabla f(x_0) \cdot (x - x_0) < 0$.

THEOREM 2.5.4

Let f be a C^1 function of n variables defined on an open convex set S in $\mathbb{R}^n$. Then f is quasiconcave on S if and only if for all x and x^0 in S

$$f(x) \geq f(x^0) \implies \nabla f(x^0) \cdot (x - x^0) = \sum_{i=1}^n \frac{\partial f(x^0)}{\partial x_i} (x_i - x_i^0) \geq 0 \quad (8)$$

Proof: We prove that if f is quasiconcave then (8) is valid. (The reverse implication is also true, but less useful.) Let $x, x^0 \in S$ and define the function g on $[0, 1]$ by $g(t) = f(x^0 + t(x - x^0))$. Then $g'(t) = \nabla f(x^0 + t(x - x^0)) \cdot (x - x^0)$. (See the argument for (2.1.7).) Suppose $f(x) \geq f(x^0)$. Then by (5) in Theorem 2.5.1, $g(t) \geq g(0)$ for all t in $[0, 1]$. This implies that $g'(0) = \nabla f(x^0) \cdot (x - x^0) \geq 0$. ■

A stronger property than quasiconcavity is strict quasiconcavity:

DEFINITION OF STRICT QUASICONCAVITY/QUASICONVEXITY

A function f defined on a convex set $S \subseteq \mathbb{R}^n$ is called **strictly quasiconcave** if

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) > \min\{f(\mathbf{x}), f(\mathbf{y})\} \quad (9)$$

for all $\mathbf{x}$ and $\mathbf{y}$ in S with $\mathbf{x} \neq \mathbf{y}$ and all λ in $(0, 1)$. The function f is **strictly quasiconvex** if $-f$ is strictly quasiconcave.

It is easy to verify that any strictly concave (convex) function is strictly quasiconcave (quasiconvex).²

It follows from Theorem 2.5.1(4) that a strictly quasiconcave function is quasiconcave. (It suffices to check what happens if $\mathbf{x} = \mathbf{y}$.) We see from the definition that a strictly increasing (or decreasing) function of one variable is always strictly quasiconcave.

An important fact about strictly quasiconcave functions is that they cannot have more than one global maximum point: if $\mathbf{x}$ and $\mathbf{y}$ are two different points with $f(\mathbf{x}) = f(\mathbf{y})$, then $f(\mathbf{z}) > f(\mathbf{x})$ for every $\mathbf{z}$ on the open line segment $(\mathbf{x}, \mathbf{y})$.

A Determinant Criterion for Quasiconcavity

For C^2 functions one can check quasiconcavity by examining the signs of certain determinants, called **bordered Hessians**. The ordinary Hessians used in Section 2.3 to examine the concavity of a function are "bordered" by an extra row and column consisting of the first-order partial derivatives of the function. For the case of two variables the result is the following (see also Problem 9):

THEOREM 2.5.5

Let $f(x, y)$ be a C^2 function defined in an open, convex set S in the plane. Define the bordered Hessian determinant

$$B_2(x, y) = \begin{vmatrix} 0 & f'_1(x, y) & f'_2(x, y) \\ f'_1(x, y) & f''_{11}(x, y) & f''_{12}(x, y) \\ f'_2(x, y) & f''_{21}(x, y) & f''_{22}(x, y) \end{vmatrix} \quad (10)$$

Then:

- A necessary condition for f to be quasiconcave in S is that $B_2(x, y) \geq 0$ for all (x, y) in S .
- A sufficient condition for f to be strictly quasiconcave in S is that $f'_1(x, y) \neq 0$ and $B_2(x, y) > 0$ for all (x, y) in S .

² Some authors use a weaker definition of strict quasiconcavity and only require (9) to hold when $f(\mathbf{x}) \neq f(\mathbf{y})$. But then f is not necessarily quasiconcave. See Problem 8.

EXAMPLE 7

We already know from Example 5 that the Cobb–Douglas function defined for all $x_1 > 0, \dots, x_n > 0$ by $z = Ax_1^{a_1}x_2^{a_2}\cdots x_n^{a_n}$ (with A and $a_1, \dots, a_n$ all positive) is always quasiconcave. We can use Theorem 2.5.5 to confirm this result for $n = 2$. The first- and second-order partial derivatives can be expressed as $z'_i = a_i z/x_i$, $z''_{ii} = a_i(a_i - 1)z/x_i^2$, $z''_{ij} = a_i a_j z/x_i x_j$ (for $i \neq j$). Consider the case $n = 2$. Then the determinant $B_2 = B_2(x_1, x_2)$ is

$$B_2 = \begin{vmatrix} 0 & \frac{a_1}{x_1}z & \frac{a_2}{x_2}z \\ \frac{a_1}{x_1}z & \frac{a_1(a_1-1)}{x_1^2}z & \frac{a_1 a_2}{x_1 x_2}z \\ \frac{a_2}{x_2}z & \frac{a_2 a_1}{x_2 x_1}z & \frac{a_2(a_2-1)}{x_2^2}z \end{vmatrix} = \frac{a_1 a_2}{(x_1 x_2)^2} z^3 \begin{vmatrix} 0 & 1 & 1 \\ a_1 & a_1 - 1 & a_1 \\ a_2 & a_2 & a_2 - 1 \end{vmatrix}$$

where we have systematically removed all the common factors from each row and column. In the last determinant, subtract the first column from each of the others, and then add the last two rows to the first. Then we see that $B_2 = a_1 a_2 (a_1 + a_2) z^3 / (x_1 x_2)^2$. Because A, a_1 , and a_2 are all positive, one has $B_2 > 0$. Moreover, $z'_1 \neq 0$. We conclude from Theorem 2.5.5(b) that the Cobb–Douglas function $z = Ax_1^{a_1}x_2^{a_2}$ is strictly quasiconcave. (This argument can easily be extended to be valid for a general value of n by using the next theorem.)

We go on to consider the general case. Define

$$B_r(\mathbf{x}) = \begin{vmatrix} 0 & f'_1(\mathbf{x}) & \cdots & f'_r(\mathbf{x}) \\ f'_1(\mathbf{x}) & f''_{11}(\mathbf{x}) & \cdots & f''_{1r}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f'_r(\mathbf{x}) & f''_{r1}(\mathbf{x}) & \cdots & f''_{rr}(\mathbf{x}) \end{vmatrix}, \quad r = 1, \dots, n \quad (11)$$

THEOREM 2.5.6

Let f be a C^2 function defined in an open, convex set S in $\mathbb{R}^n$. Define the bordered Hessian determinants $B_r(\mathbf{x})$, by (11). Then:

- A necessary condition for f to be quasiconcave is that $(-1)^r B_r(\mathbf{x}) \geq 0$ for all $\mathbf{x}$ in S and all $r = 1, \dots, n$.
- A sufficient condition for f to be strictly quasiconcave is that $(-1)^r B_r(\mathbf{x}) > 0$ for all $\mathbf{x}$ in S and all $r = 1, \dots, n$.

Proof: We prove only (b). (The proof uses some results from Chapter 3.) Let $\mathbf{x}^0$ be an arbitrary point in S and consider the problem

$$\max f(\mathbf{x}) \quad \text{subject to} \quad g(\mathbf{x}) = \nabla f(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0) \leq 0, \quad \mathbf{x} \in S \quad (*)$$

Let $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \lambda \nabla f(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0) = f(\mathbf{x}) - \lambda [f'_1(\mathbf{x}^0)(x_1 - x_1^0) + \cdots + f'_n(\mathbf{x}^0)(x_n - x_n^0)]$. Then $\mathcal{L}'(\mathbf{x}) = \nabla f(\mathbf{x}) - \lambda \nabla f(\mathbf{x}^0)$, and, in particular, $\mathcal{L}'(\mathbf{x}^0) = 0$ for $\lambda = 1$, so for this value of λ the

first-order conditions for a local maximum at $\mathbf{x}^0$ are satisfied. Moreover, $g_i'(\mathbf{x}^0) = f_i'(\mathbf{x}^0)$ for all i and $\mathcal{L}_{ij}''(\mathbf{x}^0) = f_{ij}''(\mathbf{x}^0)$ for all i and j . Thus the condition in (b) evidently implies that the sufficient conditions for a strict local maximum in problem (*) are satisfied at $\mathbf{x}^0$ (see Theorem 3.6.4). Then for $\mathbf{x} \neq \mathbf{x}^0$ with $\mathbf{x}$ close to $\mathbf{x}^0$, one has $f(\mathbf{x}) < f(\mathbf{x}^0)$ when $\nabla f(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0) \leq 0$. Equivalently,

$$\text{for } \mathbf{x} \neq \mathbf{x}^0 \text{ close to } \mathbf{x}^0, \quad f(\mathbf{x}) \geq f(\mathbf{x}^0) \Rightarrow \nabla f(\mathbf{x}^0) \cdot (\mathbf{x} - \mathbf{x}^0) > 0 \quad (**)$$

Let $\mathbf{x}' \neq \mathbf{x}''$ belong to S and assume that $f(\mathbf{x}^0) \leq f(\mathbf{x}')$. Define $\mathbf{h}(t) = t\mathbf{x}' + (1-t)\mathbf{x}''$ for t in $[0, 1]$. Consider the problem

$$\min_{t \in [0, 1]} f(\mathbf{h}(t))$$

A minimum point t' must exist, and it can be assumed to belong to $(0, 1)$ because $f(\mathbf{x}^0) \leq f(\mathbf{x}')$. For $t' > 0$ and t close to t' , we have $f(\mathbf{h}(t)) \geq f(\mathbf{h}(t'))$, and so from (**),

$$\begin{aligned} \nabla f(\mathbf{h}(t')) \cdot (\mathbf{h}(t) - \mathbf{h}(t')) &= (t - t') \nabla f(\mathbf{h}(t')) \cdot (\mathbf{x}' - \mathbf{x}'') \\ &= (t - t') [(d/dt)f(\mathbf{h}(t))]_{t=t'} > 0 \quad \text{if } f(\mathbf{h}(t)) \geq f(\mathbf{h}(t')) \end{aligned}$$

So there can be no interior minimum point t' , and the only possibility is $f(\mathbf{h}(t)) > f(\mathbf{x}^0)$ for all t in $(0, 1)$, and by definition f is strictly quasiconcave. ■

PROBLEMS FOR SECTION 2.5

- Use (6) and (7) to classify as (quasi)concave or (quasi)convex each of the functions $z = F(x, y)$ defined for all $x > 0, y > 0$ by:
 - $z = 100x^{1/3}y^{1/4}$
 - $z = x^2y^3$
 - $z = 250x^{0.02}y^{0.98}$
 - $z = \sqrt{x^2 + y^2}$
 - $z = (x^{1/3} + y^{1/3})^3$
 - $z = (x^{-1/4} + y^{-1/4})^{-1/4}$
- Determine if the following functions are quasiconcave:
 - $f(x) = 3x + 4$
 - $f(x, y) = ye^x, \quad y > 0$
 - $f(x, y) = -x^2y^3$
 - $f(x) = x^3 + x^2 + 1$ if $x < 0, f(x) = 1$ if $x \geq 0$
- If $f(x)$ is concave, for what values of the constants a and b can one be sure that $af(x) + b$ is concave?
 - If $f(x)$ is concave and positive valued, determine if the functions $g(x) = \ln f(x)$ and $h(x) = e^{f(x)}$ are concave/quasiconcave.
- Consider the function $f(x) = -x^2/(1 + x^2)$. Sketch the graph of f and prove that f is quasiconcave by using the definition. Does Theorem 2.5.6 give the same result?
- What does Theorem 2.5.6(b) say about C^2 functions of one variable?
- Show that, although the two functions $f(x) = -x$ and $g(x) = x^3$ are both quasiconcave and quasiconvex, their sum is neither.
- Say that the function $f: \mathbb{R} \rightarrow \mathbb{R}$ is *single-peaked* if there exists an x^* in $\mathbb{R}$ such that f is strictly increasing for $x \leq x^*$ and strictly decreasing for $x \geq x^*$.
 - Show that such a function is strictly quasiconcave.
 - Suppose f is also concave. Must it be strictly concave?

HARDER PROBLEMS

- Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be defined by $f(0) = 0$ and $f(x) = 1$ if $x \neq 0$. Verify that f satisfies inequality (9) for all λ in $(0, 1)$ and all x and y with $f(x) \neq f(y)$. But show that f is still not quasiconcave.
- Suppose $F(x, y)$ is a C^2 function with $F_{xx}(x, y) > 0$. Let $y = \varphi(x)$ be defined implicitly by the equation $F(x, y) = C$. Prove that $\varphi(x)$ is convex if and only if $B_2(x, y) \geq 0$ (see (10)). (Convexity of $\varphi(x)$ is equivalent to F being quasiconcave.) (Hint: You might use the formula for y'' given in Problem 2.1.9, but an alternative argument is simpler.)
- Let $f_1, \dots, f_m$ be concave functions defined on a convex set S in $\mathbb{R}^d$ and let $F(u_1, \dots, u_m)$ be quasiconcave and increasing in each variable. Prove that $g(x) = F(f_1(x), \dots, f_m(x))$ is quasiconcave.
- Modify the proof of Theorem 2.5.3 to show that if f is strictly quasiconcave and homogeneous of degree $q \in (0, 1)$, then f is strictly concave. Use this result to prove part (c) of the properties of the generalized CES function in display (7) above.

2.6 Taylor's Formula

When studying the behaviour of a complicated function f near a particular point of its domain, it is often useful to approximate f by a much simpler function. For functions of one variable you have probably seen the approximation:³

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n \quad (x \text{ close to } a) \quad (1)$$

The function f and the polynomial on the right-hand side have the same value and the same first n derivatives at a . Thus they have such a high degree of contact at $x = a$ that we can expect the approximation in (1) to be good over some (possibly small) interval centred at $x = a$. (See e.g. EMEA, Section 7.5 for details.) Nevertheless, the usefulness of such polynomial approximations is unclear unless something is known about the error that results. Taylor's formula helps remedy this deficiency (see e.g. EMEA, Section 7.6):

THEOREM 2.6.1 (TAYLOR'S FORMULA)

If f is $n+1$ times differentiable in an interval that contains a and x , then

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x-a)^{n+1} \quad (2)$$

for some number c between a and x .

³ Recall that if n is a natural number, $n!$ (read as "n factorial") is defined as $n! = 1 \cdot 2 \cdot 3 \cdots (n-1) \cdot n$. By convention, $0! = 1$.

NOTE 1 The remainder term at the end of (2) resembles the preceding terms in the sum. The only difference is that in the formula for the remainder, $f^{(n+1)}$ is evaluated at a point c , where c is some unspecified number between a and x , whereas in all the other terms, the derivatives are evaluated at a . The number c depends, in general, on x , a , and n , as well as the function f .

Putting $n = 1$, $a = 0$ in formula (2) gives

$$f(x) = f(0) + f'(0)x + \frac{1}{2}f''(c)x^2 \quad \text{for some } c \text{ between } 0 \text{ and } x \quad (3)$$

This formula tells us that $\frac{1}{2}f''(c)x^2$ is the error that results if we replace $f(x)$ with its linear approximation at $x = 0$.

Taylor's Formula for Functions of Two Variables

Let $z = f(x, y)$ be defined in a neighbourhood of (x_0, y_0) and let (h, k) be a given pair of numbers. With t as a real number, define the function g by $g(t) = f(x_0 + th, y_0 + tk)$. The function g records how $f(x, y)$ behaves along the straight line through (x_0, y_0) in the direction determined by the vector (h, k) .

We see that $g(1) = f(x_0 + h, y_0 + k)$ and $g(0) = f(x_0, y_0)$. According to formula (2), there exists a number c in $(0, 1)$ such that

$$g(1) = g(0) + \frac{1}{1!}g'(0) + \cdots + \frac{1}{n!}g^{(n)}(0) + \frac{1}{(n+1)!}g^{(n+1)}(c) \quad (*)$$

If we find $g'(0)$, $g''(0)$, etc., and insert the results into $(*)$, we obtain the general Taylor formula in two variables.

The formula is particularly useful for the case $n = 1$, when $(*)$ reduces to

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(c) \quad (**)$$

Using the chain rule we find that $g'(t) = f'_1(x_0 + th, y_0 + tk)h + f'_2(x_0 + th, y_0 + tk)k$ and $g''(t) = f''_{11}(x_0 + th, y_0 + tk)h^2 + 2f''_{12}(x_0 + th, y_0 + tk)hk + f''_{22}(x_0 + th, y_0 + tk)k^2$. Thus $g'(0) = f'_1(x_0, y_0)h + f'_2(x_0, y_0)k$ and $g''(c) = f''_{11}(\bar{x}, \bar{y})h^2 + 2f''_{12}(\bar{x}, \bar{y})hk + f''_{22}(\bar{x}, \bar{y})k^2$, where $\bar{x} = x_0 + ch$, $\bar{y} = y_0 + ck$. Inserting these expressions into $(*)$ gives:

THEOREM 2.6.2 (TAYLOR'S FORMULA: LINEAR TERMS AND REMAINDER)

If f is C^2 inside a circular disk around (x_0, y_0) that contains $(x_0 + h, y_0 + k)$, then

$$f(x_0 + h, y_0 + k) = f(x_0, y_0) + f'_1(x_0, y_0)h + f'_2(x_0, y_0)k + \frac{1}{2}[f''_{11}(\bar{x}, \bar{y})h^2 + 2f''_{12}(\bar{x}, \bar{y})hk + f''_{22}(\bar{x}, \bar{y})k^2]$$

where $\bar{x} = x_0 + ch$, $\bar{y} = y_0 + ck$ for some number c in $(0, 1)$.

In formula $(*)$, let us put $n = 2$, $x_0 = y_0 = 0$, and $h = x$, $k = y$. Disregarding the remainder, this gives the following quadratic approximation to $f(x, y)$ around $(0, 0)$:

$$f(x, y) \approx f(0, 0) + f'_1(0, 0)x + f'_2(0, 0)y + \frac{1}{2}(f''_{11}(0, 0)x^2 + 2f''_{12}(0, 0)xy + f''_{22}(0, 0)y^2) \quad (4)$$

EXAMPLE 1

Find the quadratic approximation around $(0, 0)$ for $f(x, y) = e^x \ln(1 + y)$.

Solution: We find that $f'_1(x, y) = e^x \ln(1 + y)$, $f'_2(x, y) = e^x/(1 + y)$. Moreover, $f''_{11}(x, y) = e^x \ln(1 + y)$, $f''_{12}(x, y) = e^x/(1 + y)$, and $f''_{22}(x, y) = -e^x/(1 + y)^2$. It follows that $f(0, 0) = 0$, $f'_1(0, 0) = 0$, $f'_2(0, 0) = 1$, $f''_{11}(0, 0) = 0$, $f''_{12}(0, 0) = 1$, and $f''_{22}(0, 0) = -1$. From (4) we get

$$e^x \ln(1 + y) \approx y + xy - \frac{1}{2}y^2$$

Taylor's Formula with n Variables

We briefly explain how to derive Taylor's formula for a function of many variables.

Suppose we want to approximate $z = f(\mathbf{x}) = f(x_1, \dots, x_n)$ near $\mathbf{x}^0 = (x_1^0, \dots, x_n^0)$. Let $\mathbf{h} = (h_1, \dots, h_n)$ and define the function g by $g(t) = f(x_1^0 + th_1, \dots, x_n^0 + th_n) = f(\mathbf{x}^0 + t\mathbf{h})$. We use formula $(**)$ once again. To do so, note how (2.1.7) implies that

$$g'(t) = \sum_{i=1}^n f'_i(\mathbf{x}^0 + t\mathbf{h}) \cdot h_i = f'_1(\mathbf{x}^0 + t\mathbf{h}) \cdot h_1 + \cdots + f'_n(\mathbf{x}^0 + t\mathbf{h}) \cdot h_n \quad (5)$$

Differentiating w.r.t. t and using (2.1.7) once more, we obtain $g''(t) = \sum_{i=1}^n \frac{d}{dt} f'_i(\mathbf{x}^0 + t\mathbf{h}) h_i$. Here, for each $i = 1, 2, \dots, n$, using summation notation,

$$\frac{d}{dt} f'_i(\mathbf{x}^0 + t\mathbf{h}) = \sum_{j=1}^n f''_{ij}(\mathbf{x}^0 + t\mathbf{h}) h_j$$

It follows that

$$g''(t) = \sum_{i=1}^n \sum_{j=1}^n f''_{ij}(\mathbf{x}^0 + t\mathbf{h}) h_i h_j \quad (6)$$

Now use the formula $g(1) = g(0) + g'(0) + \frac{1}{2}g''(c)$ with $0 < c < 1$, and insert $t = 0$ into the expression (5) for $g'(t)$ and $t = c$ into the expression (6) for $g''(t)$. The result is:

THEOREM 2.6.3 (TAYLOR'S FORMULA FOR FUNCTIONS OF n VARIABLES)

Suppose f is C^2 in an open set containing the line segment $[\mathbf{x}^0, \mathbf{x}^0 + \mathbf{h}]$. Then

$$f(\mathbf{x}^0 + \mathbf{h}) = f(\mathbf{x}^0) + \sum_{i=1}^n f'_i(\mathbf{x}^0) h_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n f''_{ij}(\mathbf{x}^0 + c\mathbf{h}) h_i h_j \quad (7)$$

for some c in $(0, 1)$.

If we let $\mathbf{x}^0 = (x_1^0, \dots, x_n^0) = (0, \dots, 0)$ and $\mathbf{h} = (h_1, \dots, h_n) = (x_1, \dots, x_n)$, we obtain the formula

$$f(\mathbf{x}) = f(\mathbf{0}) + \sum_{i=1}^n f'_{i1}(\mathbf{0})x_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n f''_{ij}(\mathbf{0})x_i x_j + R_3 \quad (8)$$

The remainder R_3 can be expressed as a triple sum involving third-order derivatives.

PROBLEMS FOR SECTION 2.6

1. Find the quadratic approximations at $(0, 0)$ for
 (a) $f(x, y) = e^{xy}$ (b) $f(x, y) = e^{x^2 - y^2}$ (c) $f(x, y) = \ln(1 + x + 2y)$
2. Find the quadratic approximations at $(0, 0)$ for
 (a) $f(x, y) = e^{x+y}(xy - 1)$ (b) $f(x, y) = e^{xy}$ (c) $f(x, y) = \ln(1 + x^2 + y^2)$
3. Write out formula (8) for $U(x_1, \dots, x_n) = e^{-x_1} + \dots + e^{-x_n}$.
4. z is defined implicitly as a function of x and y around $(0, 0)$ by the equation

$$\ln z = x^3 y - xz + y$$

Find the Taylor polynomial for z of order 2 at $(0, 0)$.

2.7 Implicit and Inverse Function Theorems

In economics one often considers problems of the following kind: if a system of equations defines some endogenous variables as functions of the remaining exogenous variables, what are the partial derivatives of these functions? This section addresses the question whether these functions exist and, if they do exist, whether they are differentiable.

Consider first the simplest case, with one equation of the form

$$f(x, y) = 0 \quad (*)$$

(Note that the equation of any level curve $F(x, y) = C$ can be written as $(*)$ by putting $f(x, y) = F(x, y) - C$.) Assuming that f is C^1 and that $(*)$ defines y as a differentiable function of x , implicit differentiation yields $f'_1(x, y) + f'_2(x, y)y' = 0$. If $f'_2(x, y) \neq 0$, then $y' = -f'_1(x, y)/f'_2(x, y)$.

Geometrically, $(*)$ represents a curve in the xy plane, which could be the curve illustrated in Fig. 1. Studying the curve we observe that for $x > x_1$ there is no y such that (x, y) satisfies the equation. If $x_2 < x < x_1$, there are two values of y for which (x, y) satisfies the equation. (The curve has $x = x_2$ as a vertical asymptote.) Finally, for $x \leq x_2$, there is only one corresponding y . Note that the equation defines y as a function of x in any interval

contained in $(-\infty, x_2]$. Consider on the other hand an interval contained in (x_2, x_1) . The range of variation of y must be restricted in order for the equation to define y as a function of x in that interval. Now, consider the point (x_0, y_0) . If the rectangle R is as in Fig. 1, the equation *does* define y as a function of x in this rectangle. The graph of this function is given in Fig. 2. The size of the rectangle R is constrained by the requirement that each straight line through R parallel to the y -axis must intersect the curve in only one point.

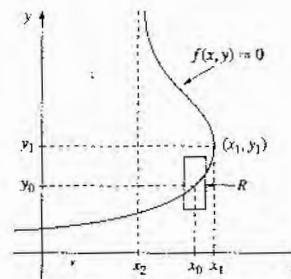


Figure 1 The graph of $f(x, y) = 0$

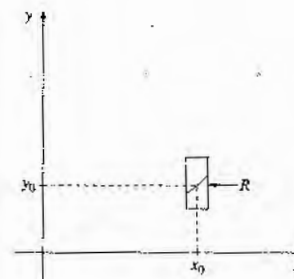


Figure 2 $f(x, y) = 0$ defines y as a function of x in the rectangle R .

Similar rectangles and corresponding solutions of the equation can be constructed for all other points on the curve, except the extreme right point (x_1, y_1) . Regardless of the size of the chosen rectangle around (x_1, y_1) (with (x_1, y_1) as an interior point), for those x close to x_1 on the left there will be two values of y satisfying the equation. For those x to the right of x_1 , there is no suitable y at all. Hence, the equation does not define y as a function of x in a neighbourhood of the point (x_1, y_1) . Note that $f'_2(x_1, y_1) = 0$. (The curve $f(x, y) = 0$ is a level curve for $z = f(x, y)$, and at each point on the curve, the gradient of f , i.e. the vector (f'_1, f'_2) , is orthogonal to the level curve. At (x_1, y_1) , the gradient is clearly parallel to the x -axis, and thus its y -component f'_2 is equal to 0.)

This example illustrates how the crucial condition for $f(x, y) = 0$ to define y as a function of x around (x_0, y_0) is that $f'_2(x_0, y_0) \neq 0$. In general, there are *implicit function theorems* which state when an equation or a system of equations defines some of the variables as functions of the remaining variables. For equation $(*)$, sufficient conditions for $f(x, y) = 0$ to define y as a function of x are briefly indicated in the following:

If $f(x_0, y_0) = 0$ and $f'_2(x_0, y_0) \neq 0$, then the equation $f(x, y) = 0$ defines y as an "implicit" function $y = \phi(x)$ of x near x_0 , with $y_0 = \phi(x_0)$, and with its derivative given by $y' = -f'_1(x, y)/f'_2(x, y)$.

For a proof of the following theorem see Marsden and Hoffman (1993) or Munkres (1991):

THEOREM 2.7.1 (THE IMPLICIT FUNCTION THEOREM FOR TWO VARIABLES)

Suppose $f(x, y)$ is C^1 in an open set A containing (x_0, y_0) , with $f(x_0, y_0) = 0$ and $f'_2(x_0, y_0) \neq 0$. Then there exist an interval $I_1 = (x_0 - \delta, x_0 + \delta)$ and an interval $I_2 = (y_0 - \varepsilon, y_0 + \varepsilon)$ (with $\delta > 0$ and $\varepsilon > 0$) such that $I_1 \times I_2 \subseteq A$ and:

- (a) for every x in I_1 the equation $f(x, y) = 0$ has a unique solution in I_2 which defines y as a function $y = \varphi(x)$ in I_1 ;
 (b) φ is C^1 in $I_1 = (x_0 - \delta, x_0 + \delta)$, with derivative

$$\varphi'(x) = -\frac{f'_1(x, \varphi(x))}{f'_2(x, \varphi(x))} \quad (1)$$

EXAMPLE 1 Show that the equation

$$x^2 e^y - 2y + x = 0$$

defines y as a function of x in an interval around the point $(-1, 0)$. Find the derivative of this function at $x = -1$. (Observe that the equation cannot be solved explicitly for y .)

Solution: Put $f(x, y) = x^2 e^y - 2y + x$. Then $f'_1(x, y) = 2xe^y + 1$, $f'_2(x, y) = x^2 e^y - 2$, and f is C^1 everywhere. Furthermore, $f(-1, 0) = 0$ and $f'_2(-1, 0) = -1 \neq 0$. According to Theorem 2.7.1, the equation therefore defines y as a C^1 function of x in an interval around $(-1, 0)$. Because $f'_1(-1, 0) = -1$, equation (2) implies that $y' = -1$ at $x = -1$. ■

Theorem 2.7.1 gives *sufficient conditions* for $f(x, y) = 0$ to define y as a function of x in a neighbourhood of (x_0, y_0) . The crucial condition is that $f'_2(x_0, y_0) \neq 0$. The next example shows that this condition is not *necessary*.

EXAMPLE 2 Consider the equation $f(x, y) = y^3 - x = 0$. With $(x_0, y_0) = (0, 0)$, the conditions in Theorem 2.7.1 are not satisfied because $f'_2(x, y) = 3y^2$, so that $f'_2(0, 0) = 0$. However, the equation is equivalent to $y^3 = x$, or $y = \sqrt[3]{x}$, and this function is defined for all x . Note that it has no derivative at $x = 0$. ■

To get an idea why Theorem 2.7.1 is true, consider Fig. 3, which shows the graph of $z = f(x, y)$ over a rectangle around (x_0, y_0) . The graph intersects the xy -plane at the point (x_0, y_0) , because $f(x_0, y_0) = 0$. Part (a) of the theorem implies that the surface cuts the xy -plane in a curve which is the graph of a function $y = \varphi(x)$ in a neighbourhood of (x_0, y_0) . The figure illustrates the case when $f'_2(x_0, y_0) > 0$. Because $f'_2(x, y)$ is continuous, $f'_2(x, y)$ is then positive in an open rectangle $(a, b) \times (c, d)$ around (x_0, y_0) , and so $f(x, y)$ becomes a strictly increasing function of y . Hence, $f(x_0, c) < f(x_0, y_0) = 0 < f(x_0, d)$. Since f is continuous, we can also assume $f(x, c) < 0 < f(x, d)$ for all x in the interval (a, b) (just make sure that a and b are sufficiently close to x_0). According to the intermediate value theorem, for each such x there exists a number y in (c, d) with $f(x, y) = 0$. Because $f(x, y)$ is strictly increasing w.r.t. y in the rectangle $(a, b) \times (c, d)$, the solution is unique. The solution y is a function $\varphi(x)$ of x and part (b) of the theorem claims that φ is C^1 .

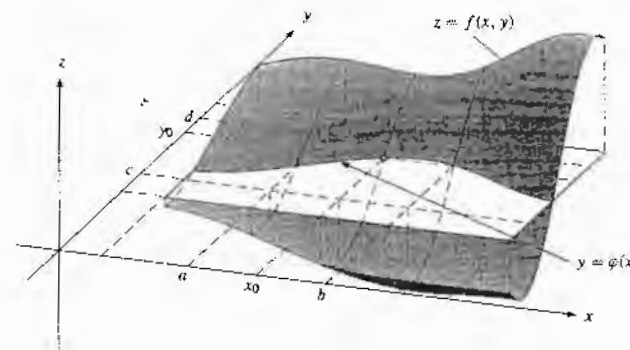


Figure 3 Part of the graph of $z = f(x, y)$ over a neighbourhood of (x_0, y_0) .

The General Case

Theorem 2.7.1 can be generalized to higher-order systems of equations of the form

$$\begin{aligned} f_1(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m) &= 0 \\ &\dots \dots \dots \text{or, in vector notation, } \mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad (2) \\ f_m(x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m) &= 0 \end{aligned}$$

with $\mathbf{f} = (f_1, \dots, f_m)'$, $\mathbf{x} = (x_1, \dots, x_n)$, and $\mathbf{y} = (y_1, \dots, y_m)$. Here there are $n + m$ variables and m equations. Let $(\mathbf{x}^0, \mathbf{y}^0) = (x_1^0, \dots, x_n^0, y_1^0, \dots, y_m^0)$ be an "equilibrium" solution of (2). If $x_1, \dots, x_n$ are the *exogenous* variables and $y_1, \dots, y_m$ are the *endogenous* variables, then the problem is: under what conditions will (2) define the endogenous variables as C^1 functions of the exogenous variables in a neighbourhood of $(\mathbf{x}^0, \mathbf{y}^0)$, and what happens to the endogenous variables when the exogenous variables are slightly changed? More specifically, what are the derivatives of $y_1, \dots, y_m$ with respect to $x_1, \dots, x_n$? Or, in vector form, what is $\mathbf{y}'_x$ (or $\partial \mathbf{y} / \partial \mathbf{x}$)?

If we assume that (2) defines $y_1, \dots, y_m$ as C^1 functions of $x_1, \dots, x_n$, then for each $j = 1, 2, \dots, n$, implicit differentiation of (2) w.r.t. x_j yields the m equations

$$\begin{aligned} \frac{\partial f_1(\mathbf{x}, \mathbf{y})}{\partial x_j} + \frac{\partial f_1(\mathbf{x}, \mathbf{y})}{\partial y_1} \frac{\partial y_1}{\partial x_j} + \dots + \frac{\partial f_1(\mathbf{x}, \mathbf{y})}{\partial y_m} \frac{\partial y_m}{\partial x_j} &= 0 \\ &\dots \dots \dots \\ \frac{\partial f_m(\mathbf{x}, \mathbf{y})}{\partial x_j} + \frac{\partial f_m(\mathbf{x}, \mathbf{y})}{\partial y_1} \frac{\partial y_1}{\partial x_j} + \dots + \frac{\partial f_m(\mathbf{x}, \mathbf{y})}{\partial y_m} \frac{\partial y_m}{\partial x_j} &= 0 \end{aligned} \quad (3)$$

If we move the first term in each equation to the right-hand side, all these mn equations can be written in the following matrix form:

$$\begin{pmatrix} \frac{\partial f_1(\mathbf{x}, \mathbf{y})}{\partial y_1} & \cdots & \frac{\partial f_1(\mathbf{x}, \mathbf{y})}{\partial y_m} \\ \vdots & & \vdots \\ \frac{\partial f_m(\mathbf{x}, \mathbf{y})}{\partial y_1} & \cdots & \frac{\partial f_m(\mathbf{x}, \mathbf{y})}{\partial y_m} \end{pmatrix} \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix} = - \begin{pmatrix} \frac{\partial f_1(\mathbf{x}, \mathbf{y})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x}, \mathbf{y})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\mathbf{x}, \mathbf{y})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x}, \mathbf{y})}{\partial x_n} \end{pmatrix} \quad (4)$$

This can be written even more compactly as $\mathbf{f}'_y(\mathbf{x}, \mathbf{y})\mathbf{y}'_x = -\mathbf{f}'_x(\mathbf{x}, \mathbf{y})$, where $\mathbf{f}'_y(\mathbf{x}, \mathbf{y})$ and $\mathbf{f}'_x(\mathbf{x}, \mathbf{y})$ are the $m \times m$ and $m \times n$ Jacobian matrices of $\mathbf{f}(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{y}$ and $\mathbf{x}$, respectively. From (4) we can obviously find the partials of $y_1, \dots, y_m$ with respect to $x_1, \dots, x_n$ provided $\mathbf{f}'_y(\mathbf{x}, \mathbf{y})$ is invertible.

The matrix $\mathbf{f}'_y(\mathbf{x}, \mathbf{y})$ is square. Its determinant is called the **Jacobian determinant** of $\mathbf{f}(\mathbf{x}, \mathbf{y})$ w.r.t. $\mathbf{y}$ and is commonly denoted by $\frac{\partial(f_1, \dots, f_m)}{\partial(y_1, \dots, y_m)}$.

In the argument leading to (4) we assumed that (2) defines $y_1, \dots, y_m$ as differentiable functions of $x_1, \dots, x_n$. The following theorem, one of the most important results in mathematical analysis, gives sufficient conditions for this to be the case:

THEOREM 2.7.2 (THE IMPLICIT FUNCTION THEOREM, GENERAL VERSION)

Suppose $\mathbf{f} = (f_1, \dots, f_m)$ is a C^1 function of $(\mathbf{x}, \mathbf{y})$ in an open set A in $\mathbb{R}^n \times \mathbb{R}^m$, and consider the m -dimensional vector equation $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$. Let $(\mathbf{x}^0, \mathbf{y}^0)$ be an interior point of A satisfying $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$. Suppose that the Jacobian determinant of $\mathbf{f}$ w.r.t. $\mathbf{y}$ is different from 0 at $(\mathbf{x}^0, \mathbf{y}^0)$, i.e.

$$|\mathbf{f}'_y(\mathbf{x}, \mathbf{y})| = \frac{\partial(f_1, \dots, f_m)}{\partial(y_1, \dots, y_m)} \neq 0 \quad \text{at } (\mathbf{x}, \mathbf{y}) = (\mathbf{x}^0, \mathbf{y}^0) \quad (5)$$

Then there exist open balls B_1 in $\mathbb{R}^n$ and B_2 in $\mathbb{R}^m$ around $\mathbf{x}^0$ and $\mathbf{y}^0$, respectively, with $B_1 \times B_2 \subseteq A$, such that $|\mathbf{f}'_y(\mathbf{x}, \mathbf{y})| \neq 0$ in $B_1 \times B_2$, and such that for each $\mathbf{x}$ in B_1 there is a unique $\mathbf{y}$ in B_2 with $\mathbf{f}(\mathbf{x}, \mathbf{y}) = \mathbf{0}$. In this way $\mathbf{y}$ is defined "implicitly" on B_1 as a C^1 function $\mathbf{g}(\mathbf{x})$ of $\mathbf{x}$. The Jacobian matrix $\mathbf{y}'_x = \mathbf{g}'(\mathbf{x}) = (\partial g_i(\mathbf{x})/\partial x_j)$ is

$$\mathbf{g}'(\mathbf{x}) = -(\mathbf{f}'_y(\mathbf{x}, \mathbf{y}))^{-1} \mathbf{f}'_x(\mathbf{x}, \mathbf{y}) \quad (6)$$

NOTE 1 Suppose $f_1, \dots, f_m$ are C^r functions of $(\mathbf{x}, \mathbf{y})$. Then the elements of the matrices in (6) are all C^{r-1} functions, and it follows that $\mathbf{g} = (g_1, \dots, g_m)$ is C^r .

Transformations and Their Inverses

Many economic applications involve functions that map points (vectors) in $\mathbb{R}^n$ to points (vectors) in $\mathbb{R}^m$. Such functions are often called **transformations** or **mappings**. For example, we are often interested in how an m -vector $\mathbf{y}$ of endogenous variables depends on an n -vector $\mathbf{x}$ of exogenous variables, as in the implicit function theorem.

Consider a transformation $\mathbf{f}: A \rightarrow B$ where $A \subseteq \mathbb{R}^n$ and $B \subseteq \mathbb{R}^m$. Suppose the range of $\mathbf{f}$ is the whole of B . Then we say that $\mathbf{f}$ maps A **onto** B . Recall that $\mathbf{f}$ is *one-to-one* if

$\mathbf{x}_1 \neq \mathbf{x}_2 \Rightarrow \mathbf{f}(\mathbf{x}_1) \neq \mathbf{f}(\mathbf{x}_2)$. In this case, for each point $\mathbf{y}$ in B there is exactly one point $\mathbf{x}$ in A such that $\mathbf{f}(\mathbf{x}) = \mathbf{y}$, and the **inverse** of $\mathbf{f}$ is the transformation $\mathbf{f}^{-1}: B \rightarrow A$ (note the order!) that maps each $\mathbf{y}$ in B to precisely that point $\mathbf{x}$ in A for which $\mathbf{f}(\mathbf{x}) = \mathbf{y}$.

When does a transformation $\mathbf{f}$ have an inverse? By definition, it has an inverse if and only if $\mathbf{f}$ is one-to-one, but this is often difficult to check directly. The problem is then to find useful conditions on $\mathbf{f}$ which ensure that the inverse exists.

Theorem 2.7.3 below gives a local solution to this problem for transformations from $\mathbb{R}^n$ into $\mathbb{R}^n$. Global solutions are much harder to come by. See Section 2.10 for some results in this direction.

THEOREM 2.7.3 (INVERSE FUNCTION THEOREM)

Consider a transformation $\mathbf{f} = (f_1, \dots, f_n)$ from $A \subseteq \mathbb{R}^n$ into $\mathbb{R}^n$ and assume that $\mathbf{f}$ is C^k ($k \geq 1$) in an open set containing $\mathbf{x}^0 = (x_1^0, \dots, x_n^0)$. Furthermore, suppose that the Jacobian determinant

$$|\mathbf{f}'(\mathbf{x})| = \frac{\partial(f_1, \dots, f_n)}{\partial(x_1, \dots, x_n)} \neq 0 \quad \text{at } \mathbf{x} = \mathbf{x}^0$$

Let $\mathbf{y}^0 = \mathbf{f}(\mathbf{x}^0)$. Then there exists an open set U around $\mathbf{x}^0$ such that $\mathbf{f}$ maps U one-to-one onto an open set V around $\mathbf{y}^0$, and there is an inverse mapping $\mathbf{g} = \mathbf{f}^{-1}: V \rightarrow U$ which is also C^k . Moreover, for all $\mathbf{y}$ in V , we have

$$\mathbf{g}'(\mathbf{y}) = (\mathbf{f}'(\mathbf{x}))^{-1}, \quad \text{where } \mathbf{x} = \mathbf{g}(\mathbf{y}) \in U$$

Proofs can be found in e.g. Marsden and Hoffman (1993) or Munkres (1991). Because $\mathbf{f}$, in general, is one-to-one only in a (possibly small) neighbourhood of $\mathbf{x}^0$, we say that Theorem 2.7.3 gives sufficient conditions for the existence of a *local* inverse. Briefly formulated:

A C^k transformation $\mathbf{f}$ from $\mathbb{R}^n$ into $\mathbb{R}^n$ with nonzero Jacobian determinant at $\mathbf{x}^0$ has a local inverse transformation around $\mathbf{f}(\mathbf{x}^0)$, and this inverse is also C^k . (7)

One implication of the theorem is that if $\mathbf{f}$ is C^1 in an open set around $\mathbf{x}^0$ and $|\mathbf{f}'(\mathbf{x}^0)| \neq 0$, then $\mathbf{f}$ is one-to-one in an open ball around $\mathbf{x}^0$. If the Jacobian determinant is different from 0 for all $\mathbf{x}$ in a set A , will $\mathbf{f}$ then be one-to-one in the whole of A ? If $n = 1$ and A is an interval in $\mathbb{R}$, this is true, but in general the answer is no, as shown by the next example.

EXAMPLE 3 Define the transformation $\mathbf{f}$ from $A = \{(x_1, x_2) : x_1^2 + x_2^2 \geq 1\} \subseteq \mathbb{R}^2$ into $\mathbb{R}^2$ by

$$\mathbf{f}(x_1, x_2) = (y_1, y_2), \quad \text{where } y_1 = x_1^2 - x_2^2, \quad y_2 = x_1 x_2$$

- Compute the Jacobian determinant $|\mathbf{f}'|$ of $\mathbf{f}$ and show that it is $\neq 0$ in the whole of A .
- What does $\mathbf{f}$ do to the points $(1, 1)$ and $(-1, -1)$?
- Comment on the results in (a) and (b).

Solution:

- (a) $|f'(x_1, x_2)| = \begin{vmatrix} \partial y_1 / \partial x_1 & \partial y_1 / \partial x_2 \\ \partial y_2 / \partial x_1 & \partial y_2 / \partial x_2 \end{vmatrix} = \begin{vmatrix} 2x_1 & -2x_2 \\ x_2 & x_1 \end{vmatrix} = 2(x_1^2 + x_2^2) \neq 0$ for all (x_1, x_2) in A .
- (b) Both $(x_1, x_2) = (1, 1)$ and $(x_1, x_2) = (-1, -1)$ are mapped to $(y_1, y_2) = (0, 1)$.
- (c) Even though $|f'| \neq 0$ in all of A , f is not one-to-one in all of A .

A More General Case

The theorem on inverse functions deals with transformations from a subset of $\mathbb{R}^n$ into $\mathbb{R}^n$. Now, consider a more general situation, in which f is a C^1 transformation from a subset of $\mathbb{R}^n$ into $\mathbb{R}^m$ with $m \leq n$, and let U be a neighbourhood of a point $\mathbf{x}^0$ in $\mathbb{R}^n$. Then U will contain a ball B centred at $\mathbf{x}^0$ with a positive radius r . The set $f(U) = \{f(\mathbf{x}) : \mathbf{x} \in U\}$ will contain $f(\mathbf{x}^0) = \mathbf{y}^0$. Is $\mathbf{y}^0$ an interior point of $f(U)$? Not necessarily. For example, if $f: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is defined by $f(x, y, z) = (x + y + z, x + y + z)$, then f maps $\mathbb{R}^3$ into a straight line through $\mathbf{y}^0 = (0, 0)$, which certainly does not contain a neighbourhood of $\mathbf{y}^0$. Thus the mapping collapses $\mathbb{R}^3$ into a set of lower dimension than the target set $\mathbb{R}^2$ (the image $f(\mathbb{R}^3)$ is a straight line in this case). The final theorem of this section tells us that such a collapse is impossible if the Jacobian matrix of f has maximal rank at $\mathbf{x}^0$. (Note that the Jacobian of $f(x, y) = (x + y + z, x + y + z)$ is the matrix $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$, which has rank 1.)

THEOREM 2.7.4 (GENERAL CASE)

Suppose f is a transformation from an open subset A of $\mathbb{R}^n$ into $\mathbb{R}^m$, and $m \leq n$. If f is C^1 in a neighbourhood of a point $\mathbf{x}^0$ in A , and the Jacobian matrix $F'(\mathbf{x})$ has rank m at $\mathbf{x}^0$, then $f(\mathbf{x}^0)$ is an interior point of $f(U)$ for any open neighbourhood U of $\mathbf{x}^0$.

A proof can be found in e.g. Marsden and Hoffman (1993) or Munkres (1991). In fact, this theorem and the implicit function theorem (Theorem 2.7.2) are intimately related. Once either of them is proved, the other follows quite easily.

Linear transformations

Linear transformations constitute a simple but very important class of transformations. A transformation $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is called linear if

$$f(\mathbf{x}_1 + \mathbf{x}_2) = f(\mathbf{x}_1) + f(\mathbf{x}_2), \quad f(\alpha \mathbf{x}_1) = \alpha f(\mathbf{x}_1) \quad (8)$$

for all $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbb{R}^n$ and all scalars α . A well-known result from linear algebra states that for every linear transformation $f: \mathbb{R}^n \rightarrow \mathbb{R}^m$ there is a unique $m \times n$ matrix A such that

$f(\mathbf{x}) = A\mathbf{x}$ for all $\mathbf{x}$ in $\mathbb{R}^n$. Indeed, the j th column $\mathbf{a}_j = (a_{1j}, \dots, a_{mj})'$ of A must be $f(\mathbf{e}_j)$, where $\mathbf{e}_j = (0, \dots, 1, \dots, 0)'$ is the j th standard unit vector in $\mathbb{R}^n$, that is, the n -vector with 1 in the j th component and 0 elsewhere.

Conversely, it is clear that for any $m \times n$ matrix A , the mapping $\mathbf{x} \mapsto A\mathbf{x}$ is a linear transformation $\mathbb{R}^n \rightarrow \mathbb{R}^m$. Indeed, the rules for matrix multiplication show that

$$A(\mathbf{x}_1 + \mathbf{x}_2) = A\mathbf{x}_1 + A\mathbf{x}_2, \quad A(\alpha \mathbf{x}_1) = \alpha A\mathbf{x}_1 \quad (9)$$

This demonstrates a one-to-one correspondence between linear transformations $\mathbb{R}^n \rightarrow \mathbb{R}^m$ and $m \times n$ matrices. (When vectors are considered as matrices, they are usually taken to be column vectors—that is, matrices with a single column. The matrix product $A\mathbf{x}$ then makes sense, and yields a column vector.)

PROBLEMS FOR SECTION 2.7

- Show that the following equations define y implicitly as a function of x in an interval around x_0 . Find y' when $x = x_0$.
(a) $f(x, y) = y^3 + y - x^3 = 0, \quad x_0 = 0$ (b) $f(x, y) = x^2 + y + \sin(xy) = 0, \quad x_0 = 0$
- Check if the following equations can be represented in the form $z = g(x, y)$ in a neighbourhood of the given point (x_0, y_0, z_0) . Compute $g'_1(x_0, y_0)$ and $g'_2(x_0, y_0)$.
(a) $F(x, y, z) = x^3 + y^3 + z^3 - xyz - 1 = 0, \quad (x_0, y_0, z_0) = (0, 0, 1)$
(b) $F(x, y, z) = e^z - z^2 - x^2 - y^2 = 0, \quad (x_0, y_0, z_0) = (1, 0, 0)$
- The point $P = (x, y, z, u, v, w) = (1, 1, 0, -1, 0, 1)$ satisfies all the equations

$$\begin{aligned} y^2 - z + u - v - w^3 &= -1 \\ -2x + y - z^2 + u + v^3 - w &= -3 \\ x^2 + z - u - v + w^3 &= 3 \end{aligned}$$

These equations define u, v, w as C^1 functions of x, y, z around P . Find u'_x, v'_x , and w'_x at P .

- Suppose the functions f and g are defined in $\mathbb{R}^2$ by $f(u, v) = e^u \cos v, g(u, v) = e^u \sin v$. Show that the Jacobian determinant $\partial(f, g)/\partial(u, v)$ of this transformation is different from 0 everywhere. How many solutions are there to the following two systems of equations?
(a) $e^u \cos v = 0, \quad e^u \sin v = 0$ (b) $e^u \cos v = 1, \quad e^u \sin v = 1$
- Suppose (x_0, y_0, u_0, v_0) satisfies the two equations

$$\begin{aligned} F(x, y, u, v) &= x^2 - y^2 + uv - v^2 + 3 = 0 \\ G(x, y, u, v) &= x + y^2 + u^2 + uv - 2 = 0 \end{aligned}$$

State conditions that are sufficient for this system to be represented by two equations $u = f(x, y), v = g(x, y)$, where f and g are C^1 , in a neighbourhood of this point. Show that

such a representation is possible when $(x_0, y_0, u_0, v_0) = (2, 1, -1, 2)$, and compute $f'_x(2, 1)$, $f'_y(2, 1)$, $g'_x(2, 1)$, and $g'_y(2, 1)$.

- 5M 6. Consider the transformation from $\mathbb{R}^2$ to $\mathbb{R}^2$ determined by $(x_1, x_2) \mapsto (y_1, y_2)$, where

$$y_1 = x_1 - x_1x_2, \quad y_2 = x_1x_2 \quad (*)$$

Compute the Jacobian determinant of this transformation, and find the inverse (where it exists) by solving the system of equations in $(*)$ for x_1 and x_2 . Examine what the transformation does to the rectangle determined by $1 \leq x_1 \leq 2$, $1/2 \leq x_2 \leq 2/3$. Draw a figure!

7. Consider the linear transformation $T: (x, y) \mapsto (u, v)$ from $\mathbb{R}^2$ to $\mathbb{R}^2$ determined by $u = ax + by$, $v = cx + dy$, where a, b, c , and d are constants, not all equal to 0. Suppose the Jacobian determinant of T is 0. Then, show that T maps the whole of $\mathbb{R}^2$ onto a straight line through the origin of the uv -plane.

- 5M 8. Consider the transformation $T: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined by $T(r, \theta) = (r \cos \theta, r \sin \theta)$.

(a) Compute the Jacobian determinant J of T .

(b) Let A be the domain in the (r, θ) plane determined by $1 \leq r \leq 2$ and $\theta \in [0, k]$, where $k > 2\pi$. Show that $J \neq 0$ in the whole of A , yet T is not one-to-one in A .

9. Give sufficient conditions on f and g to ensure that the equations

$$u = f(x, y), \quad v = g(x, y)$$

can be solved for x and y locally. Show that if the solutions are $x = F(u, v)$, $y = G(u, v)$, and if f, g, F , and G are C^1 , then

$$\frac{\partial F}{\partial u} = \frac{1}{J} \frac{\partial g}{\partial y}, \quad \frac{\partial G}{\partial u} = -\frac{1}{J} \frac{\partial g}{\partial x}$$

where J denotes the Jacobian determinant of f and g w.r.t. x and y .

- 5M 10. (a) Consider the system of equations

$$\begin{aligned} 1 + (x + y)u - (2 + u)^{1+u} &= 0 \\ 2u - (1 + xy)e^{u(x-1)} &= 0 \end{aligned}$$

Use Theorem 2.7.2 to show that the system defines u and v as functions of x and y in an open ball around $(x, y, u, v) = (1, 1, 1, 0)$. Find the values of the partial derivatives of the two functions w.r.t. x when $x = 1$, $y = 1$, $u = 1$, $v = 0$.

- (b) Let a and b be arbitrary numbers in the interval $[0, 1]$. Use the intermediate value theorem (see e.g. EMEA, Section 7.10) to show that the equation

$$u - ae^{u(b-1)} = 0$$

has a solution in the interval $[0, 1]$. Is the solution unique?

- (c) Show by using (b) that for any point (x, y) , $x \in [0, 1]$, $y \in [0, 1]$, there exist solutions u and v of the system. Are u and v uniquely determined?

2.8 Degrees of Freedom and Functional Dependence

A system of equations with more variables than equations will in general have many solutions. Usually, the larger the difference between the number of variables and the number of equations, the larger the set of solutions. In general, a system of equations in n variables is said to have k degrees of freedom if there is a set of k variables that can be freely chosen, while the remaining $n - k$ variables are uniquely determined once the k free variables have been assigned specific values. Thus, the system must define $n - k$ of the variables as functions of the remaining k free variables. If the n variables are restricted to vary in a set A in $\mathbb{R}^n$, we say that the system has k degrees of freedom in A .

For an equation system to have k degrees of freedom, it suffices that there exist k variables that can be freely chosen. We do not require that any set of k variables can be chosen freely.

A rough rule can be used for a preliminary estimate of the number of degrees of freedom for a system of equation. This is called the "counting rule":

ROUGH COUNTING RULE

To find the number of degrees of freedom for a system of equations, count the number of variables, n , and the number of "independent" equations, m . If $n > m$, there are $n - m$ degrees of freedom in the system.

(1)

This rule lies behind the following economic proposition: "The number of independent targets that a government can pursue cannot possibly exceed the number of available policy instruments." For example, assuming that the targets of price stability, low unemployment, and stable exchange rates are independent, a government seeking to meet all three simultaneously needs at least three independent policy instruments.

It is easy to give examples where the counting rule fails, and it is obvious that the word "independent" cannot be dropped from the statement of the counting rule. For instance, if we just add one equation which repeats one that has appeared before, the number of degrees of freedom will certainly not be reduced.

NOTE 1 If we have a linear system $Ax = b$ of m equations in n unknowns, then according to Theorem 1.4.1, the system has a solution iff the rank of the coefficient matrix A is equal to the rank of the augmented matrix A_b . In this case, the counting rule gives the correct result iff the m row vectors in A are linearly independent, because then both A and A_b have rank m . (See Theorem 1.4.2(b).) So for the counting rule to apply in the case of linear systems, the row vectors of the coefficient matrix must be linearly independent.

The implicit function theorem gives us a precise local counting rule for the system

$$f_1(x_1, \dots, x_n) = 0$$

$$\dots \dots \dots \text{or, in vector notation } \mathbf{f}(\mathbf{x}) = 0$$

(2)

$$f_m(x_1, \dots, x_n) = 0$$

in the case $n > m$. Suppose that $f_1, \dots, f_m$ are C^1 functions in a neighbourhood of a solution $\mathbf{x}^0$, and suppose that the Jacobian matrix $\mathbf{f}'(\mathbf{x})$ has rank m at $\mathbf{x}^0$. This implies that for some selection of m of the variables x_i , which we denote by $x_{i_1}, x_{i_2}, \dots, x_{i_m}$, the Jacobian determinant

$$\frac{\partial(f_1, f_2, \dots, f_m)}{\partial(x_{i_1}, x_{i_2}, \dots, x_{i_m})}$$

is not 0 at $\mathbf{x}^0$. Then an easy modification of Theorem 2.7.2 shows that there exists a ball B around $\mathbf{x}^0$ in which system (1) defines $x_{i_1}, x_{i_2}, \dots, x_{i_m}$ as functions of the other $n - m$ variables. Then system (1) has, by definition, $n - m$ degrees of freedom in B . Thus we have the following result:

CORRECT COUNTING RULE

If $\mathbf{x}^0$ is a solution of system (2) and the Jacobian matrix $\mathbf{f}'(\mathbf{x})$ has rank m at $\mathbf{x}^0$, then there exists a ball B around $\mathbf{x}^0$ such that the system has $n - m$ degrees of freedom in B . In this case the counting rule yields the correct result. (3)

Functional Dependence

In formulating the counting rule we assumed that the equations were independent. Consider the system of equations (2) and suppose there exists a function G such that $f_m(\mathbf{x}) \equiv G(f_1(\mathbf{x}), \dots, f_{m-1}(\mathbf{x}))$. Then $f_m(\mathbf{x})$ is *functionally dependent* on $f_1, \dots, f_{m-1}$. More symmetrically, for the functions $f_1, \dots, f_m$ to be functionally dependent in a set A , it is required that there exists a function F of m variables such that

$$F(f_1(\mathbf{x}), \dots, f_m(\mathbf{x})) = 0 \quad \text{for all } \mathbf{x} \text{ in } A \quad (4)$$

If $F \equiv 0$, then (4) is satisfied regardless of the functions $f_1, \dots, f_m$, so some additional requirements on F are needed. The following definition is the standard one:

DEFINITION OF FUNCTIONAL DEPENDENCE

The functions $f_1, \dots, f_m$ are **functionally dependent** in A if there exists a C^1 function F on $\mathbb{R}^m$ that satisfies (4) and, for some j , has $F'_j \neq 0$ everywhere. (5)

With these conditions imposed on F , it is always possible (at least locally) to solve the equation $F(y_1, \dots, y_m) = 0$ for y_j and get y_j as a function of the other variables y_i , i.e. $f_j(\mathbf{x})$ can be expressed as a function of the other $f_i(\mathbf{x})$ for $\mathbf{x} \in A$.

The concept of functional dependence helps determine when there are superfluous equations. Suppose F is a C^1 function and that $f_1, \dots, f_m$ and F together satisfy (4) with

$A = \mathbb{R}^n$. Suppose, in addition, that $F(0, \dots, 0) = 0$ and $F'_j \neq 0$ everywhere. Then the equation $F(0, \dots, 0, f_j(\mathbf{x}), 0, \dots, 0) = 0$ implies that $f_j(\mathbf{x}) = 0$, i.e. $f_j(\mathbf{x}) = 0$ for all $i \neq j$ implies that $f_j(\mathbf{x}) = 0$ also holds. In other words: if $\mathbf{x}$ satisfies all the equations except the j th, then $\mathbf{x}$ automatically also satisfies the j th equation. In this case, the j th equation is superfluous. Hence:

Suppose the equation system (2) has solutions and that $f_1(\mathbf{x}), \dots, f_m(\mathbf{x})$ are functionally dependent. Then the system contains at least one superfluous equation.

NOTE 2 Consider system (2) and suppose the C^1 functions $f_1, \dots, f_m$ are functionally dependent in $\mathbb{R}^n$ according to the definition (5). Because there is a superfluous equation, the counting rule fails. Then, according to (3), the Jacobian matrix $\mathbf{f}'(\mathbf{x})$ cannot have rank m at any solution point. By differentiating (4) w.r.t. $\mathbf{x}$ it follows from the chain rule in matrix form that $\nabla F(\mathbf{f}(\mathbf{x}))\mathbf{f}'(\mathbf{x}) = 0$. Because $\nabla F(\mathbf{f}(\mathbf{x})) \neq 0$, the rows in the $m \times m$ Jacobian matrix $\mathbf{f}'(\mathbf{x})$ are linearly dependent, so the rank of $\mathbf{f}'(\mathbf{x})$ must be less than m .

EXAMPLE 1 Show that $f(x, y) = e^y(1+x^2)$ and $g(x, y) = \ln(1+x^2) + y$ are functionally dependent in $\mathbb{R}^2$.

Solution: Put $F(y_1, y_2) = \ln y_1 - y_2$. Then $F(f(x, y), g(x, y)) = y + \ln(1+x^2) - \ln(1+x^2) - y \equiv 0$ for all x, y . Moreover, $F'_2(y_1, y_2) = -1 \neq 0$, so f and g are functionally dependent. $\square$

Local Functional Dependence

A property implied by functional dependence is local functional dependence, defined as follows:

DEFINITION OF LOCAL FUNCTIONAL DEPENDENCE

The functions $f_1, \dots, f_m$ are **locally functionally dependent** in an open set A if for each $\mathbf{x}_0$ in A there exists an open ball $B(\mathbf{x}_0; \varepsilon) \subseteq A$ such that $f_1, \dots, f_m$ are functionally dependent in $B(\mathbf{x}_0; \varepsilon)$. (6)

With this concept of local functional dependence we record the following two theorems, whose proofs are given on the book's website.

THEOREM 2.8.1 (LOCAL FUNCTIONAL DEPENDENCE)

Let $\mathbf{f} = (f_1, \dots, f_m)$ be a C^1 transformation defined in an open set A in $\mathbb{R}^n$. If the Jacobian matrix $\mathbf{f}'(\mathbf{x})$ has constant rank $r < m$ in A , then $f_1, \dots, f_m$ are locally functionally dependent.

THEOREM 2.8.2 (THE RANK THEOREM)

Suppose that the m functions $f_1, \dots, f_m$ are defined and C^1 in an open set A in $\mathbb{R}^n$. Assume that the Jacobian matrix of these functions has constant rank $r < m$ in A , and let $\mathbf{x}^0$ be a point in A . Then there exist an open ball $B(\mathbf{x}^0; \varepsilon) \subseteq A$ together with r functions $f_{i_1}, \dots, f_{i_r}$ selected from $f_1, \dots, f_m$, and $m - r$ functions H_j , $j \notin \{i_1, \dots, i_r\}$, which are all C^1 on a suitable subset of $\mathbb{R}^r$, such that $f_j(\mathbf{x}) = H_j(f_{i_1}(\mathbf{x}), \dots, f_{i_r}(\mathbf{x}))$ for all $j \notin \{i_1, \dots, i_r\}$ and all $\mathbf{x}$ in $B(\mathbf{x}^0; \varepsilon)$.

PROBLEMS FOR SECTION 2.8

501. (a) Consider the macroeconomic model described by the system of equations

$$(i) Y = C + I + G, \quad (ii) C = f(Y - T), \quad (iii) I = h(r), \quad (iv) r = m(M)$$

where f , h , and m are given C^1 functions. According to the counting rule, how many degrees of freedom has this system?

(b) Give sufficient conditions for the system to determine Y , C , I , and r as functions of the exogenous policy variables M , T , and G in a neighbourhood of an equilibrium point.

502. (a) Consider the two pairs of functions $u = f(x, y)$, $v = g(x, y)$ given by

$$(i) u = e^{x+y}, \quad v = 2x^2 + 4xy + 2y^2 - x - y \quad (ii) u = \frac{x}{y}, \quad v = \frac{y-x}{y+x}$$

Show that for each pair the Jacobian determinant $\partial(u, v)/\partial(x, y) \neq 0$ for all (x, y) where u and v are defined.

(b) Find a functional dependence between u and v in each case. (Hint: Solve the equation $u = f(x, y)$ for x and put the result into $v = g(x, y)$.)

503. Let $u = f(x, y)$, $v = g(x, y)$ and suppose that $\partial(u, v)/\partial(x, y) \neq 0$ for all (x, y) in A . Furthermore, suppose $\partial f/\partial x \neq 0$ at $(x_0, y_0) \in A$. Show that under suitable continuity conditions, f and g are functionally dependent in a ball around (x_0, y_0) . (Hint: $u = f(x, y)$ yields $x = \varphi(y, u)$ because $\partial f/\partial x \neq 0$. Hence, $v = g(\varphi(y, u), y)$. Show that $\partial v/\partial y = 0$, so that $g(\varphi(y, u), y)$ is independent of y , and hence $v = g(\varphi(y, u), y) = \psi(u)$.)

4. Suppose that u , v , and w are defined as C^1 functions of x , y , and z by the three equations

$$u = x + y - z, \quad v = x - y + z, \quad w = x^2 + y^2 + z^2 - 2yz$$

(a) Show that $\partial(u, v, w)/\partial(x, y, z) = 0$ for all (x, y, z) .

(b) Show that u , v , and w are functionally dependent.

2.9 Differentiability

Recall that if a one-variable function f is differentiable at a point a , then the linear approximation to f around a is given by

$$f(a+h) \approx f(a) + f'(a)h \quad (\text{for small values of } h)$$

This approximation is useful because the approximation error defined by

$$R(h) = \text{true value} - \text{approximate value} = f(a+h) - f(a) - f'(a)h$$

becomes negligible for sufficiently small h . Of course, $R(h)$ becomes small in the trivial sense that $R(h) \rightarrow 0$ as $h \rightarrow 0$. More importantly, however, $R(h)$ also becomes small in comparison with h —that is,

$$\lim_{h \rightarrow 0} \frac{R(h)}{h} = \lim_{h \rightarrow 0} \left(\frac{f(a+h) - f(a)}{h} - f'(a) \right) = 0$$

In fact, f is differentiable at a if and only if there exists a number c such that

$$\lim_{h \rightarrow 0} \frac{f(a+h) - f(a) - ch}{h} = 0$$

If such a c exists, it is unique and $c = f'(a)$.

These one-dimensional concepts admit straightforward generalizations to many dimensions. In particular, a transformation $\mathbf{f}$ is differentiable at a point $\mathbf{a}$ if it admits a linear approximation around $\mathbf{a}$:

DEFINITION OF DIFFERENTIABILITY AND DERIVATIVES

If $\mathbf{f} : A \rightarrow \mathbb{R}^m$ is a transformation defined on a subset A of $\mathbb{R}^n$ and $\mathbf{a}$ is an interior point of A , then $\mathbf{f}$ is said to be differentiable at $\mathbf{a}$ if there exists an $m \times n$ matrix $\mathbf{C}$ such that

$$\lim_{\mathbf{h} \rightarrow 0} \frac{\|\mathbf{f}(\mathbf{a} + \mathbf{h}) - \mathbf{f}(\mathbf{a}) - \mathbf{C}\mathbf{h}\|}{\|\mathbf{h}\|} = 0 \quad (1)$$

If such a matrix $\mathbf{C}$ exists, it is called the (total) derivative of $\mathbf{f}$ at $\mathbf{a}$, and is denoted by $\mathbf{f}'(\mathbf{a})$. An alternative notation for $\mathbf{f}'(\mathbf{a})$ is $D\mathbf{f}(\mathbf{a})$.

If $\mathbf{f}$ has a derivative, then the derivative is unique (in fact, it must equal the Jacobian matrix of $\mathbf{f}$ at $\mathbf{a}$, see below), so we are justified in speaking of the derivative. As in the one-dimensional case, if $\mathbf{f}$ is differentiable at $\mathbf{a}$, then the linear transformation $\mathbf{h} \mapsto \mathbf{f}'(\mathbf{a})\mathbf{h}$ is a good approximation to $\mathbf{h} \mapsto \mathbf{f}(\mathbf{a} + \mathbf{h}) - \mathbf{f}(\mathbf{a})$ for sufficiently small $\mathbf{h}$, and the approximation $\mathbf{f}(\mathbf{a} + \mathbf{h}) \approx \mathbf{f}(\mathbf{a}) + \mathbf{f}'(\mathbf{a})\mathbf{h}$, or, equivalently,

$$\mathbf{f}(\mathbf{x}) \approx \mathbf{f}(\mathbf{a}) + \mathbf{f}'(\mathbf{a})(\mathbf{x} - \mathbf{a})$$

is called the **linear (or first-order) approximation to f around $\mathbf{a}$** .

Several questions arise: how can we tell whether a transformation f is differentiable at a point $\mathbf{x}$; how do we find $f'(\mathbf{x})$; and what properties do derivatives have?

Consider first the special case $m = 1$, so f is an "ordinary" (one-dimensional) function of n variables. It turns out that if f is differentiable at $\mathbf{x}$, then f has a derivative $f'_a(\mathbf{x})$ along every vector $\mathbf{a}$, and these derivatives are all determined by the derivative $f'(\mathbf{x})$ of f at $\mathbf{x}$.

THEOREM 2.9.1

If $f : A \rightarrow \mathbb{R}$ is defined on a subset A of $\mathbb{R}^n$ and f is differentiable at an interior point $\mathbf{x}$ of A , then f has a derivative $f'_a(\mathbf{x})$ along every n -vector $\mathbf{a}$, and $f'_a(\mathbf{x}) = f'(\mathbf{x})\mathbf{a}$. (Remember that $f'(\mathbf{x})$ is a $1 \times n$ matrix.)

Proof: From the definition in Section 2.1, the derivative along $\mathbf{a}$ is

$$f'_a(\mathbf{x}) = \lim_{h \rightarrow 0} \left(\frac{f(\mathbf{x} + h\mathbf{a}) - f(\mathbf{x}) - f'(\mathbf{x})\mathbf{a}h}{h} + f'(\mathbf{x})\mathbf{a} \right) = 0 + f'(\mathbf{x})\mathbf{a} \quad \blacksquare$$

In particular, if $\mathbf{e}_j = (0, \dots, 1, \dots, 0)^T$ is the j th standard unit vector in $\mathbb{R}^n$, then $f'_{\mathbf{e}_j}(\mathbf{x}) = f'(\mathbf{x})\mathbf{e}_j$ is the partial derivative $f'_j(\mathbf{x})$ of f with respect to the j th variable. On the other hand, $f'(\mathbf{x})\mathbf{e}_j$ is the j th component of $f'(\mathbf{x})$. Hence, $f'(\mathbf{x})$ is the row vector

$$f'(\mathbf{x}) = (f'(\mathbf{x})\mathbf{e}_1, \dots, f'(\mathbf{x})\mathbf{e}_n) = (f'_1(\mathbf{x}), \dots, f'_n(\mathbf{x}))$$

which we recognize as the gradient $\nabla f(\mathbf{x})$ of f at $\mathbf{x}$.⁴ (See Section 2.1.)

We are now prepared to tackle the case of transformations into $\mathbb{R}^m$.

THEOREM 2.9.2

A transformation $\mathbf{f} = (f_1, \dots, f_m)$ from a subset A of $\mathbb{R}^n$ into $\mathbb{R}^m$ is differentiable at an interior point $\mathbf{x}$ of A if and only if each component function $f_i : A \rightarrow \mathbb{R}$, $i = 1, \dots, m$, is differentiable at $\mathbf{x}$. Moreover,

$$\mathbf{f}'(\mathbf{x}) = \begin{pmatrix} f'_1(\mathbf{x}) \\ \vdots \\ f'_m(\mathbf{x}) \end{pmatrix}$$

is the $m \times n$ matrix whose i th row is $f'_i(\mathbf{x}) = \nabla f_i(\mathbf{x})$.

⁴ The gradient $\nabla f(\mathbf{x})$ of a function is thus an exception to the rule that vectors (regarded as matrices) are usually taken to be column vectors.

Proof: Let $\mathbf{C}$ be an $m \times n$ matrix and let $\mathbf{R}(\mathbf{h}) = \mathbf{f}(\mathbf{x} + \mathbf{h}) - \mathbf{f}(\mathbf{x}) - \mathbf{C}\mathbf{h}$. For each $i = 1, \dots, m$, the i th component of $\mathbf{R}(\mathbf{h})$ is $R_i(\mathbf{h}) = f_i(\mathbf{x} + \mathbf{h}) - f_i(\mathbf{x}) - C_i\mathbf{h}$, where C_i is the i th row of $\mathbf{C}$. But

$$|R_i(\mathbf{h})| \leq \|\mathbf{R}(\mathbf{h})\| \leq |R_1(\mathbf{h})| + \dots + |R_m(\mathbf{h})|$$

for each i . It follows that

$$\lim_{\mathbf{h} \rightarrow 0} \frac{\|\mathbf{R}(\mathbf{h})\|}{\|\mathbf{h}\|} = 0 \iff \lim_{\mathbf{h} \rightarrow 0} \frac{|R_i(\mathbf{h})|}{\|\mathbf{h}\|} = 0 \text{ for all } i = 1, \dots, m$$

Hence, $\mathbf{f}$ is differentiable at $\mathbf{x}$ if and only if each f_i is differentiable at $\mathbf{x}$. Also, the i th row of the matrix $\mathbf{C} = \mathbf{f}'(\mathbf{x})$ is the derivative of f_i , that is $C_i = \nabla f_i(\mathbf{x})$. $\blacksquare$

We see that if $\mathbf{f}$ is differentiable at $\mathbf{x}$, then its derivative is the matrix

$$\mathbf{f}'(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

We have previously called $\mathbf{f}'(\mathbf{x})$ the **Jacobian matrix of $\mathbf{f}$ at $\mathbf{x}$** . Its rows are the gradients of the component functions of $\mathbf{f}$.

We know that if a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is differentiable at a point a , then it is continuous at a . A similar result holds in the multidimensional case.

THEOREM 2.9.3

If a transformation $\mathbf{f}$ from $A \subseteq \mathbb{R}^n$ into $\mathbb{R}^m$ is differentiable at an interior point $\mathbf{a}$ of A , then $\mathbf{f}$ is continuous at $\mathbf{a}$.

Proof: Let $\mathbf{C} = \mathbf{f}'(\mathbf{a})$. Then for small but nonzero $\mathbf{h}$, the triangle inequality yields

$$\begin{aligned} \|\mathbf{f}(\mathbf{a} + \mathbf{h}) - \mathbf{f}(\mathbf{a})\| &\leq \|\mathbf{f}(\mathbf{a} + \mathbf{h}) - \mathbf{f}(\mathbf{a}) - \mathbf{C}\mathbf{h}\| + \|\mathbf{C}\mathbf{h}\| \\ &= \|\mathbf{h}\| \left(\frac{\|\mathbf{f}(\mathbf{a} + \mathbf{h}) - \mathbf{f}(\mathbf{a}) - \mathbf{C}\mathbf{h}\|}{\|\mathbf{h}\|} \right) + \|\mathbf{C}\mathbf{h}\| \end{aligned}$$

Because $\mathbf{C} = \mathbf{f}'(\mathbf{a})$, the fraction in the parentheses tends to 0 as $\mathbf{h} \rightarrow 0$, and the term $\|\mathbf{C}\mathbf{h}\|$ also tends to 0. Hence, $\mathbf{f}(\mathbf{a} + \mathbf{h}) \rightarrow \mathbf{f}(\mathbf{a})$ as $\mathbf{h} \rightarrow 0$. $\blacksquare$

If $\mathbf{f}$ and $\mathbf{g}$ are transformations from $A \subseteq \mathbb{R}^n$ into $\mathbb{R}^m$, and if they are both differentiable at a point $\mathbf{x}$ in A , then the following rules hold (α is a constant scalar):

$$(\alpha \mathbf{f})'(\mathbf{x}) = \alpha \mathbf{f}'(\mathbf{x}), \quad (\mathbf{f} + \mathbf{g})'(\mathbf{x}) = \mathbf{f}'(\mathbf{x}) + \mathbf{g}'(\mathbf{x})$$

There is also a *chain rule* for transformations:

THEOREM 2.9.4 (THE CHAIN RULE)

Suppose $f: A \rightarrow \mathbb{R}^m$ and $g: B \rightarrow \mathbb{R}^p$ are defined on $A \subseteq \mathbb{R}^n$ and $B \subseteq \mathbb{R}^m$, with $f(A) \subseteq B$. Suppose also that f and g are differentiable at $\mathbf{x}$ and $f(\mathbf{x})$, respectively. Then the composite transformation $g \circ f: A \rightarrow \mathbb{R}^p$ defined by $(g \circ f)(\mathbf{x}) = g(f(\mathbf{x}))$ is differentiable at $\mathbf{x}$, and

$$(g \circ f)'(\mathbf{x}) = g'(f(\mathbf{x})) f'(\mathbf{x}) \quad (2)$$

Proof: A heuristic derivation of formula (2) using linear approximations is

$$\begin{aligned} (g \circ f)(\mathbf{x} + \mathbf{h}) - (g \circ f)(\mathbf{x}) &= g(f(\mathbf{x} + \mathbf{h})) - g(f(\mathbf{x})) \\ &\approx g'(f(\mathbf{x})) [f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x})] \approx g'(f(\mathbf{x})) f'(\mathbf{x}) \mathbf{h} \end{aligned}$$

To go beyond a heuristic explanation and prove the theorem rigorously, one must show that the error $\mathbf{e}(\mathbf{h}) = (g \circ f)(\mathbf{x} + \mathbf{h}) - (g \circ f)(\mathbf{x}) - g'(f(\mathbf{x})) f'(\mathbf{x}) \mathbf{h}$ involved in this approximation satisfies $\|\mathbf{e}(\mathbf{h})\|/\|\mathbf{h}\| \rightarrow 0$ as $\mathbf{h} \rightarrow \mathbf{0}$.

Define $\mathbf{k}(\mathbf{h}) = f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) = f'(\mathbf{x})\mathbf{h} + \mathbf{e}_f(\mathbf{h})$, where $\|\mathbf{e}_f(\mathbf{h})\|/\|\mathbf{h}\| \rightarrow 0$ as $\mathbf{h} \rightarrow \mathbf{0}$ because f is differentiable at $\mathbf{x}$. Similarly $g(f(\mathbf{x} + \mathbf{k})) - g(f(\mathbf{x})) = g'(f(\mathbf{x}))\mathbf{k} + \mathbf{e}_g(\mathbf{k})$, where $\|\mathbf{e}_g(\mathbf{k})\|/\|\mathbf{k}\| \rightarrow 0$ as $\mathbf{k} \rightarrow \mathbf{0}$. Note that $\|\mathbf{k}(\mathbf{h})\| \leq \|f'(\mathbf{x})\mathbf{h}\| + \|\mathbf{e}_f(\mathbf{h})\| \leq K\|\mathbf{h}\|$ for all small $\mathbf{h}$, with K some fixed constant. Observe also that for all $\varepsilon > 0$, $\|\mathbf{e}_g(\mathbf{k})\| < \varepsilon\|\mathbf{k}\|$ for $\mathbf{k}$ small, so $\|\mathbf{e}_g(\mathbf{k}(\mathbf{h}))\| < \varepsilon\|\mathbf{k}(\mathbf{h})\| \leq \varepsilon K\|\mathbf{h}\|$ when $\mathbf{h}$ is small. Hence, $\|\mathbf{e}_g(\mathbf{k}(\mathbf{h}))\|/\|\mathbf{h}\| \rightarrow 0$ as $\mathbf{h} \rightarrow \mathbf{0}$. Then

$$\begin{aligned} \mathbf{e}(\mathbf{h}) &= g(f(\mathbf{x} + \mathbf{k})) - g(f(\mathbf{x})) - g'(f(\mathbf{x})) f'(\mathbf{x}) \mathbf{h} \\ &= g'(f(\mathbf{x})) \mathbf{k}(\mathbf{h}) + \mathbf{e}_g(\mathbf{k}(\mathbf{h})) - g'(f(\mathbf{x})) f'(\mathbf{x}) \mathbf{h} \\ &= g'(f(\mathbf{x})) \mathbf{e}_f(\mathbf{h}) + \mathbf{e}_g(\mathbf{k}(\mathbf{h})) \end{aligned}$$

Hence $\|\mathbf{e}(\mathbf{h})\|/\|\mathbf{h}\| \leq \|g'(f(\mathbf{x}))\mathbf{e}_f(\mathbf{h})\|/\|\mathbf{h}\| + \|\mathbf{e}_g(\mathbf{k}(\mathbf{h}))\|/\|\mathbf{h}\|$. The right-hand side converges to zero as $\mathbf{h} \rightarrow \mathbf{0}$. ■

The Jacobian matrices $g'(f(\mathbf{x}))$, $f'(\mathbf{x})$, and $(g \circ f)'(\mathbf{x}_0)$ are $p \times m$, $m \times n$, and $p \times n$ matrices, respectively. Note that the chain rule relates composition of functions to multiplication of the Jacobian matrices representing their derivatives, and thus to compositions of the linear transformations given by these derivatives.

The chain rule (2) is written in a very compact form. The following example shows that it actually represents familiar formulas from calculus written in matrix form.

EXAMPLE 1 Suppose $f: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ and $g: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ are defined by

$$\begin{aligned} y_1 &= f_1(x_1, x_2, x_3), & z_1 &= g_1(y_1, y_2) \\ y_2 &= f_2(x_1, x_2, x_3), & z_2 &= g_2(y_1, y_2) \end{aligned}$$

Then $\mathbf{h} = g \circ f$ is defined by

$$\begin{aligned} z_1 &= h_1(x_1, x_2, x_3) = g_1(f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3)) \\ z_2 &= h_2(x_1, x_2, x_3) = g_2(f_1(x_1, x_2, x_3), f_2(x_1, x_2, x_3)) \end{aligned}$$

According to the chain rule (2),

$$\begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \frac{\partial h_1}{\partial x_3} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \frac{\partial h_2}{\partial x_3} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial y_1} & \frac{\partial g_1}{\partial y_2} \\ \frac{\partial g_2}{\partial y_1} & \frac{\partial g_2}{\partial y_2} \end{pmatrix} \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} \end{pmatrix}$$

Evaluating the matrix product on the right, we get the familiar formula

$$\frac{\partial h_i}{\partial x_j} = \frac{\partial g_i}{\partial y_1} \frac{\partial f_1}{\partial x_j} + \frac{\partial g_i}{\partial y_2} \frac{\partial f_2}{\partial x_j}, \quad i = 1, 2, \quad j = 1, 2, 3$$

(The partial derivatives of g_i are evaluated at $\mathbf{y} = (y_1, y_2) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$.) ■

We now know that if the derivative of $\mathbf{f} = (f_1, \dots, f_m)$ at $\mathbf{x}$ exists, it must equal the Jacobian matrix $f'(\mathbf{x})$, and we know some of the properties of the derivative. But when does it exist? It is not sufficient that all the first-order partial derivatives $\partial f_i / \partial x_j$ exist. In fact, Problem 1 shows that $\mathbf{f}$ need not be differentiable at a point even if $\mathbf{f}$ has directional derivatives in all directions at that point. But it turns out that if the partial derivatives $\partial f_i / \partial x_j$ are all continuous at a point $\mathbf{x}$, then $\mathbf{f}$ is differentiable at $\mathbf{x}$.

Recall that a function $f: \mathbb{R} \rightarrow \mathbb{R}$ is said to be of class C^k ($k = 1, 2, \dots$) if all of its partial derivatives of order up to and including k exist and are continuous. Similarly a transformation $\mathbf{f} = (f_1, \dots, f_m)$ from (a subset of) $\mathbb{R}^n$ into $\mathbb{R}^m$ is said to be of class C^k if each of its component functions $f_1, \dots, f_m$ is C^k .

THEOREM 2.9.5 (C¹ FUNCTIONS ARE DIFFERENTIABLE)

If $\mathbf{f}$ is a C^1 transformation from an open set $A \subseteq \mathbb{R}^n$ into $\mathbb{R}^m$, then $\mathbf{f}$ is differentiable at every point $\mathbf{x}$ in A .

Proof: By Theorem 2.9.2, $\mathbf{f}$ is differentiable if each component f_i is differentiable. Hence, in the proof we can assume that $\mathbf{f}$ is real-valued, denoted by f . Let $B(\mathbf{x}; \varepsilon)$ be an open ball small enough so that $B(\mathbf{x}; \varepsilon) \subseteq A$. For each $\mathbf{h}$ such that $\mathbf{x} + \mathbf{h} \in B(\mathbf{x}; \varepsilon)$ define the error term by

$$R(\mathbf{h}) = f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - \sum_{i=1}^n f'_i(\mathbf{x}) h_i$$

and let $\mathbf{e}_j$ denote the j th standard unit vector in $\mathbb{R}^n$. Note that

$$f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) = \sum_{i=1}^n \left[f\left(\mathbf{x} + \sum_{j=1}^i h_j \mathbf{e}_j\right) - f\left(\mathbf{x} + \sum_{j=1}^{i-1} h_j \mathbf{e}_j\right) \right] = \sum_{i=1}^n \left[f'_i(\mathbf{x} + \sum_{j=1}^{i-1} h_j \mathbf{e}_j + \theta_i h_i \mathbf{e}_i) h_i \right]$$

with $\theta_i \in (0, 1)$ for $i = 1, 2, \dots, n$, where the mean value theorem is used to obtain the last equality. Hence

$$\frac{R(\mathbf{h})}{\|\mathbf{h}\|} = \sum_{i=1}^n \left[f'_i\left(\mathbf{x} + \sum_{j=1}^{i-1} h_j \mathbf{e}_j + \theta_i h_i \mathbf{e}_i\right) - f'_i(\mathbf{x}) \right] \frac{h_i}{\|\mathbf{h}\|} \rightarrow 0 \text{ as } \mathbf{h} \rightarrow \mathbf{0}$$

because the derivatives $f'_i(\mathbf{x})$ are assumed to be continuous. ■

We next derive two useful inequalities. First, let f be a C^1 function from $\mathbb{R}^n$ into $\mathbb{R}^m$. Then for all $\mathbf{x}$ and $\mathbf{y}$,

$$\|f(\mathbf{y}) - f(\mathbf{x})\| \leq \max_{\mathbf{z} \in [\mathbf{x}, \mathbf{y}]} \|f'(\mathbf{z})(\mathbf{y} - \mathbf{x})\| \quad (3)$$

where $f'(\mathbf{z})$ is the Jacobian matrix of f at $\mathbf{z}$, and $[\mathbf{x}, \mathbf{y}]$ is the closed line segment from $\mathbf{x}$ to $\mathbf{y}$.

The inequality in (3) is obviously true if $f(\mathbf{y}) = f(\mathbf{x})$. To prove (3) in the case when $f(\mathbf{y}) \neq f(\mathbf{x})$, let $\mathbf{b} = \frac{f(\mathbf{y}) - f(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} (\mathbf{y} - \mathbf{x})$. Then $\|\mathbf{b}\| = 1$ and $\|f(\mathbf{y}) - f(\mathbf{x})\| = \mathbf{b} \cdot (f(\mathbf{y}) - f(\mathbf{x})) = g(\mathbf{y}) - g(\mathbf{x})$, where g is the function defined by $g(\mathbf{z}) = \mathbf{b} \cdot f(\mathbf{z})$.

An easy calculation shows that the gradient of g is $\nabla g(\mathbf{z}) = \mathbf{b} f'(\mathbf{z})$, where $\mathbf{b}$ is interpreted as a $1 \times m$ matrix. By the mean value theorem (Theorem 2.1.2), there exists a $\mathbf{z}$ in $[\mathbf{x}, \mathbf{y}]$ such that $g(\mathbf{y}) - g(\mathbf{x}) = \nabla g(\mathbf{z}) \cdot (\mathbf{y} - \mathbf{x}) = \mathbf{b} \cdot (f'(\mathbf{z})(\mathbf{y} - \mathbf{x}))$. The Cauchy-Schwarz inequality, (1.1.38), then implies $|g(\mathbf{y}) - g(\mathbf{x})| \leq \|\mathbf{b}\| \cdot \|f'(\mathbf{z})(\mathbf{y} - \mathbf{x})\| = \|f'(\mathbf{z})(\mathbf{y} - \mathbf{x})\|$, and (3) follows.

By applying the inequality (3) to the new function $\mathbf{y} \mapsto f(\mathbf{y}) - f'(\mathbf{w})\mathbf{y}$, with $\mathbf{w}$ a fixed vector, the following inequality is obtained for all $\mathbf{x}, \mathbf{y}$, and $\mathbf{w}$ in $\mathbb{R}^n$:

$$\|f(\mathbf{y}) - f(\mathbf{x}) - f'(\mathbf{w})(\mathbf{y} - \mathbf{x})\| \leq \max_{\mathbf{z} \in [\mathbf{x}, \mathbf{y}]} \|f'(\mathbf{z}) - f'(\mathbf{w})(\mathbf{y} - \mathbf{x})\| \quad (4)$$

PROBLEMS FOR SECTION 2.9

- (a) Let f be defined for all (x, y) by $f(x, y) = \frac{xy^2}{x^2 + y^4}$ and $f(0, 0) = 0$. Show that $f'_1(x, y)$ and $f'_2(x, y)$ exist for all (x, y) .
(b) Show that f has a directional derivative in every direction at every point.
(c) Show that f is not continuous at $(0, 0)$. (Hint: Consider the behaviour of f along the curve $x = y^2$.) Is f differentiable at $(0, 0)$?

2.10 Existence and Uniqueness of Solutions of Systems of Equations

This section is concerned with the system of n equations in n unknowns, of the form

$$f_1(x_1, \dots, x_n) = y_1, \dots, f_n(x_1, \dots, x_n) = y_n \quad \text{or in vector form} \quad \mathbf{f}(\mathbf{x}) = \mathbf{y} \quad (1)$$

For given values of $y_1, \dots, y_n$, when will system (1) have a solution $x_1, \dots, x_n$? Also, when is the solution unique?

The inverse function theorem (Theorem 2.7.3) tells us that if system (1) has a solution $\mathbf{y}^0 = \mathbf{f}(\mathbf{x}^0)$, and the Jacobian determinant is not 0 at $\mathbf{x}^0$, then there exist open balls B_1 around $\mathbf{x}^0$ and B_2 around $\mathbf{y}^0$ such that (1) has a unique solution $\mathbf{x}$ in B_1 for each $\mathbf{y}$ in B_2 . This result tells us about (local) uniqueness; it says nothing about the *existence* of a solution $\mathbf{x}^0$ of (1), for $\mathbf{y} = \mathbf{y}^0$.

General theorems on the existence and uniqueness of solutions to $\mathbf{f}(\mathbf{x}) = \mathbf{y}$ must involve strong restrictions on f . This is clear even in the case $n = 1$. One can hardly claim that the

equation $f(x) = 0$ usually has a unique solution. Think about the case where $f(x)$ is a quadratic polynomial, or more generally a polynomial of degree n with n real zeros.

Suppose f is a continuous function from $\mathbb{R}$ to $\mathbb{R}$ where either $f(x) \rightarrow \infty$ as $x \rightarrow \infty$ and $f(x) \rightarrow -\infty$ as $x \rightarrow -\infty$, or $f(x) \rightarrow -\infty$ as $x \rightarrow \infty$ and $f(x) \rightarrow \infty$ as $x \rightarrow -\infty$. Then by the intermediate value theorem, for any number y the equation $f(x) = y$ has at least one solution. Of course, this solution will not necessarily be unique. However, suppose the following condition is satisfied:

$$\text{There exists a positive number } \gamma \text{ such that } f'(x) > \gamma \text{ for all } x \quad (2)$$

Then $f(x) \rightarrow \infty$ as $x \rightarrow \infty$, and $f(x) \rightarrow -\infty$ as $x \rightarrow -\infty$, so there is a solution, and in addition, $f(x)$ is strictly increasing, so the solution is unique.

NOTE 1 Condition (2) secures that $f(x) = y$ has a solution for any choice of y . It cannot be weakened by assuming that $f'(x) > 0$ for all x . For example, the function $f(x) = e^x$ has $f'(x) > 0$ for all x , but $e^x = -1$ has no solution.

The problem of existence and uniqueness of solutions to (1) becomes more complicated when $n \geq 2$. Let us present some arguments and results that sometimes are useful. We refer to Parthasarathy (1983) for proofs and more details.

For $n = 2$, we consider

$$f_1(x_1, x_2) = y_1, \quad f_2(x_1, x_2) = y_2 \quad \iff \quad \mathbf{f}(\mathbf{x}) = \mathbf{y} \quad (3)$$

where f_1 and f_2 are C^1 functions. We seek sufficient conditions for system (3) to be uniquely solvable for x_1 and x_2 , so that $x_1 = \varphi(y_1, y_2)$ and $x_2 = \psi(y_1, y_2)$.

Define

$$f_2(x_1, \infty) = \lim_{x_2 \rightarrow \infty} f_2(x_1, x_2), \quad f_2(x_1, -\infty) = \lim_{x_2 \rightarrow -\infty} f_2(x_1, x_2)$$

where we implicitly assume that the corresponding limit exists or is $\pm\infty$. Suppose that for all x_1 either $f_2(x_1, \pm\infty) = \pm\infty$ or $f_2(x_1, \pm\infty) = \mp\infty$. Then for each x_1 the equation

$$f_2(x_1, x_2) = y_2$$

has a solution $x_2 = \tilde{x}_2(x_1, y_2)$, with $\partial \tilde{x}_2(x_1, y_2) / \partial x_1 = -(\partial f_2 / \partial x_1) / (\partial f_2 / \partial x_2)$. Suppose that $\tilde{x}_2(x_1, y_2)$ is uniquely determined as a C^1 function of x_1 and y_2 . Insert this value of x_2 into the first equation in (3) to obtain

$$f_1(x_1, \tilde{x}_2(x_1, y_2)) = y_1 \quad (4)$$

Suppose that $\lim_{x_1 \rightarrow \pm\infty} f_1(x_1, \tilde{x}_2(x_1, y_2)) = f_1(\pm\infty, \tilde{x}_2(\pm\infty, y_2)) = \pm\infty$ (or $\mp\infty$). Then for all (y_1, y_2) equation (4) has a solution $x_1 = x_1(y_1, y_2)$. If we define $x_2(y_1, y_2) = \tilde{x}_2(x_1(y_1, y_2), y_2)$, then $\mathbf{x}(\mathbf{y}) = (x_1(y_1, y_2), x_2(y_1, y_2))$ is a solution of $\mathbf{f}(\mathbf{x}) = \mathbf{y}$.

If there exists a constant $\alpha > 0$ such that the function $H(x_1, y_2) = f_1(x_1, \bar{x}_2(x_1, y_2))$ has derivative $\partial H/\partial x_1 \geq \alpha > 0$ everywhere, then $H(\pm\infty, y_2) = \pm\infty$, which was a property we used above. Now,

$$\begin{aligned}\frac{\partial H}{\partial x_1} &= \frac{\partial f_1(x_1, \bar{x}_2(x_1, y_2))}{\partial x_1} + \frac{\partial f_1(x_1, \bar{x}_2(x_1, y_2))}{\partial x_2} \frac{\partial \bar{x}_2(x_1, y_2)}{\partial x_1} \\ &= \frac{\partial f_1}{\partial x_1} + \frac{\partial f_1}{\partial x_2} \left(-\frac{\partial f_2/\partial x_1}{\partial f_2/\partial x_2} \right) = \frac{1}{\partial f_2/\partial x_2} \begin{vmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 \end{vmatrix}\end{aligned}$$

So if there exist positive constants k and h such that $0 < \partial f_2/\partial x_2 \leq k$ and the determinant is $\geq h$, then $\partial H/\partial x_1 \geq h/k > 0$. This is a loose motivation for the next theorem in the two-dimensional case.

THEOREM 2.10.1 (HADAMARD)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a C^1 function, and suppose that there exist numbers h and k such that for all $\mathbf{x}$ and all $i, j = 1, \dots, n$,

$$|f'(\mathbf{x})| \geq h > 0 \text{ and } |\partial f_i(\mathbf{x})/\partial x_j| \leq k \quad (5)$$

Then f has an inverse which is defined and C^1 on the whole of $\mathbb{R}^n$.

The theorem implies that for all $\mathbf{y}$ the equation $\mathbf{y} = f(\mathbf{x})$ has a unique solution $\mathbf{x} = \mathbf{x}(\mathbf{y})$, and $\mathbf{x}'(\mathbf{y})$ is continuous.

In the two-dimensional case discussed above we postulated that when $|x_2|$ is large, so is $|f_2(x_1, x_2)|$, and hence also $\|f(x_1, x_2)\|$. Furthermore, we also postulated that when $|x_1|$ is large, so is $|f_1(x_1, \bar{x}_2(x_1, y_2))|$, and hence $\|f(x_1, \bar{x}_2(x_1, y_2))\|$. Provided $|f'(\mathbf{x})| \neq 0$, we do get solutions of the equation $f(\mathbf{x}) = \mathbf{y}$ if we require that $\|f(\mathbf{x})\|$ is large when $\|\mathbf{x}\|$ is large. (For the definition of inf, see Section A.4.)

THEOREM 2.10.2

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a C^1 function and suppose that $|f'(\mathbf{x})| \neq 0$ for all $\mathbf{x}$. Then $f(\mathbf{x})$ has an inverse which is defined and C^1 in all of $\mathbb{R}^n$ if and only if

$$\inf\{\|f(\mathbf{x})\| : \|\mathbf{x}\| \geq n\} \rightarrow \infty \text{ as } n \rightarrow \infty \quad (6)$$

For proofs of the last two theorems see Ortega and Rheinboldt (1970).

The results referred to so far deal with the existence and uniqueness of solutions to equations. We conclude with two results that are only concerned with uniqueness.

THEOREM 2.10.3 (GALE-NIKAIDO)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be C^1 and let Ω be the rectangle $\Omega = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}$, where $\mathbf{a}$ and $\mathbf{b}$ are given vectors in $\mathbb{R}^n$. Then f is one-to-one in Ω if one of the following conditions is satisfied for all $\mathbf{x}$:

- (a) $f'(\mathbf{x})$ has only positive principal minors.
- (b) $f'(\mathbf{x})$ has only negative principal minors.

The last theorem in this section gives sufficient conditions for a function $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ to be one-to-one on an arbitrarily given convex set Ω in $\mathbb{R}^n$. We need the following definition:

QUASIDEFINITE MATRICES

An $n \times n$ matrix A (not necessarily symmetric) is called **positive quasidefinite** in $S \subseteq \mathbb{R}^n$ if $\mathbf{x}'A\mathbf{x} > 0$ for every n -vector $\mathbf{x} \neq \mathbf{0}$ in S . The matrix A is **negative quasidefinite** if $-A$ is positive quasidefinite. (7)

NOTE 2 If A is symmetric, A is obviously positive quasidefinite if and only if A is positive semidefinite. A general quadratic matrix is positive quasidefinite if and only if the (symmetric) matrix $A + A'$ is positive semidefinite.

THEOREM 2.10.4 (GALE-NIKAIDO)

Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a C^1 function and suppose that the Jacobian matrix $f'(\mathbf{x})$ is either positive quasidefinite everywhere in a convex set Ω , or negative quasidefinite everywhere in Ω . Then f is one-to-one in Ω .

Proof: Let $\mathbf{a} \neq \mathbf{b}$ be arbitrary points of Ω . Define $g(t) = t\mathbf{a} + (1-t)\mathbf{b}$, $t \in [0, 1]$. Let $\mathbf{h} = \mathbf{a} - \mathbf{b} \neq \mathbf{0}$, and define $w(t) = \mathbf{h}' \cdot f(g(t)) = h_1 f_1(g(t)) + \dots + h_n f_n(g(t))$. Then $w'(t) = [h_1 f'_1(g(t)) + \dots + h_n f'_n(g(t))] \cdot g'(t) = [h_1 f'_1(g(t)) + \dots + h_n f'_n(g(t))] \cdot \mathbf{h} = \mathbf{h}' \cdot f'(g(t)) \cdot \mathbf{h}$. If $f'(\mathbf{x})$ is positive quasidefinite, then $w'(t) = \mathbf{h}' \cdot f'(g(t)) \cdot \mathbf{h} > 0$ for $\mathbf{h} \neq \mathbf{0}$ and so $w(1) > w(0)$. On the other hand, if $f'(\mathbf{x})$ is negative quasidefinite, then $w(1) < w(0)$. In either case, therefore, $f(\mathbf{a}) \neq f(\mathbf{b})$, so f is one-to-one. ■

NOTE 3 Theorem 1.7.1(a) states that when A is symmetric, A is positive (quasi)definite iff the leading principal minors are all positive. Note that the Jacobian matrix $f'(\mathbf{x})$ is not, in general, symmetric. Nevertheless, Theorem 2.10.3 uses sign conditions that, if applied to a symmetric matrix, are equivalent to its definiteness.

NOTE 4 Theorems about global uniqueness (univalence) are useful in several economic applications. For example, suppose that a national economy has n different industries each

producing a positive amount of a single output under constant returns to scale, using other goods and scarce primary factors as inputs. Suppose the country is small, and faces a fixed price vector $\mathbf{p}$ in $\mathbb{R}_+^n$ at which it can import or export the n goods it produces. Suppose there are n primary factors whose prices are given by the vector $\mathbf{w}$ in $\mathbb{R}_+^n$. Equilibrium requires that $p_i = c_i(\mathbf{w})$ for each $i = 1, 2, \dots, n$, where $c_i(\mathbf{w})$ is the minimum cost at prices $\mathbf{w}$ of producing one unit of good i . Then the vector equation $\mathbf{p} = \mathbf{c}(\mathbf{w})$, if it has a unique solution, will determine the factor price vector $\mathbf{w}$ as a function of $\mathbf{p}$. When different countries have the same unit cost functions, this implies **factor price equalization**—because $\mathbf{p}$ is the same for all countries that trade freely, so is the factor price vector $\mathbf{w}$. See Parthasarathy (1983), Chapter IX, and the references there.

PROBLEMS FOR SECTION 2.10

1. Show that $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}$ has positive leading principal minors, but is not positive quasidefinite.
2. Suppose the national economy in Note 4 has two industries, whose unit cost functions take the Cobb–Douglas form $c_1(w_1, w_2) = \gamma_1 w_1^\alpha w_2^{1-\alpha}$ and $c_2(w_1, w_2) = \gamma_2 w_1^\beta w_2^{1-\beta}$, respectively. Show that, provided $\alpha \neq \beta$, the vector equation $\mathbf{p} = \mathbf{c}(\mathbf{w})$ determines w_1/w_2 uniquely as a function of p_1/p_2 , and comment on the solution.

3

STATIC OPTIMIZATION

If, then, in Political Economy we have to deal with quantities and complicated relations of quantities, we must reason mathematically; we do not render the science less mathematical by avoiding the symbols of algebra ...

—Jevons (1871)

Much of economic analysis relies on static optimization problems. For example, producers seek those input combinations that maximize profits or minimize costs, whereas consumers seek commodity bundles that maximize utility subject to their budget constraints.

In most static optimization problems there is an **objective function** $f(x_1, \dots, x_n) = f(\mathbf{x})$, a real-valued function of n variables whose value is to be optimized, i.e., maximized or minimized. There is also an **admissible set** (or **feasible set**) S that is some subset of $\mathbb{R}^n$, the n -dimensional Euclidean space. Then the problem is to find maximum or minimum points of f in S :

$$\max(\min) f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in S$$

where $\max(\min)$ indicates that we want to maximize or minimize f .

Depending on the set S , several different types of optimization problem can arise. If the optimum occurs at an interior point of S , we talk about the **classical case**. In Sections 3.1 and 3.2 some basic facts are reviewed. In particular, Section 3.1 discusses an envelope theorem for unconstrained maxima, and Section 3.2 deals with second-order conditions for local extrema.

If S is the set of all points $\mathbf{x}$ that satisfy a given system of equations, we have the **Lagrange problem** of maximizing (or minimizing) a function subject to equality constraints. Such problems are discussed in Sections 3.3 and 3.4. In addition to the standard results for the general Lagrange problem, sensitivity results are discussed in some detail.

The general **nonlinear programming problem** arises when S consists of all points $\mathbf{x}$ in $\mathbb{R}^n$ that satisfy a system of inequality constraints. Bounds on available resources typically lead to such constraints. Section 3.5 presents the basic facts, along with a proof of the necessary Kuhn–Tucker conditions assuming that the value function is differentiable. Section 3.6 goes on to discuss sufficient conditions, in particular quasiconcave programming, of particular importance to economists. Section 3.7 deals with comparative statics for nonlinear programming problems.

Many economic optimization problems have nonnegativity constraints on the variables. Section 3.8 shows how to handle such problems in an efficient way. It also considers problems with

equality as well as inequality constraints (mixed constraints).

Section 3.9 on concave programming deals with results that do not require differentiability. Section 3.10 gives precise theorems on envelope results, and the final Section 3.11 contains a general proof of the existence of Lagrange multipliers in constrained optimization problems for the general case of mixed constraints.

3.1 Extreme Points

We begin by recalling some basic definitions and results. Let f be a function of n variables $x_1, \dots, x_n$ defined on a set S in $\mathbb{R}^n$. Suppose that the point $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ belongs to S and that the value of f at $\mathbf{x}^*$ is greater than or equal to the values attained by f at all other points $\mathbf{x} = (x_1, \dots, x_n)$ of S . Thus, in symbols,

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) \quad \text{for all } \mathbf{x} \text{ in } S \quad (*)$$

Then $\mathbf{x}^*$ is called a (global) **maximum point** for f in S and $f(\mathbf{x}^*)$ is called the **maximum value**. If the inequality in $(*)$ is strict for all $\mathbf{x} \neq \mathbf{x}^*$, then $\mathbf{x}^*$ is a **strict maximum point** for f in S . We define (strict) **minimum point** and **minimum value** by reversing the inequality sign in $(*)$. As collective names, we use **extreme points** and **extreme values** to indicate both maxima or minima.

A **stationary point** of f is a point where all the first-order partial derivatives are 0. We have the following well-known theorem:

THEOREM 3.1.1 (NECESSARY FIRST-ORDER CONDITIONS)

Let f be defined on a set S in $\mathbb{R}^n$ and let $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ be an interior point in S at which f has partial derivatives. A necessary condition for $\mathbf{x}^*$ to be a maximum or minimum point for f is that $\mathbf{x}^*$ is a stationary point for f —that is, it satisfies the equations

$$f'_i(\mathbf{x}) = 0, \quad i = 1, \dots, n \quad (1)$$

Interior stationary points for concave or convex functions are automatically extreme points:

THEOREM 3.1.2 (SUFFICIENT CONDITIONS WITH CONCAVITY/CONVEXITY)

Suppose that the function $f(\mathbf{x})$ is defined in a convex set S in $\mathbb{R}^n$ and let $\mathbf{x}^*$ be an interior point of S . Assume also that f is C^1 in an open ball around $\mathbf{x}^*$.

- If f is concave in S , then $\mathbf{x}^*$ is a (global) maximum point for f in S if and only if $\mathbf{x}^*$ is a stationary point for f .
- If f is convex in S , then $\mathbf{x}^*$ is a (global) minimum point for f in S if and only if $\mathbf{x}^*$ is a stationary point for f .

Proof: If f has a maximum or minimum at $\mathbf{x}^*$, then, according to Theorem 3.1.1, $\mathbf{x}^*$ must be a stationary point. Suppose on the other hand that $\mathbf{x}^*$ is a stationary point for a concave function f . Apply Theorem 2.4.1 with $\mathbf{x}^0 = \mathbf{x}^*$. Since $f'_i(\mathbf{x}^*) = 0$ for $i = 1, \dots, n$, inequality (1) in Theorem 2.4.1 implies that $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ for all $\mathbf{x}$ in S . This means that $\mathbf{x}^*$ is a maximum point.

To prove (b), apply (a) to $-f$, which is concave. ■

This important theorem is illustrated for the case of functions of two variables in Fig. 1.

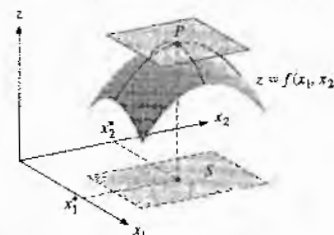


Figure 1• The concave function $f(x_1, x_2)$ has a maximum at the stationary point (x_1^*, x_2^*) . The horizontal tangent plane at the corresponding point P lies on top of the graph.

EXAMPLE 1

Find all (global) extreme points of $f(x, y, z) = x^2 + 2y^2 + 3z^2 + 2xy + 2xz$.

Solution: The only stationary point is $(0, 0, 0)$. The Hessian matrix is

$$f''(x, y, z) = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 4 & 0 \\ 2 & 0 & 6 \end{pmatrix}$$

The leading principal minors are $D_1 = 2$, $D_2 = 4$, and $D_3 = 8$. Hence, according to Theorem 2.3.2(a), f is (strictly) convex, and we conclude from Theorem 3.1.2(b) that $(0, 0, 0)$ is a (global) minimum point, and the only one. ■

EXAMPLE 2

Let $x = F(\mathbf{v}) = F(v_1, \dots, v_n)$ denote a firm's production function, which is assumed to be differentiable. If the positive price of output is p and $q_1, \dots, q_n$ are the positive prices of the factors of production $v_1, \dots, v_n$, then the firm's profit is given by

$$\pi = pF(v_1, \dots, v_n) - q_1v_1 - \dots - q_nv_n \quad (*)$$

The first-order conditions for maximum profit are

$$\frac{\partial \pi}{\partial v_i} = pF'_i(v_1, \dots, v_n) - q_i = 0, \quad i = 1, \dots, n \quad (**)$$

Suppose that $(**)$ has a solution $\mathbf{v}^* = (v_1^*, \dots, v_n^*)$, with $v_1^* > 0, \dots, v_n^* > 0$. If F is concave, then π is also concave as the sum of the concave function $pF(v_1, \dots, v_n)$ and the linear, hence concave, function $-q_1v_1 - \dots - q_nv_n$. It follows from Theorem 3.1.2(a) that the stationary point $(v_1^*, \dots, v_n^*)$ really does maximize profit.

Suppose that F is the Cobb–Douglas function $x = F(v_1, \dots, v_n) = Av_1^{a_1} \cdots v_n^{a_n}$, where A and $a_1, \dots, a_n$ are positive, with $a = a_1 + \cdots + a_n < 1$. Then F is (strictly) concave (see (2.5.6)). In this case (***) reduces to

$$pAv_1^{a_1} \cdots v_i^{a_i-1} \cdots v_n^{a_n} = q_i, \quad i = 1, \dots, n \quad (***)$$

Multiplying the i th equation in (***) by v_i/a_i gives $px = pAv_1^{a_1} \cdots v_n^{a_n} = q_i v_i/a_i$. Hence, $v_i = a_i px/q_i$. Substituting these into $x = Av_1^{a_1} \cdots v_n^{a_n}$ gives

$$x = A \left(\frac{a_1 px}{q_1} \right)^{a_1} \cdots \left(\frac{a_n px}{q_n} \right)^{a_n} = A(px)^a \left(\frac{a_1}{q_1} \right)^{a_1} \cdots \left(\frac{a_n}{q_n} \right)^{a_n}$$

Solving this equation for x gives $x = A^{1/(1-a)} p^{a/(1-a)} (a_1/q_1)^{a_1/(1-a)} \cdots (a_n/q_n)^{a_n/(1-a)}$, so we conclude that

$$v_i = \left(\frac{a_i}{q_i} \right) (Ap)^{1/(1-a)} \left(\frac{a_1}{q_1} \right)^{a_1/(1-a)} \left(\frac{a_2}{q_2} \right)^{a_2/(1-a)} \cdots \left(\frac{a_n}{q_n} \right)^{a_n/(1-a)}, \quad i = 1, \dots, n$$

This is the profit maximizing choice of input quantity for each factor of production. $\blacksquare$

Theorem 3.1.2 requires concavity for a maximum, or convexity for a minimum. Recall that maximizing (minimizing) a function $f(x)$ is equivalent to maximizing (minimizing) $F(f(x))$, for any given strictly increasing function F . (See e.g. EMEA, Section 13.6.) If $f(x)$ is not concave (convex), then the transformed function $F(f(x))$ may be concave (convex) for a suitably chosen strictly increasing F . Such a transformation makes it possible to apply Theorem 3.1.2.

EXAMPLE 3 Show that the function g defined for all x, y , and z by

$$g(x, y, z) = (x^2 + 2y^2 + 3z^2 + 2xy + 2xz - 5)^3$$

has a minimum at $(0, 0, 0)$.

Solution: Theorem 3.1.2 does not apply to g , because g is not convex. Nevertheless, note that $g(x, y, z) = [f(x, y, z) - 5]^3$, where f is the convex function studied in Example 1. Since g is a strictly increasing transformation of f , it too has $(0, 0, 0)$ as its unique minimum point. $\blacksquare$

The following theorem is important in optimization theory (see Section 13.3):

THEOREM 3.1.3 (EXTREME VALUE THEOREM)

Let $f(x)$ be a continuous function on a closed, bounded set S . Then f has both a maximum point and a minimum point in S .

NOTE 1 In most economic applications the set S referred to in Theorem 3.1.3 is specified using one or more inequalities. If the functions $g_j(x)$, $j = 1, \dots, m$ are all continuous and $b_1, \dots, b_m$ are given numbers, then the set $S = \{x : g_j(x) \leq b_j, j = 1, \dots, m\}$ is closed.

If some (or all) of the inequalities are replaced by $\geq$ or $=$, the set is still closed. The set S is bounded if it is contained in some ball around the origin. (See Section 13.1 for general definitions and results on open sets, closed sets, and related concepts.)

Suppose $f(x)$ is a C^1 function defined on a set S in $\mathbb{R}^n$, and suppose too that we know that f has a maximum point x^* in S —because of Theorem 3.1.3 or for other reasons. If x^* is an interior point of S , it must be a stationary point for f . If x^* is not an interior point of S , it must belong to the boundary of S . (If the set in question is of the form $S = \{x : g_j(x) \leq b_j, j = 1, \dots, m\}$, then an interior point will often (but not always!) be one for which all the inequalities are strict.) The following procedure can therefore be used to locate the maximum point:

- Record all interior stationary points of S . They are candidates for maximum.
- Find all maximum points for f restricted to the boundary of S . They are also candidates.
- Compute the value of f at each of the points found in (A) and (B). Those that give f its largest value are the maximum points.

If we know that f has a minimum point, a completely analogous procedure will give us the minimum point or points. In EMEA this procedure was used to find extreme points for functions of two variables. Later we shall give more efficient methods for finding extreme points of such functions.

Envelope Theorem for Unconstrained Maxima

The objective function in economic optimization problems usually involves parameters like prices in addition to choice variables like quantities. Consider an objective function with a parameter vector r of the form $f(x, r) = f(x_1, \dots, x_n, r_1, \dots, r_k)$, where $x \in S \subseteq \mathbb{R}^n$ and $r \in \mathbb{R}^k$. For each fixed r suppose we have found the maximum of $f(x, r)$ when x varies in S . The maximum value of $f(x, r)$ usually depends on r . We denote this value by $f^*(r)$ and call f^* the **value function**. Thus,

$$f^*(r) = \max_{x \in S} f(x, r) \quad (\text{the value function}) \quad (2)$$

The vector x that maximizes $f(x, r)$ depends on r and is denoted by $x^*(r)$.¹ Then $f^*(r) = f(x^*(r), r)$.

How does $f^*(r)$ vary as the j th parameter r_j changes? Provided that f^* is differentiable we have the following so-called **envelope result**:

$$\frac{\partial f^*(r)}{\partial r_j} = \left[\frac{\partial f(x, r)}{\partial r_j} \right]_{x=x^*(r)}, \quad j = 1, \dots, k \quad (3)$$

Note that on the right-hand side we differentiate f w.r.t. its $(n + j)$ th argument, which is r_j , and evaluate the derivative at $(x^*(r), r)$.

When the parameter r_j changes, $f^*(r)$ changes for two reasons. First, a change in r_j changes $x^*(r)$. Second, $f(x^*(r), r)$ changes directly because the variable r_j changes.

¹ There may be several choices of x that maximize $f(x, r)$ for a given parameter vector r . Then we let $x^*(r)$ denote one of these choices, and try to select x for different values of r so that $x^*(r)$ is a differentiable function of r .

Formula (3) claims that the first effect is zero. To see why, assume an interior solution and that f^* is differentiable. Then, because $\mathbf{x} = \mathbf{x}^*(r)$ maximizes $f(\mathbf{x}, r)$ w.r.t. $\mathbf{x}$, all the partial derivatives $\partial f(\mathbf{x}^*(r), r)/\partial x_i$ must be 0. Hence:

$$\begin{aligned}\frac{\partial f^*(r)}{\partial r_j} &= \frac{\partial}{\partial r_j} (f(\mathbf{x}^*(r), r)) = \sum_{i=1}^n \frac{\partial f(\mathbf{x}^*(r), r)}{\partial x_i} \frac{\partial x_i^*(r)}{\partial r_j} + \left[\frac{\partial f(\mathbf{x}, r)}{\partial r_j} \right]_{\mathbf{x}=\mathbf{x}^*(r)} \\ &= \left[\frac{\partial f(\mathbf{x}, r)}{\partial r_j} \right]_{\mathbf{x}=\mathbf{x}^*(r)}\end{aligned}$$

EXAMPLE 4 The profit function π in Example 2 depends on the input vector $\mathbf{v}$ and the parametric prices p and $\mathbf{q} = (q_1, \dots, q_n)$. Specifically,

$$\pi = \pi(\mathbf{v}, p, \mathbf{q}) = pF(\mathbf{v}) - q_1 v_1 - \dots - q_n v_n$$

Let $\pi^*(p, \mathbf{q})$ denote the value function in the problem of maximizing π w.r.t. $\mathbf{v}$, and let $\mathbf{v}^* = \mathbf{v}^*(p, \mathbf{q})$ be the associated $\mathbf{v}$ vector. Then according to (3),

$$\frac{\partial \pi^*(p, \mathbf{q})}{\partial p} = \frac{\partial \pi(\mathbf{v}^*, p, \mathbf{q})}{\partial p} = F(\mathbf{v}^*), \quad \frac{\partial \pi^*(p, \mathbf{q})}{\partial q_j} = \frac{\partial \pi(\mathbf{v}^*, p, \mathbf{q})}{\partial q_j} = -v_j^* \quad (*)$$

This is intuitively understandable: when the price of output increases by Δp , the optimal profit increases by approximately $F(\mathbf{v}^*)\Delta p$, since $F(\mathbf{v}^*)$ is the optimal number of units produced. If the price of the j th input factor q_j increases by Δq_j , the optimal profit decreases by about $v_j^* \Delta q_j$ units, since v_j^* is the amount of factor j used at the optimum. The equations in $(*)$ are known as **Hotelling's lemma**.

We next formulate the envelope theorem in a slightly more precise manner and give an alternative proof which is easier to generalize to more complicated constrained optimization problems. Note, however, that we still *assume that the value function is differentiable*. In Section 3.10 we discuss sufficient conditions for differentiability.

THEOREM 3.1.4 (ENVELOPE THEOREM)

In the problem $\max_{\mathbf{x} \in S} f(\mathbf{x}, \mathbf{r})$, where $S \subseteq \mathbb{R}^n$ and $\mathbf{r} = (r_1, \dots, r_k)$, suppose that there is a maximum point $\mathbf{x}^*(\mathbf{r})$ in S for every $\mathbf{r}$ in some ball $B(\bar{\mathbf{r}}; \delta)$, with $\delta > 0$. Furthermore, assume that the mappings $\mathbf{r} \mapsto f(\mathbf{x}^*(\bar{\mathbf{r}}), \mathbf{r})$ (with $\bar{\mathbf{r}}$ fixed) and $\mathbf{r} \mapsto f^*(\mathbf{r})$ (defined in (2)) are both differentiable at $\bar{\mathbf{r}}$. Then

$$\frac{\partial f^*(\bar{\mathbf{r}})}{\partial r_j} = \left[\frac{\partial f(\mathbf{x}, \mathbf{r})}{\partial r_j} \right]_{(\mathbf{x}=\mathbf{x}^*(\bar{\mathbf{r}}), \mathbf{r}=\bar{\mathbf{r}})} \quad j = 1, \dots, k \quad (4)$$

Proof: Define the function $\varphi(\mathbf{r}) = f(\mathbf{x}^*(\bar{\mathbf{r}}), \mathbf{r}) - f^*(\mathbf{r})$. Because $\mathbf{x}^*(\bar{\mathbf{r}})$ is a maximum point of $f(\mathbf{x}, \mathbf{r})$ when $\mathbf{r} = \bar{\mathbf{r}}$, one has $\varphi(\bar{\mathbf{r}}) = 0$. Also the definition of f^* implies that $\varphi(\mathbf{r}) \leq 0$ for all $\mathbf{r}$ in $B(\bar{\mathbf{r}}; \delta)$. Hence φ has an interior maximum at $\mathbf{r} = \bar{\mathbf{r}}$. The equations in (4) follow from the fact that $\mathbf{r} = \bar{\mathbf{r}}$ must satisfy the first-order conditions $\varphi_j'(\mathbf{r}) = 0$ for $j = 1, \dots, k$. ■

A Geometric Illustration of the Envelope Theorem

Figure 2 illustrates (3) (or (4)) in the case where there is only one parameter r . For each fixed value of $\mathbf{x}$ there is a curve $K_{\mathbf{x}}$ in the ry -plane, given by the equation $y = f(\mathbf{x}, r)$. The figure shows some of these curves together with the graph of f^* —that is, the curve $y = f^*(r)$.

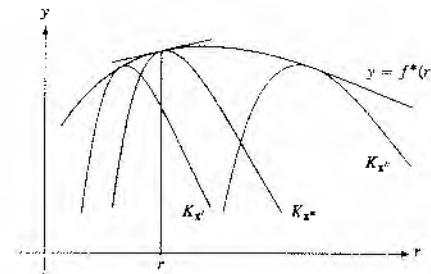


Figure 2 The curve $y = f^*(r)$ is the envelope of all the curves $y = f(\mathbf{x}, r)$.

For all $\mathbf{x}$ and all r we have $f(\mathbf{x}, r) \leq \max_{\mathbf{x}} f(\mathbf{x}, r) = f^*(r)$. It follows that none of the $K_{\mathbf{x}}$ -curves can ever lie above the curve $y = f^*(r)$. On the other hand, for each value of r there is at least one value $\mathbf{x}^*$ of $\mathbf{x}$ such that $f(\mathbf{x}^*, r) = f^*(r)$, namely the choice of $\mathbf{x}^*$ that solves the maximization problem for the given value of r . The curve $K_{\mathbf{x}^*}$ will then just touch the curve $y = f^*(r)$ at the point $(\mathbf{x}^*, f^*(r)) = (\mathbf{x}^*, f(\mathbf{x}^*, r))$, and so must have exactly the same tangent as the graph of f^* at this point. Moreover, the slope of this common tangent must be both df^*/dr , the slope of the tangent to the graph of f^* , and $\partial f(\mathbf{x}^*, r)/\partial r$, the slope of the tangent to the curve $K_{\mathbf{x}^*}$, which is the graph of $f(\mathbf{x}^*, r)$ when $\mathbf{x}^*$ is fixed.

As Fig. 2 suggests, the graph of $y = f^*(r)$ is the lowest curve with the property that it lies on or above all the curves $K_{\mathbf{x}}$. So its graph is like an envelope that is used to “wrap” all these curves; that is why we call the graph of f^* the **envelope** of the family of $K_{\mathbf{x}}$ -curves.

PROBLEMS FOR SECTION 3.1

1. Show that

$$g(x, y) = x^3 + y^3 - 3x - 2y \quad \text{defined for } x > 0, y > 0$$

is strictly convex, and find its (global) minimum value.

2. A firm produces two output goods, denoted by A and B . The cost per day is

$$C(x, y) = 0.04x^2 - 0.01xy + 0.01y^2 + 4x + 2y + 500$$

when x units of A and y units of B are produced ($x > 0, y > 0$). The firm sells all it produces at prices 13 per unit of A and 8 per unit of B . Find the profit function $\pi(x, y)$ and the values of x and y that maximize profit.

3. (a) Referring to Example 2, solve the problem $\max p v_1^{1/3} v_2^{2/3} - q_1 v_1 - q_2 v_2$.

(b) Let $\pi^*(p, q_1, q_2)$ denote the value function. Verify the equalities in $(*)$ in Example 4 for this case.

4. Find the functions $x^*(r)$ and $y^*(r)$ such that $x = x^*(r)$ and $y = y^*(r)$ solve the problem

$$\max_{x,y} f(x, y, r) = \max_{x,y} (-x^2 - xy - 2y^2 + 2rx + 2ry)$$

where r is a parameter. Verify equation (3).

5. Find the solutions $x^*(r, s)$ and $y^*(r, s)$ of the problem

$$\max_{x,y} f(x, y, r, s) = \max_{x,y} (r^2x + 3s^2y - x^2 - 8y^2)$$

where r and s are parameters. Verify equation (3).

6. (a) Suppose the production function in Example 2 is

$$F(v_1, \dots, v_n) = a_1 \ln(v_1 + 1) + \dots + a_n \ln(v_n + 1)$$

where $a_1, \dots, a_n$ are positive constants and $p > q_i/a_i$ for $i = 1, \dots, n$. Find the profit maximizing choice of input quantities.

- (b) Verify the envelope result (3) w.r.t. p , each q_i and each a_i .

3.2 Local Extreme Points

Suppose one is trying to find the maximum of a function that is not concave, or a minimum of a function that is not convex. Then Theorem 3.1.2 cannot be used. Instead, one possible procedure is to identify *local* extreme points, and then compare the values of the function at different local extreme points in the hope of finding a global maximum (or minimum).

The point $\mathbf{x}^*$ is a **local maximum point** of f in S if $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ for all $\mathbf{x}$ in S sufficiently close to $\mathbf{x}^*$. More precisely, the requirement is that there exists a positive number r such that

$$f(\mathbf{x}) \leq f(\mathbf{x}^*) \quad \text{for all } \mathbf{x} \text{ in } S \text{ with } \|\mathbf{x} - \mathbf{x}^*\| < r \quad (*)$$

(Equivalently, if we let $B(\mathbf{x}^*; r) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{x}^*\| < r\}$ denote the open n -ball centred at $\mathbf{x}^*$ and with radius r , then $\mathbf{x}^*$ is a local maximum point for f in S if there exists a positive number r such that $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ for all $\mathbf{x}$ in $B(\mathbf{x}^*; r) \cap S$.) If the first inequality in $(*)$ is strict for $\mathbf{x} \neq \mathbf{x}^*$, then $\mathbf{x}^*$ is a **strict local maximum point** for f in S .

A (strict) **local minimum point** is defined in the obvious way, and it should be clear what is meant by **local maximum and minimum values**, **local extreme points**, and **local extreme values**. Of course, a global extreme point is also a local extreme point, but the converse is not always true.

In searching for maximum and minimum points, Theorem 3.1.1 on necessary first-order conditions is very useful. The same result applies to local extreme points as well: *a local extreme point in the interior of the domain of a differentiable function must be a stationary point.* (This observation follows because the proof of Theorem 3.1.1 considers the behaviour of the function only in a small neighbourhood of the optimal point.) A stationary point $\mathbf{x}^*$

of f that is neither a local maximum point nor a local minimum point is called a **saddle point** of f . Thus, arbitrarily close to a saddle point, there are points with both higher and lower values than the function value at the saddle point. Figure 1 illustrates these concepts in the case of a function of two variables.

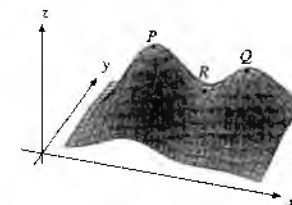


Figure 1 P is a maximum, Q is a local maximum, and R is a saddle point.

We next study conditions that allow the stationary points of a function of n variables to be classified as local maximum points, local minimum points, and saddle points. First recall the second-order conditions for local extreme points for functions of two variables. If $f(x, y)$ is a C^2 function with (x^*, y^*) as an interior stationary point, then

$$f''_{11}(x^*, y^*) > 0 \text{ \& } \begin{vmatrix} f''_{11}(x^*, y^*) & f''_{12}(x^*, y^*) \\ f''_{21}(x^*, y^*) & f''_{22}(x^*, y^*) \end{vmatrix} > 0 \implies \text{local min. at } (x^*, y^*) \quad (1)$$

$$f''_{11}(x^*, y^*) < 0 \text{ \& } \begin{vmatrix} f''_{11}(x^*, y^*) & f''_{12}(x^*, y^*) \\ f''_{21}(x^*, y^*) & f''_{22}(x^*, y^*) \end{vmatrix} > 0 \implies \text{local max. at } (x^*, y^*) \quad (2)$$

$$\begin{vmatrix} f''_{11}(x^*, y^*) & f''_{12}(x^*, y^*) \\ f''_{21}(x^*, y^*) & f''_{22}(x^*, y^*) \end{vmatrix} < 0 \implies (x^*, y^*) \text{ is a saddle point} \quad (3)$$

In order to generalize these results to functions of n variables, we need to consider the *n* leading principal minors of the Hessian matrix $\mathbf{f}''(\mathbf{x}) = (f''_{ij}(\mathbf{x}))_{n \times n}$:

$$D_k(\mathbf{x}) = \begin{vmatrix} f''_{11}(\mathbf{x}) & f''_{12}(\mathbf{x}) & \dots & f''_{1k}(\mathbf{x}) \\ f''_{21}(\mathbf{x}) & f''_{22}(\mathbf{x}) & \dots & f''_{2k}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ f''_{k1}(\mathbf{x}) & f''_{k2}(\mathbf{x}) & \dots & f''_{kk}(\mathbf{x}) \end{vmatrix}, \quad k = 1, \dots, n \quad (4)$$

THEOREM 3.2.1 (LOCAL SECOND-ORDER CONDITIONS)

Suppose that $f(\mathbf{x}) = f(x_1, \dots, x_n)$ is defined on a set S in $\mathbb{R}^n$ and that $\mathbf{x}^*$ is an interior stationary point. Assume also that f is C^2 in an open ball around $\mathbf{x}^*$. Let $D_k(\mathbf{x})$ be defined by (4). Then:

- (a) $D_k(\mathbf{x}^*) > 0$, $k = 1, \dots, n \implies \mathbf{x}^*$ is a local minimum point.
- (b) $(-1)^k D_k(\mathbf{x}^*) > 0$, $k = 1, \dots, n \implies \mathbf{x}^*$ is a local maximum point.
- (c) $D_n(\mathbf{x}^*) \neq 0$ and neither (a) nor (b) is satisfied $\implies \mathbf{x}^*$ is a saddle point.

Proof: (a) A determinant is a continuous function of its elements. If $D_k(\mathbf{x}^*) > 0$ for all k , it is possible to find a ball $B(\mathbf{x}^*; r)$ with radius $r > 0$ so small that $D_k(\mathbf{x}) > 0$ for all $\mathbf{x}$ in $B(\mathbf{x}^*; r)$ and all $k = 1, \dots, n$. Hence, according to Theorem 1.7.1(a), the quadratic form $\sum_{i=1}^n \sum_{j=1}^n f''_{ij}(\mathbf{x})h_i h_j$ is positive definite for all $\mathbf{x}$ in $B(\mathbf{x}^*; r)$. It follows from (2.3.4) that f is (strictly) convex in $B(\mathbf{x}^*; r)$. But then Theorem 3.1.2 shows that the stationary point $\mathbf{x}^*$ is a minimum point for f in $B(\mathbf{x}^*; r)$, and therefore a local minimum point for f in S .

(b) This follows from (a) by replacing f with $-f$ and using rule (1.1.20) for evaluating determinants.

(c) A proof is given in Note 1. ■

Theorem 3.2.1 is often referred to as the **second-derivative test**. Check to see that for $n = 2$ it reduces to (1)–(3).

An alternative way of formulating (a) and (b) in Theorem 3.2.1 is as follows:

A sufficient condition for an interior stationary point $\mathbf{x}^*$ of $f(\mathbf{x})$ to be a local minimum point is that the Hessian matrix $\mathbf{f}''(\mathbf{x}^*)$ is positive definite at $\mathbf{x}^*$. (5)

A sufficient condition for an interior stationary point $\mathbf{x}^*$ of $f(\mathbf{x})$ to be a local maximum point is that the Hessian matrix $\mathbf{f}''(\mathbf{x}^*)$ is negative definite at $\mathbf{x}^*$. (6)

By Theorem 3.2.1, stationary points $\mathbf{x}^*$ for f where $D_n(\mathbf{x}^*) \neq 0$ are now fully classified as either local maximum points, local minimum points, or saddle points. If $D_n(\mathbf{x}^*) = 0$, a closer examination is necessary in order to classify the stationary point. (Note the analogy with the one-variable case in which “anything can happen” at a stationary point where $f''(\mathbf{x}^*) = 0$.)

EXAMPLE 1 The following function has stationary points $(-2, -2, -2)$ and $(0, 0, 0)$:

$$f(x, y, z) = x^3 + 3xy + 3xz + y^3 + 3yz + z^3$$

Classify these points by using Theorem 3.2.1.

Solution: The Hessian matrix is
$$\begin{pmatrix} f''_{11} & f''_{12} & f''_{13} \\ f''_{21} & f''_{22} & f''_{23} \\ f''_{31} & f''_{32} & f''_{33} \end{pmatrix} = \begin{pmatrix} 6x & 3 & 3 \\ 3 & 6y & 3 \\ 3 & 3 & 6z \end{pmatrix}.$$

At $(-2, -2, -2)$ the leading principal minors are

$$6(-2) = -12, \quad \begin{vmatrix} -12 & 3 \\ 3 & -12 \end{vmatrix} = 135, \quad \begin{vmatrix} -12 & 3 & 3 \\ 3 & -12 & 3 \\ 3 & 3 & -12 \end{vmatrix} = -1350$$

According to Theorem 3.2.1(b), $(-2, -2, -2)$ is a local maximum point.

At $(0, 0, 0)$ the leading principal minors are 0, -9 , and 54 . In this case neither the conditions in part (a) of the theorem nor the conditions in part (b) are satisfied. Moreover, $D_3(0, 0, 0) = 54 \neq 0$. According to part (c), $(0, 0, 0)$ is a saddle point. ■

Necessary Second-Order Conditions for Local Extrema

Suppose $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ is an interior stationary point for a C^2 function $f(\mathbf{x})$. Then, according to (6), the assumption that the Hessian matrix $\mathbf{f}''(\mathbf{x}^*)$ is negative definite is sufficient for $\mathbf{x}^*$ to be a local maximum point. But the condition is not necessary. A standard counterexample is $f(x, y) = -x^4 - y^4$, which has a (global) maximum at $(0, 0)$. Yet $f''_{11}(0, 0) = 0$, so $\mathbf{f}''(\mathbf{x}^*)$ is not negative definite. However, we claim that $\mathbf{f}''(\mathbf{x}^*)$ has to be negative semidefinite in order for $\mathbf{x}^*$ to be a local maximum point.

Consider for each $i = 1, \dots, n$ the function $g(x_i) = f(x_1^*, \dots, x_{i-1}^*, x_i, x_{i+1}^*, \dots, x_n^*)$ of one variable. It has a stationary point at x_i^* , because stationarity of f at $\mathbf{x}^*$ implies that $g'(x_i^*) = f'_i(\mathbf{x}^*) = 0$. In order for $g(x_i)$ to have a local maximum at x_i^* , one must have $g''(x_i^*) = f''_{ii}(\mathbf{x}^*) \leq 0$. (If not, g would have a local minimum at x_i^* .) Thus $f''_{11}(\mathbf{x}^*) \leq 0, \dots, f''_{nn}(\mathbf{x}^*) \leq 0$ are necessary for f to have a local maximum at $\mathbf{x}^*$ in the direction of each coordinate axis, and thus necessary for $\mathbf{x}^*$ to be a local maximum point for f . But these conditions do not say much about whether f has a local maximum or minimum in directions through $\mathbf{x}^*$ other than those parallel to one of the coordinate axes.

To study the behaviour of f in an arbitrary direction, define the function g by

$$g(t) = f(\mathbf{x}^* + t\mathbf{h}) = f(x_1^* + th_1, \dots, x_n^* + th_n) \quad (7)$$

where $\mathbf{h} = (h_1, \dots, h_n)$ is an arbitrary fixed vector in $\mathbb{R}^n$ with length 1, so $\|\mathbf{h}\| = 1$. The function g describes the behaviour of f along the straight line through $\mathbf{x}^*$ parallel to the vector $\mathbf{h}$, as suggested in Fig. 1.

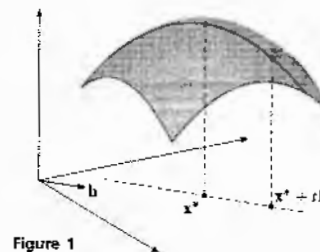


Figure 1

Suppose that $\mathbf{x}^*$ is an interior local maximum point for f . Then if $r > 0$ is small enough, the ball $B(\mathbf{x}^*; r) \subseteq S$. If $t \in (-r, r)$, then $\mathbf{x}^* + t\mathbf{h} \in B(\mathbf{x}^*; r)$ because $\|(\mathbf{x}^* + t\mathbf{h}) - \mathbf{x}^*\| = \|t\mathbf{h}\| = |t| < r$. But then for all t in $(-r, r)$, we have $\mathbf{x}^* + t\mathbf{h} \in S$ and so $f(\mathbf{x}^* + t\mathbf{h}) \leq f(\mathbf{x}^*)$, or $g(t) \leq g(0)$. Thus the function g has an interior maximum at $t = 0$. From the theory of functions of one variable, $g'(0) = 0$ and $g''(0) \leq 0$ are necessary conditions for a maximum at $t = 0$. By (2.6.5) and (2.6.6),

$$g'(t) = \sum_{i=1}^n f'_i(\mathbf{x}^* + t\mathbf{h})h_i, \quad g''(t) = \sum_{i=1}^n \sum_{j=1}^n f''_{ij}(\mathbf{x}^* + t\mathbf{h})h_i h_j \quad (8)$$

Putting $t = 0$, it follows that $g'(0) = 0$ and $g''(0) \leq 0$. The condition $g'(0) = 0$ for each $\mathbf{h}$ with $\|\mathbf{h}\| = 1$ affirms that f must be stationary. The condition $g''(0) \leq 0$ yields

$$\sum_{i=1}^n \sum_{j=1}^n f''_{ij}(\mathbf{x}^*) h_i h_j \leq 0 \quad \text{for all } \mathbf{h} = (h_1, \dots, h_n) \quad \text{with } \|\mathbf{h}\| = 1 \quad (9)$$

This is an equality if $\mathbf{h} = \mathbf{0}$, and if $\mathbf{h} = (h_1, \dots, h_n)$ is a vector with length $\neq 0$, the inequality holds for $\mathbf{h}/\|\mathbf{h}\|$, and so also for $\mathbf{h}$. Thus, in the terminology of Section 1.7, after noting that the quadratic form in (9) involves the Hessian matrix of f , we have:

A necessary condition for $f(\mathbf{x})$ to have a local minimum (maximum) at $\mathbf{x}^*$ is that the Hessian matrix of f at $\mathbf{x}^*$ is positive (negative) semidefinite. (10)

The results in (10) can be formulated by using Theorem 1.7.1. Recall that the *principal minors* of order k of the Hessian matrix are obtained by deleting $n - k$ rows in $[f''(\mathbf{x})] = [f''_{ij}(\mathbf{x})]_{n \times n}$ and the $n - k$ columns with the same numbers.

THEOREM 3.2.2 (NECESSARY CONDITIONS FOR LOCAL EXTREME POINTS)

Suppose that $f(\mathbf{x}) = f(x_1, \dots, x_n)$ is defined on a set S in $\mathbb{R}^n$, and $\mathbf{x}^*$ is an interior stationary point in S . Assume that f is C^2 in a ball around $\mathbf{x}^*$. Let $\Delta_k(\mathbf{x})$ denote an arbitrary principal minor of order k of the Hessian matrix. Then:

- (a) $\mathbf{x}^*$ is a local minimum point $\implies \begin{cases} \Delta_k(\mathbf{x}^*) \geq 0 \text{ for all principal minors} \\ \text{of order } k = 1, \dots, n. \end{cases}$
- (b) $\mathbf{x}^*$ is a local maximum point $\implies \begin{cases} (-1)^k \Delta_k(\mathbf{x}^*) \geq 0 \text{ for all principal} \\ \text{minors of order } k = 1, \dots, n. \end{cases}$

NOTE 1 The definiteness of a quadratic form can also be determined from the signs of the eigenvalues of the corresponding matrix. Suppose that neither of the conditions in Theorem 3.2.1(a) and (b) are satisfied. Then the Hessian matrix $f''(\mathbf{x}^*)$ is neither positive nor negative definite. According to Theorem 1.7.2, this means that it cannot have only positive or only negative eigenvalues. If, in addition, $D_n(\mathbf{x}^*) = |f''(\mathbf{x}^*)| \neq 0$, then 0 is not an eigenvalue of the Hessian matrix $f''(\mathbf{x}^*)$, so it must have both positive and negative eigenvalues. According to Theorem 1.7.2, the matrix $f''(\mathbf{x}^*)$ cannot then be either positive or negative semidefinite, so (10) shows that $\mathbf{x}^*$ is neither a local maximum nor a local minimum point. That is, $\mathbf{x}^*$ must be a saddle point. This proves the saddle point result in Theorem 3.2.1.

NOTE 2 Conclusions about local (or global) extreme points of functions of several variables cannot always be based only on what happens to the function along each straight line through a stationary point. For instance, $f(x, y) = (y - x^2)(y - 2x^2)$ has a saddle point at $(0, 0)$ even though the function has a local minimum along each straight line through the origin. (See Problem 13.3.6 in EMEA.)

PROBLEMS FOR SECTION 3.2

- SM** 1. The function $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + 3x_3^2 - x_1x_2 + 2x_1x_3 + x_2x_3$ defined on $\mathbb{R}^3$ has only one stationary point. Show that it is a local minimum point.
2. (a) Let f be defined for all (x, y) by $f(x, y) = x^3 + y^3 - 3xy$. Show that $(0, 0)$ and $(1, 1)$ are the only stationary points, and compute the quadratic form in (9) for f at the stationary points.
(b) Check the definiteness of the quadratic form at the stationary points.
(c) Classify the stationary points by using (1)–(3).
- SM** 3. Classify the stationary points of
(a) $f(x, y, z) = x^2 + x^2y + y^2z + z^2 - 4z$
(b) $f(x_1, x_2, x_3, x_4) = 20x_2 + 48x_3 + 6x_4 + 8x_1x_2 - 4x_1^2 - 12x_3^2 - x_4^2 - 4x_2^3$

HARDER PROBLEMS

4. Suppose $f(x, y)$ has only one stationary point (x^*, y^*) which is a local minimum point. Is (x^*, y^*) necessarily a global minimum point? It may be surprising that the answer is no. Prove this by examining the function defined for all (x, y) by $f(x, y) = (1 + y)^3 x^2 + y^2$. (Hint: Look at $f(x, -2)$ as $x \rightarrow \infty$.)

3.3 Equality Constraints: The Lagrange Problem

A general optimization problem with equality constraints is of the form

$$\max(\min) f(x_1, \dots, x_n) \quad \text{subject to} \quad \begin{cases} g_1(x_1, \dots, x_n) = b_1 \\ \dots \dots \dots \\ g_m(x_1, \dots, x_n) = b_m \end{cases} \quad (m < n) \quad (1)$$

where the b_j are constants. We assume that $m < n$ because otherwise there are usually no degrees of freedom.

In vector formulation, the problem is

$$\max(\min) f(\mathbf{x}) \quad \text{subject to} \quad g_j(\mathbf{x}) = b_j, \quad j = 1, \dots, m \quad (m < n) \quad (2)$$

(If we define the vector function $\mathbf{g} = (g_1, g_2, \dots, g_m)$ and let $\mathbf{b} = (b_1, b_2, \dots, b_m)$, the constraints can be expressed more simply as the vector equality $\mathbf{g}(\mathbf{x}) = \mathbf{b}$.)

The standard procedure for solving this problem is first to define the **Lagrange function**, or **Lagrangian**,

$$\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \lambda_1(g_1(\mathbf{x}) - b_1) - \dots - \lambda_m(g_m(\mathbf{x}) - b_m) \quad (3)$$

where $\lambda_1, \dots, \lambda_m$ are called **Lagrange multipliers**. The **necessary first-order conditions** for optimality are then:

$$\frac{\partial \mathcal{L}(\mathbf{x})}{\partial x_i} = \frac{\partial f(\mathbf{x})}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g_j(\mathbf{x})}{\partial x_i} = 0, \quad i = 1, \dots, n \quad (4)$$

The n equations in (4) and the m equations in (1) are to be solved simultaneously for the $n + m$ variables $x_1, \dots, x_n$ and $\lambda_1, \dots, \lambda_m$. The resulting solution vectors $(x_1, \dots, x_n)$ are then the candidates for optimality.

NOTE 1 Sometimes the Lagrangian is written in the alternative form $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \lambda_1 g_1(\mathbf{x}) - \dots - \lambda_m g_m(\mathbf{x})$, without the constants b_j . This makes no difference to the partial derivatives $\partial \mathcal{L}(\mathbf{x}) / \partial x_i$, of course.

An illuminating argument for condition (4) can be based on studying the (optimal) value function for problem (1). In the maximization case it is defined as

$$f^*(\mathbf{b}) = \max \{ f(\mathbf{x}) : g_j(\mathbf{x}) = b_j, \quad j = 1, \dots, m \} \quad (m < n) \quad (5)$$

If f denotes profit, and $\mathbf{b} = (b_1, \dots, b_m)$ denotes a resource vector, then $f^*(\mathbf{b})$ is the maximum profit obtainable given the available resource vector $\mathbf{b}$. In the following argument we assume that $f^*(\mathbf{b})$ is differentiable.

Fix a vector $\mathbf{b} = \bar{\mathbf{b}}$, and let $\bar{\mathbf{x}}$ be the corresponding optimal solution. Then $f(\bar{\mathbf{x}}) = f^*(\bar{\mathbf{b}})$. Obviously, for all $\mathbf{x}$ we have $f(\mathbf{x}) \leq f^*(\mathbf{g}(\mathbf{x}))$. Hence, $\varphi(\mathbf{x}) = f(\mathbf{x}) - f^*(\mathbf{g}(\mathbf{x}))$ has a maximum at $\mathbf{x} = \bar{\mathbf{x}}$, so

$$0 = \frac{\partial \varphi(\bar{\mathbf{x}})}{\partial x_i} = \frac{\partial f(\bar{\mathbf{x}})}{\partial x_i} - \sum_{j=1}^m \left[\frac{\partial f^*(\bar{\mathbf{b}})}{\partial b_j} \right]_{\mathbf{b}=\mathbf{g}(\bar{\mathbf{x}})} \frac{\partial g_j(\bar{\mathbf{x}})}{\partial x_i} \quad (*)$$

Suppose we define

$$\lambda_j = \frac{\partial f^*(\bar{\mathbf{b}})}{\partial b_j} \quad (6)$$

Then equation (*) reduces to (4).

NOTE 2 With a mild extra condition on the g_j functions, we shall see that the Lagrange multipliers obtained from the recipe are unique.

The argument for (4) assumes that $f^*(\mathbf{b})$ is differentiable, which is not always the case. To obtain (4) as a necessary condition for optimality, we need a "constraint qualification" given in the following theorem, which also gives sufficient conditions for optimality. We highly recommend that the reader studies the illuminating and simple proof of sufficiency.

THEOREM 3.3.1 (NECESSARY CONDITIONS AND SUFFICIENT CONDITIONS)

- (a) Suppose that the functions f and $g_1, \dots, g_m$ are defined on a set S in $\mathbb{R}^n$, and that $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ is an interior point of S that solves problem (1). Suppose further that f and $g_1, \dots, g_m$ are C^1 in a ball around $\mathbf{x}^*$, and that the $m \times n$ matrix of partial derivatives of the constraint functions

$$\mathbf{g}'(\mathbf{x}^*) = \begin{pmatrix} \frac{\partial g_1(\mathbf{x}^*)}{\partial x_1} & \dots & \frac{\partial g_1(\mathbf{x}^*)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial g_m(\mathbf{x}^*)}{\partial x_1} & \dots & \frac{\partial g_m(\mathbf{x}^*)}{\partial x_n} \end{pmatrix} \quad \text{has rank } m \quad (7)$$

Then there exist unique numbers $\lambda_1, \dots, \lambda_m$ such that (4) is valid.

- (b) If there exist numbers $\lambda_1, \dots, \lambda_m$ and an admissible $\mathbf{x}^*$ which together satisfy the first-order conditions (4), and if the Lagrangian $\mathcal{L}(\mathbf{x})$ defined by (3) is concave (convex) in $\mathbf{x}$, and if S is convex, then $\mathbf{x}^*$ solves the maximization (minimization) problem (1).

Proof of Theorem 3.3.1(b): Suppose that the Lagrangian $\mathcal{L}(\mathbf{x})$ is concave. Condition (4) means that the Lagrangian is stationary at $\mathbf{x}^*$. Then by Theorem 3.1.2,

$$\mathcal{L}(\mathbf{x}^*) = f(\mathbf{x}^*) - \sum_{j=1}^m \lambda_j (g_j(\mathbf{x}^*) - b_j) \geq f(\mathbf{x}) - \sum_{j=1}^m \lambda_j (g_j(\mathbf{x}) - b_j) = \mathcal{L}(\mathbf{x}) \quad \text{for all } \mathbf{x} \text{ in } S \quad (i)$$

But for all admissible $\mathbf{x}$, we have $g_j(\mathbf{x}) = b_j$ and, in particular, $g_j(\mathbf{x}^*) = b_j$ for all j . Hence, all admissible $\mathbf{x}$ satisfy $\sum_{j=1}^m \lambda_j g_j(\mathbf{x}) = \sum_{j=1}^m \lambda_j g_j(\mathbf{x}^*)$, so (i) implies that $f(\mathbf{x}^*) \geq f(\mathbf{x})$. Thus $\mathbf{x}^*$ solves problem (1). Note that part (b) of the theorem does not require the rank of $\mathbf{g}'(\mathbf{x}^*)$ to be m . ■

NOTE 3 The argument used to prove sufficiency actually uses only the fact that $\mathcal{L}(\mathbf{x}^*) \geq \mathcal{L}(\mathbf{x})$ for all $\mathbf{x}$. So even when $\mathcal{L}$ is not concave, $\mathbf{x}^*$ solves (1) if it is a global (unconstrained) maximum point for $\mathcal{L}$ that happens to satisfy $\mathbf{g}(\mathbf{x}^*) = \mathbf{b}$ for the given $\lambda_1, \dots, \lambda_m$. But it is a long lived misconception in the economic literature that if $\mathbf{x}^*$ solves problem (1), then $\mathbf{x}^*$ necessarily maximizes the Lagrangian. Such maximization is sufficient, but by no means necessary. (See EMEA, Problem 14.4.1 for a counter-example.)

NOTE 4 A standard proof of Theorem 3.3.1 appeals to the implicit function theorem, using the rank condition to "solve" the constraint for m of the variables as functions of the $n - m$ remaining variables. Inserting these m variables into the criterion function, the problem is reduced to seeking first-order conditions for an unconstrained maximum of the new objective function of $n - m$ variables. A full proof along these lines is given on the website for the book. (The case $n = 2, m = 1$ is discussed in EMEA, Section 14.4.)

Theorem 3.3.1 is actually a special case of Theorem 3.8.3. It is still valid if $\mathbf{x}$ is (also) restricted to some open set A , provided the sufficient condition of part (b) is strengthened by adding the requirement that A is convex.

NOTE 5 The condition on the rank of $\mathbf{g}'(\mathbf{x}^*)$ is called a *constraint qualification*. Using the gradient concept from Section 2.1, condition (4) can be expressed as:

$$\nabla f(\mathbf{x}^*) = \lambda_1 \nabla g_1(\mathbf{x}^*) + \cdots + \lambda_m \nabla g_m(\mathbf{x}^*) \quad (8)$$

The gradients $\nabla g_1(\mathbf{x}^*), \dots, \nabla g_m(\mathbf{x}^*)$ are the row vectors of $\mathbf{g}'(\mathbf{x}^*)$. The constraint qualification is therefore equivalent to the condition that the gradients $\nabla g_1(\mathbf{x}^*), \dots, \nabla g_m(\mathbf{x}^*)$ are linearly independent. (See Section 1.2.)

We now offer one possible interpretation of (8) in the problem

$$\max f(x, y, z) \text{ subject to } g_1(x, y, z) = b_1, \quad g_2(x, y, z) = b_2 \quad (*)$$

assuming that $f(x, y, z)$ measures the temperature at the point (x, y, z) in space. Each constraint represents (in general) a surface in 3-space, and the admissible set is, therefore, typically, a curve K in 3-space, the intersection of the two surfaces. See Fig. 1.

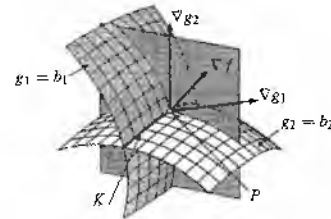


Figure 1 At P , ∇f is a linear combination of ∇g_1 and ∇g_2 .

At each point (x, y, z) on K , $f(x, y, z)$ records the temperature at that point. Problem (*) is therefore to find the hottest point $P = (x^*, y^*, z^*)$ on the curve K . The two gradients ∇g_1 and ∇g_2 are normals to their respective surfaces, and therefore both are normals to K at P . If the maximum temperature is T^* at P , then the level surface $f(x, y, z) = T^*$ must be tangent to the curve K at P . If this were not the case, the curve K would cut through the level surface of f and presumably would intersect level surfaces of f that correspond to higher as well as lower temperatures than T^* . Thus the gradient ∇f must be a normal to K at P . Therefore, the three vectors ∇f , ∇g_1 , and ∇g_2 all lie in the (two-dimensional) plane of normals to K at P . Provided that ∇g_1 and ∇g_2 are linearly independent, ∇f can be expressed as a linear combination $\nabla f = \lambda_1 \nabla g_1 + \lambda_2 \nabla g_2$ of ∇g_1 and ∇g_2 , for appropriate numbers λ_1 and λ_2 . This is exactly what (8) states.

EXAMPLE 1

Use Theorem 3.3.1 to solve the problem

$$\max f(x, y, z) = x + 2z \text{ subject to } \begin{cases} g_1(x, y, z) = x + y + z = 1 \\ g_2(x, y, z) = x^2 + y^2 + z = 7/4 \end{cases}$$

(Hint: Eliminate the Lagrange multipliers from the first-order conditions to show that one gets $y = 2x - 1/2$.)

Solution: Here $\mathcal{L}(x, y, z) = x + 2z - \lambda_1(x + y + z - 1) - \lambda_2(x^2 + y^2 + z - 7/4)$, so the first-order conditions are

$$(i) \mathcal{L}'_1 = 1 - \lambda_1 - 2\lambda_2 x = 0 \quad (ii) \mathcal{L}'_2 = -\lambda_1 - 2\lambda_2 y = 0 \quad (iii) \mathcal{L}'_3 = 2 - \lambda_1 - \lambda_2 = 0$$

From equation (iii) we get $\lambda_2 = 2 - \lambda_1$, which inserted into (ii) gives $-\lambda_1 - 4y + 2\lambda_1 y = 0$, or $\lambda_1(2y - 1) = 4y$. This equation implies that $y \neq 1/2$, so $\lambda_1 = 4y/(2y - 1)$. Inserting this into (i) with $\lambda_2 = 2 - \lambda_1$ eventually yields $y = 2x - 1/2$. The constraints then reduce to $3x + z = 3/2$ and $5x^2 - 2x + z = 3/2$. The first of these equations yields $z = -3x + 3/2$, which inserted into the second gives $5x(x - 1) = 0$. Hence, $x = 0$ or $x = 1$. For $x = 0$, we have $y = -1/2$ and $z = 3/2$. For $x = 1$, we have $y = 3/2$ and $z = -3/2$. Inserting these into the maximand shows that $f(0, -1/2, 3/2) = 3$ and $f(1, 3/2, -3/2) = -2$. We conclude that the only possible solution candidate is $(0, -1/2, 3/2)$. The associated values of the multipliers are $\lambda_1 = \lambda_2 = 1$.

When $\lambda_1 = \lambda_2 = 1$ the Lagrangian is $-x^2 - y^2 - y + 11/4$, which is a concave function of (x, y, z) . We conclude that $(0, -1/2, 3/2)$ is a maximum point. $\square$

Interpreting the Lagrange Multipliers

Equation (6) can be written as

$$\frac{\partial f^*(\mathbf{b})}{\partial b_j} = \lambda_j(\mathbf{b}), \quad j = 1, \dots, m \quad (9)$$

This tells us that the *Lagrange multiplier* $\lambda_j = \lambda_j(\mathbf{b})$ for the j th constraint is the rate at which the optimal value of the objective function changes w.r.t. changes in the constant b_j .

Suppose, for instance, that $f^*(\mathbf{b})$ is the maximum profit that a firm can obtain from a production process when $b_1, \dots, b_m$ are the available quantities of m different resources. Then $\partial f^*(\mathbf{b})/\partial b_j$ is the marginal profit that the firm can earn per extra unit of resource j , which is therefore the firm's marginal willingness to pay for this resource. Equation (9) tells us that this marginal willingness to pay is equal to the Lagrange multiplier λ_j for the corresponding resource constraint whose right-hand side in (1) is b_j . If the firm could buy more of this resource at a price below λ_j per unit, it could earn more profit by doing so; but if the price exceeds λ_j , the firm could increase its overall profit by selling a small enough quantity of the resource at this price because the revenue from the sale would exceed the profit from production that is sacrificed by giving up the resource.

In economics, the number $\lambda_j(\mathbf{b})$ is referred to a **shadow price** (or demand price) of the resource j . It is a shadow rather than an actual price because it need not correspond to a market price. Indeed, the resource may be something unique like a particular entrepreneur's time for which there is not even a market that could determine its actual price.

Note that the Lagrange multipliers for problem (1) may well be negative, so that an increase in b_j can lead to a decrease in the value function.

If $db_1, \dots, db_m$ are small in absolute value, then, according to the linear approximation formula, $f^*(\mathbf{b} + d\mathbf{b}) - f^*(\mathbf{b}) \approx (\partial f^*(\mathbf{b})/\partial b_1) db_1 + \cdots + (\partial f^*(\mathbf{b})/\partial b_m) db_m$ and, using (9),

$$f^*(\mathbf{b} + d\mathbf{b}) - f^*(\mathbf{b}) \approx \lambda_1(\mathbf{b}) db_1 + \cdots + \lambda_m(\mathbf{b}) db_m \quad (10)$$

This formula makes it possible to estimate the change in the value function when one or more components of the resource vector are slightly changed.

EXAMPLE 2. Consider Example 1 and suppose we change the first constraint to $x + y + z = 0.98$ and the second constraint to $x^2 + y^2 + z = 1.80$. Estimate the corresponding change in the value function by using (10). Then solve the constrained optimization problem with the new right-hand sides, and find the corresponding (exact) value of the value function.

Solution: Using the notation introduced above and the results in Example 1, we have $b_1 = 1$, $b_2 = 1.75$, $db_1 = -0.02$, $db_2 = 0.05$, $\lambda_1(1, 1.75) = \lambda_2(1, 1.75) = 1$, and $f^*(b_1, b_2) = f^*(1, 1.75) = 0 + 2(3/2) = 3$. Then (10) yields $f^*(1 - 0.02, 1.75 + 0.05) - f^*(1, 1.75) \approx \lambda_1(1, 1.75) db_1 + \lambda_2(1, 1.75) db_2 = 1 \cdot (-0.02) + 1 \cdot (0.05) = 0.03$. Thus, $f^*(0.98, 1.80) = f^*(1 - 0.02, 1.75 + 0.05) \approx 3 + 0.03 = 3.03$.

In the new optimization problem with the right-hand sides adjusted, equation (iv) in Example 1 can be derived exactly as before. Thus, we end up with the three equations $x + y + z = 0.98$, $x^2 + y^2 + z = 1.8$, and $y = 2x - 0.5$. The solutions are $(x_1, y_1, z_1) \approx (-0.0138, -0.5276, 1.5214)$ and $(x_2, y_2, z_2) \approx (1.0138, 1.5276, -1.5614)$. The first solution is optimal and the value function is 3.029. So using the approximation in (10) gives a very good estimate of the change in the value function. $\blacksquare$

EXAMPLE 3 An economy consists of two consumers with labels $i = 1, 2$. They exchange two goods, labelled $j = 1, 2$. Suppose there is a fixed total endowment e_j of each good to be distributed between the two consumers. Let c_j^i denote i 's consumption of good j . Suppose that each consumer i has preferences represented by the utility function

$$U^i(c_1^i, c_2^i) = \alpha_i \ln c_1^i + \alpha_2 \ln c_2^i$$

where the parameters α_j are positive, and independent of i , with $\alpha_1 + \alpha_2 = 1$. Suppose the goods are to be distributed in order to maximize social welfare in the form of the weighted sum $W = \beta_1 U^1 + \beta_2 U^2$, where the weights β_i are positive, and $\beta_1 + \beta_2 = 1$.

- Formulate the welfare maximization problem with one equality constraint for each of the goods.
- Write down the Lagrangian, where λ_j denotes the Lagrange multiplier associated with the constraint for good j . Find the welfare maximizing distribution of the goods.
- Verify that $\lambda_j = \partial W^* / \partial e_j$, where W^* denotes the maximum value of W .

Solution: (a) The problem is to find consumer 1's consumption of the two goods, c_1^1 and c_2^1 , as well as consumer 2's consumption of the two goods, c_1^2 and c_2^2 , which together solve the problem:

$$\max \beta_1 (\alpha_1 \ln c_1^1 + \alpha_2 \ln c_2^1) + \beta_2 (\alpha_1 \ln c_1^2 + \alpha_2 \ln c_2^2) \quad \text{subject to} \quad \begin{cases} c_1^1 + c_1^2 = e_1 \\ c_2^1 + c_2^2 = e_2 \end{cases}$$

(b) The Lagrangian is

$$\mathcal{L} = \beta_1 (\alpha_1 \ln c_1^1 + \alpha_2 \ln c_2^1) + \beta_2 (\alpha_1 \ln c_1^2 + \alpha_2 \ln c_2^2) - \lambda_1 (c_1^1 + c_1^2 - e_1) - \lambda_2 (c_2^1 + c_2^2 - e_2)$$

The first-order conditions take the form

$$\partial \mathcal{L} / \partial c_1^1 = \alpha_1 \beta_1 / c_1^1 - \lambda_1 = 0, \quad \partial \mathcal{L} / \partial c_2^1 = \alpha_2 \beta_1 / c_2^1 - \lambda_1 = 0 \quad (*)$$

$$\partial \mathcal{L} / \partial c_1^2 = \alpha_1 \beta_2 / c_1^2 - \lambda_2 = 0, \quad \partial \mathcal{L} / \partial c_2^2 = \alpha_2 \beta_2 / c_2^2 - \lambda_2 = 0 \quad (**)$$

The first equalities in (*) and (**) imply respectively $c_1^1 = \alpha_1 \beta_1 / \lambda_1$ and $c_2^1 = \alpha_2 \beta_1 / \lambda_1$. From the first constraint, $e_1 = c_1^1 + c_1^2 = \alpha_1 \beta_1 / \lambda_1 + \alpha_1 \beta_2 / \lambda_1 = \alpha_1 (\beta_1 + \beta_2) / \lambda_1 = \alpha_1 / \lambda_1$, because $\beta_1 + \beta_2 = 1$. So $\lambda_1 = \alpha_1 / e_1$. This implies in turn that $c_1^1 = \alpha_1 \beta_1 e_1 / \alpha_1 = \beta_1 e_1$ and $c_2^1 = \alpha_2 \beta_1 e_1 / \alpha_1 = \beta_2 e_1$. In the same way, we find $c_1^2 = \beta_1 e_2$ and $c_2^2 = \beta_2 e_2$.

Thus, each individual i receives a share β_i of the total endowment of each good, which is equal to the weight given to i 's utility in the welfare sum $W = \beta_1 U^1 + \beta_2 U^2$.

(c) The maximum value of W is

$$W^* = \beta_1 [\alpha_1 \ln(\beta_1 e_1) + \alpha_2 \ln(\beta_2 e_2)] + \beta_2 [\alpha_1 \ln(\beta_2 e_1) + \alpha_2 \ln(\beta_2 e_2)]$$

So $\partial W^* / \partial e_1 = \beta_1 (\alpha_1 (1/\beta_1 e_1) \beta_1 + \beta_2 \alpha_1 (1/\beta_2 e_1) \beta_2) = (\alpha_1 / e_1) (\beta_1 + \beta_2) = \alpha_1 / e_1 = \lambda_1$. In the same way we see that $\partial W^* / \partial e_2 = \lambda_2$. $\blacksquare$

An Economic Interpretation

Consider a firm producing some final product by using n different intermediate goods. Those intermediate goods are themselves produced using as inputs m different resources whose total supplies are $b_1, \dots, b_m$. Given the quantities $x_1, \dots, x_n$ of the intermediate goods, let $f(\mathbf{x}) = f(x_1, \dots, x_n)$ denote the number of units of the final output, and let $g_j(\mathbf{x})$ be the corresponding number of units of resource number j required, $j = 1, \dots, m$. Problem (1) can then be formulated as follows:

Find the amounts $x_1, \dots, x_n$ of the intermediate goods that give the largest possible output of the final good, while making full use of all the resources.

Suppose the price of output is 1, and let p_j denote the price per unit of resource j , $j = 1, \dots, m$. Then the net profit at these prices is

$$P(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^m p_j g_j(\mathbf{x}) \quad (11)$$

Suppose that $P(\mathbf{x})$ is concave. A sufficient condition for P to have a maximum point with $x_1 > 0, \dots, x_n > 0$, is that

$$\frac{\partial P(\mathbf{x})}{\partial x_i} = \frac{\partial f(\mathbf{x})}{\partial x_i} - \sum_{j=1}^m p_j \frac{\partial g_j(\mathbf{x})}{\partial x_i} = 0, \quad i = 1, \dots, n \quad (12)$$

For each resource price vector $\mathbf{p} = (p_1, \dots, p_m)$, this system of equations determines optimal values of $x_1, \dots, x_n$, and, in turn, the quantities $g_1(\mathbf{x}), \dots, g_m(\mathbf{x})$ of the m resources needed.

Depending on the price vector $\mathbf{p}$, the quantity of resource j used at the optimum will be less than, equal to, or larger than the stock b_j available of that resource. Is it possible to choose prices for the resources so that the quantities actually used at the optimum are precisely equal to the available resources? Since (12) gives the first-order conditions for problem (1), we

see that the answer is yes: make each resource price equal to the corresponding Lagrange multiplier obtained from the first-order conditions for problem (1):

So, when the Lagrange multipliers are used as resource prices, maximizing the net profit in (11) leads to the available stock of each resource being used in full. The resulting quantities $x_1, \dots, x_n$ are those that maximize production subject to the resource constraints.

Assume now that the firm has solved its internal production problem. Then the value function $f^*(\mathbf{b})$ is determined. If the firm were to buy the amounts b_j at prices $\bar{\lambda}_j$, then it would want to maximize

$$\pi(\mathbf{b}) = f^*(\mathbf{b}) - \sum_{j=1}^m \bar{\lambda}_j b_j \quad (*)$$

Here $\bar{\lambda}_j$ are the Lagrange multipliers associated with a given resource vector $\mathbf{b} = \bar{\mathbf{b}}$. The first-order conditions for maximum profit at $\bar{\mathbf{b}}$ are

$$\partial f^*(\bar{\mathbf{b}})/\partial b_j = \bar{\lambda}_j, \quad j = 1, \dots, m \quad (**)$$

This equality accords with (9). If $\pi(\mathbf{b})$ is concave, then $\pi(\mathbf{b})$ does have a maximum at $\bar{\mathbf{b}}$.

Envelope Result

Consider the following version of the general Lagrange problem (1):

$$\max_{\mathbf{x}} f(\mathbf{x}, \mathbf{r}) \quad \text{subject to} \quad g_j(\mathbf{x}, \mathbf{r}) = 0, \quad j = 1, \dots, m \quad (13)$$

where $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{r} = (r_1, \dots, r_k)$ is a vector of parameters. Note that we can absorb each constant b_j in (1) by including it as a component of the parameter vector $\mathbf{r}$ and by including a term $-b_j$ in the corresponding function $g_j(\mathbf{x}, \mathbf{r})$. If we put $\mathbf{g} = (g_1, \dots, g_m)$, the m constraints can then be written as a vector equality, $\mathbf{g}(\mathbf{x}, \mathbf{r}) = \mathbf{0}$. Note that in problem (13) we maximize w.r.t. $\mathbf{x}$, with $\mathbf{r}$ held constant.

The value function for problem (13) is

$$f^*(\mathbf{r}) = \max_{\mathbf{x}} \{ f(\mathbf{x}, \mathbf{r}) : \mathbf{g}(\mathbf{x}, \mathbf{r}) = \mathbf{0} \} \quad (14)$$

Let the Lagrangian be defined as

$$\mathcal{L}(\mathbf{x}, \mathbf{r}) = f(\mathbf{x}, \mathbf{r}) - \sum_{j=1}^m \lambda_j g_j(\mathbf{x}, \mathbf{r})$$

We want to find an expression for $\partial f^*(\mathbf{r})/\partial r_i$ at a given point $\bar{\mathbf{r}}$, assuming there is a unique optimal choice $\mathbf{x}^*(\bar{\mathbf{r}})$ for $\mathbf{x}$. Let $\lambda_1, \dots, \lambda_m$ be the associated Lagrange multipliers. Under certain conditions (see Theorem 3.10.4), we also have the following relationship:

ENVELOPE RESULT

$$\frac{\partial f^*(\bar{\mathbf{r}})}{\partial r_i} = \left(\frac{\partial \mathcal{L}(\mathbf{x}, \mathbf{r})}{\partial r_i} \right)_{\mathbf{x}=\mathbf{x}^*(\bar{\mathbf{r}})}, \quad i = 1, \dots, k \quad (15)$$

The expression on the right-hand side is obtained by differentiating $\mathcal{L}(\mathbf{x}, \mathbf{r})$ w.r.t. its $(n+i)$ th argument, which is r_i , and then evaluating the derivative at $\mathbf{x}^*(\bar{\mathbf{r}})$. This result generalizes the envelope results from Section 3.1. For a proof based on (9), see Problem 10. The result is a special case of Theorem 3.10.4. For a more standard proof of the latter, assuming differentiability, see the website.

EXAMPLE 4* Consider the standard utility maximization problem (see Example 2.2.2):

$$\max U(\mathbf{x}) \quad \text{subject to} \quad \mathbf{p} \cdot \mathbf{x} = m, \quad \mathbf{x} \geq \mathbf{0}, \quad \mathbf{p} \gg \mathbf{0} \quad (16)$$

The maximum value of U will depend on the price vector $\mathbf{p}$ and income m : $U^* = U^*(\mathbf{p}, m)$, which is called the **indirect utility function**. Find expressions for $\partial U^*/\partial m$ and $\partial U^*/\partial p_i$.

Solution: The Lagrangian is $\mathcal{L}(\mathbf{x}, m) = U(\mathbf{x}) - \lambda(\mathbf{p} \cdot \mathbf{x} - m)$ and, since $\partial \mathcal{L}/\partial m = \lambda$, (15) gives

$$\frac{\partial U^*}{\partial m} = \lambda \quad (17)$$

In this case λ measures the limiting increase in maximal utility per unit increase in income. Therefore, λ is often called the **marginal utility of income**. Moreover, equation (15) gives

$$\frac{\partial U^*}{\partial p_i} = \frac{\partial \mathcal{L}}{\partial p_i} = -\lambda x_i^*, \quad i = 1, \dots, n \quad (\text{Roy's identity}) \quad (18)$$

This formula has a nice interpretation: the marginal disutility of a price increase is the marginal utility of income (λ) multiplied by the quantity demanded (x_i^*). Intuitively, this is because, for a small price change, the loss of real income is approximately equal to the change in price multiplied by the quantity demanded.

PROBLEMS FOR SECTION 3.3

- (a) Solve the problem $\max 100 - x^2 - y^2 - z^2$ subject to $x + 2y + z = a$.
(b) Compute the optimal value function $f^*(a)$ and verify that (9) holds.

Ⓜ 2. (a) Solve the problem

$$\max x + 4y + z \quad \text{subject to} \quad x^2 + y^2 + z^2 = 216 \quad \text{and} \quad x + 2y + 3z = 0$$

- Change the first constraint to $x^2 + y^2 + z^2 = 215$ and the second to $x + 2y + 3z = 0.1$. Estimate the corresponding change in the maximum value by using (10).

- (a) Solve the problem $\max e^x + y + z$ subject to $\begin{cases} x + y + z = 1 \\ x^2 + y^2 + z^2 = 1 \end{cases}$

- Replace the constraints by $x + y + z = 1.02$ and $x^2 + y^2 + z^2 = 0.98$. What is the approximate change in optimal value of the objective function?

Ⓜ 4. (a) Solve the utility maximizing problem (assuming $m \geq 4$)

$$\max U(x_1, x_2) = \frac{1}{2} \ln(1 + x_1) + \frac{1}{4} \ln(1 + x_2) \quad \text{subject to} \quad 2x_1 + 3x_2 = m$$

- With $U^*(m)$ as the indirect utility function, show that $dU^*/dm = \lambda$.

5. (a) Solve the problem $\max 1 - rx^2 - y^2$ subject to $x + y = m$, with $r > 0$.
 (b) Find the value function $f^*(r, m)$, compute $\partial f^*/\partial r$ and $\partial f^*/\partial m$ and verify (15).

6. (a) Solve the problem

$$\max x^2 + y^2 + z^2 \quad \text{subject to} \quad x^2 + y^2 + 4z^2 = 1 \quad \text{and} \quad x + 3y + 2z = 0$$

- (b) Suppose we change the first constraint to $x^2 + y^2 + 4z^2 = 1.05$ and the second constraint to $x + 3y + 2z = 0.05$. Estimate the corresponding change in the value function.
7. (a) In Example 4 let $U(\mathbf{x}) = \sum_{j=1}^n \alpha_j \ln(x_j - a_j)$, where α_j , a_j , p_j , and m are all positive constants with $\sum_{j=1}^n \alpha_j = 1$, and with $m > \sum_{j=1}^n p_j a_j$. Show that if $\mathbf{x}^*$ solves problem (16), then the expenditure on good j is the following linear function of prices and income

$$p_j x_j^* = \alpha_j m + p_j a_j - \alpha_j \sum_{i=1}^n p_i a_i, \quad j = 1, 2, \dots, n$$

This is called the **linear expenditure system**.

- (b) Let $U^*(\mathbf{p}, m) = U(\mathbf{x}^*)$ denote the indirect utility function. Verify Roy's identity.

HARDER PROBLEMS

8. (a) Find the solution of the following problem by solving the constraints for x and y :

$$\text{minimize } x^2 + (y - 1)^2 + z^2 \quad \text{subject to} \quad x + y = \sqrt{2} \quad \text{and} \quad x^2 + y^2 = 1$$

- (b) Note that there are three variables and two constraints (z does not appear in the constraints). Show that the conditions in Theorem 3.3.1 are not satisfied, and that there are no Lagrange multipliers for which the Lagrangian is stationary at the solution point.

9. Let

$$Q(x_1, \dots, x_n) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j, \quad S = \{(x_1, \dots, x_n) : x_1^2 + \dots + x_n^2 = 1\}$$

Assume that the coefficient matrix $\mathbf{A} = (a_{ij})$ of the quadratic form Q is symmetric and prove that Q attains maximum and minimum values over the set S which are equal to the largest and smallest eigenvalues of $\mathbf{A}$. (Hint: Consider first the case $n = 2$. Write $Q(\mathbf{x})$ as $Q(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x}$. The first-order conditions give $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$.)

10. Consider the problem

$$\max_{\mathbf{x}, \mathbf{r}} f(\mathbf{x}, \mathbf{r}) \quad \text{subject to} \quad \begin{cases} g_j(\mathbf{x}, \mathbf{r}) < 0, & j = 1, \dots, m \\ r_i = b_{m+i}, & i = 1, \dots, k \end{cases} \quad (*)$$

where f and $g_1, \dots, g_m$ are given functions and $b_{m+1}, \dots, b_{m+k}$ are fixed parameters. (We maximize f w.r.t. both $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{r} = (r_1, \dots, r_k)$, but with $r_1, \dots, r_k$ completely fixed.) Define $\mathbf{b} = (0, \dots, 0, b_{m+1}, \dots, b_{m+k})$ (there are m zeros). Prove (15) by using (9) for $i = m + 1, \dots, m + k$ and those first-order conditions for problem (*) that refer to the variables r_j .

3.4 Local Second-Order Conditions

This section deals with local second-order conditions for the general optimization problem (3.3.1) with equality constraints. We begin by considering briefly the case with only one constraint, since this is the one that occurs most commonly in economics:

$$\text{local max (min)} \quad f(\mathbf{x}) = f(x_1, \dots, x_n) \quad \text{subject to} \quad g(\mathbf{x}) = g(x_1, \dots, x_n) = b \quad (1)$$

The Lagrangian is $\mathcal{L} = f(\mathbf{x}) - \lambda(g(\mathbf{x}) - b)$. Suppose $\mathbf{x}^*$ satisfies the first-order conditions in Theorem 3.3.1, so there exists a λ such that the Lagrangian is stationary at $\mathbf{x}^*$. (The constraint qualification is that the gradient of g at $\mathbf{x}^*$ is not 0.) For each $r = 2, \dots, n$, define the **bordered Hessian determinant**²

$$B_r(\mathbf{x}^*) = \begin{vmatrix} 0 & g'_1(\mathbf{x}^*) & \dots & g'_r(\mathbf{x}^*) \\ g'_1(\mathbf{x}^*) & \mathcal{L}''_{11}(\mathbf{x}^*) & \dots & \mathcal{L}''_{1r}(\mathbf{x}^*) \\ \vdots & \vdots & \ddots & \vdots \\ g'_r(\mathbf{x}^*) & \mathcal{L}''_{r1}(\mathbf{x}^*) & \dots & \mathcal{L}''_{rr}(\mathbf{x}^*) \end{vmatrix} \quad (2)$$

Then we have the following results:

$$B_r(\mathbf{x}^*) < 0 \text{ for } r = 2, \dots, n \implies \mathbf{x}^* \text{ solves the local min. problem in (1).} \quad (3)$$

$$(-1)^r B_r(\mathbf{x}^*) > 0 \text{ for } r = 2, \dots, n \implies \mathbf{x}^* \text{ solves the local max. problem in (1).} \quad (4)$$

Consider next the general optimization problem with several equality constraints,

$$\text{local max (min)} \quad f(\mathbf{x}) \quad \text{subject to} \quad g_j(\mathbf{x}) = b_j, \quad j = 1, \dots, m \quad (m < n) \quad (5)$$

The Lagrangian is $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^m \lambda_j(g_j(\mathbf{x}) - b_j)$. The general results that extend (3) and (4) use the following determinants, for $r = m + 1, \dots, n$:

$$B_r(\mathbf{x}^*) = \begin{vmatrix} 0 & \dots & 0 & \frac{\partial g_1(\mathbf{x}^*)}{\partial x_1} & \dots & \frac{\partial g_1(\mathbf{x}^*)}{\partial x_r} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{\partial g_m(\mathbf{x}^*)}{\partial x_1} & \dots & \frac{\partial g_m(\mathbf{x}^*)}{\partial x_r} \\ \frac{\partial g_1(\mathbf{x}^*)}{\partial x_1} & \dots & \frac{\partial g_m(\mathbf{x}^*)}{\partial x_1} & \mathcal{L}''_{11}(\mathbf{x}^*) & \dots & \mathcal{L}''_{1r}(\mathbf{x}^*) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1(\mathbf{x}^*)}{\partial x_r} & \dots & \frac{\partial g_m(\mathbf{x}^*)}{\partial x_r} & \mathcal{L}''_{r1}(\mathbf{x}^*) & \dots & \mathcal{L}''_{rr}(\mathbf{x}^*) \end{vmatrix} \quad (6)$$

NOTE 1 In order to apply the following theorem you may have to renumber the variables in order to make the first m columns in the matrix $(\partial g_j(\mathbf{x}^*)/\partial x_j)$ linearly independent. (The

² It is called the bordered Hessian because it is the determinant of the Hessian matrix of $\mathcal{L}$ with an extra row and column added as "borders".

constraint qualification in Theorem 3.3.1 implies that this matrix must have rank m . So by Theorem 1.3.1 such a renumbering is possible.)

THEOREM 3.4.1 (SECOND-DERIVATIVE TEST: GENERAL CASE)

Suppose the functions f and $g_1, \dots, g_m$ are defined on a set S in $\mathbb{R}^n$, and let $\mathbf{x}^*$ be an interior point in S satisfying the necessary conditions in Theorem 3.3.1. Suppose that f and $g_1, \dots, g_m$ are C^2 in a ball around $\mathbf{x}^*$. Define the determinant $B_r(\mathbf{x}^*)$ by (6). Then:

- If $(-1)^m B_r(\mathbf{x}^*) > 0$ for $r = m+1, \dots, n$, then $\mathbf{x}^*$ solves the local minimization problem in (5).
- If $(-1)^r B_r(\mathbf{x}^*) > 0$ for $r = m+1, \dots, n$, then $\mathbf{x}^*$ solves the local maximization problem in (5).

Check to see that if $m = 1$, the conditions in (a) and (b) reduce to those in (3) and (4).

Note that the sign factor $(-1)^m$ is the same for all r in (a), while the sign factor in (b) varies with r . The conditions (a) and (b) on the signs of the determinants are referred to as the (local) **second-order conditions**. Note that these determinants are the last $n - m$ leading principal minors of the "full" determinant $B_n(\mathbf{x}^*)$ that we get in (6) when $r = n$. Note too that $n - m$ is the number of degrees of freedom remaining when m independent constraints are imposed on n variables.

EXAMPLE 1 It is easy to see that the only point that satisfies the first-order conditions for the problem

$$\text{local max (min)} \quad f(x, y, z) = x^2 + y^2 + z^2 \quad \text{s.t.} \quad \begin{cases} g_1(x, y, z) = x + 2y + z = 30 \\ g_2(x, y, z) = 2x - y - 3z = 10 \end{cases}$$

is $P = (10, 10, 0)$. What has Theorem 3.4.1 to say about this point?

Solution: With $m = 2$ and $n = 3$, the conditions in (a) reduce to $(-1)^2 B_r(P) > 0$ for $r = 3$, i.e. $B_3(P) > 0$. On the other hand, (b) reduces to $(-1)^r B_r(P) > 0$ for $r = 3$, i.e. $B_3(P) < 0$. Thus, only the sign of $B_3(P)$ must be checked. This determinant is

$$B_3(P) = \begin{vmatrix} 0 & 0 & \partial g_1 / \partial x & \partial g_1 / \partial y & \partial g_1 / \partial z \\ 0 & 0 & \partial g_2 / \partial x & \partial g_2 / \partial y & \partial g_2 / \partial z \\ \partial g_1 / \partial x & \partial g_2 / \partial x & \mathcal{L}''_{xx} & \mathcal{L}''_{xy} & \mathcal{L}''_{xz} \\ \partial g_1 / \partial y & \partial g_2 / \partial y & \mathcal{L}''_{yx} & \mathcal{L}''_{yy} & \mathcal{L}''_{yz} \\ \partial g_1 / \partial z & \partial g_2 / \partial z & \mathcal{L}''_{zx} & \mathcal{L}''_{zy} & \mathcal{L}''_{zz} \end{vmatrix} = \begin{vmatrix} 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 2 & -1 & -3 \\ 1 & 2 & 2 & 0 & 0 \\ 2 & -1 & 0 & 2 & 0 \\ 1 & -3 & 0 & 0 & 2 \end{vmatrix}$$

This determinant is equal to 150, so P is a local minimum point.

Motivation and Proof of Theorem 3.4.1

To give some explanation for the conditions in the theorem above, suppose for the moment that $\mathbf{x}^*$ is a local extreme point for the Lagrange function $\mathcal{L}$ itself. If in addition $\mathbf{x}^*$ satisfies the m equality constraints, then $\mathbf{x}^*$ obviously solves problem (5). Hence, if $\mathbf{x}^*$ is a stationary point for the Lagrangian, then

$$\sum_{i=1}^n \sum_{j=1}^n \mathcal{L}''_{ij}(\mathbf{x}^*) h_i h_j \text{ is negative (positive) definite} \quad (*)$$

is a sufficient condition for $\mathcal{L}$ to have a local maximum (minimum) at $\mathbf{x}^*$. Therefore, when $(*)$ holds and $\mathbf{x}^*$ is a stationary point that satisfies the constraints in (5), then $\mathbf{x}^*$ solves the problem. Note that the sufficient condition $(*)$ is "unnecessarily strong" in the sense that it considers every vector $\mathbf{h} = (h_1, \dots, h_n) \neq \mathbf{0}$, while it is enough that the quadratic form is negative (positive) only for variations in $\mathbf{h}$ that (roughly speaking) satisfy the restrictions imposed by the constraints. It turns out that it suffices to consider variations in $h_1, \dots, h_n$ that cause $\mathbf{x}^* + \mathbf{h}$ to vary within the intersection of the tangent planes of the graphs of $g_1, \dots, g_m$ at $\mathbf{x}^*$.

SUFFICIENT CONDITIONS FOR LOCAL OPTIMALITY

Suppose that $\mathbf{x}^*$ is a stationary point of the Lagrangian which satisfies the constraints, and that the quadratic form in $(*)$ is negative (positive) for all the $(h_1, \dots, h_n) \neq (0, \dots, 0)$ that satisfy the m linearly independent equations

$$\frac{\partial g_j(\mathbf{x}^*)}{\partial x_1} h_1 + \dots + \frac{\partial g_j(\mathbf{x}^*)}{\partial x_n} h_n = 0, \quad j = 1, 2, \dots, m \quad (7)$$

Then $\mathbf{x}^*$ is a solution to problem (5).

Proof of Theorem 3.4.1: To prove part (a) we must show that $f(\mathbf{x}^* + \mathbf{h}) - f(\mathbf{x}^*) \geq 0$ for all vectors $\mathbf{x}^* + \mathbf{h}$ that are sufficiently close to $\mathbf{x}^*$ and satisfy $g_i(\mathbf{x}^* + \mathbf{h}) = b_i$, $i = 1, \dots, m$.

We begin by expanding the Lagrangian $\mathcal{L}$ about $\mathbf{x}^*$ using Taylor's formula (as in Theorem 2.6.8), and including terms up to the second order. With $\mathbf{h} = (h_1, \dots, h_n)$, we get

$$\mathcal{L}(\mathbf{x}^* + \mathbf{h}) = \mathcal{L}(\mathbf{x}^*) + \sum_{i=1}^n \mathcal{L}'_i(\mathbf{x}^*) h_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathcal{L}''_{ij}(\mathbf{x}^* + c\mathbf{h}) h_i h_j, \quad c \in (0, 1) \quad (i)$$

Because $\mathbf{x}^*$ satisfies all the constraints, $\mathcal{L}(\mathbf{x}^*) = f(\mathbf{x}^*)$. Moreover, $\mathcal{L}$ is stationary at $\mathbf{x}^*$, so (i) can be written as

$$f(\mathbf{x}^* + \mathbf{h}) - f(\mathbf{x}^*) = \sum_{k=1}^m \lambda_k (g_k(\mathbf{x}^* + \mathbf{h}) - b_k) + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \mathcal{L}''_{ij}(\mathbf{x}^* + c\mathbf{h}) h_i h_j \quad (ii)$$

The first sum on the right-hand side is 0 when $\mathbf{x}^* + \mathbf{h}$ satisfies the constraints. Therefore, if the double sum in (ii) is positive for all such $\mathbf{x}^* + \mathbf{h} \neq \mathbf{x}^*$ sufficiently close to $\mathbf{x}^*$, then $\mathbf{x}^*$ is a local minimum point for problem (1).

We proceed by expanding each g_k about $\mathbf{x}^*$, this time retaining only terms of the first order:

$$g_k(\mathbf{x}^* + \mathbf{h}) - b_k = \sum_{j=1}^n \frac{\partial g_k(\mathbf{x}^* + c_k \mathbf{h})}{\partial x_j} h_j, \quad c_k \in (0, 1), \quad k = 1, \dots, m \quad (\text{iii})$$

where we have used the fact that $g_k(\mathbf{x}^*) = b_k$, $k = 1, \dots, m$.

Now consider the $(m+n) \times (m+n)$ bordered Hessian matrix

$$B(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^m) = \begin{pmatrix} 0 & G(\mathbf{x}^1, \dots, \mathbf{x}^m) \\ G(\mathbf{x}^1, \dots, \mathbf{x}^m)^T & \mathcal{L}''(\mathbf{x}^0) \end{pmatrix}$$

where $G(\mathbf{x}^1, \dots, \mathbf{x}^m) = (\partial g_j(\mathbf{x}^1)/\partial x_j)_{m \times n}$ for arbitrary vectors $\mathbf{x}^1, \dots, \mathbf{x}^m$ in some open ball around $\mathbf{x}^*$, and $\mathcal{L}''(\mathbf{x}^0)$ is the Hessian matrix of $\mathcal{L}$ evaluated at $\mathbf{x}^0$. For $r = m+1, \dots, n$, let $\bar{B}_r(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^m)$ be the $(m+r) \times (m+r)$ leading principal minor of this matrix. The determinants $\bar{B}_1, \dots, \bar{B}_n$ are continuous functions of the collection of vectors $(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^m)$, viewed as a point of $\mathbb{R}^{(m+1)n}$. Moreover, $\bar{B}_r(\mathbf{x}^*, \mathbf{x}^*, \dots, \mathbf{x}^*) = \bar{B}_r(\mathbf{x}^*)$. So, under the hypothesis that $(-1)^m \bar{B}_r(\mathbf{x}^*) > 0$ for $r = m+1, \dots, n$, there is an open ball U in $\mathbb{R}^n$ with its centre at $\mathbf{x}^*$ such that $(-1)^m \bar{B}_r(\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^m) > 0$ for $r = m+1, \dots, n$ whenever the $m+1$ vectors $\mathbf{x}^0, \mathbf{x}^1, \dots, \mathbf{x}^m$ all belong to U . Now, if $\mathbf{x}^* + \mathbf{h}$ belongs to U , then so do all the vectors $\mathbf{x}^* + c_k \mathbf{h}$ and $\mathbf{x}^* + c_k \mathbf{h}$, $k = 1, \dots, m$. By Theorem 1.8.1, if $\|\mathbf{h}\|$ is sufficiently small, then with $\mathbf{r} = (r_1, \dots, r_n)$,

$$\sum_{i=1}^n \sum_{j=1}^n \mathcal{L}''_{ij}(\mathbf{x}^* + c_k \mathbf{h}) r_i r_j > 0 \text{ for all } \mathbf{r} \neq \mathbf{0} \text{ such that } \sum_{j=1}^n \frac{\partial g_k(\mathbf{x}^* + c_k \mathbf{h})}{\partial x_j} r_j = 0, \quad k = 1, \dots, m \quad (\text{iv})$$

Suppose now that $\mathbf{x}^* + \mathbf{h}$ is a point of U that satisfies all the constraints. Then, according to (iii), $\sum_{j=1}^n (\partial g_k(\mathbf{x}^* + c_k \mathbf{h})/\partial x_j) h_j = 0$, for $k = 1, \dots, m$. If we put $r_j = h_j$, $j = 1, \dots, n$, it follows from (iv) that the double sum in (ii) is > 0 .

Part (b) is shown in a similar way, reversing appropriate inequalities, especially the one in (iv). ■

PROBLEMS FOR SECTION 3.4

1. (a) Find the four points that satisfy the first-order conditions for the problem

$$\max (\min) \quad x^2 + y^2 \quad \text{subject to} \quad 4x^2 + 2y^2 = 4$$

- (b) Compute $B_2(x, y)$ in (2) at the four points found in (a). What can you conclude?
(c) Can you give a geometric interpretation of the problem?

2. Compute B_2 and B_3 in (2) for the problem

$$\max (\min) \quad x^2 + y^2 + z^2 \quad \text{subject to} \quad x + y + z = 1$$

Show that the second-order conditions for a local minimum are satisfied.

3. Use Theorem 3.4.1 to classify the candidates for optimality in the problem

$$\text{local max} (\min) \quad x + y + z \quad \text{subject to} \quad x^2 + y^2 + z^2 = 1 \text{ and } x - y - z = 1$$

3.5 Inequality Constraints: Nonlinear Programming

A more general form of constrained optimization problem arises when we replace the equality constraints in the Lagrange problem (3.3.1) with inequality constraints. The result is the following **nonlinear programming problem**, which we will call the **standard problem**:

$$\max f(x_1, \dots, x_n) \quad \text{subject to} \quad \begin{cases} g_1(x_1, \dots, x_n) \leq b_1 \\ \dots \\ g_m(x_1, \dots, x_n) \leq b_m \end{cases} \quad (1)$$

where $b_1, \dots, b_m$ are all constants. A vector $\mathbf{x} = (x_1, \dots, x_n)$ that satisfies all the constraints is called **admissible** (or **feasible**). The set of all admissible vectors is called the **admissible set** (or the **feasible set**). We assume that f and all the g_j functions are C^1 .

Whereas in the Lagrange problem the number of constraints were assumed to be strictly less than the number of variables, in the present problem this restriction is not necessary. In fact, there can be many more constraints than variables.

Note that minimizing $f(\mathbf{x})$ is equivalent to maximizing $-f(\mathbf{x})$. Moreover, an inequality constraint of the form $g_j(\mathbf{x}) \geq b_j$ can be rewritten as $-g_j(\mathbf{x}) \leq -b_j$, whereas an equality constraint $g_j(\mathbf{x}) = b_j$ is equivalent to the pair of constraints $g_j(\mathbf{x}) \leq b_j$ and $-g_j(\mathbf{x}) \leq -b_j$. In this way, most constrained optimization problems can be expressed in the form (1).

The standard procedure for solving problem (1) is similar to the recipe used to solve the corresponding problem with equality constraints in Section 3.3.³ We define the Lagrangian exactly as before. That is, $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \lambda_1(g_1(\mathbf{x}) - b_1) - \dots - \lambda_m(g_m(\mathbf{x}) - b_m)$, where $\lambda_1, \dots, \lambda_m$ are the Lagrange multipliers. Again the first-order partial derivatives of the Lagrangian are equated to 0:

$$\frac{\partial \mathcal{L}(\mathbf{x})}{\partial x_i} = \frac{\partial f(\mathbf{x})}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g_j(\mathbf{x})}{\partial x_i} = 0, \quad i = 1, \dots, n \quad (2)$$

In addition, and this is the vitally important new feature, we introduce the **complementary slackness conditions**

$$\lambda_j \geq 0, \quad \text{with} \quad \lambda_j = 0 \text{ if } g_j(\mathbf{x}) < b_j, \quad j = 1, \dots, m \quad (3)$$

Finally, of course, the inequality constraints themselves have to be satisfied. Conditions (2) and (3) are often called the **Kuhn-Tucker conditions**.

Condition (3) is rather tricky. It requires that, for each j , the number λ_j is nonnegative, and moreover that $\lambda_j = 0$ if $g_j(\mathbf{x}) < b_j$. Thus, if $\lambda_j > 0$, we must have $g_j(\mathbf{x}) = b_j$. An alternative formulation of this condition is:

$$\lambda_j \geq 0, \quad \text{with} \quad \lambda_j [g_j(\mathbf{x}) - b_j] = 0, \quad j = 1, \dots, m \quad (4)$$

³ If you have not studied nonlinear programming before, it might be a good idea to study a somewhat more elementary treatment first, starting with the case $n = 2$, $m = 1$ as in, say, EMBA, Section 14.8.

The two inequalities $\lambda_j \geq 0$ and $g_j(\mathbf{x}) \leq b_j$ are **complementarily slack** in the sense that at most one can be "slack"—that is, at most one can hold with strict inequality. Equivalently, at least one must be an equality.

If $g_j(\mathbf{x}^*) = b_j$, we say that the constraint $g_j(\mathbf{x}) \leq b_j$ is **active** or **binding** at $\mathbf{x}^*$.

Warning: It is possible to have both $\lambda_j = 0$ and $g_j(\mathbf{x}) = b_j$ at the same time in (3). See Problem 3.6.1.

Let us see these conditions in action in a simple case.

EXAMPLE 1 Check what the conditions (2) and (3) give for the problem

$$\max f(x, y) = -(x-2)^2 - (y-3)^2 \quad \text{subject to} \quad x \leq 1, \quad y \leq 2$$

Solution: The solution is obviously $x^* = 1$ and $y^* = 2$, because $f(1, 2) = -2$ and, if we choose any other point (x, y) with $x \leq 1$ and $y \leq 2$, then $f(x, y)$ has a value less than -2 . So $(x^*, y^*) = (1, 2)$ solves the problem.

Let us see how to derive this solution by using the Kuhn–Tucker conditions. With the Lagrangian $\mathcal{L} = -(x-2)^2 - (y-3)^2 - \lambda_1(x-1) - \lambda_2(y-2)$, the conditions (2)–(3) for (x^*, y^*) to solve the problem become

$$\mathcal{L}'_x = -2(x^* - 2) - \lambda_1 = 0 \quad (\text{i})$$

$$\mathcal{L}'_y = -2(y^* - 3) - \lambda_2 = 0 \quad (\text{ii})$$

$$\lambda_1 \geq 0, \quad \text{with } \lambda_1 = 0 \text{ if } x^* < 1 \quad (\text{iii})$$

$$\lambda_2 \geq 0, \quad \text{with } \lambda_2 = 0 \text{ if } y^* < 2 \quad (\text{iv})$$

If $x^* < 1$, then from (iii) $\lambda_1 = 0$, and then (i) yields $x^* = 2$, contradicting the constraint $x \leq 1$. So $x^* = 1$. In the same way we see that $y^* < 2$ leads to $\lambda_2 = 0$ and so from (ii) $y^* = 3$, contradicting the constraint $y \leq 2$. But then $x^* = 1$ and $y^* = 2$, and from (ii) and (iv) we find $\lambda_1 = \lambda_2 = 2$. So we get the same solution as with the direct argument.

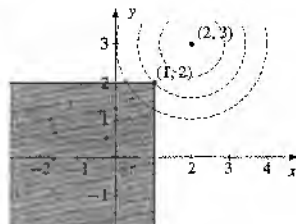


Figure 1

Another argument is this: maximizing $-(x-2)^2 - (y-3)^2$ subject to the constraints must have the same solutions for x and y as minimizing $(x-2)^2 + (y-3)^2$ subject to the same constraints. But this latter problem is geometrically to find the point in the shaded constraint set in Fig. 1 that is closest to $(2, 3)$. This is obviously the point $(1, 2)$. In Fig. 1

we have also indicated some level curves for the objective function. They are circles centered at $(2, 3)$. The optimum is found at the point where one of the level curves just touches the admissible set.

The Value Function

An ingenious argument for the Kuhn–Tucker conditions is based on studying the **(optimal) value function** for problem (1). It is defined as⁴

$$f^*(\mathbf{b}) = \max_{\mathbf{x}} \{f(\mathbf{x}) : g_j(\mathbf{x}) \leq b_j, \quad j = 1, \dots, m\} \quad (5)$$

In the following argument we assume that $f^*(\mathbf{b})$ is differentiable. Notice that the value function $f^*(\mathbf{b})$ must be nondecreasing in each variable $b_1, \dots, b_m$. This is because as b_j increases with all the other variables held fixed, the admissible set becomes larger; hence $f^*(\mathbf{b})$ cannot decrease.

Fix a vector $\mathbf{b} = \bar{\mathbf{b}}$, and let $\bar{\mathbf{x}}$ be a corresponding optimal solution, which must therefore satisfy $f(\bar{\mathbf{x}}) = f^*(\bar{\mathbf{b}})$. For any $\mathbf{x}$, we have $f(\mathbf{x}) \leq f^*(\mathbf{g}(\mathbf{x}))$ because $\mathbf{x}$ obviously satisfies the constraints in (5) when each $\bar{b}_j$ is replaced by $g_j(\mathbf{x})$. But then $f^*(\mathbf{g}(\mathbf{x})) \leq f^*(\mathbf{g}(\mathbf{x}) + \mathbf{b} - \mathbf{g}(\bar{\mathbf{x}}))$ since $\mathbf{g}(\bar{\mathbf{x}}) \leq \bar{\mathbf{b}}$ and f^* is nondecreasing. It follows that the function $\varphi(\mathbf{x}) = f(\mathbf{x}) - f^*(\mathbf{g}(\mathbf{x}) + \bar{\mathbf{b}} - \mathbf{g}(\bar{\mathbf{x}})) \leq 0$ for all $\mathbf{x}$. Since $\varphi(\bar{\mathbf{x}}) = 0$, $\varphi(\mathbf{x})$ has a maximum at $\bar{\mathbf{x}}$, so

$$0 = \frac{\partial \varphi(\bar{\mathbf{x}})}{\partial x_i} = \frac{\partial f(\bar{\mathbf{x}})}{\partial x_i} - \sum_{j=1}^m \frac{\partial f^*(\bar{\mathbf{b}})}{\partial b_j} \frac{\partial g_j(\bar{\mathbf{x}})}{\partial x_i}, \quad i = 1, \dots, n \quad (*)$$

If we define

$$\lambda_j = \frac{\partial f^*(\bar{\mathbf{b}})}{\partial b_j} \quad (6)$$

then the n equations $(*)$ reduce to (2). We thus have a counterpart to equations (3.3.6) and (3.3.9), with a corresponding interpretation of the Lagrange multiplier.

Since $f^*(\mathbf{b})$ is nondecreasing, from (6) we have $\lambda_j \geq 0$. Suppose $g_j(\bar{\mathbf{x}}) < \bar{b}_j$. We want to prove that $\lambda_j = 0$. Choose $\mathbf{b}' = (\bar{b}_1, \dots, \bar{b}_{j-1}, b_j, \bar{b}_{j+1}, \dots, \bar{b}_m)$, where $b_j \in (g_j(\bar{\mathbf{x}}), \bar{b}_j)$. Then $\mathbf{g}(\bar{\mathbf{x}}) \leq \mathbf{b}' \leq \bar{\mathbf{b}}$. Once again, because f^* is nondecreasing, we have $f^*(\mathbf{g}(\bar{\mathbf{x}})) \leq f^*(\mathbf{b}') \leq f^*(\bar{\mathbf{b}})$. But optimality of $\bar{\mathbf{x}}$ implies that $f(\bar{\mathbf{x}}) = f^*(\bar{\mathbf{b}})$. So $f^*(\mathbf{b}')$ is squeezed between two equal numbers, implying that $f^*(\mathbf{b}') = f^*(\bar{\mathbf{b}})$ whenever $b_j \in (g_j(\bar{\mathbf{x}}), \bar{b}_j)$. Under the assumption that $f^*(\mathbf{b})$ is differentiable at $\bar{\mathbf{b}}$, it follows that $\partial f^*(\bar{\mathbf{b}})/\partial b_j = 0$, so by (6), $\lambda_j = 0$. Thus the complementary slackness condition holds.

A Precise Result

In order to show that the Kuhn–Tucker conditions (2) and (3) are truly necessary for optimality, we should not assume that the value function f^* is differentiable. To avoid this unsatisfactory assumption, however, we need a restriction on the constraint functions made precise in the following theorem:

⁴ We assume that the maximum value always exists. This will be true, for example, in the common situation where the admissible set is bounded for all possible $\mathbf{b}$. Where the maximum does not exist, we have to replace $\max$ in (5) by $\sup$.

THEOREM 3.5.1 (KUHN-TUCKER NECESSARY CONDITIONS)

Suppose that $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ solves problem (1) where f and $g_1, \dots, g_m$ are C^1 functions on a set S in $\mathbb{R}^n$ and $\mathbf{x}^*$ is an interior point of S . Suppose furthermore that the following **constraint qualification** is satisfied:

CQ: The gradient vectors $\nabla g_j(\mathbf{x}^*)$, $1 \leq j \leq m$, corresponding to those constraints that are active at $\mathbf{x}^*$ are linearly independent.

Then there exist unique numbers $\lambda_1, \dots, \lambda_m$ such that the Kuhn-Tucker conditions (2)–(3) hold at $\mathbf{x} = \mathbf{x}^*$.

An alternative formulation of the CQ is this: delete all rows in the Jacobian matrix $\mathbf{g}'(\mathbf{x}^*)$ (see (3.3.7)) that correspond to constraints that are inactive at $\mathbf{x}^*$. Then the remaining matrix should have rank equal to the number of rows.

The theorem gives *necessary* conditions for an admissible vector to solve problem (1). In general, they are definitely not sufficient on their own. Indeed, suppose one can find a point $\mathbf{x}^*$ at which f is stationary and $g_j(\mathbf{x}^*) < b_j$ for all j . Then the Kuhn-Tucker conditions will automatically be satisfied by $\mathbf{x}^*$ together with all the Lagrange multipliers $\lambda_j = 0$. Yet while $\mathbf{x}^*$ could be a local or global maximum, it could also be a minimum or some kind of saddle point.

In the next section we shall see that, with proper concavity conditions, the Kuhn-Tucker conditions are sufficient.

How to Handle the Constraint Qualification

If the CQ fails at an optimal point, it may happen that this point does not satisfy the Kuhn-Tucker conditions. See Problem 5 for an example. In fact, since some textbooks are unclear about how to use the CQ, let us explain carefully how to apply Theorem 3.5.1. To find all possible solution candidates, you need to carry out the following two steps:

- (I) Find all admissible points where the Kuhn-Tucker conditions are satisfied.
- (II) Find also all the admissible points where the CQ fails.

If values of the objective function are calculated for all these candidates, the “best” candidates can be singled out: those giving the objective function the highest value among all candidates. If the problem has optimal points, then these are the same as the best candidates.

An erroneous procedure is sometimes encountered: when using the Kuhn-Tucker conditions to find a unique candidate $\mathbf{x}^*$, the CQ is checked only at $\mathbf{x} = \mathbf{x}^*$. However, the CQ may fail at other admissible points. These points will also be candidates. See Problem 5.

In the next example we take the constraint qualification seriously. However, it is more complicated than the previous problem and in such cases it is sometimes hard to know where to begin “attacking” the necessary conditions. A general method for finding the candidates for optimality in a nonlinear programming problem can be formulated as follows: first, examine the case in which all the constraints are active; then examine all cases in which all

but one of the constraints are active; then all cases in which all but two are active; and so on. Last, examine the case in which none of the constraints is active. At each step, we find all the vectors $\mathbf{x}$, with associated values of the Lagrange multipliers, that satisfy all the relevant conditions—if there are any. Then we calculate the value of the objective function for these values of $\mathbf{x}$, and retain those $\mathbf{x}$ with the highest values. The procedure should become clearer once you have seen it in action.

An alternative way to cover all cases starts by examining the case in which all the Lagrange multipliers are 0. Then, examine all cases where one of the multipliers is 0, while all the others are positive, etc. This method can be seen in action in the suggested answer to Problem 3.6.1.

EXAMPLE 2 Solve the problem:

$$\max f(x, y) = xy + x^2 \quad \text{subject to} \quad g_1(x, y) = x^2 + y \leq 2, \quad g_2(x, y) = -y \leq -1$$

Solution: Note that the second constraint is equivalent to $y \geq 1$. The Lagrangian is $\mathcal{L} = xy + x^2 - \lambda_1(x^2 + y - 2) - \lambda_2(-y + 1)$. So the Kuhn-Tucker conditions reduce to

$$\mathcal{L}'_x = y + 2x - 2\lambda_1 x = 0 \quad (i)$$

$$\mathcal{L}'_y = x - \lambda_1 + \lambda_2 = 0 \quad (ii)$$

$$\lambda_1 \geq 0, \quad \text{with } \lambda_1 = 0 \text{ if } x^2 + y < 2 \quad (iii)$$

$$\lambda_2 \geq 0, \quad \text{with } \lambda_2 = 0 \text{ if } y > 1 \quad (iv)$$

We start the systematic procedure:

- (I) *Both constraints are active.* Then $x^2 + y = 2$ and $y = 1$, and so $x = \pm 1$, $y = 1$. When $x = y = 1$, (i) and (ii) yield $\lambda_1 = 3/2$ and $\lambda_2 = 1/2$. Thus $(x, y) = (1, 1)$ with $\lambda_1 = 3/2$ and $\lambda_2 = 1/2$ is a solution candidate.
When $x = -1$, $y = 1$, (i) and (ii) yield $\lambda_1 = 1/2$ and $\lambda_2 = 3/2$. Thus $(x, y) = (-1, 1)$ with $\lambda_1 = 1/2$ and $\lambda_2 = 3/2$ is a solution candidate.
- (II) *Constraint 1 is active, 2 is inactive.* Then $x^2 + y = 2$ and $y > 1$. From (iv), $\lambda_2 = 0$, and (ii) yields $\lambda_1 = x$. Inserted into (i) this yields $y + 2x - 2x^2 = 0$. Since $y = 2 - x^2$, we get $3x^2 - 2x - 2 = 0$. The solutions are $x = \frac{1}{3}(1 \pm \sqrt{7})$. But $x = \lambda_1 \geq 0$, so only $x = \frac{1}{3}(1 + \sqrt{7})$ is admissible. But then $y = 2 - x^2 = \frac{2}{3}(5 - \sqrt{7})$, which we easily see is less than 1. So there is no solution candidate in this case.
- (III) *Constraint 1 is inactive, 2 is active.* Then $x^2 + y < 2$ and $y = 1$. But then from (iii), $\lambda_1 = 0$. Then (i) gives $x = -1/2$, and (ii) gives $\lambda_2 = 1/2$. Thus $(x, y) = (-1/2, 1)$ with $\lambda_1 = 0$ and $\lambda_2 = 1/2$ is a solution candidate.
- (IV) *Both constraints are inactive.* Then $x^2 + y > 2$ and $y > 1$, so (iii) and (iv) yield $\lambda_1 = \lambda_2 = 0$. Then from (i) and (ii) we have $y = 0$, which contradicts $y \geq 1$. So there is no solution candidate in this case.

The three solution candidates are $f(1, 1) = 2$, $f(-1, 1) = 0$, and $f(-1/2, 1) = -1/4$. Among these, the objective function is highest at $(1, 1)$. Since the objective function is

continuous and the admissible set is closed and bounded (why?), the extreme value theorem ensures that there is a solution. It remains only to check the constraint qualification.

The gradients of the two constraint functions are $\nabla g_1(x, y) = (2x, 1)$ and $\nabla g_2(x, y) = (0, -1)$. In case (I), when both constraints are active, only two points satisfy the constraints, which are already candidates for optimality. In case (II), only the first constraint is active, and we need only look at $\nabla g_1(x, y) = (2x, 1)$, which is linearly independent since it is not the zero vector.⁵ In case (III), only the second constraint is active, and we need only look at $\nabla g_2(x, y) = (0, -1)$, which is linearly independent since it is not the zero vector. Finally, in case (IV), the CQ holds trivially. So there are no admissible points at which the CQ fails. We conclude that (1, 1) solves the problem. $\square$

PROBLEMS FOR SECTION 3.5

1. Solve the problem $\max 1 - x^2 - y^2$ subject to $x \geq 2$ and $y \geq 3$ by a direct argument, and then see what the Kuhn–Tucker conditions have to say about the problem.

2. (a) Consider the nonlinear programming problem (where c is a positive constant)

$$\text{maximize } \ln(x+1) + \ln(y+1) \quad \text{subject to} \quad \begin{cases} x+2y \leq c \\ x+y \leq 2 \end{cases}$$

Write down the necessary Kuhn–Tucker conditions for a point (x, y) to be a solution of the problem.

- (b) Solve the problem for $c = 5/2$. (Theorem 3.6.1 will secure that the optimum is attained.)
(c) Let $V(c)$ denote the value function. Find the value of $V'(5/2)$.

3. Solve the following problem (assuming it has a solution)

$$\text{minimize } 4 \ln(x^2 + 2) + y^2 \quad \text{subject to} \quad x^2 + y \geq 2, \quad x \geq 1$$

(Hint: Reformulate it as a standard Kuhn–Tucker maximization problem.)

4. Solve the problem $\max -(x-a)^2 - (y-b)^2$ subject to $x \leq 1$, $y \leq 2$, for all possible values of the constants a and b . (A good check of the results is to use a geometric interpretation of the problem. See Example 1.)
5. Consider the problem $\max f(x, y) = xy$ subject to $g(x, y) = (x+y-2)^2 \leq 0$. Explain why the solution is $(x, y) = (1, 1)$. Verify that the Kuhn–Tucker conditions are not satisfied for any λ , and that the CQ does not hold at $(1, 1)$.

6. (a) Find the only possible solution to the nonlinear programming problem

$$\text{maximize } x^5 - y^3 \quad \text{subject to} \quad x \leq 1, \quad x \leq y$$

- (b) Solve the problem by using iterated optimization: Find first the maximum value $f(x)$ in the problem of maximizing $x^5 - y^3$ subject to $x \leq y$, where x is fixed and y varies. Then maximize $f(x)$ subject to $x \leq 1$.

⁵ A single vector $\mathbf{a}$ is linearly independent if and only if it is not the zero vector.

3.6 Sufficient Conditions

The Kuhn–Tucker conditions for the nonlinear programming problem

$$\max_{\mathbf{x}} f(\mathbf{x}) \quad \text{subject to} \quad g_j(\mathbf{x}) \leq b_j, \quad j = 1, \dots, m \quad (1)$$

are by themselves far from sufficient for optimality. However, as in the Lagrange problem, the conditions are sufficient if the Lagrangian is concave.

THEOREM 3.6.1 (SUFFICIENT CONDITIONS)

Consider the standard problem (1) with associated Lagrangian $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^m \lambda_j (g_j(\mathbf{x}) - b_j)$. Suppose that $\mathbf{x}^*$ is admissible and, in combination with the vector $\lambda = (\lambda_1, \dots, \lambda_m)$, satisfies conditions (3.5.2)–(3.5.3). Provided that the Lagrangian is concave, then $\mathbf{x}^*$ is optimal.

Proof: Since $\mathcal{L}(\mathbf{x})$ is concave and $\partial \mathcal{L}(\mathbf{x}^*) / \partial x_i = 0$ for $i = 1, \dots, n$, then according to Theorem 3.1.2(a), $\mathbf{x} = \mathbf{x}^*$ maximizes $\mathcal{L}(\mathbf{x})$. Hence, writing $\mathbf{g} = (g_1, \dots, g_m)$ and $\lambda = (\lambda_1, \dots, \lambda_m)$, one has for all $\mathbf{x}$, $f(\mathbf{x}^*) - \lambda \cdot (\mathbf{g}(\mathbf{x}^*) - \mathbf{b}) \geq f(\mathbf{x}) - \lambda \cdot (\mathbf{g}(\mathbf{x}) - \mathbf{b})$. Rearranging, we obtain the equivalent inequality

$$f(\mathbf{x}^*) - f(\mathbf{x}) \geq \lambda \cdot (\mathbf{g}(\mathbf{x}^*) - \mathbf{g}(\mathbf{x})) \quad (*)$$

It suffices to show that the sum on the right-hand side $\sum_{j=1}^m \lambda_j (g_j(\mathbf{x}^*) - g_j(\mathbf{x}))$ is ≥ 0 for all admissible $\mathbf{x}$, because this will imply that $\mathbf{x}^*$ solves problem (1).

Suppose that $g_j(\mathbf{x}^*) < b_j$. Then (3.5.3) shows that $\lambda_j = 0$. For those terms in the sum in (*) where $g_j(\mathbf{x}^*) = b_j$, we have $\lambda_j (g_j(\mathbf{x}^*) - g_j(\mathbf{x})) = \lambda_j (b_j - g_j(\mathbf{x})) \geq 0$, since $\mathbf{x}$ is admissible and $\lambda_j \geq 0$. The sum on the right-hand side of (*) therefore consists partly of terms that are 0 (since $\lambda_j = 0$), and partly of terms that are ≥ 0 . All in all, the sum is ≥ 0 . $\blacksquare$

NOTE 1 The proof actually shows that if $\mathbf{x}^*$ maximizes the Lagrangian, is admissible, and satisfies the complementary slackness condition (3.5.3), then $\mathbf{x}^*$ solves problem (1), even if $\mathcal{L}$ is not concave.

EXAMPLE 1

Find the maximum of $\frac{1}{2}x - y$ subject to $x + e^{-x} + z^2 \leq y$ and $x \geq 0$.

Solution: It is important first to write the problem in exactly the same form as (1), with all constraints as $\leq$ inequalities:

$$\max_{x, y, z} f(x, y, z) = \frac{1}{2}x - y \quad \text{subject to} \quad \begin{cases} g_1(x, y, z) = x + e^{-x} - y + z^2 \leq 0 \\ g_2(x, y, z) = -x \leq 0 \end{cases}$$

The Lagrangian is $\mathcal{L}(x, y, z) = \frac{1}{2}x - y - \lambda_1(x + e^{-x} - y + z^2) - \lambda_2(-x)$. Then the Kuhn–Tucker conditions take the form

$$\mathcal{L}'_x = \frac{1}{2} - \lambda_1(1 - e^{-x}) + \lambda_2 = 0 \quad (\text{i})$$

$$\mathcal{L}'_y = -1 + \lambda_1 = 0 \quad (\text{ii})$$

$$\mathcal{L}'_z = -2\lambda_2 z = 0 \quad (\text{iii})$$

$$\lambda_1 \geq 0, \text{ with } \lambda_1 = 0 \text{ if } x + e^{-x} + z^2 < y \quad (\text{iv})$$

$$\lambda_2 \geq 0, \text{ with } \lambda_2 = 0 \text{ if } x > 0 \quad (\text{v})$$

From (ii) we obtain $\lambda_1 = 1$, and then (iii) gives $z = 0$. Moreover, from (iv) and $z = 0$ we see that $x + e^{-x} = y$. Also (i) with $\lambda_1 = 1$ implies that $e^{-x} = \frac{1}{2} - \lambda_2 \leq \frac{1}{2}$, so $x \geq \ln 2 > 0$. Then (v) gives $\lambda_2 = 0$ and so (i) gives $x = \ln 2$, which implies that $y = x + e^{-x} = \ln 2 + \frac{1}{2}$.

Thus only the point $(x, y, z) = (\ln 2, \ln 2 + \frac{1}{2}, 0)$ is admissible and also satisfies the Kuhn-Tucker conditions with $\lambda_1 = 1, \lambda_2 = 0$. The Lagrangian is therefore $\mathcal{L}(x, y, z) = -\frac{1}{2}x - e^{-x} - z^2$, which is concave as the sum of the concave functions $-\frac{1}{2}x, -e^{-x}$, and $-z^2$. Theorem 3.6.1 then tells us that $(x, y, z) = (\ln 2, \ln 2 + \frac{1}{2}, 0)$ solves the problem. ■

In the next example it is convenient to use the systematic method first used in Example 3.5.2.

EXAMPLE 2 Solve the problem

$$\max f(x, y) = x^2 + 2y \quad \text{subject to} \quad x^2 + y^2 \leq m, \quad y \geq 0 \quad (m \text{ positive constant})$$

Solution: Rewriting the constraint $y \geq 0$ as $-y \leq 0$, the Lagrangian is $\mathcal{L} = x^2 + 2y - \lambda_1(x^2 + y^2 - m) - \lambda_2(-y)$. So the Kuhn-Tucker conditions reduce to

$$\mathcal{L}'_x = 2x - 2\lambda_1 x = 0 \quad (\text{i})$$

$$\mathcal{L}'_y = 2 - 2\lambda_1 y + \lambda_2 = 0 \quad (\text{ii})$$

$$\lambda_1 \geq 0, \text{ with } \lambda_1 = 0 \text{ if } x^2 + y^2 < m \quad (\text{iii})$$

$$\lambda_2 \geq 0, \text{ with } \lambda_2 = 0 \text{ if } y > 0 \quad (\text{iv})$$

We start the systematic procedure:

(I) *Both constraints are active.* Then $x^2 + y^2 = m$ and $y = 0$. But with $y = 0$, (ii) gives $\lambda_2 = -2$, which contradicts (iv). So there are no solution candidates in this case.

(II) *Constraint 1 is active, 2 is inactive.* In this case $x^2 + y^2 = m$ and $y > 0$. From (iv) we obtain $\lambda_2 = 0$, and (ii) gives $\lambda_1 y = 1$ while (i) implies $x(1 - \lambda_1) = 0$. From the last equality we conclude that *either* $x = 0$ *or* $\lambda_1 = 1$, *or both*.

If $x = 0$, then $x^2 + y^2 = m$ yields $y = \pm\sqrt{m}$. Since $y > 0$, only $y = \sqrt{m}$ is possible. Then $\lambda_1 = 1/\sqrt{m}$ and $\lambda_2 = 0$. Hence we have found that $(x, y) = (0, \sqrt{m})$, with $\lambda_1 = 1/\sqrt{m}$ and $\lambda_2 = 0$, satisfies (i)-(iv) and thus is one solution candidate.

If $\lambda_1 = 1$, then $\lambda_1 y = 1$ implies $y = 1$ and from $x^2 + y^2 = m$ we obtain $x = \pm\sqrt{m-1}$. Thus, provided $m \geq 1$, $(\sqrt{m-1}, 1)$ and $(-\sqrt{m-1}, 1)$ are two more solution candidates, with $\lambda_1 = 1$ and $\lambda_2 = 0$.

(III) *Constraint 1 is inactive, 2 is active.* Then $x^2 + y^2 < m$ and $y = 0$. But then from (ii) it follows that $\lambda_2 = -2$, a contradiction. So no candidate arises in this case.

(IV) *Both constraints are inactive.* Then $x^2 + y^2 < m, y > 0$, and $\lambda_1 = \lambda_2 = 0$. This contradicts (ii). So no candidate solution appears in this case.

Thus, if $0 < m < 1$, the only candidate satisfying all the Kuhn-Tucker conditions is $(x, y) = (0, \sqrt{m})$, with $\lambda_1 = 1/\sqrt{m}, \lambda_2 = 0$, and the objective function is $f(0, \sqrt{m}) = 2\sqrt{m}$.

If $m \geq 1$, in addition to $(x, y) = (0, \sqrt{m})$, the points $(\sqrt{m-1}, 1)$ and $(-\sqrt{m-1}, 1)$ also satisfy the Kuhn-Tucker conditions. For the two latter candidates the objective function is $f(\pm\sqrt{m-1}, 1) = m + 1$. Note that $m + 1 \geq 2\sqrt{m}$ if and only if $(\sqrt{m-1})^2 \geq 0$, which obviously is satisfied.

It does not look promising to apply Theorem 3.6.1 in this case, because $f(x, y) = x^2 + 2y$ is convex. Still, in the case $0 < m < 1$, with $\lambda_1 = 1/\sqrt{m}, \lambda_2 = 0$, the Lagrangian is

$$\mathcal{L} = x^2 + 2y - \frac{1}{\sqrt{m}}(x^2 + y^2 - 5) - 0 \cdot (-y) = \left(1 - \frac{1}{\sqrt{m}}\right)x^2 - \frac{y^2}{\sqrt{m}} + 2y + \frac{5}{\sqrt{m}}$$

This is actually concave in (x, y) since $1 - 1/\sqrt{m} < 0$ when $0 < m < 1$. So Theorem 3.6.1 shows that $(0, \sqrt{m})$ solves the problem when $m \in (0, 1)$. For $m \geq 1$, the two points $(\pm\sqrt{m-1}, 1)$ both solve the maximization problem because $\lambda_1 = 1$ and $\lambda_2 = 0$, so the Lagrangian in this case is

$$\mathcal{L} = x^2 + 2y - (x^2 + y^2 - 5) - 0 \cdot (-y) = -y^2 + 2y + 5$$

which is also concave. ■

The Lagrangian $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \lambda_1(g_1(\mathbf{x}) - b_1) - \dots - \lambda_m(g_m(\mathbf{x}) - b_m)$ is concave if $f(\mathbf{x})$ is concave and $\lambda_1 g_1(\mathbf{x}), \dots, \lambda_m g_m(\mathbf{x})$ are all convex, since a sum of concave functions is concave. The next theorem gives an interesting generalization.

THEOREM 3.6.2 (SUFFICIENT CONDITIONS II)

In Theorem 3.6.1 concavity of $\mathcal{L}(\mathbf{x})$ can be replaced by the following condition:

$f(\mathbf{x})$ is concave and $\lambda_j g_j(\mathbf{x}), j = 1, \dots, m$, are all quasiconvex

Proof: We want to show that for all admissible $\mathbf{x}$, $f(\mathbf{x}) - f(\mathbf{x}^*) \leq 0$. Since $f(\mathbf{x})$ is concave, then, according to Theorem 2.4.1,

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \nabla f(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) = \sum_{j=1}^m \lambda_j \nabla g_j(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \quad (\text{i})$$

where we also used (3.5.2) in the vector form that had been introduced in (3.3.8). It suffices, therefore, to show that for all $j = 1, \dots, m$, and all admissible $\mathbf{x}$,

$$\lambda_j \nabla g_j(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \leq 0 \quad (\text{ii})$$

The inequality in (ii) is satisfied for those j such that $g_j(\mathbf{x}^*) < b_j$, because then $\lambda_j = 0$. For those j such that $g_j(\mathbf{x}^*) = b_j$, we have $g_j(\mathbf{x}) \leq g_j(\mathbf{x}^*)$ because $\mathbf{x}$ is admissible. Therefore $-\lambda_j g_j(\mathbf{x}) \geq -\lambda_j g_j(\mathbf{x}^*)$. Since the function $-\lambda_j g_j$ is quasiconcave, it follows from Theorem 2.5.4 that we have the inequality $\nabla(-\lambda_j g_j(\mathbf{x}^*)) \cdot (\mathbf{x} - \mathbf{x}^*) \geq 0$, and hence $\lambda_j \nabla g_j(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \leq 0$. ■

NOTE 2 It is easy to see that the requirement that $\lambda_j g_j(\mathbf{x})$ is quasiconvex in Theorem 3.6.2 can be replaced by the weaker requirement that $\sum_{j=1}^m \lambda_j g_j(\mathbf{x})$ is quasiconvex. (Remember that a sum of quasiconvex functions is not necessarily quasiconvex.)

Quasiconcave Programming

The following theorem is important for economists, because in many economic optimization problems the objective function is assumed to be quasiconcave, rather than concave.

THEOREM 3.6.3 (SUFFICIENT CONDITIONS FOR QUASICONCAVE PROGRAMS)

Consider the standard problem (1), where the objective function f is C^1 and quasiconcave. Assume that there exist numbers $\lambda_1, \dots, \lambda_m$ and a vector $\mathbf{x}^*$ such that

- (a) $\mathbf{x}^*$ is admissible and satisfies the Kuhn–Tucker conditions (3.5.2)–(3.5.3);
- (b) $\nabla f(\mathbf{x}^*) \neq 0$;
- (c) $\lambda_j g_j(\mathbf{x})$ is quasiconvex for $j = 1, \dots, m$.

Then $\mathbf{x}^*$ is optimal.

Proof: We prove first that, for all $\mathbf{x}$,

$$f(\mathbf{x}) > f(\mathbf{x}^*) \Rightarrow \nabla f(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) > 0 \quad (*)$$

Suppose $f(\mathbf{x}) > f(\mathbf{x}^*)$ and choose $\alpha > 0$ so small that $f(\mathbf{x} - \alpha \nabla f(\mathbf{x}^*)) \geq f(\mathbf{x}^*)$. Then from Theorem 2.5.4, $\nabla f(\mathbf{x}^*) \cdot (\mathbf{x} - \alpha \nabla f(\mathbf{x}^*) - \mathbf{x}^*) \geq 0$, or $\nabla f(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \geq \alpha (\nabla f(\mathbf{x}^*))^2 > 0$ because of (b).

Let $\mathbf{x}$ be any vector such that $g_j(\mathbf{x}) \leq b_j$ for $j = 1, \dots, m$. Let $J = \{j : g_j(\mathbf{x}^*) = b_j\}$. If $j \in J$, then $\lambda_j g_j(\mathbf{x}) \leq \lambda_j g_j(\mathbf{x}^*)$, or $-\lambda_j g_j(\mathbf{x}) \geq -\lambda_j g_j(\mathbf{x}^*)$. This last inequality is also valid if $j \notin J$, because then $\lambda_j = 0$. Since each $-\lambda_j g_j(\mathbf{x})$ is quasiconcave, Theorem 2.5.4 implies that $\lambda_j \nabla g_j(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*) \leq 0$, and so $0 \geq \sum_{j=1}^m \lambda_j \nabla g_j = \nabla f(\mathbf{x}^*) \cdot (\mathbf{x} - \mathbf{x}^*)$. Because of (*), this implies that $f(\mathbf{x}) \leq f(\mathbf{x}^*)$, so $\mathbf{x}^*$ is optimal. ■

NOTE 3 Condition (b) cannot be dropped. Consider the problem $\max f(x) = x^3$ subject to $g(x) = -x \leq 0$. With the Lagrangian $\mathcal{L} = x^3 + \lambda x$, condition (a) in Theorem 3.6.3 reduces to $3(x^*)^2 + \lambda = 0$ and $\lambda \geq 0$ with $\lambda = 0$ if $x^* > 0$. Obviously, these conditions are satisfied by $x^* = 0$, $\lambda = 0$, which is definitely not a solution to the given problem. ($\max x^3$ subject to $x \geq 0$ has no solution.) Here f is quasiconcave and λg is quasiconvex. But condition (b) in Theorem 3.6.3 is not satisfied because $\nabla f(0) = f'(0) = 0$.

EXAMPLE 3 Consider the following problem in consumer theory (see Example 3.3.4),

$$\max U(\mathbf{x}) \quad \text{subject to} \quad \mathbf{p} \cdot \mathbf{x} \leq m, \quad \mathbf{x} \geq 0, \quad \mathbf{p} \gg 0$$

assuming that the utility function U is C^1 and quasiconcave. Suppose $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ is admissible and satisfies (a) and (b) in Theorem 3.6.3:

$$U_i'(\mathbf{x}^*) \leq \lambda p_i, \quad \text{with } U_i'(\mathbf{x}^*) = \lambda p_i \text{ if } x_i^* > 0, \quad i = 1, \dots, n \quad (i)$$

$$\lambda \geq 0, \quad \text{with } \lambda = 0 \text{ if } \mathbf{p} \cdot \mathbf{x}^* < m \quad (ii)$$

Suppose too that $\nabla U(\mathbf{x}^*) \neq 0$, i.e. not all the partial derivatives $U_1'(\mathbf{x}^*), \dots, U_n'(\mathbf{x}^*)$ are zero. Then $\mathbf{x}^*$ solves the problem. The usual assumption in economics is that $U_i'(\mathbf{x}^*) \geq 0$ for all i . Then at least one $U_i'(\mathbf{x}^*) > 0$. Hence, (i) implies that $\lambda > 0$, so $\mathbf{p} \cdot \mathbf{x}^* = m$, i.e. all income is spent. ■

In Section 3.4 we formulated and proved sufficient conditions for local optimality in optimization problems with equality constraints. Here is the corresponding result for nonlinear programming. For a proof, see the book's website.

THEOREM 3.6.4 (SUFFICIENT CONDITION FOR LOCAL MAXIMUM)

Assume that an admissible vector $\mathbf{x}^*$ and multipliers $\lambda_1, \dots, \lambda_m$ satisfy the necessary Kuhn–Tucker conditions (3.5.2)–(3.5.3) for problem (1). Let $J = \{j : g_j(\mathbf{x}^*) = b_j\}$ denote the set of active constraints, and assume that $\lambda_j > 0$ for all j in J . Consider the Lagrange problem

$$\max f(\mathbf{x}) \quad \text{subject to} \quad g_j(\mathbf{x}) = b_j, \quad j \in J \quad (2)$$

Evidently, $\mathbf{x}^*$ satisfies the necessary first-order conditions for this problem, for the given multipliers λ_j . If $\mathbf{x}^*$ also satisfies the sufficient second-order conditions of Theorem 3.4.1 for problem (2) with these same λ_j , then $\mathbf{x}^*$ is a strict local maximum point for problem (1).

PROBLEMS FOR SECTION 3.6

1. Solve the problem $\max 1 - (x-1)^2 - e^{x^2}$ subject to $x^2 + y^2 \leq 1$.
2. Solve the problem $\max xy + x + y$ subject to $x^2 + y^2 \leq 2$, $x + y \leq 1$.

3.7 Comparative Statics

For the standard nonlinear programming problem (3.5.1) we defined the value function as

$$f^*(\mathbf{b}) = \max \{ f(\mathbf{x}) : g_j(\mathbf{x}) \leq b_j, \quad j = 1, \dots, m \} \quad (1)$$

According to (3.5.6), provided $\partial f^*(\mathbf{b})/\partial b_j$ exists,

$$\frac{\partial f^*(\mathbf{b})}{\partial b_j} = \lambda_j(\mathbf{b}), \quad j = 1, \dots, m \quad (2)$$

The value function f^* is not necessarily C^1 . (See Problem 2 and Example 3.8.2(d).) In Section 3.10 sufficient conditions for (2) to hold (and for $f^*(\mathbf{b})$ to be differentiable) are given.

EXAMPLE 1 A firm has L units of labour available and produces three goods which it sells at prices a , b , and c per unit, respectively. Producing x , y , and z units of the goods requires αx^2 , βy^2 , and γz^2 units of labour, respectively. Solve the problem

$$\max f(x, y, z) = ax + by + cz \quad \text{subject to} \quad g(x, y, z) = \alpha x^2 + \beta y^2 + \gamma z^2 \leq L$$

of maximizing the value of output that can be produced using L units of labour, where the coefficients a , b , c , α , β , and γ are all positive constants. Find the value function and verify (2) in this case.

Solution: The Lagrangian is $\mathcal{L}(x, y, z) = ax + by + cz - \lambda(\alpha x^2 + \beta y^2 + \gamma z^2)$. Necessary conditions for (x^*, y^*, z^*) to solve the problem are

$$a - 2\lambda\alpha x^* = 0, \quad b - 2\lambda\beta y^* = 0, \quad c - 2\lambda\gamma z^* = 0$$

$$\lambda \geq 0 \quad \text{with} \quad \lambda = 0 \text{ if } \alpha(x^*)^2 + \beta(y^*)^2 + \gamma(z^*)^2 < L$$

Here λ , x^* , y^* , and z^* must all be positive, and $\lambda = a/2\alpha x^* = b/2\beta y^* = c/2\gamma z^*$. So

$$x^* = a/2\alpha\lambda, \quad y^* = b/2\beta\lambda, \quad z^* = c/2\gamma\lambda. \quad (*)$$

Because $\lambda > 0$, the complementary slackness condition implies that $\alpha(x^*)^2 + \beta(y^*)^2 + \gamma(z^*)^2 = L$. Inserting the expressions in (*) for x^* , y^* , and z^* into the resource constraint yields

$$\frac{a^2}{4\alpha\lambda^2} + \frac{b^2}{4\beta\lambda^2} + \frac{c^2}{4\gamma\lambda^2} = L$$

It follows that

$$\lambda = \frac{1}{2} L^{-1/2} \sqrt{a^2/\alpha + b^2/\beta + c^2/\gamma} \quad (**)$$

The suggestion for a solution of the problem is therefore given by (*), with λ as in (**). The Lagrangian $\mathcal{L}$ is obviously concave, so we have found the solution.

The value function is

$$f^*(L) = ax^* + by^* + cz^* = (a^2/\alpha + b^2/\beta + c^2/\gamma)/2\lambda = \sqrt{L} \sqrt{a^2/\alpha + b^2/\beta + c^2/\gamma}$$

But then $df^*(L)/dL = \frac{1}{2} L^{-1/2} \sqrt{a^2/\alpha + b^2/\beta + c^2/\gamma}$, so (2) is confirmed. $\square$

EXAMPLE 2 In Example 3.6.2 we found the following value function:

$$f^*(m) = \begin{cases} 2\sqrt{m} & \text{if } 0 < m < 1 \\ m+1 & \text{if } m \geq 1 \end{cases}$$

We see that $df^*/dm = 1/\sqrt{m} = \lambda_1$ for $0 < m < 1$ and $df^*/dm = 1 = \lambda_1$ for $m > 1$, so (2) is confirmed. The value function is graphed in Fig. 1. It is differentiable for all $m > 0$ ($f'(1) = \lim_{m \rightarrow 1^-} f'(m) = \lim_{m \rightarrow 1^+} f'(m) = 1$), and concave. $\square$

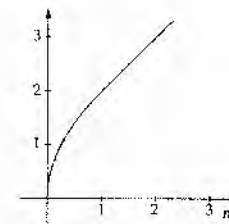


Figure 1

Here is a general result about concavity of the value function:

THEOREM 3.7.1 (CONCAVITY OF THE VALUE FUNCTION)

If $f(\mathbf{x})$ is concave and $g_1(\mathbf{x}), \dots, g_m(\mathbf{x})$ are convex, then the value function $f^*(\mathbf{b})$ defined in (1) is concave.

Proof: Let $\mathbf{b}'$ and $\mathbf{b}''$ be two arbitrary right-hand side vectors, and let $\mathbf{x}^*(\mathbf{b}')$, $\mathbf{x}^*(\mathbf{b}'')$ be corresponding optimal solutions. Let $t \in [0, 1]$. Corresponding to the right-hand side vector $t\mathbf{b}' + (1-t)\mathbf{b}''$ there exists an optimal solution $\mathbf{x}^*(t\mathbf{b}' + (1-t)\mathbf{b}'')$, and

$$f^*(t\mathbf{b}' + (1-t)\mathbf{b}'') = f(\mathbf{x}^*(t\mathbf{b}' + (1-t)\mathbf{b}''))$$

Define $\hat{\mathbf{x}} = t\mathbf{x}^*(\mathbf{b}') + (1-t)\mathbf{x}^*(\mathbf{b}'')$. Then convexity of g_j for $j = 1, \dots, m$ implies that

$$g_j(\hat{\mathbf{x}}) \leq tg_j(\mathbf{x}^*(\mathbf{b}')) + (1-t)g_j(\mathbf{x}^*(\mathbf{b}'')) \leq tb'_j + (1-t)b''_j$$

Thus $\hat{\mathbf{x}}$ is admissible in the problem where the right-hand side vector is $t\mathbf{b}' + (1-t)\mathbf{b}''$, and in that problem $\mathbf{x}^*(t\mathbf{b}' + (1-t)\mathbf{b}'')$ is optimal. It follows that

$$f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*(t\mathbf{b}' + (1-t)\mathbf{b}'')) = f^*(t\mathbf{b}' + (1-t)\mathbf{b}'') \quad (*)$$

But concavity of f implies that

$$f(\hat{\mathbf{x}}) \geq tf(\mathbf{x}^*(\mathbf{b}')) + (1-t)f(\mathbf{x}^*(\mathbf{b}'')) = tf^*(\mathbf{b}') + (1-t)f^*(\mathbf{b}'') \quad (**)$$

From the inequalities (*) and (**) we conclude that $f^*(\mathbf{b})$ is concave. $\blacksquare$

Envelope Result

Consider the following more general nonlinear programming problem with parameters

$$\max_{\mathbf{x}} f(\mathbf{x}, \mathbf{r}) \quad \text{subject to} \quad g_j(\mathbf{x}, \mathbf{r}) \leq 0, \quad j = 1, \dots, m \quad (3)$$

where $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{r} = (r_1, \dots, r_k)$ is a vector of parameters. If we put $\mathbf{g} = (g_1, \dots, g_m)$, the m constraints can then be written as a vector inequality, $\mathbf{g}(\mathbf{x}, \mathbf{r}) \leq \mathbf{0}$. Note that in problem (3) we maximize w.r.t. $\mathbf{x}$, with $\mathbf{r}$ held constant.

The optimal value of the objective function in problem (3) is (again) called the **value function**:

$$f^*(\mathbf{r}) = \max_{\mathbf{x}} \{f(\mathbf{x}, \mathbf{r}) : \mathbf{g}(\mathbf{x}, \mathbf{r}) \leq \mathbf{0}\} \quad (4)$$

assuming that there is a unique maximum.

Let the Lagrangian be defined as $\mathcal{L}(\mathbf{x}, \mathbf{r}) = f(\mathbf{x}, \mathbf{r}) - \sum_{j=1}^m \lambda_j g_j(\mathbf{x}, \mathbf{r})$. We want to find an expression for $\partial f^*(\mathbf{r})/\partial r_i$ at a given point $\bar{\mathbf{r}}$. The corresponding optimal choice for $\mathbf{x}$ is $\mathbf{x}^*(\bar{\mathbf{r}})$, and we let $\lambda_1, \dots, \lambda_m$ be the associated Lagrange multipliers. Under certain conditions (see Theorem 3.10.4), we also have the following relationship:

ENVELOPE RESULT

$$\frac{\partial f^*(\bar{\mathbf{r}})}{\partial r_i} = \left(\frac{\partial \mathcal{L}(\mathbf{x}, \mathbf{r})}{\partial r_i} \right)_{\mathbf{x}=\mathbf{x}^*(\bar{\mathbf{r}})}, \quad i = 1, \dots, k \quad (5)$$

The interpretation of the right-hand side of this formula is analogous to the interpretation of formula (3.3.15).

PROBLEMS FOR SECTION 3.7

1. (a) Solve the nonlinear programming problem (a and b are constants)

$$\text{maximize } 100 - e^{-x} - e^{-y} - e^{-z} \quad \text{subject to } x + y + z \leq a, \quad x \leq b$$

- (b) Let $f^*(a, b)$ be the (optimal) value function. Compute the partial derivatives of f^* with respect to a and b , and relate them to the Lagrange multipliers.
(c) Put $b = 0$, and show that $F^*(a) = f^*(a, 0)$ is concave in a .

2. For $r = 0$ the problem

$$\max_{x \in [-1, 0]} (x - r)^2$$

has two solutions, $x = \pm 1$. For $r \neq 0$, there is only one solution. Show that the value function $f^*(r)$ is not differentiable at $r = 0$.

3. (a) Consider the problem

$$\max(\min) \quad x^2 + y^2 \quad \text{subject to} \quad r^2 \leq 2x^2 + 4y^2 \leq s^2$$

where $0 < r < s$. Solve the maximization problem and verify (5) in this case.

- (b) Reformulate the minimization problem as a maximization problem, solve it, and verify (5) in this case.
(c) Can you give a geometric interpretation of the problem and its solution?

HARDER PROBLEMS

4. Prove that $f^*(\mathbf{r})$ defined in (4) is concave if f is concave and $g_1, \dots, g_m$ are convex in $(\mathbf{x}, \mathbf{r})$. (This generalizes Theorem 3.7.1.)

3.8 Nonnegativity Constraints

Often the variables involved in economic optimization problems are inherently nonnegative. Thus we frequently encounter the standard nonlinear programming problem with **nonnegativity constraints**:

$$\max f(\mathbf{x}) \quad \text{subject to} \quad g_j(\mathbf{x}) \leq b_j, \quad j = 1, \dots, m, \quad x_i \geq 0, \dots, x_n \geq 0 \quad (1)$$

We introduce n new constraints in addition to the m original ones:

$$g_{m+1}(\mathbf{x}) = -x_1 \leq 0, \dots, g_{m+n}(\mathbf{x}) = -x_n \leq 0 \quad (2)$$

This converts (1) into a problem of the form (3.5.1). We introduce Lagrange multipliers $\mu_1, \dots, \mu_n$ to go with the new constraints and form the extended Lagrangian

$$\mathcal{L}_1(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^m \lambda_j (g_j(\mathbf{x}) - b_j) - \sum_{i=1}^n \mu_i (-x_i) \quad (3)$$

According to (3.5.2) and (3.5.3) the necessary conditions for $\mathbf{x}^*$ to solve the problem are

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} + \mu_i = 0, \quad i = 1, \dots, n \quad (i)$$

$$\lambda_j \geq 0, \quad \text{with } \lambda_j = 0 \text{ if } g_j(\mathbf{x}^*) < b_j, \quad j = 1, \dots, m \quad (ii)$$

$$\mu_i \geq 0, \quad \text{with } \mu_i = 0 \text{ if } x_i > 0, \quad i = 1, \dots, n \quad (iii)$$

To reduce this collection of $m+n$ constraints and $m+n$ Lagrange multipliers, the necessary conditions for problem (1) are often formulated slightly differently, as in Theorem 3.8.1 below. In fact, it follows from (i) that $\partial f(\mathbf{x}^*)/\partial x_i - \sum_{j=1}^m \lambda_j \partial g_j(\mathbf{x}^*)/\partial x_i = -\mu_i$. Since $\mu_i \geq 0$ and $-\mu_i = 0$ if $x_i > 0$, we see that (i) and (iii) together are equivalent to the condition

$$\frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} \leq 0 \quad (= 0 \text{ if } x_i^* > 0), \quad i = 1, \dots, n$$

As in (3.5.4), we can say that the two inequalities $\frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} \leq 0$ and $x_i^* \geq 0$

are **complementarily slack**.

THEOREM 3.8.1 (KUHN-TUCKER NECESSARY CONDITIONS)

Suppose that $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ solves problem (1). Suppose further that the gradient vectors $\nabla g_j(\mathbf{x}^*)$, $j = 1, \dots, m+n$, corresponding to those constraints that are active at $\mathbf{x}^*$, are linearly independent. Then there exist unique numbers $\lambda_1, \dots, \lambda_m$ such that with the Lagrangian $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^m \lambda_j (g_j(\mathbf{x}) - b_j)$,

$$(a) \quad \frac{\partial \mathcal{L}(\mathbf{x}^*)}{\partial x_i} = \frac{\partial f(\mathbf{x}^*)}{\partial x_i} - \sum_{j=1}^m \lambda_j \frac{\partial g_j(\mathbf{x}^*)}{\partial x_i} \leq 0 \quad (= 0 \text{ if } x_i^* > 0), \quad i = 1, \dots, n$$

$$(b) \quad \lambda_j \geq 0, \quad \text{with } \lambda_j = 0 \text{ if } g_j(\mathbf{x}^*) < b_j, \quad j = 1, \dots, m$$

Note that in the new formulation of the necessary/sufficient conditions we use the ordinary Lagrangian $\mathcal{L}$, not the extended Lagrangian $\mathcal{L}_1$ used in (3).

THEOREM 3.8.2 (SUFFICIENT CONDITIONS)

Consider problem (1) and suppose that $\mathbf{x}^*$ is admissible and, together with $\lambda_1, \dots, \lambda_m$, satisfies conditions (a) and (b) in Theorem 3.8.1. If the Lagrangian $\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^m \lambda_j (g_j(\mathbf{x}) - b_j)$ is concave, then $\mathbf{x}^*$ is optimal.

The proof of Theorem 3.8.2 is a simple application of Theorem 3.6.1 to the extended Lagrangian $\mathcal{L}_1$.

EXAMPLE 1 Solve the following problem:

$$\text{maximize } f(x, y) = \frac{2}{3}x - \frac{1}{2}x^2 + \frac{1}{12}y \quad \text{s. t.} \quad \begin{cases} x \leq 5 \\ -x + y \leq 1 \end{cases}, \quad x \geq 0, y \geq 0$$

Solution: With the Lagrangian $\mathcal{L} = \frac{2}{3}x - \frac{1}{2}x^2 + \frac{1}{12}y - \lambda_1(x - 5) - \lambda_2(-x + y - 1)$, the Kuhn-Tucker conditions for (x^*, y^*) to solve the problem are:

$$\mathcal{L}'_1 = \frac{2}{3} - x^* - \lambda_1 + \lambda_2 \leq 0 \quad (= 0 \text{ if } x^* > 0) \quad (\text{i})$$

$$\mathcal{L}'_2 = \frac{1}{12} - \lambda_2 \leq 0 \quad (= 0 \text{ if } y^* > 0) \quad (\text{ii})$$

$$\lambda_1 \geq 0, \text{ and } \lambda_1 = 0 \text{ if } x^* < 5 \quad (\text{iii})$$

$$\lambda_2 \geq 0, \text{ and } \lambda_2 = 0 \text{ if } -x^* + y^* < 1 \quad (\text{iv})$$

From (ii) we see that $\lambda_2 > 0$. Then (iv) and $-x^* + y^* \leq 1$ imply $-x^* + y^* = 1$. It follows that $y^* = x^* + 1 > 0$, since $x^* \geq 0$. Then (ii) implies $\lambda_2 = \frac{1}{12}$.

Suppose $\lambda_1 > 0$. Then from (iii) and $x^* \leq 5$, it follows that $x^* = 5$. Inserting $\lambda_2 = \frac{1}{12}$ and $x^* = 5$ into (i) yields a negative value for λ_1 .

So $\lambda_1 = 0$. Then (i) yields $x^* \geq \frac{2}{3} - \lambda_1 + \lambda_2 = \frac{2}{3} + \frac{1}{12} > 0$. From (i) we find that $\frac{2}{3} - x^* + \frac{1}{12} = 0$, so $x^* = \frac{5}{4}$. Then $y^* = 1 + x^* = \frac{9}{4}$. Conclusion: $(x^*, y^*) = (\frac{5}{4}, \frac{9}{4})$, with $\lambda_1 = 0$ and $\lambda_2 = \frac{1}{12}$, satisfies all the conditions. The Lagrangian is easily seen to be concave, so we have found the solution. $\square$

EXAMPLE 2 An aircraft manufacturing firm can operate plants in either of two countries. In country A, its cost as a function of output $x \geq 0$ is $C_A(x) = \ln(1 + 3x/100)$. In country B, its cost as a function of output $y \geq 0$ is $C_B(y) = 2 \ln(1 + y/100)$. The firm allocates production between the two plants in order to minimize the total cost of producing at least q units of output worldwide, where $q > 0$.

- Show that the firm's cost-minimizing choices of x and y must solve a particular constrained optimization problem with non-negativity constraints.
- Use the Lagrange multiplier method to show that there are one, two, or three solution candidates satisfying the Kuhn-Tucker conditions, depending on the value of q .

- Show that the firm uses only the plant in country A for levels of output below some critical level q^* , and only the plant in country B when $q > q^*$.
- Find the firm's minimum cost as a function of q , and show that it is not differentiable at q^* .

Solution: (a) The firm will choose (x, y) to solve the problem

$$\text{min } C(x, y) = \ln(1 + 3x/100) + 2 \ln(1 + y/100) \quad \text{s. t.} \quad x + y \geq q, \quad x \geq 0, y \geq 0$$

(b) Because this is a minimization problem, we write the Lagrangian as

$$\mathcal{L}(x, y) = -[\ln(1 + 3x/100) + 2 \ln(1 + y/100)] + \lambda(x + y - q)$$

The Kuhn-Tucker conditions from Theorem 3.8.1 are

$$-3(100 + 3x)^{-1} + \lambda \leq 0 \quad (= 0 \text{ if } x > 0) \quad (\text{i})$$

$$-2(100 + y)^{-1} + \lambda \leq 0 \quad (= 0 \text{ if } y > 0) \quad (\text{ii})$$

$$\lambda \geq 0, \quad \text{with } \lambda = 0 \text{ if } x + y > q \quad (\text{iii})$$

If $\lambda = 0$, then from (i) and (ii) we get $x = y = 0$, which contradicts $x + y \geq q > 0$. Thus $\lambda > 0$, and $x + y = q$.

Suppose $x > 0$ and $y > 0$. From (i) and (ii) we get $-3(100 + 3x)^{-1} = -2(100 + y)^{-1}$. It follows that $6x - 3y = 100$ and because $x + y = q$, we have the unique solution candidate

$$(x, y) = (q/3 + 100/9, 2q/3 - 100/9) \quad \text{with } \lambda = 9(400 + 3q)^{-1} \quad (q > 50/3) \quad (*)$$

This interior solution candidate is valid provided that $q > 50/3$, which ensures that $y > 0$. The associated cost is $\ln[(400 + 3q)/300] + 2 \ln[(400 + 3q)/450]$.

Suppose $x > 0$ and $y = 0$. The solution candidate is then $(x, y) = (q, 0)$, with $\lambda = 3(100 + 3q)^{-1}$, and an associated cost $C(q, 0) = \ln(1 + 3q/100)$. This is valid provided that (ii) is also satisfied, that is $\lambda = 3(100 + 3q)^{-1} \leq 2(100 + 0)^{-1} = 2/100$. This is true for all $q \geq 50/3$.

Suppose $x = 0$ and $y > 0$. Then the solution candidate is $(x, y) = (0, q)$, with $\lambda = 2(100 + q)^{-1}$, and an associated cost $C(0, q) = 2 \ln(1 + q/100)$. This is valid provided that (i) is also satisfied, that is $\lambda = 2(100 + q)^{-1} \leq 3(100 + 0)^{-1} = 3/100$. This is obviously true for all $q > 0$.

Thus, for $0 < q < 50/3$ there is only one solution candidate, $(0, q)$. For $q = 50/3$, $(q, 0)$ and $(0, q)$ are both candidates. Finally, for $q > 50/3$, $(q, 0)$, $(0, q)$, and (x, y) given in $(*)$ are solution candidates.

(c) Comparing these different solution candidates, note that when the interior solution candidate exists (for $q > 50/3$), it is always the worst. In fact, it is a global maximum of the concave function $C(x, y)$ subject to the constraints $x + y = q$ and $x \geq 0, y \geq 0$.

So the cost-minimizing solution is a choice between $(x, y) = (q, 0)$ with the associated cost $\ln(1 + 3q/100)$, and $(x, y) = (0, q)$ with cost $2 \ln(1 + q/100)$. Note that $\ln(1 + 3q/100) > 2 \ln(1 + q/100) = \ln(1 + q/100)^2 \iff (1 + 3q/100) > (1 + q/100)^2$.

which reduces to $q < 100$. It follows that $(q, 0)$ is cheaper when $q > q^* = 100$, but $(0, q)$ is cheaper when $q < 100$. When $q = 100$, both extreme solutions are equally good. (When $q < 50/3$, the corner solution $(q, 0)$ is actually a global maximum.)

(d) The minimum cost function is $C^*(q) = 2\ln(1 + q/100)$ if $q < 100$, and $C^*(q) = \ln(1 + 3q/100)$ if $q > 100$. Provided that $q \neq 100$, the derivative $C^{**}(q) = \lambda$ (which accords with (1.8.2)). But C^* is not differentiable at $q = 100$. In fact, the left-hand derivative of C^* at $q = 100$ is $1/100$, and the right-hand derivative is $3/400$.

NOTE 1 It is actually easier to solve this problem by putting $y = q - x$ and then minimizing $C(x, q - x)$ w.r.t. the single variable x over the interval $[0, q]$. Nevertheless, we have presented it as an example of the Lagrange multiplier method, because the Lagrange multiplier gives useful information about the firm's marginal cost.

Mixed Constraints

Some optimization problems in economics include both equality constraints and inequality constraints. Thus, they take the form

$$\max f(\mathbf{x}) \quad \text{subject to} \quad \begin{cases} g_j(\mathbf{x}) = b_j, & j = 1, \dots, r \\ h_k(\mathbf{x}) \leq c_k, & k = 1, \dots, s \end{cases} \quad (4)$$

The basic conditions for solving such problems should now be obvious. Associate a Lagrange multiplier λ_j with each of the r equality constraints and a multiplier μ_k with each of the s inequality constraints, then form the Lagrangian

$$\mathcal{L} = f(\mathbf{x}) - \sum_{j=1}^r \lambda_j (g_j(\mathbf{x}) - b_j) - \sum_{k=1}^s \mu_k (h_k(\mathbf{x}) - c_k)$$

Equate each partial derivative of the Lagrangian w.r.t. x_i to 0. The Lagrange multipliers associated with the equality constraints have no sign restrictions. The Lagrange multipliers associated with the inequality constraints must satisfy complementary slackness conditions. The precise result is as follows:

THEOREM 3.8.3 (MIXED CONSTRAINTS)

Suppose $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ solves problem (4), where f , along with $g_1, \dots, g_r$ and $h_1, \dots, h_s$ are C^1 functions, and $r < n$. Suppose further that the CQ in Theorem 3.5.1 holds for $g_1, \dots, g_r, h_1, \dots, h_s$. Then there exist unique numbers $\lambda_1, \dots, \lambda_r$ and $\mu_1, \dots, \mu_s$ such that

$$(a) \quad \nabla f(\mathbf{x}^*) = \sum_{j=1}^r \lambda_j \nabla g_j(\mathbf{x}^*) + \sum_{k=1}^s \mu_k \nabla h_k(\mathbf{x}^*)$$

$$(b) \quad \mu_k \geq 0, \quad \text{and} \quad \mu_k = 0 \quad \text{if} \quad h_k(\mathbf{x}^*) < c_k, \quad k = 1, \dots, s$$

(c) If the Lagrangian is concave in $\mathbf{x}$, an admissible $\mathbf{x}^*$ that satisfies (a) and (b) solves problem (4).

The necessary conditions in this theorem, i.e. parts (a) and (b), follow from Theorem 3.11.1 in the last section of this chapter. It covers, in particular, the necessary conditions in Theorem 3.3.1 (existence of Lagrange multipliers in the Lagrange problem), Theorem 3.5.1 (the standard Kuhn–Tucker conditions), and Theorem 3.8.1 (explicit nonnegativity constraints). Part (c) can be shown just as for Theorem 3.3.1(b) and Theorem 3.6.1.

PROBLEMS FOR SECTION 3.8

1. Solve the problem $\max 1 - x^2 - y^2$ subject to $x \geq 0, y \geq 0$, by (a) a direct argument and (b) using the Kuhn–Tucker conditions.

SM 2. Solve the following nonlinear programming problems:

(a) $\max xy$ subject to $x + 2y \leq 2, x \geq 0, y \geq 0$

(b) $\max x^\alpha y^\beta$ subject to $x + 2y \leq 2, x > 0, y > 0$, where $\alpha > 0, \beta > 0$, and $\alpha + \beta \leq 1$.

SM 3. (a) Solve the following problem for all values of the constant c :

$$\max f(x, y) = cx + y \quad \text{subject to} \quad g(x, y) = x^2 + 3y^2 \leq 2, \quad x \geq 0, \quad y \geq 0$$

(b) Let $f^*(c)$ denote the value function. Verify that it is continuous. Check if (3.7.5) holds.

4. (a) Write down the necessary Kuhn–Tucker conditions for the problem

$$\max \ln(1+x) + y \quad \text{subject to} \quad px + y \leq m, \quad x \geq 0, \quad y \geq 0$$

(b) Find the solution whenever $p \in (0, 1]$ and $m > 1$.

SM 5. A model for studying the export of gas from Russia to the rest of Europe involves the following optimization problem:

$$\max \left\{ x + y - \frac{1}{2}(x+y)^2 - \frac{1}{4}x - \frac{1}{2}y \right\} \quad \text{subject to} \quad x \leq 5, \quad y \leq 3, \quad -x + 2y \leq 2, \quad x \geq 0, \quad y \geq 0$$

Sketch the admissible set S in the xy -plane, and show that the maximum cannot occur at an interior point of S . Solve the problem.

HARDER PROBLEMS

SM 6. With reference to problem (1), define $\widehat{\mathcal{L}}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \sum_{j=1}^r \lambda_j (g_j(\mathbf{x}) - b_j)$. We say that $\widehat{\mathcal{L}}$ has a saddle point at $(\mathbf{x}^*, \lambda^*)$, with $\mathbf{x}^* \geq 0, \lambda^* \geq 0$, if

$$\widehat{\mathcal{L}}(\mathbf{x}, \lambda^*) \leq \widehat{\mathcal{L}}(\mathbf{x}^*, \lambda^*) \leq \widehat{\mathcal{L}}(\mathbf{x}^*, \lambda) \quad \text{for all } \mathbf{x} \geq 0 \text{ and all } \lambda \geq 0 \quad (*)$$

(a) Show that if $\widehat{\mathcal{L}}$ has a saddle point at $(\mathbf{x}^*, \lambda^*)$, then $\mathbf{x}^*$ solves problem (1). (Hint: Use the second inequality in $(*)$ to show that $g_j(\mathbf{x}^*) \leq b_j$ for $j = 1, \dots, r$. Show next that $\sum_{j=1}^r \lambda_j^* (g_j(\mathbf{x}^*) - b_j) = 0$. Then use the first inequality in $(*)$ to finish the proof.)

(b) Suppose that there exist $\mathbf{x}^* \geq 0$ and $\lambda^* \geq 0$ satisfying both $g_j(\mathbf{x}^*) \leq b_j$ and $g_j(\mathbf{x}^*) = b_j$ whenever $\lambda_j^* > 0$ for $j = 1, \dots, r$, as well as $\widehat{\mathcal{L}}(\mathbf{x}, \lambda^*) \leq \widehat{\mathcal{L}}(\mathbf{x}^*, \lambda^*)$ for all $\mathbf{x} \geq 0$. Show that $\widehat{\mathcal{L}}(\mathbf{x}, \lambda)$ has a saddle point at $(\mathbf{x}^*, \lambda^*)$ in this case.

3.9 Concave Programming

The nonlinear programming problem (3.5.1) is said to be a **concave programming problem** (or just a **concave program**) in the case when f is concave and each g_j is a convex function. In this case, the set of admissible vectors satisfying the m constraints is convex. From now on, we write the concave program in the vector form

$$\max f(\mathbf{x}) \quad \text{subject to} \quad \mathbf{g}(\mathbf{x}) \leq \mathbf{b} \quad (1)$$

where $\mathbf{g} = (g_1, \dots, g_m)$ and $\mathbf{b} = (b_1, \dots, b_m)$. When each component function g_j is convex, we also say that the vector function $\mathbf{g}$ is convex. The Lagrangian in vector notation is

$$\mathcal{L}(\mathbf{x}) = f(\mathbf{x}) - \lambda \cdot (\mathbf{g}(\mathbf{x}) - \mathbf{b}) \quad (2)$$

In the following results no differentiability requirements are imposed at all. Instead, however, we make use of the following constraint qualification:

THE SLATER CONDITION

There exists a vector $\mathbf{z}$ in $\mathbb{R}^n$ such that $\mathbf{g}(\mathbf{z}) \ll \mathbf{b}$, i.e. $g_j(\mathbf{z}) < b_j$ for all j . (3)

So at least one vector in the admissible set simultaneously satisfies all the constraints with strict inequality.

THEOREM 3.9.1 (NECESSARY CONDITIONS FOR CONCAVE PROGRAMMING)

Suppose that (1) is a concave program satisfying the Slater condition. Let the optimal value function f^* be defined for all $\mathbf{c}$ such that $\{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq \mathbf{c}\} \neq \emptyset$. Then f^* has a supergradient at $\mathbf{b}$. Furthermore, if λ is any supergradient of f^* at $\mathbf{b}$, then $\lambda \geq 0$, any solution $\mathbf{x}^*$ of problem (1) is an unconstrained maximum point of the Lagrangian $\mathcal{L}(\mathbf{x})$, and $\lambda \cdot (\mathbf{g}(\mathbf{x}^*) - \mathbf{b}) = 0$ (complementary slackness).

Proof: We consider only the (usual) special case where, for all $\mathbf{c}$ in $\mathbb{R}^m$, the set of admissible points $\mathbf{x}$ that satisfy $\mathbf{g}(\mathbf{x}) \leq \mathbf{c}$ is bounded, and so compact because of the assumption that the functions g_j are C^1 . By assumption there exists a point $\mathbf{z}$ such that $\mathbf{g}(\mathbf{z}) \ll \mathbf{b}$. The function $f^*(\mathbf{c})$ is defined as a maximum value whenever there exists at least one $\mathbf{x}$ satisfying $\mathbf{g}(\mathbf{x}) \leq \mathbf{c}$, which is certainly true when $\mathbf{c} \geq \mathbf{g}(\mathbf{z})$. (For a full proof allowing for the possibility that $f^*(\mathbf{c})$ may only be defined as a supremum for some values of $\mathbf{c}$, one first has to prove that f^* is concave in this case as well.) According to Theorem 3.7.1, f^* must be concave. Moreover, $f^*(\mathbf{g}(\mathbf{x}^*)) = f^*(\mathbf{b})$, by definition. Because of the Slater condition, $\mathbf{b}$ is an interior point in the domain of f^* . By Theorem 2.4.5, the concave function $f^*(\mathbf{c})$ has a supergradient λ at $\mathbf{c} = \mathbf{b}$ for which

$$f^*(\mathbf{c}) - f^*(\mathbf{b}) \leq \lambda \cdot (\mathbf{c} - \mathbf{b})$$

For any such supergradient, if $\mathbf{c} \geq \mathbf{b}$, then $f^*(\mathbf{c}) \geq f^*(\mathbf{b})$, so $\lambda \cdot (\mathbf{c} - \mathbf{b}) \geq 0$. Hence $\lambda \cdot \mathbf{d} \geq 0$ for all $\mathbf{d} \geq 0$, which implies that $\lambda \geq 0$.

Now, if $\mathbf{x}^*$ solves (1), then $f(\mathbf{x}^*) = f^*(\mathbf{b})$ and for every $\mathbf{x}$ in $\mathbb{R}^n$,

$$f(\mathbf{x}) \leq f^*(\mathbf{g}(\mathbf{x})) \leq f^*(\mathbf{b}) + \lambda \cdot (\mathbf{g}(\mathbf{x}) - \mathbf{b}), \quad \text{so} \quad f(\mathbf{x}) - \lambda \cdot \mathbf{g}(\mathbf{x}) \leq f^*(\mathbf{b}) - \lambda \cdot \mathbf{b} \leq f(\mathbf{x}^*) - \lambda \cdot \mathbf{g}(\mathbf{x}^*) \quad (*)$$

Thus $\mathbf{x}^*$ maximizes $f(\mathbf{x}) - \lambda \cdot \mathbf{g}(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$. Also, for $\mathbf{x} = \mathbf{x}^*$, when $f(\mathbf{x}) = f^*(\mathbf{b})$, the last pair of inequalities in (*) become equalities, so $\lambda \cdot (\mathbf{g}(\mathbf{x}^*) - \mathbf{b}) = 0$, which shows complementary slackness. ■

NOTE 1 The complementary slackness condition $\lambda \cdot (\mathbf{g}(\mathbf{x}^*) - \mathbf{b}) = 0$ together with $\lambda \geq 0$ is equivalent to the complementary slackness condition (3.5.4). In fact, $0 = \lambda \cdot (\mathbf{g}(\mathbf{x}^*) - \mathbf{b}) = \sum_{i=1}^m \lambda_j (g_j(\mathbf{x}^*) - b_j)$. Each term in this sum is nonnegative, and so the terms add up to 0 only if each term is 0.

An Economic Interpretation

A general economic interpretation of (1) can be given in line with the interpretation of the Lagrange problem in Section 3.3. The only difference is that in the present case the inequalities $g_j(\mathbf{x}) \leq b_j$ reflect the fact that we no longer insist that all the resources are fully utilized. Thus Problem (1) can then be formulated as follows:

Find activity levels at which to operate the production processes in order to obtain the largest possible output of the produced commodity, taking into account the impossibility of using more of any resource than its total supply.

For each resource j , specify a shadow price of λ_j per unit. To produce $f(\mathbf{x})$ units of the commodity requires $g_j(\mathbf{x})$ units of resource j at a shadow cost of $\lambda_j g_j(\mathbf{x})$. If we let the shadow price per unit of the produced commodity be 1, then the function $\pi(\mathbf{x})$ defined by

$$\pi(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^m \lambda_j g_j(\mathbf{x}) \quad (4)$$

indicates the *shadow profit* from running the processes at the vector $\mathbf{x}$ of activity levels. Suppose that we find an activity vector $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ and nonnegative shadow prices $\lambda_1, \dots, \lambda_m$ such that:

- (A) $\mathbf{x} = \mathbf{x}^*$ maximizes shadow profit among all activity levels $\mathbf{x}$.
- (B) $\mathbf{x}^*$ satisfies each resource constraint $g_j(\mathbf{x}^*) \leq b_j$, $j = 1, \dots, m$.
- (C) If the j th resource is not fully used because $g_j(\mathbf{x}^*) < b_j$, then the shadow price λ_j of that resource is 0.

Under these conditions $\mathbf{x}^*$ solves problem (1). For the proof, see Note 3.6.1. It follows from (C) that

$$\sum_{i=1}^n \lambda_j g_j(\mathbf{x}^*) = \sum_{i=1}^n \lambda_j b_j \quad (5)$$

Thus, at the given shadow prices for the resources, the total value of the resources used at the optimum $\mathbf{x}^*$ is equal to the total shadow value of the initial stocks.

The conditions (A)–(C) are not, in general, necessary for optimality, i.e. the appropriate prices do not necessarily exist. However, if the function π in (4) is concave, and if we impose the Slater condition on the admissible set, then Theorem 3.9.1 shows that $\mathbf{x}^*$ maximizes profit.

PROBLEMS FOR SECTION 3.9

1. Suppose that $\mathbf{x}^* = (x_1^*, \dots, x_n^*) \geq \mathbf{0}$ and $\lambda = (\lambda_1, \dots, \lambda_m) \geq \mathbf{0}$ satisfy the sufficient conditions (A)–(C), so that $\mathbf{x}^*$ solves problem (I). Suppose that $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_n) \geq \mathbf{0}$ also solves the problem. Prove that, for the same $\lambda_1, \dots, \lambda_m$ as those associated with $\mathbf{x}^*$, the vector $\hat{\mathbf{x}}$ will also satisfy (A)–(C), but with $\mathbf{x}^*$ replaced by $\hat{\mathbf{x}}$.

3.10 Precise Comparative Statics Results

So far the arguments related to the value function and the envelope theorems have assumed a priori that the functions are differentiable. It is about time to relax this assumption and give sufficient conditions for differentiability.

We begin with the unconstrained case. The first result is this:

THEOREM 3.10.1 (ENVELOPE THEOREM A)

Suppose $f(\mathbf{x}, \mathbf{r})$ is a C^2 function for all $\mathbf{x}$ in an open convex set $S \subseteq \mathbb{R}^n$ and for each $\mathbf{r}$ in an open ball $B(\bar{\mathbf{r}}; \delta) \subseteq \mathbb{R}^k$. Assume that for each fixed $\mathbf{r}$ in $B(\bar{\mathbf{r}}; \delta)$, the function $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{r})$ is concave, and that when $\mathbf{r} = \bar{\mathbf{r}}$ it satisfies the sufficient second-order conditions for strict concavity in Theorem 2.3.2(b). Moreover, assume that $\mathbf{x}^*$ is a maximum point for $\mathbf{x} \mapsto f(\mathbf{x}, \bar{\mathbf{r}})$ in S . Then $f^*(\mathbf{r}) = \max_{\mathbf{x} \in S} f(\mathbf{x}, \mathbf{r})$ is defined for all $\mathbf{r}$ in an open ball around $\bar{\mathbf{r}}$. Moreover, f^* is C^1 at $\bar{\mathbf{r}}$, and

$$\frac{\partial f^*(\bar{\mathbf{r}})}{\partial r_j} = \left[\frac{\partial f(\mathbf{x}, \mathbf{r})}{\partial r_j} \right]_{(\mathbf{x}=\mathbf{x}^*(\bar{\mathbf{r}}), \mathbf{r}=\bar{\mathbf{r}})} \quad j = 1, \dots, k \quad (1)$$

Proof: The first-order conditions for maximizing $f(\mathbf{x}, \mathbf{r})$ w.r.t. $\mathbf{x}$ can be written in the form $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{r}) = \mathbf{0}$, where $\nabla_{\mathbf{x}} f$ denotes the partial gradient vector w.r.t. $\mathbf{x}$, holding $\mathbf{r}$ fixed. The Jacobian matrix J of the mapping $\mathbf{x} \mapsto \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{r})$ evaluated at $(\mathbf{x}^*, \bar{\mathbf{r}})$ is the Hessian matrix $\mathbf{f}_{\mathbf{xx}}''(\mathbf{x}^*, \bar{\mathbf{r}})$. By the sufficient conditions for strict concavity, this Hessian matrix is negative definite, hence nonsingular. Because f is a C^2 function, the Hessian matrix $\mathbf{f}_{\mathbf{xx}}''(\mathbf{x}, \mathbf{r})$ must still be negative definite in some open ball of $\mathbb{R}^{n+k}$ centred at $(\mathbf{x}^*, \bar{\mathbf{r}})$. By the implicit

function theorem (Theorem 2.7.2), it follows that the equation system $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{r}) = \mathbf{0}$ in the unknown vector $\mathbf{x}$ has a unique solution $\mathbf{x}(\mathbf{r})$ which is a C^1 function of $\mathbf{r}$ in some ball $B(\bar{\mathbf{r}}; \varepsilon)$, and moreover $\mathbf{x}(\bar{\mathbf{r}}) = \mathbf{x}^*$. Provided that $\mathbf{r}$ lies in $B(\bar{\mathbf{r}}; \varepsilon) \cap B(\bar{\mathbf{r}}; \delta)$, the function $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{r})$ is concave, so $\mathbf{x}(\mathbf{r})$ is a maximum point of $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{r})$ for $\mathbf{x}$ in S . Because $\mathbf{x}(\mathbf{r})$ is differentiable at $\mathbf{r} = \bar{\mathbf{r}}$, so is $f^*(\mathbf{r}) = f(\mathbf{x}(\mathbf{r}), \mathbf{r})$. In particular, Theorem 3.1.4 applies. ■

A crucial assumption in the previous theorem is that $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{r})$ is concave. The next theorem replaces this with the weaker assumption that $\mathbf{r} \mapsto f^*(\mathbf{r})$ is concave, but drops from Theorem 3.10.1 the second-order conditions for strict concavity.

THEOREM 3.10.2 (ENVELOPE THEOREM B)

Suppose that $f^*(\mathbf{r}) = \sup_{\mathbf{x} \in S} f(\mathbf{x}, \mathbf{r})$ is finite and concave in $\mathbf{r} \in A$, where A is an open convex set in $\mathbb{R}^k$, and $S \subseteq \mathbb{R}^n$. Assume that the point $(\mathbf{x}^*, \bar{\mathbf{r}}) \in S \times A$ satisfies $f(\mathbf{x}^*, \bar{\mathbf{r}}) = f^*(\bar{\mathbf{r}})$ and that the gradient vector $\nabla_{\mathbf{r}} f$ exists at $(\mathbf{x}^*, \bar{\mathbf{r}})$. Then $f^*(\mathbf{r})$ is differentiable at $\bar{\mathbf{r}}$ and $\nabla f^*(\bar{\mathbf{r}}) = \nabla_{\mathbf{r}} f(\mathbf{x}^*, \bar{\mathbf{r}})$, i.e. (1) holds.

Proof: Because A is open, Theorem 2.4.5 implies that f^* has a supergradient at $\bar{\mathbf{r}}$, which we will denote by $\mathbf{a}$. From the definition of f^* , it follows that

$$f(\mathbf{x}^*, \mathbf{r}) - f(\mathbf{x}^*, \bar{\mathbf{r}}) \leq f^*(\mathbf{r}) - f^*(\bar{\mathbf{r}}) \leq \mathbf{a} \cdot (\mathbf{r} - \bar{\mathbf{r}}) \quad \text{for all } \mathbf{r} \text{ in } A \quad (*)$$

This implies that $\mathbf{a}$ is a supergradient of $\mathbf{r} \mapsto f(\mathbf{x}^*, \mathbf{r})$ at $\bar{\mathbf{r}}$. By Note 2.4.3, we conclude that $\mathbf{a} = \nabla_{\mathbf{r}} f(\mathbf{x}^*, \bar{\mathbf{r}})$. But (*) implies that

$$\frac{f(\mathbf{x}^*, \mathbf{r}) - f(\mathbf{x}^*, \bar{\mathbf{r}}) - \mathbf{a} \cdot (\mathbf{r} - \bar{\mathbf{r}})}{\|\mathbf{r} - \bar{\mathbf{r}}\|} \leq \frac{f^*(\mathbf{r}) - f^*(\bar{\mathbf{r}}) - \mathbf{a} \cdot (\mathbf{r} - \bar{\mathbf{r}})}{\|\mathbf{r} - \bar{\mathbf{r}}\|} \leq 0 \quad \text{for all } \mathbf{r} \neq \bar{\mathbf{r}}$$

The first expression here $\rightarrow 0$ as $\mathbf{r} \rightarrow \bar{\mathbf{r}}$. So $[f^*(\mathbf{r}) - f^*(\bar{\mathbf{r}}) - \mathbf{a} \cdot (\mathbf{r} - \bar{\mathbf{r}})]/\|\mathbf{r} - \bar{\mathbf{r}}\| \rightarrow 0$ also, which confirms that f^* is differentiable at $\bar{\mathbf{r}}$, with $\nabla f^*(\bar{\mathbf{r}}) = \mathbf{a} = \nabla_{\mathbf{r}} f(\mathbf{x}^*, \bar{\mathbf{r}})$. ■

Envelope Theorems for Mixed Constraints

In Section 3.8 we introduced a problem with mixed constraints. When formulating a precise envelope result for such problems it is convenient to represent the problem in this way:

$$\max_{\mathbf{x}} f(\mathbf{x}, \mathbf{r}) \quad \text{s.t.} \quad \begin{cases} g_j(\mathbf{x}, \mathbf{r}) \leq b_j, & j = 1, \dots, m' \\ g_j(\mathbf{x}, \mathbf{r}) = b_j, & j = m' + 1, \dots, m \end{cases} \quad (2)$$

where $\mathbf{r} = (r_1, \dots, r_k)$ is a vector of parameters. Note that in problem (2) we maximize w.r.t. $\mathbf{x}$, with $\mathbf{r}$ held constant.

The maximum value of $f(\mathbf{x}, \mathbf{r})$ will depend on $\mathbf{r}$, and we denote it by $f^*(\mathbf{r})$. If we let $\Gamma(\mathbf{r})$ denote the set of admissible points in (2), i.e.

$$\Gamma(\mathbf{r}) = \{\mathbf{x} : g_j(\mathbf{x}, \mathbf{r}) \leq b_j, \quad j = 1, \dots, m', \quad g_j(\mathbf{x}, \mathbf{r}) = b_j, \quad j = m' + 1, \dots, m\}$$

we define the (maximum) value function by

$$f^*(\mathbf{r}) = \sup_{\mathbf{x} \in \Gamma(\mathbf{r})} f(\mathbf{x}, \mathbf{r}) \quad (3)$$

We use sup (supremum) to cover the case where the maximum value does not exist. The domain of f^* is the set of all $\mathbf{r}$ for which $\Gamma(\mathbf{r})$ is nonempty.

The values of $x_1, \dots, x_n$ that solve problem (2) will be functions of $\mathbf{r}$. (We assume for the moment a unique solution.) If we denote them by $x_1^*(\mathbf{r}), \dots, x_n^*(\mathbf{r})$, then

$$f^*(\mathbf{r}) = f(x_1^*(\mathbf{r}), \dots, x_n^*(\mathbf{r})) \quad (4)$$

Suppose that $\lambda_i = \lambda_i(\bar{\mathbf{r}})$, $i = 1, \dots, m$, are the Lagrange multipliers in the first-order conditions for the problem (2) when $\mathbf{r}$ equals a particular vector $\bar{\mathbf{r}}$, and let $\mathcal{L}(\mathbf{x}, \mathbf{r}) = f(\mathbf{x}, \mathbf{r}) - \sum_{j=1}^m \lambda_j(g_j(\mathbf{x}, \mathbf{r}) - b_j)$ be the Lagrangian. Under certain conditions (see Theorem 3.10.4 below), we have the following generalization of the envelope results in Sections 3.1 and 3.3:

ENVELOPE RESULT

$$\frac{\partial f^*(\bar{\mathbf{r}})}{\partial r_i} = \left[\frac{\partial \mathcal{L}(\mathbf{x}, \mathbf{r})}{\partial r_i} \right]_{\mathbf{x}=\mathbf{x}^*(\bar{\mathbf{r}}), \mathbf{r}=\bar{\mathbf{r}}} \quad i = 1, \dots, k \quad (5)$$

We state more precise results for two different cases: first when the vector $\mathbf{b} = (b_1, \dots, b_m)$ varies with $\mathbf{r}$ fixed, and second when $\mathbf{r}$ varies with $\mathbf{b}$ fixed. Because each constraint $g_j(\mathbf{x}, \mathbf{r}) \leq b_j$ (or $g_j(\mathbf{x}, \mathbf{r}) = b_j$) is equivalent to the constraint $\bar{g}_j(\mathbf{x}, \mathbf{r}, b_j) \leq 0$ (or $\bar{g}_j(\mathbf{x}, \mathbf{r}, b_j) = 0$) where $\bar{g}_j(\mathbf{x}, \mathbf{r}, b_j)$ is defined as $g_j(\mathbf{x}, \mathbf{r}) - b_j$, the second case actually includes the first.

THEOREM 3.10.3 (INTERPRETATION OF LAGRANGE MULTIPLIERS)

Consider the problem

$$\max f(\mathbf{x}) \text{ s.t. } g_j(\mathbf{x}) \leq b_j, j = 1, \dots, m', \quad g_j(\mathbf{x}) = b_j, j = m' + 1, \dots, m$$

and let $f^*(\mathbf{b})$ be maximum value function for the problem. Suppose that:

- For $\mathbf{b} = \bar{\mathbf{b}}$ the problem has a unique solution $\mathbf{x}^* = \mathbf{x}(\bar{\mathbf{b}})$.
- There exist an open ball $B(\bar{\mathbf{b}}; \alpha)$ and a constant K such that for every $\mathbf{b}$ in $B(\bar{\mathbf{b}}; \alpha)$, the problem has an optimal solution $\hat{\mathbf{x}}$ in $B(\mathbf{x}^*; K)$.
- The functions f and $g_1, \dots, g_m$ are C^1 in a ball around $\mathbf{x}(\bar{\mathbf{b}})$.
- The gradient vectors $\nabla g_j(\mathbf{x}^*)$ corresponding to those constraints that are active when $\mathbf{b} = \bar{\mathbf{b}}$, are linearly independent.

Then $f^*(\mathbf{b})$ is differentiable at $\bar{\mathbf{b}}$ and $\partial f^*(\bar{\mathbf{b}})/\partial b_i = \lambda_i(\bar{\mathbf{b}})$, $i = 1, \dots, m$.

THEOREM 3.10.4 (A GENERAL ENVELOPE THEOREM)

Consider problem (2) and suppose:

- For $\mathbf{r} = \bar{\mathbf{r}}$ the problem has a unique solution $\mathbf{x}^* = \mathbf{x}(\bar{\mathbf{r}})$.
- There exist an open ball $B(\bar{\mathbf{r}}; \alpha)$ and a constant K such that for every $\mathbf{r}$ in $B(\bar{\mathbf{r}}; \alpha)$, problem (2) has at least one solution $\hat{\mathbf{x}}$ in $B(\mathbf{x}^*; K)$.
- The functions f and $g_1, \dots, g_m$ are C^1 in some open ball around $(\mathbf{x}(\bar{\mathbf{r}}), \bar{\mathbf{r}})$.
- The gradient vectors $\nabla_{\mathbf{x}} g_j(\mathbf{x}^*, \bar{\mathbf{r}})$ corresponding to those constraints that are active when $\mathbf{r} = \bar{\mathbf{r}}$, are linearly independent.

Then $f^*(\mathbf{r})$ is differentiable at $\bar{\mathbf{r}}$ and (5) is valid.

NOTE 1 The proof of Theorem 3.10.4 (which implies 3.10.3, as explained above) can be found on the book's website.

NOTE 2 If the appropriate Lagrangian is concave, condition (b) can be deleted in Theorems 3.10.3 and 3.10.4, respectively. (Formally, "max" must be replaced by sup, and f^* must be given the value $-\infty$ if the supremum is taken over an empty set.)

NOTE 3 Conditions (c) and (d) alone imply that $\bar{\mathbf{r}}$ is an interior point of the domain of f^* .

NOTE 4 The conditions stated in Theorem 3.10.4 guarantee that the function $f^*(\mathbf{r})$ is defined for $\mathbf{r}$ in a neighbourhood of $\bar{\mathbf{r}}$. Moreover, $f^*(\mathbf{r})$ is C^1 near $\bar{\mathbf{r}}$ if the solution $\hat{\mathbf{x}}$ is unique for all $\mathbf{r}$ in $B(\bar{\mathbf{r}}; \alpha)$.

PROBLEMS FOR SECTION 3.10

- 50 1. (a) Solve the problem $\max x^2 + y^2 + z^2$ subject to $\begin{cases} 2x^2 + y^2 + z^2 \leq a^2 \\ x + y + z = 0 \end{cases}$
- (b) Verify (5) in this case.

3.11 Existence of Lagrange Multipliers

In this section we prove a theorem that implies the necessary conditions in Theorem 3.8.3 as a special case. Consider the problem

$$\max f(\mathbf{x}) \text{ subject to } \begin{cases} g_j(\mathbf{x}) = 0, j = 1, \dots, r \\ h_k(\mathbf{x}) \leq 0, k = 1, \dots, s \end{cases} \quad (1)$$

To simplify notation we assume that the right-hand side variables b_j and c_k in the constraints have been absorbed into the functions g_j and h_k . We also allow r or s or both to be zero, in which case one just ignores the corresponding sums in the discussion below. (If both r and s are 0, we get the problem of finding an unconstrained local maximum of f .) The proof of Theorem 3.11.1 is adapted from an argument in Smirnov (2002).

THEOREM 3.11.1 (EXISTENCE OF LAGRANGE MULTIPLIERS)

Suppose that $f, g_1, \dots, g_r, h_1, \dots, h_s$ are all C^1 in some open set A in $\mathbb{R}^n$, and suppose that $\mathbf{x}^*$ is a local maximum point in the problem (1) over A . Then there exist numbers $\alpha, \lambda_1, \dots, \lambda_r, \mu_1, \dots, \mu_s$ that are not all 0, such that

- (a) $\alpha \geq 0$
 (b) $\alpha \nabla f(\mathbf{x}^*) = \sum_{j=1}^r \lambda_j \nabla g_j(\mathbf{x}^*) + \sum_{k=1}^s \mu_k \nabla h_k(\mathbf{x}^*)$
 (c) For each $k = 1, \dots, s$ one has $\mu_k \geq 0$, and $\mu_k = 0$ if $h_k(\mathbf{x}^*) < 0$.

Proof: Let $\varphi(\mathbf{x}; \gamma) = \max\{\gamma - f(\mathbf{x}), 0\}$, where γ is a real parameter, let $h_k^+(\mathbf{x}) = \max\{h_k(\mathbf{x}), 0\}$, $k = 1, \dots, s$, and define

$$\Phi(\mathbf{x}; \gamma) = \varphi(\mathbf{x}; \gamma)^2 + \sum_{j=1}^r (g_j(\mathbf{x}))^2 + \sum_{k=1}^s (h_k^+(\mathbf{x}))^2$$

It is clear that $\Phi(\mathbf{x}; \gamma) = 0$ if and only if $\mathbf{x}$ is admissible in problem (1) and $\gamma \leq f(\mathbf{x})$. Therefore $\Phi(\mathbf{x}; \gamma) > 0$ for all $\gamma > f(\mathbf{x}^*)$ and all $\mathbf{x}$ in A . Furthermore, $\mathbf{x} \mapsto \Phi(\mathbf{x}; \gamma)$ is C^1 and its gradient vector w.r.t. $\mathbf{x}$ satisfies

$$\nabla \Phi(\mathbf{x}; \gamma) = -2\varphi(\mathbf{x}; \gamma) \nabla f(\mathbf{x}) + \sum_{j=1}^r 2g_j(\mathbf{x}) \nabla g_j(\mathbf{x}) + \sum_{k=1}^s 2h_k^+(\mathbf{x}) \nabla h_k(\mathbf{x})$$

(See Problem 1 below.)

Finally define

$$F(\mathbf{x}; \gamma) = \sqrt{\Phi(\mathbf{x}; \gamma)} + \|\mathbf{x} - \mathbf{x}^*\|^2$$

When $\Phi(\mathbf{x}; \gamma) \neq 0$, we get

$$\begin{aligned} \nabla F(\mathbf{x}; \gamma) &= \frac{1}{2\sqrt{\Phi(\mathbf{x}; \gamma)}} \nabla \Phi(\mathbf{x}; \gamma) + 2(\mathbf{x} - \mathbf{x}^*) \\ &= -\alpha^\gamma(\mathbf{x}) \nabla f(\mathbf{x}) + \sum_{j=1}^r \lambda_j^\gamma(\mathbf{x}) \nabla g_j(\mathbf{x}) + \sum_{k=1}^s \mu_k^\gamma(\mathbf{x}) \nabla h_k(\mathbf{x}) + 2(\mathbf{x} - \mathbf{x}^*) \end{aligned}$$

where

$$\alpha^\gamma(\mathbf{x}) = \frac{\varphi(\mathbf{x}; \gamma)}{\sqrt{\Phi(\mathbf{x}; \gamma)}}, \quad \lambda_j^\gamma(\mathbf{x}) = \frac{g_j(\mathbf{x})}{\sqrt{\Phi(\mathbf{x}; \gamma)}}, \quad \mu_k^\gamma(\mathbf{x}) = \frac{h_k^+(\mathbf{x})}{\sqrt{\Phi(\mathbf{x}; \gamma)}}$$

Since $\mathbf{x}^*$ is a local maximum point for f over the admissible set, there exists an $r > 0$ such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all admissible $\mathbf{x}$ in the closed ball $K = \overline{B(\mathbf{x}^*; r)}$. For each value of γ let $\mathbf{x}^\gamma$ be a minimum point for $F(\mathbf{x}; \gamma)$ over the closed and bounded set K . Then

$$F(\mathbf{x}^\gamma; \gamma) \leq F(\mathbf{x}^*; \gamma) = \sqrt{\Phi(\mathbf{x}^*; \gamma)} = \varphi(\mathbf{x}^*; \gamma)$$

Now let γ be a number in the interval $(f(\mathbf{x}^*), f(\mathbf{x}^*) + r^2)$. Then

$$\|\mathbf{x}^\gamma - \mathbf{x}^*\|^2 \leq F(\mathbf{x}^\gamma; \gamma) \leq \varphi(\mathbf{x}^*; \gamma) = \gamma - f(\mathbf{x}^*) < r^2 \quad (i)$$

so $\mathbf{x}^\gamma$ lies in the interior of K . Since $\gamma > f(\mathbf{x}^*)$, we also know that $\Phi(\mathbf{x}^\gamma; \gamma) > 0$. Hence F is differentiable at $\mathbf{x}^\gamma$, and $\nabla F(\mathbf{x}^\gamma; \gamma) = 0$. It follows that

$$\alpha^\gamma(\mathbf{x}^\gamma) \nabla f(\mathbf{x}^\gamma) = \sum_{j=1}^r \lambda_j^\gamma(\mathbf{x}^\gamma) \nabla g_j(\mathbf{x}^\gamma) + \sum_{k=1}^s \mu_k^\gamma(\mathbf{x}^\gamma) \nabla h_k(\mathbf{x}^\gamma) + 2(\mathbf{x}^\gamma - \mathbf{x}^*) \quad (ii)$$

Also,

$$(\alpha^\gamma(\mathbf{x}^\gamma))^2 + \sum_{j=1}^r (\lambda_j^\gamma(\mathbf{x}^\gamma))^2 + \sum_{k=1}^s (\mu_k^\gamma(\mathbf{x}^\gamma))^2 = 1$$

so the point $\mathbf{v}^\gamma = (\alpha^\gamma(\mathbf{x}^\gamma), \lambda_1^\gamma(\mathbf{x}^\gamma), \dots, \lambda_r^\gamma(\mathbf{x}^\gamma), \mu_1^\gamma(\mathbf{x}^\gamma), \dots, \mu_s^\gamma(\mathbf{x}^\gamma))$ lies on the unit sphere S in $\mathbb{R}^{1+m+p}$.

Now choose a sequence $\{\gamma_i\}_{i=1}^\infty$ of numbers in $(f(\mathbf{x}^*), f(\mathbf{x}^*) + r^2)$ such that $\gamma_i \rightarrow f(\mathbf{x}^*)$ as $i \rightarrow \infty$. This gives rise to a sequence $\{\mathbf{v}^i\}$ of points on S . By the Bolzano-Weierstrass theorem (Theorem 13.2.5) this sequence has a convergent subsequence. Replacing $\{\gamma_i\}$ with a subsequence, if necessary, we can therefore assume that $\{\mathbf{v}^i\}$ is itself convergent, with a limit $(\alpha, \lambda_1, \dots, \lambda_r, \mu_1, \dots, \mu_s)$, which lies on S .

Since $\gamma_i \rightarrow f(\mathbf{x}^*)$ and $\|\mathbf{x}^i - \mathbf{x}^*\| \leq \gamma_i - f(\mathbf{x}^*)$, it is clear that $\mathbf{x}^i \rightarrow \mathbf{x}^*$ as $i \rightarrow \infty$. Taking limits in (ii), we get equation (b) in the theorem. The definition of Φ implies that the inequalities in (a) and (c) are all satisfied, and if $h_k(\mathbf{x}^*) < 0$, then $\mu_k = 0$. ■

NOTE 1 The numbers α, λ_j , and μ_k in the theorem are not all 0. Hence, the gradient of f and the gradients of the functions corresponding to constraints that are active at $\mathbf{x}^*$ are linearly dependent. Furthermore, if the constraint qualification is satisfied, then the “active” gradients are linearly independent, so the coefficient α in part (b) of the theorem must be different from 0, hence positive. If we divide the equation in (b) by α , then let $\tilde{\lambda}_j = \lambda_j/\alpha$ and $\tilde{\mu}_k = \mu_k/\alpha$, we get

$$\nabla f(\mathbf{x}^*) = \sum_{j=1}^r \tilde{\lambda}_j \nabla g_j(\mathbf{x}^*) + \sum_{k=1}^s \tilde{\mu}_k \nabla h_k(\mathbf{x}^*)$$

just as needed for the necessary conditions (a) and (b) of Theorem 3.8.3.

PROBLEMS FOR SECTION 3.11

- 1.** For a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, the positive part f^+ of f is the function $f^+(\mathbf{x}) = \max\{f(\mathbf{x}), 0\}$. Show that if f is C^1 in an open set A , then $(f^+)^2$ is also C^1 in A , and $\nabla((f^+)^2)(\mathbf{x}) = 2f^+(\mathbf{x}) \nabla f(\mathbf{x})$. (Hint: If $f(\mathbf{x}_0) > 0$, then $f^+(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x}$ in an open ball around $\mathbf{x}_0$, and if $f(\mathbf{x}_0) < 0$, then $f^+(\mathbf{x}) = 0$ for all $\mathbf{x}$ in an open ball around $\mathbf{x}_0$.)

4

TOPICS IN INTEGRATION

I don't know mathematics, therefore I have to think.

—Joan Robinson

This chapter considers some topics in the theory of integration. We presume that the reader has previously studied the elementary theory for functions of one variable, for instance in EMEA. Section 4.1 briefly reviews some of this material, and provides several problems that may help the reader recall material that is supposed to have been learned previously.

Leibniz's rule for differentiating definite integrals with respect to parameters is discussed in Section 4.2. Thereafter Section 4.3 contains a brief treatment of the gamma function. The main topic of this chapter is, however, multiple integration. In particular, the rule for changing variables in multiple integrals is considered in some detail.

4.1 Review of One-Variable Integration

Let $f(x)$ be a continuous function on an interval I . Recall that an **indefinite integral** of $f(x)$ is a function $F(x)$ whose derivative is equal to $f(x)$ for all x in I . In symbols,

$$\int f(x) dx = F(x) + C \quad \text{where} \quad F'(x) = f(x)$$

For instance, if $a \neq -1$, then

$$\int x^a dx = \frac{1}{a+1} x^{a+1} + C \quad \text{because} \quad \frac{d}{dx} \left(\frac{1}{a+1} x^{a+1} \right) = x^a$$

Two other important indefinite integrals are

$$(a) \int \frac{1}{x} dx = \ln |x| + C, \quad (b) \int e^{ax} dx = \frac{1}{a} e^{ax} + C \quad (a \neq 0)$$

Note that (a) has been expressed in a form that makes it valid even when x is negative

Two useful ways to transform an integral involve **integration by parts**,

$$\int f(x)g'(x) dx = f(x)g(x) - \int f'(x)g(x) dx \quad (1)$$

and **integration by substitution**, or **by change of variable**,

$$\int f(x) dx = \int f(g(u))g'(u) du \quad (\text{where } x = g(u)) \quad (2)$$

The **definite integral** of a continuous function $f(x)$ is given by

$$\int_a^b f(x) dx = \left[F(x) \right]_a^b = F(b) - F(a), \quad \text{where } F'(x) = f(x) \text{ for all } x \text{ in } (a, b) \quad (3)$$

Recall that if $f(x) \geq 0$ in the interval $[a, b]$, then $\int_a^b f(x) dx$ is the area under the graph of f over $[a, b]$.

For a definite integral the formula for integration by substitution is

$$\int_a^b f(x) dx = \int_{u_1}^{u_2} f(g(u))g'(u) du \quad (x = g(u), g(u_1) = a, g(u_2) = b) \quad (4)$$

Note also the following implications of (3):

$$\frac{d}{dx} \int_a^x f(t) dt = f(x), \quad \frac{d}{dx} \int_x^b f(t) dt = -f(x) \quad (5)$$

MS FOR SECTION 4.1

Find the integrals in Problems 1–5.

1. (a) $\int (1 - 3x^2) dx$ (b) $\int x^{-4} dx$ (c) $\int (1 - x^2)^2 dx$
2. (a) $\int_0^{10} (10t^2 - t^3) dt$ (b) $\int_0^{10} 4te^{-2t} dt$ (c) $\int_0^{10} \frac{10t^2 - t^3}{t + 1} dt$
3. (a) $\int_0^1 \frac{4x^3}{\sqrt{4 - x^2}} dx$ (b) $\int_1^8 \frac{1}{3 + \sqrt{t + 8}} dt$ (c) $\int_1^e \sqrt{x} \ln x dx$
4. (a) $\int \frac{(x^n - x^m)^2}{\sqrt{x}} dx$ (b) $\int_0^{1/3} \frac{dx}{e^x + 1}$ (c) $\int_1^5 x\sqrt{x - 1} dx$
5. (a) $\int_4^9 \frac{(\sqrt{x} - 1)^2}{x} dx$ (b) $\int_0^1 \ln(1 + \sqrt{x}) dx$ (c) $\int_0^{27} \frac{x^{1/3}}{1 + x^{1/3}} dx$

4.2 Leibniz's Formula

Integrals appearing in economics often depend on parameters. How does the value of the integral change if the parameters change? We consider first a simple case.

Differentiation under the Integral Sign

Let f be a function of two variables and consider the function F defined by

$$F(x) = \int_c^d f(x, t) dt$$

where c and d are constants. We want to find $F'(x)$. Since the limits of integration do not depend on x , it is natural to guess that we have the following result:

$$F(x) = \int_c^d f(x, t) dt \implies F'(x) = \int_c^d \frac{\partial f(x, t)}{\partial x} dt \quad (1)$$

Thus we *differentiate the integral with respect to a parameter that occurs only under the integral sign, by differentiating under the integral sign*.

In order to prove (1) we have to rely on the definition of the derivative. We get

$$\begin{aligned} F'(x) &= \lim_{h \rightarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \rightarrow 0} \int_c^d \frac{f(x+h, t) - f(x, t)}{h} dt \\ &= \int_c^d \lim_{h \rightarrow 0} \frac{f(x+h, t) - f(x, t)}{h} dt = \int_c^d f'_x(x, t) dt \end{aligned}$$

The only non-obvious step here is moving the limit inside the integral sign. (See Protter and Morrey (1991), Theorem 11.1.) A more precise result (for a more general case) is given in Theorem 4.2.1.

EXAMPLE 1 The present value of a continuous flow of income $f(t)$, $t \in [0, T]$, at interest rate r , is

$$K = \int_0^T f(t)e^{-rt} dt$$

Find dK/dr . (The limits of integration are independent of r .)

Solution: Formula (1) implies that

$$\frac{dK}{dr} = \int_0^T f(t)(-t)e^{-rt} dt = - \int_0^T tf(t)e^{-rt} dt$$

The General Case

The general problem can be formulated as follows: let $f(x, t)$, $u(x)$, and $v(x)$ be given functions, and define the function F by the formula

$$F(x) = \int_{u(x)}^{v(x)} f(x, t) dt \quad (2)$$

If x changes, then the limits of integration $v(x)$ and $u(x)$ both change, and in addition the integrand $f(x, t)$ changes for each t . What is the total effect on $F(x)$ from such a change in x ? In particular, what is $F'(x)$?

The answer is given in Theorem 4.2.1.¹ (Recall that a function of n variables is a C^k function if it and all its partial derivatives up to and including order k are continuous.)

EM 4.2.1 (LEIBNIZ'S FORMULA)

Suppose that $f(x, t)$ and $f'_x(x, t)$ are continuous over the rectangle determined by $a \leq x \leq b$, $c \leq t \leq d$. Suppose too that $u(x)$ and $v(x)$ are C^1 functions over $[a, b]$, and that the ranges of u and v are contained in $[c, d]$. Then

$$F(x) = \int_{u(x)}^{v(x)} f(x, t) dt$$

$$\implies F'(x) = f(x, v(x))v'(x) - f(x, u(x))u'(x) + \int_{u(x)}^{v(x)} \frac{\partial f(x, t)}{\partial x} dt \quad (3)$$

Proof: Let H be the following function of three variables:

$$H(x, u, v) = \int_u^v f(x, t) dt$$

Then $F(x) = H(x, u(x), v(x))$ and, according to the chain rule,

$$F'(x) = H'_x + H'_u u'(x) + H'_v v'(x) \quad (*)$$

where H'_x is the partial derivative of H w.r.t. x with u and v as constants. Because of (1), $H'_x = \int_u^v f'_x(x, t) dt$. Moreover, according to (4.1.5), $H'_u = f(x, u)$ and $H'_v = f(x, v)$. Inserting these results into (*) yields (3). ■

LE 2 Use (3) to compute $F'(x)$ when $F(x) = \int_x^{x^2} \frac{1}{2} t^2 x dt$. Check the answer by calculating the integral first and then differentiating.

Solution: We obtain

$$F'(x) = \frac{1}{2}(x^2)^2 x \cdot 2x - \frac{1}{2}x^2 x \cdot 1 + \int_x^{x^2} \frac{1}{2} t^2 dt$$

$$= x^6 - \frac{1}{2}x^3 + \left[\frac{1}{6} t^3 \right]_x^{x^2} = x^6 - \frac{1}{2}x^3 + \frac{1}{6}(x^2)^3 - \frac{1}{6}x^3 = \frac{7}{6}x^6 - \frac{2}{3}x^3$$

¹ In Richard Feynman's *Surely You're Joking, Mr. Feynman!* (Bantam Books, New York, 1986), the late Nobel laureate vividly describes the usefulness of this result to physicists; it is equally useful to economists.

In this case, the integral $F(x)$ is easy to calculate explicitly:

$$F(x) = \frac{1}{2}x \int_x^{x^2} t^2 dt = \frac{1}{2}x \left[\frac{1}{3} t^3 \right]_x^{x^2} = \frac{1}{6}(x^7 - x^4)$$

Differentiating w.r.t. x gives the same expression for $F'(x)$ as before. ■

EXAMPLE 3

In a growth model studied by N. Kaldor and J. A. Mirrlees, a function N is defined by

$$N(t) = \int_{t-T(t)}^t n(\tau) e^{-\delta(t-T(t))} d\tau$$

where $T = T(t)$ is a given function. Compute $\dot{N}(t)$ under appropriate conditions on the functions n and T .

Solution: If n is continuous and T is C^1 , Leibniz's formula gives

$$\dot{N}(t) = n(t) e^{-\delta(t-T(t))} - n(t-T(t)) e^{-\delta(t-T(t))} (1 - \dot{T}(t))$$

$$+ \int_{t-T(t)}^t n(\tau) (-\delta) (1 - \dot{T}(t)) e^{-\delta(t-T(t))} d\tau$$

$$= [n(t) - (1 - \dot{T}(t)) n(t-T(t))] e^{-\delta(t-T(t))} - \delta(1 - \dot{T}(t)) N(t) \quad \blacksquare$$

EXAMPLE 4

Suppose that a small business earns a net profit stream $y(t)$ for $t \in [0, T]$. At time $s \in [0, T]$, the discounted value (DV) of future profits is

$$V(s, r) = \int_s^T y(t) e^{-r(t-s)} dt$$

where r is the constant rate of discount. Compute $V'_s(s, r)$ by means of Leibniz's rule.

Solution: We get

$$V'_s(s, r) = -y(s) + \int_s^T y(t) r e^{-r(t-s)} dt = -y(s) + r V(s, r) \quad (*)$$

where the last equality was obtained by moving the constant r outside the integral sign.

Solving equation (*) for r yields

$$r = \frac{y(s) + V'_s(s, r)}{V(s, r)} \quad (**)$$

This has an important interpretation. At time s , the business owner earns $y(s)$, and the DV of future profits is increasing at the instantaneous rate $V'_s(s, r)$. The ratio on the right-hand side of (**) is known as the *instantaneous proportional rate of return* of the investment. Equation (**) requires this ratio to be equal to r . In fact, if r were the instantaneous proportional rate of return on a (relatively) safe asset like government bonds, and if the left-hand side of (**) were higher than the right-hand side, then the business owner would be better off selling the business for the amount $V(s, r)$, which it is worth at time s , and holding bonds instead. But if the left-hand side of (**) were lower than the right-hand side, then existing bondholders would do better to sell their bonds and set up replicas of this small business.

Infinite Intervals of Integration

Leibniz's formula can be generalized to integrals with unbounded intervals of integration.

EM 4.2.2

Suppose that $f(x, t)$ and $f'_x(x, t)$ are continuous for all $t \geq c$ and all x in $[a, b]$, and suppose that the integral

$$\int_c^\infty f(x, t) dt \quad (4)$$

converges for each x in $[a, b]$. Suppose further that $f'_x(x, t)$ is **integrably bounded** in the sense that there exists a function $p(t)$, independent of x , for which $\int_c^\infty p(t) dt$ converges and $|f'_x(x, t)| \leq p(t)$ for all $t \geq c$ and all x in $[a, b]$. Then

$$\frac{d}{dx} \int_c^\infty f(x, t) dt = \int_c^\infty f'_x(x, t) dt \quad (5)$$

The existence of $p(t)$ can be replaced by the weaker condition that $\int_c^\infty f'_x(x, t) dt$ converges uniformly on $[a, b]$. We refer to Chapter 11 of Protter and Morrey (1991) for the definition of uniform convergence, and for the proof of Theorem 4.2.2.

Obvious changes to Theorem 4.2.2 yield similar theorems for integrals of the type $\int_{-\infty}^d f(x, t) dt$, and also of the type $\int_{-\infty}^{+\infty} f(x, t) dt$. Combining these results with Leibniz's formula gives conditions ensuring that the formula applies to integrals like $\int_{-\infty}^{u(x)} f(x, t) dt$ and $\int_{u(x)}^\infty f(x, t) dt$ over variable infinite intervals.

LE 5 Let $K(t)$ denote the capital stock of some firm at time t , and $p(t)$ the purchase price per unit of capital. Let $R(t)$ denote the rental price per unit of capital. In capital theory, one principle for determining the acquisition value $V(t)$ of the firm's capital implies that

$$V(t) = p(t)K(t) = \int_t^\infty R(\tau)K(\tau)e^{-r(\tau-t)} d\tau \quad (\text{for all } t) \quad (*)$$

This says that $V(t)$ should equal the discounted present value of the returns from using the firm's capital. Find an expression for $R(t)$ by differentiating $(*)$ w.r.t. t .

Solution: We get $\dot{p}(t)K(t) + p(t)\dot{K}(t) = -R(t)K(t) + \int_t^\infty R(\tau)K(\tau)re^{-r(\tau-t)} d\tau$. The last integral is simply $r p(t)K(t)$, so solving the equation for $R(t)$ yields

$$R(t) = \left(r - \frac{\dot{K}(t)}{K(t)}\right)p(t) - \dot{p}(t)$$

Thus the rental price is equal to $r p(t)$, the interest cost of each unit of capital, minus $p\dot{K}/K$, which represents the loss from depreciation, minus $\dot{p}$, because increases in the price of the

PROBLEMS FOR SECTION 4.2

1. Find an expression for $F'(x)$ when

$$(a) F(x) = \int_1^2 \frac{e^{xt}}{t^2} dt \quad (x \neq 0) \quad (b) F(x) = \int_1^e \ln(xt) dt \quad (x > 0)$$

$$(c) F(x) = \int_0^1 \frac{e^{-t} dt}{1+xt} \quad (x > -1) \quad (d) F(x) = \int_3^8 \frac{t^2}{(1-xt)^2} dt \quad (x > \frac{1}{3})$$

(Do not try to evaluate the integrals you get in (c) and (d).)

2. Use formula (1) to find $F'(\alpha)$ when $F(\alpha) = \int_0^1 x e^{\alpha x^2} dx$. Check the result by finding an explicit expression for $F(\alpha)$ and differentiating.

3. Use (3) to find an expression for $F'(x)$ when

$$(a) F(x) = \int_0^{2x} t^3 dt \quad (b) F(x) = \int_0^x (x^2 + t^3)^2 dt \quad (c) F(x) = \int_{\sqrt{x}}^{x^2} \cos(t^2 - x^4) dt$$

4. Let f and g be C^1 functions. Find an expression for $I = \frac{d}{d\rho} \int_0^{\kappa(\rho)} e^{-\rho t} f(t) dt$.

5. The **moment generating function** of a random variable X with density function f is $M(t) = \int_{-\infty}^\infty e^{tx} f(x) dx$. Prove (under suitable conditions on f and t) that $M'(0) = \int_{-\infty}^\infty x f(x) dx$, the expectation of X , and generally that the n th derivative $M^{(n)}(0) = \int_{-\infty}^\infty x^n f(x) dx$, which is the n th moment.

6. Find $\dot{x}(t)$ if $x(t) = \int_{-\infty}^t e^{-\delta(t-\tau)} y(\tau) d\tau$.

7. A model by J. Tobin involves the function $F(\sigma_k) = \int_{-\infty}^{+\infty} U(\mu_k + \sigma_k z) f(z, 0, 1) dz$, where μ_k is a function of σ_k . Under suitable restrictions on the functions U , μ_k , and f , find an expression for $dF(\sigma_k)/d\sigma_k$.

HARDER PROBLEMS

8. A vintage growth model due to L. Johansen involves the following definitions:

$$K(t) = \int_{-\infty}^t f(t-\tau)k(\tau) d\tau, \quad T(0) = \int_0^\infty f(\xi) d\xi, \quad V(t) = \frac{1}{T(0)} \int_{-\infty}^t G(\tau, t) d\tau$$

where $G(\tau, t) = k(\tau) \int_{-\infty}^\infty f(\xi) d\xi$. With suitable restrictions on the functions involved, prove that $\dot{V}(t) = k(t) - K(t)/T(0)$.

9. Define

$$z(t) = \int_t^{2t} x(\tau) \exp\left(-\int_t^\tau r(s) ds\right) d\tau, \quad p(t) = \exp\left(-\int_t^{2t} r(s) ds\right)$$

where the functions $x(\tau)$ and $r(s)$ are both differentiable. Show that

$$\dot{z}(t) - r(t)z(t) = 2p(t)x(2t) - x(t)$$

10. A firm faces uncertain demand D and has existing inventory I . There are different costs per unit of having too much or too little stock. So the firm wants to choose its stock level Q to minimize the function

$$g(Q) = c(Q - I) + h \int_0^Q (Q - D)f(D) dD + p \int_Q^a (D - Q)f(D) dD$$

where c, I, h, p , and a are positive constants, $p \geq c$, and f is a given continuous and non-negative function that satisfies $\int_0^a f(D) dD = 1$ (so f can be interpreted as a probability density function).

- (a) Compute $g'(Q)$ and $g''(Q)$, and show that g is convex.
 (b) Define $F(Q^*) = \int_0^{Q^*} f(D) dD$, where Q^* is the minimum point of $g(Q)$. Use the first-order conditions for minimization of g to find an equation for $F(Q^*)$, the probability that demand D does not exceed Q^* . Use this equation to find the value of $F(Q^*)$.

4.3 The Gamma Function

Around 1725 the Swiss mathematician L. Euler asked the following question: is there a natural way to extend the definition of the factorial function $n! = 1 \cdot 2 \cdot \dots \cdot n$ to noninteger values of n ? Euler thereby discovered one of the most studied functions in the whole of mathematical analysis, the **gamma function**. It is defined by

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt, \quad x > 0 \quad (1)$$

(Recall that Γ is the upper case Greek letter “gamma”.) This function crops up in several areas of application, and there is a vast literature investigating its mathematical properties.² We shall just mention a few simple properties.

For definition (1) to make sense, it must be shown that the integral exists for each $x > 0$. Not only is the interval of integration unbounded, but for each x in $(0, 1)$ the integrand $e^{-t} t^{x-1}$ tends to ∞ as $t \rightarrow 0$. In order to show that the integral converges, we partition the interval $(0, \infty)$ into two parts to obtain:

$$\int_0^\infty e^{-t} t^{x-1} dt = \int_0^1 e^{-t} t^{x-1} dt + \int_1^\infty e^{-t} t^{x-1} dt \quad (*)$$

Concerning the first integral on the right-hand side, note that $0 \leq e^{-t} \leq 1$ for $t \geq 0$, so $0 \leq e^{-t} t^{x-1} \leq t^{x-1}$ for all $t > 0$. Because $\int_0^1 t^{x-1} dt$ converges even when $0 < x < 1$ (to $1/x$), it follows that $\int_0^1 e^{-t} t^{x-1} dt$ converges. As for the second integral, because $e^{-t} t^b \rightarrow 0$ as $t \rightarrow \infty$ for every b , there exists a number t_0 such that $t \geq t_0$ implies $e^{-t} t^{x+1} < 1$. Hence, $e^{-t} t^{x-1} < 1/t^2$ for $t \geq t_0$. But $\int_{t_0}^\infty (1/t^2) dt$ converges (to $1/t_0$), so the second integral on the right-hand side of $(*)$ converges. Thus $\Gamma(x)$ is well-defined for all $x > 0$.

² For example N. Nielsen: *Handbuch der Theorie der Gammafunktion*. Leipzig (1906) 376 pages

Let us compute some values of $\Gamma(x)$. For $x = 1$ it is easy:

$$\Gamma(1) = \int_0^\infty e^{-t} dt = 1$$

Further, integration by parts gives $\Gamma(x+1) = \int_0^\infty e^{-t} t^x dt = -\int_0^\infty e^{-t} t^x dt + \int_0^\infty e^{-t} x t^{x-1} dt = x \Gamma(x)$. This implies the **functional equation**

$$\Gamma(x+1) = x \Gamma(x) \quad \text{for } x > 0 \quad (2)$$

for the gamma function. It implies that $\Gamma(2) = \Gamma(1+1) = 1 \cdot \Gamma(1) = 1$, that $\Gamma(3) = 2 \cdot \Gamma(2) = 2 \cdot 1$, and that $\Gamma(4) = 3 \cdot \Gamma(3) = 3 \cdot 2 \cdot 1$. By induction,

$$\Gamma(n) = (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1 = (n-1)!$$

for every natural number n .

It is more difficult to compute $\Gamma(x)$ if x is not a natural number. In order to compute $\Gamma(\frac{1}{2})$, for instance, we need the **Poisson integral formula**

$$\int_0^\infty e^{-t^2} dt = \frac{1}{2} \sqrt{\pi} \quad (3)$$

This is proved in Example 4.8.2. By symmetry of the graph of e^{-t^2} about $t = 0$, it follows that $\int_{-\infty}^{+\infty} e^{-t^2} dt = \sqrt{\pi}$. Substituting $u = \sqrt{\lambda} t$ leads to

$$\int_{-\infty}^{+\infty} e^{-\lambda t^2} dt = \sqrt{\frac{\pi}{\lambda}} \quad (\lambda > 0) \quad (4)$$

Now (3) allows us to compute $\Gamma(\frac{1}{2})$, using the substitution $t = u^2$. In fact, $\Gamma(\frac{1}{2}) = \int_0^\infty e^{-t} t^{-1/2} dt = 2 \int_0^\infty e^{-u^2} du = \sqrt{\pi}$.

Once the values of Γ in $(0, 1]$ are known, the functional equation (2) allows us to find $\Gamma(x)$ for every positive x .

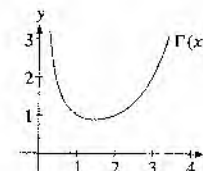


Figure 1 The gamma function

The gamma function is continuous in the interval $(0, \infty)$. It can be shown that it has a minimum ≈ 0.8856 at the point $x \approx 1.4616$. The graph is shown in Fig. 1.

The gamma function plays an important role in statistics. Problem 5 is concerned with the gamma distribution, which is used in many statistical models.

PROBLEMS FOR SECTION 4.3

1. Compute

$$(a) \int_0^{\infty} e^{-ax^2} dx \quad (a > 0)$$

$$(b) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

2. Use (2) to find $\Gamma(\frac{3}{2})$ and show by induction that, for every natural number n ,

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{2^n} \sqrt{\pi} = \frac{(2n-1)!}{2^{2n-1} (n-1)!} \sqrt{\pi}$$

3. One can show that for every $x > 0$ there exists a θ in $(0, 1)$ (where θ depends on x) such that

$$\Gamma(x) = \sqrt{2\pi} x^{x-1/2} e^{-x} e^{\theta/12x}$$

Use this formula to show that if n is a natural number, then

$$n! \approx \sqrt{2\pi n} (n/e)^n \quad (\text{Stirling's formula})$$

in the sense that the ratio between the two expressions tends to 1 as $n \rightarrow \infty$.

4. Show that $\Gamma(x) = \int_0^1 (\ln(1/z))^{x-1} dz$. (Hint: Substitute $t = -\ln z$ in (1).)

5. (a) The **gamma distribution** with parameters $\lambda > 0$ and $\alpha > 0$ is given by

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \quad \text{for } x > 0, \quad f(x) = 0 \quad \text{for } x \leq 0$$

Prove that $\int_{-\infty}^{\infty} f(x) dx = 1$.

(b) Compute the moment generating function $M(t)$ associated with f provided $t < \lambda$. (See Problem 4.2.5.) Compute also $M'(0)$, and in general, $M^{(n)}(0)$.

4.4 Multiple Integrals over Product Domains

The remainder of this chapter deals with multiple integrals. These arise in statistics when considering multidimensional continuous (probability) distributions. Double integrals also play a role in some interesting continuous time dynamic optimization problems. We start with the simplest case.

Double Integrals over Rectangles

The first topic is integration of functions of two variables defined over rectangles in the xy -plane. We begin with a geometric problem.

The Cartesian product of the two intervals $[a, b]$ and $[c, d]$ is the rectangle $R = [a, b] \times [c, d]$.

be a continuous function defined on R with $f(x, y) \geq 0$ for all (x, y) in R . Consider Fig. 1. The double integral over R will measure the volume of the "box" that has the rectangle R as its bottom and the graph of f as its curved "lid". This box consists of all points (x, y, z) such that $(x, y) \in R$ and $0 \leq z \leq f(x, y)$. This is also called the **ordinate set** of f over R .

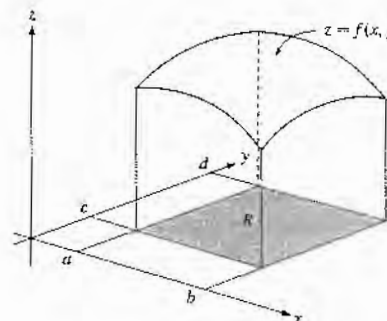


Figure 1

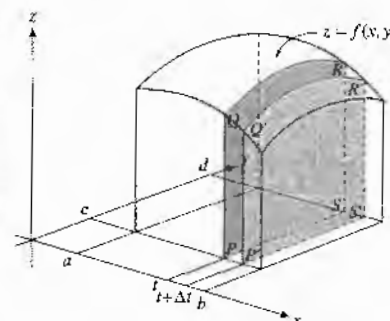


Figure 2

Let t be an arbitrary point in the interval $[a, b]$. Construct a plane parallel to the yz -plane intersecting the x -axis at $x = t$. This plane cuts the ordinate set of f into two parts. The intersection of this plane with the ordinate set is the shaded plane surface $PQRS$ in Fig. 2. The area of this shaded surface is a function of t , which we denote by $A(t)$. It is the area under the curve connecting Q to R over the interval $[c, d]$. The relevant curve is the intersection between the graph of $z = f(x, y)$ and the plane $x = t$, so its equation is $z = \phi(y) = f(t, y)$ with t fixed and $y \in [c, d]$. Hence,

$$A(t) = \int_c^d f(t, y) dy \quad (*)$$

Denote by $V(t)$ the volume of the ordinate set of f over the variable rectangle $[a, t] \times [c, d]$. In particular, $V(a) = 0$, and $V(b)$ is the total volume to be evaluated.

If we add Δt to t , the incremental volume is $V(t + \Delta t) - V(t)$. In Fig. 2 this is the volume of the slice that lies between the surfaces $PQRS$ and $P'Q'R'S'$. If Δt is small, then this volume is approximately equal to $A(t)\Delta t$. Therefore $V(t + \Delta t) - V(t) \approx A(t)\Delta t$, implying that

$$\frac{V(t + \Delta t) - V(t)}{\Delta t} \approx A(t)$$

This approximation, in general, improves as Δt gets smaller. In the limit as $\Delta t \rightarrow 0$ we can reasonably expect to obtain $V'(t) = A(t)$. Hence, $V(b) - V(a) = \int_a^b A(t) dt$. Because $V(a) = 0$, if we put $V = V(b)$ and use $(*)$, we get

$$V = \int_a^b \left(\int_c^d f(t, y) dy \right) dt \quad (**)$$

The preceding argument receives support from the next example and from Theorem 4.4.1. This makes $(**)$ a natural definition of the volume of the ordinate set of f over R .

E 1 If $f(x, y) = M$ for all (x, y) in R , where M is a positive constant, then the ordinate set of f over R is a rectangular box in the usual sense. The base area is $(b-a)(d-c)$ and its height is M , so its volume is $M(b-a)(d-c)$. Show that (**) gives the same result.

Solution: Letting $f(x, y) = M$ in (**) yields

$$\int_a^b \left(\int_c^d M dy \right) dx = \int_a^b \left(\int_c^d M dy \right) dx = \int_a^b M(d-c) dx = \int_a^b M(d-c) dx = M(d-c)(b-a)$$

Suppose we try to find the volume of the ordinate set of f over $R = [a, b] \times [c, d]$ by using the argument above, except that we now choose t in $[c, d]$ and then let the intersecting plane be parallel to the xz -plane and pass through the point $y = t$ on the y -axis. The area of the plane surface in the intersection between the ordinate set and the plane $y = t$ is $\int_a^b f(x, t) dx$, so the formula for the volume becomes

$$\int_c^d \left(\int_a^b f(x, t) dx \right) dt \quad (***)$$

Because we are computing the same volume in both cases, we should get the same answer, provided our intuitive argument above is correct. The next theorem guarantees that the numbers obtained in (**) and (***) are indeed equal if f is continuous on R . (See Protter and Morrey (1991), Chapter 8, for a proof.)

THEOREM 4.4.1

Let f be a continuous function defined over the rectangle $R = [a, b] \times [c, d]$. Then

$$\int_a^b \left(\int_c^d f(t, y) dy \right) dt = \int_c^d \left(\int_a^b f(x, t) dx \right) dt$$

Now, let f be an arbitrary continuous function over the rectangle $R = [a, b] \times [c, d]$. We then define the **double integral of f over R** , denoted by $\iint_R f(x, y) dx dy$, as

$$\iint_R f(x, y) dx dy = \int_a^b \left(\int_c^d f(x, y) dy \right) dx = \int_c^d \left(\int_a^b f(x, y) dx \right) dy \quad (1)$$

We can take either of the two last expressions as the definition of the double integral, because they are equal according to Theorem 4.4.1.

Note that we can calculate $\int_a^b \left(\int_c^d f(x, y) dy \right) dx$ in two stages as follows:

- First, keep x fixed and integrate $f(x, y)$ w.r.t. y from $y = c$ to $y = d$. This gives $\int_c^d f(x, y) dy$, a function of x .
- Then integrate $\int_c^d f(x, y) dy$ from $x = a$ to $x = b$ to obtain $\int_a^b \left(\int_c^d f(x, y) dy \right) dx$.

Notice that in (1) we do not require $f(x, y)$ to be nonnegative. It turns out therefore that double integrals need not always be interpreted as volumes, just as single integrals need not always be interpreted as areas.

Let us now consider some applications of (1).

EXAMPLE 2

Compute $\iint_R (x^2 y + xy^2 + 2x) dx dy$, where $R = [0, 1] \times [-1, 3]$.

Solution: The integrand is continuous everywhere. Consider first

$$\int_0^1 \left(\int_{-1}^3 (x^2 y + xy^2 + 2x) dy \right) dx$$

Treating x as a constant, first evaluate the inner integral:

$$\int_{-1}^3 (x^2 y + xy^2 + 2x) dy = \left| \frac{1}{2} x^2 y^2 + \frac{1}{3} x y^3 + 2xy \right|_{y=-1}^{y=3} = 4x^2 + \frac{52}{3} x$$

Integrating a second time gives

$$\int_0^1 \left(\int_{-1}^3 (x^2 y + xy^2 + 2x) dy \right) dx = \int_0^1 \left(4x^2 + \frac{52}{3} x \right) dx = \left| \frac{4}{3} x^3 + \frac{26}{3} x^2 \right|_0^1 = 10$$

Let us now perform the integration in the reverse order. Holding y constant, we get

$$\int_0^1 (x^2 y + xy^2 + 2x) dx = \left| \frac{1}{3} x^3 y + \frac{1}{2} x^2 y^2 + x^2 \right|_0^1 = \frac{1}{3} y + \frac{1}{2} y^2 + 1$$

Therefore,

$$\int_{-1}^3 \left(\int_0^1 (x^2 y + xy^2 + 2x) dx \right) dy = \int_{-1}^3 \left(\frac{1}{3} y + \frac{1}{2} y^2 + 1 \right) dy = 10$$

We reached the same result by both procedures. So Theorem 4.4.1 is confirmed in this case, and we can write with confidence

$$\iint_R (x^2 y + xy^2 + 2x) dx dy = 10 \quad \text{when } R = [0, 1] \times [-1, 3]$$

EXAMPLE 3

Compute $\int_1^b \left(\int_1^d \frac{y-x}{(y+x)^3} dy \right) dx$, where b and d are constants greater than 1.

Solution: By means of a little trick, the inner integral becomes

$$\begin{aligned} \int_1^d \frac{y-x}{(y+x)^3} dy &= \int_1^d \frac{y+x-2x}{(y+x)^3} dy = \int_1^d \frac{1}{(y+x)^2} dy - 2x \int_1^d \frac{1}{(y+x)^3} dy \\ &= \left| -\frac{1}{y+x} \right|_{y=1}^{y=d} - 2x \left| -\frac{1}{2(y+x)^2} \right|_{y=1}^{y=d} = -\frac{d}{(x+d)^2} + \frac{1}{(x+1)^2} \end{aligned}$$

Hence,

$$\begin{aligned} \int_1^b \left(\int_1^d \frac{y-x}{(y+x)^3} dy \right) dx &= -\int_1^b \frac{d}{(x+d)^2} dx + \int_1^b \frac{1}{(x+1)^2} dx \\ &= \left| \frac{d}{x+d} \right|_1^b - \left| \frac{1}{x+1} \right|_1^b = \frac{d}{b+d} - \frac{d}{1+d} - \frac{1}{b+1} + \frac{1}{1+1} \end{aligned}$$

Choosing instead to integrate w.r.t. x first, a similar trick leads to

$$\int_1^b \frac{y-x}{(y+x)^3} dx = \frac{b}{(y+b)^2} - \frac{1}{(y+1)^2}$$

Then

$$\int_1^d \left(\int_1^b \frac{y-x}{(y+x)^3} dx \right) dy = -\frac{b}{b+d} + \frac{b}{b+1} + \frac{1}{d+1} - \frac{1}{2}$$

Simple algebra now shows that the two results are equal.

Multiple Integrals

Let Ω denote the Cartesian product $[a_1, b_1] \times \cdots \times [a_n, b_n]$ of the closed intervals $[a_1, b_1], \dots, [a_n, b_n]$. It is the set of all n -vectors $(x_1, x_2, \dots, x_n)$ in $\mathbb{R}^n$ such that $a_i \leq x_i \leq b_i$ for $i = 1, 2, \dots, n$. We call Ω an **n -dimensional rectangle**.

If f is a continuous function defined over Ω , define the **multiple integral** of f over Ω as

$$\begin{aligned} \iint \cdots \int_{\Omega} f(x_1, \dots, x_{n-1}, x_n) dx_1 \cdots dx_{n-1} dx_n \\ = \int_{a_n}^{b_n} \left(\int_{a_{n-1}}^{b_{n-1}} \cdots \left(\int_{a_1}^{b_1} f(x_1, \dots, x_{n-1}, x_n) dx_1 \right) \cdots dx_{n-1} \right) dx_n \end{aligned} \quad (2)$$

The meaning of the notation on the right-hand side of (2) is that integration is to be performed first w.r.t. x_1 , all other variables being treated as constants, then w.r.t. x_2 , treating the remainder of the variables $(x_3, \dots, x_n)$ as constants, etc.

Definition (2) is a simple generalization of (1). In this general case one can still prove that the order of integration on the right-hand side is immaterial, provided that f is continuous in Ω .

IS FOR SECTION 4.4

1. Evaluate the following double integrals:

- (a) $\int_0^2 \int_0^1 (2x + 3y + 4) dx dy$ (b) $\int_0^a \int_0^b (x-a)(x-b) dx dy$
 (c) $\int_1^3 \int_1^2 (x-y)/(x+y) dx dy$ (d) $\int_0^{1/2} \int_0^{2\pi} y^3 \sin(xy^2) dx dy$

SM 2. Find $I = \int_1^a \left(\int_0^b \frac{1}{x^2} e^{y/x} dy \right) dx$ ($a > 1, b > 0$)

SM 3. Consider the function $f(x, y) = \frac{2k}{(x+y+1)^2}$, where k is a constant. Let R be the rectangle $R = [0, a] \times [0, 1]$, where $a > 0$ is a constant. Determine the value k_a of k such that

4. Compute the double integral $I = \int_0^2 \left(\int_{-2}^1 (x^2 y^3 - (y+1)^2) dy \right) dx$.

HARDER PROBLEMS

SM 5. Find $I = \iiint_{\Omega} \cdots \int_{\Omega} (x_1^2 + x_2^2 + \cdots + x_n^2) dx_1 dx_2 \cdots dx_n$ where Ω is the region in $\mathbb{R}^n$ determined by the inequalities $0 \leq x_i \leq 1$ for $i = 1, 2, \dots, n$.

4.5 Double Integrals over General Domains

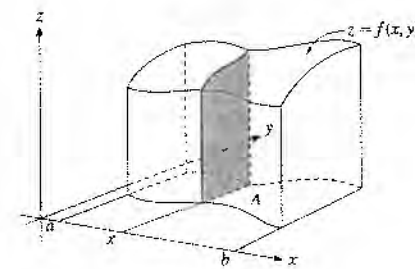
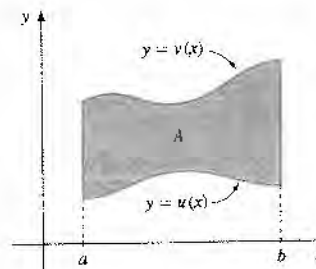
Consider the set A in the xy -plane indicated in Fig. 1. The boundary of A consists of segments of the lines $x = a$ and $x = b$ and the graphs of the continuous functions $u(x)$ and $v(x)$, where $u(x) \leq v(x)$ for all x in $[a, b]$.

Suppose that $f(x, y)$ is a continuous function defined over A , and that $f(x, y) \geq 0$ for all (x, y) in A . Then the graph of f above the set A determines a three-dimensional volume, as indicated in Fig. 2. The intersection of the solid with the plane at distance x from the yz -plane is the shaded plane region indicated in Fig. 2. The area of this region can be described as the area under the graph of $f(x, y)$ (with x fixed) over the interval $[u(x), v(x)]$. Let $F(x)$ denote the resulting function of x . Then

$$F(x) = \int_{u(x)}^{v(x)} f(x, y) dy$$

Here we have integrated w.r.t. y while keeping x fixed. One can prove that $F(x)$ is a continuous function of x . As in the case where A is rectangular, a geometrically plausible argument supports the conclusion that the volume V of the solid must be given by

$$V = \int_a^b F(x) dx = \int_a^b \left(\int_{u(x)}^{v(x)} f(x, y) dy \right) dx \quad (1)$$



Briefly formulated, the argument is this: let $V(x)$ denote the volume of that part of the solid in Fig. 2 which lies to left of the shaded region. Thus $V(a) = 0$ and $V(b) = V$, and $F(x)$ is the area of the shaded region. Let x be incremented by Δx . Then the volume of the associated slice of thickness Δx is approximately $F(x)\Delta x$. The exact volume of this slice is $V(x + \Delta x) - V(x)$, which is therefore approximately equal to $F(x)\Delta x$. This approximation will, in general, be better for smaller Δx , so in the limit we expect to have $V'(x) = F(x)$. Hence, $V(b) - V(a) = \int_a^b F(x) dx$, so that $V = \int_a^b F(x) dx$ (since $V(b) = V$, $V(a) = 0$).

Formally, we could define the volume V by (1). Note that if $u(x) = c$ and $v(x) = d$, so that A is a rectangle, then definition (1) reduces to (**) in Section 4.4. Let us illustrate with an example.

EX 1 Let A be the set in the xy -plane bounded by the straight lines $x = 0$ and $x = 1$ and the graphs of $y = x$ and $y = x^2 + 1$. The set A is indicated in Fig. 3. The function $f(x, y) = xy^2$ is continuous and ≥ 0 over A . Find the volume V under the graph of f .

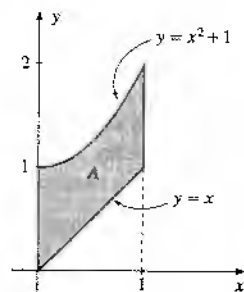


Figure 3

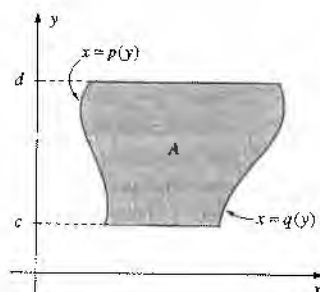


Figure 4

Solution: In this case

$$V = \int_0^1 \left(\int_x^{x^2+1} xy^2 dy \right) dx$$

Here

$$\int_x^{x^2+1} xy^2 dy = x \left[\frac{1}{3}y^3 \right]_x^{x^2+1} = \frac{1}{3}x[(x^2+1)^3 - x^3] = \frac{1}{3}x^7 + x^5 - \frac{1}{3}x^4 + x^3 + \frac{1}{3}x$$

because x is kept constant when integrating w.r.t. y . Hence,

$$V = \int_0^1 \left(\frac{1}{3}x^7 + x^5 - \frac{1}{3}x^4 + x^3 + \frac{1}{3}x \right) dx = \left[\frac{1}{24}x^8 + \frac{1}{6}x^6 - \frac{1}{15}x^5 + \frac{1}{4}x^4 + \frac{1}{6}x^2 \right]_0^1 = \frac{67}{120}$$

It is possible to derive similar expressions for volumes in space when the base A is determined in other ways. For example, if the set A is as indicated in Fig. 4, and $f(x, y) \geq 0$ in A , then the volume V under the graph of f over A is given by

$$\int_c^d \left(\int_{p(y)}^{q(y)} f(x, y) dx \right) dy \quad (2)$$

It is a worthwhile exercise to go through the argument leading to this formula following the same pattern as used for equation (1) above.

In Fig. 5, we see that every straight line parallel to the x -axis or y -axis intersects the boundary of the shaded set in at most two points. Let the functions u and v depicted in Fig. 5 be continuous. Since they are strictly increasing, they have continuous inverse functions u^{-1} and v^{-1} . If f is a continuous nonnegative function defined over this set, the volume under the graph of f can be computed in two different ways. Under the given conditions one can prove that

$$\int_0^b \left(\int_{u(x)}^{v(x)} f(x, y) dy \right) dx = \int_0^d \left(\int_{v^{-1}(y)}^{u^{-1}(y)} f(x, y) dx \right) dy \quad (3)$$

On the left-hand side we have integrated first w.r.t. y and then w.r.t. x , and on the right-hand side we have integrated in the reverse order. If the set is of the type indicated, and f is continuous, the two expressions are always equal. Nevertheless, it is sometimes important to choose the right order of integration in order to have simple integrals. (See Problem 4.)

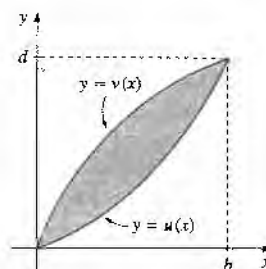


Figure 5

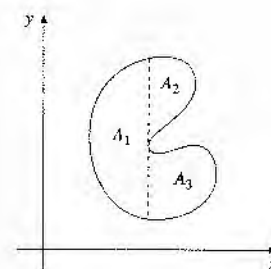


Figure 6

How do we define the double integral over more complicated domains of integration, such as the set in Fig. 6? The obvious solution is to partition the set into smaller parts, each of which is one of the types considered above. (One such partition is indicated in Fig. 6.) We then define the double integral over the entire set as the sum of the double integrals over each of its parts. If the set is a union of finitely many sets of the types we have considered, one can prove that the double integral is independent of how this subdivision is done.

Let A be an arbitrary set in the xy -plane of one of the types considered above, and f a continuous function defined on A (not necessarily ≥ 0). The **double integral of f over A** ,

$$\iint_A f(x, y) dx dy \quad (4)$$

is defined as in (1) provided A is as in Fig. 1, by (2) if A is of the form in Fig. 4, and so on. If $f(x, y) \geq 0$, the number obtained from (4) can be interpreted as the volume of a solid in space. It turns out, however, that the double integral as defined here can be given a number of other interpretations of greater interest to economists. In statistics, for example, suppose two random variables x and y have a joint probability density given by $f(x, y)$, which is always ≥ 0 . Then the probability that the random pair (x, y) belongs to the set A is given by (4). Also, in the theory of production, multiple integrals of capacity distributions are

A Useful Formula

Let $f(x, y)$ be a continuous function over the rectangle $[a, b] \times [c, d]$. We shall prove that

$$\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y) \quad \text{for all } (x, y) \text{ in } [a, b] \times [c, d] \implies \int_c^d \left(\int_a^b f(x, y) dx \right) dy = F(b, d) - F(a, d) - F(b, c) + F(a, c) \quad (5)$$

Indeed, if y in $[c, d]$ is fixed, then $\partial F / \partial y$ is a function of x whose derivative w.r.t. x is $\partial^2 F / \partial x \partial y = f(x, y)$. Hence, for each y in $[c, d]$,

$$\int_a^b f(x, y) dx = \left|_{x=a}^{x=b} \frac{\partial F(x, y)}{\partial y} \right| = \frac{\partial F(b, y)}{\partial y} - \frac{\partial F(a, y)}{\partial y}$$

This implies that

$$\begin{aligned} \int_c^d \left(\int_a^b f(x, y) dx \right) dy &= \int_c^d \left(\frac{\partial F(b, y)}{\partial y} - \frac{\partial F(a, y)}{\partial y} \right) dy = \left|_c^d (F(b, y) - F(a, y)) \right| \\ &= F(b, d) - F(a, d) - F(b, c) + F(a, c) \end{aligned}$$

MS FOR SECTION 4.5

- (a) Sketch the domain of integration and compute the integral $\int_0^1 \left(\int_{x^2}^x (x^2 + xy) dy \right) dx$.
(b) Change the order of integration and verify that you obtain the same result as in (a).
- Compute the integral in Example 1 by first integrating w.r.t. x . (Hint: The set in Fig. 3 must be subdivided into two parts.)
- What is the geometric interpretation of $\iint_A 1 dx dy$, where A is a set in the xy -plane?
- Let $f(x, y) = e^{x^2}$ be defined over the triangle $A = \{(x, y) : x \in [0, 1], 0 \leq y \leq x\}$. Find the volume V under the graph of f over A . (Hint: Integrate first w.r.t. y . If you try to integrate w.r.t. x first, there is no expression for the relevant integral in terms of elementary functions.)
- Compute the integral $\int_0^3 \left(\int_{\sqrt{3x/3}}^{\sqrt{25-x^2}} 2x dy \right) dx$ by reversing the order of integration.
- Calculate $\int_0^1 \int_0^1 |x - y| dx dy$. Can you confirm your result by a geometric argument?

HARDER PROBLEMS

- Sketch the set $A = \{(x, y) : 0 \leq x \leq 2\pi, -x \leq y \leq \sin x\}$ in the xy -plane. Then compute the double integral $\iint_A \sin x \cos y dx dy$.

- A model by J. E. Meade of savings, inheritance, and economic growth involves the double integral

$$I = \int_0^F \left(\int_0^{F-\theta} e^{a\theta} e^{bT} dT \right) d\theta, \quad a \neq 0, b \neq 0, a \neq b.$$

- Show that $I = (\varphi(a) - \varphi(b))/(a - b)$, where $\varphi(u) = (e^{uF} - 1)/u$.
 - Sketch the domain of integration in the θT -plane and write down the expression for I when we first integrate w.r.t. θ . Test the answer in (a) by computing this new double integral (if you have the energy).
- (From Johansen (1972).) For fixed positive values of q_1 and q_2 , consider the set $G(q_1, q_2)$ in the $\xi_1 \xi_2$ -plane given by (draw a sketch!)

$$G(q_1, q_2) = \{(\xi_1, \xi_2) : q_1 \xi_1 + q_2 \xi_2 \leq 1, \xi_1 \geq 0, \xi_2 \geq 0\}$$

Let $f(\xi_1, \xi_2)$ be a continuous function defined over $G(q_1, q_2)$.

- Write down the double integral of f over $G(q_1, q_2)$ when integrating first w.r.t. ξ_2 .
- Write down the corresponding expression, integrating first w.r.t. ξ_1 .
- The value of the double integral in (a) and in (b) will depend on q_1 and q_2 , denote it by $g(q_1, q_2)$. Compute $\partial g / \partial q_1$.

4.6 The Multiple Riemann Integral

We consider next how to define multiple integrals in a way that corresponds to Riemann's definition of the usual single integral. (See e.g. Chapter 9 in EMEA.) We need this definition in order to explain the rule for changing variables in multiple integrals.

Let f be a bounded function defined on a closed and bounded set A in the plane, and let R be a rectangle containing A . Subdivide the rectangle into a number of smaller rectangles as indicated in Fig. 1.

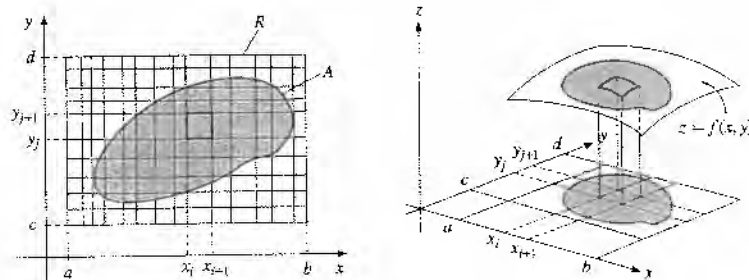


Figure 2

Let R_{ij} denote the typical subrectangle $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$. Some of these rectangles R_{ij} will lie inside A , others will be entirely or partly outside A . For each R_{ij} inside A , choose an arbitrary point (x_i^*, y_j^*) in R_{ij} . The product $f(x_i^*, y_j^*) \Delta x_i \Delta y_j$, where $\Delta x_i = x_{i+1} - x_i$, $\Delta y_j = y_{j+1} - y_j$, can be interpreted geometrically as the volume of the rectangular column suggested in Fig. 2. Form the sum of all these products corresponding to rectangles R_{ij} inside A :

$$\sum_{R_{ij} \subseteq A} f(x_i^*, y_j^*) \Delta x_i \Delta y_j \quad (1)$$

Suppose this sum tends to a limit as the number of rectangles increases in such a way that the diameter of the largest tends to 0.³ Suppose further that this limit is independent of the particular sequence of subdivisions that we choose and also independent of which points (x_i^*, y_j^*) we choose in R_{ij} . Then the limit is called the **double integral of f over A** . (The limit process here is more complicated than those we have used before. For technical details, see e.g. Protter and Morrey (1991), Chapter 8, or Munkres (1991).)

When A is one of the types of set described in Section 4.5 (such as Figs. 4.5.1, 4.5.4, and 4.5.6), one can prove that the limit in (1) exists and is equal to the double integral as defined in Section 4.5.

EX 1 Compute $\int_0^1 \int_0^1 (x + xy) dx dy$ from the definition associated with (1).

Solution: Subdivide the rectangle $R = [0, 1] \times [0, 1]$ into n^2 subrectangles by putting $x_i = i/n$, $y_j = j/n$ for $i, j = 0, \dots, n$. Then $\Delta x_i = x_{i+1} - x_i = 1/n$, $\Delta y_j = y_{j+1} - y_j = 1/n$. Put $x_i^* = i/n$, $y_j^* = j/n$. Then $(x_i^* + x_i^* y_j^*) \Delta x_i \Delta y_j = (i/n + i^2/n^2) 1/n^2 = i/n^3 + i^2/n^4$, so that the sum in (1) becomes

$$\begin{aligned} \sum_{j=0}^{n-1} \sum_{i=0}^{n-1} \left(\frac{1}{n^3} i + \frac{1}{n^4} i^2 \cdot j \right) &= \frac{1}{n^3} \sum_{j=0}^{n-1} \left(\sum_{i=0}^{n-1} i \right) + \frac{1}{n^4} \left(\sum_{j=0}^{n-1} j \right) \left(\sum_{i=0}^{n-1} i \right) \\ &= \frac{1}{n^3} n \frac{(n-1)n}{2} + \frac{1}{n^4} \frac{(n-1)n}{2} \cdot \frac{(n-1)n}{2} = \frac{1}{2} \left(1 - \frac{1}{n} \right) + \frac{1}{4} \left(1 - \frac{1}{n} \right)^2 \end{aligned}$$

As $n \rightarrow \infty$, the number of subrectangles will increase, and at the same time their (equal) diameters will tend to 0. The expression above clearly tends to $1/2 + 1/4 = 3/4$, so we finally obtain

$$\int_0^1 \int_0^1 (x + xy) dx dy = \frac{3}{4}$$

To confirm this result, compute the integral in the usual way.

The definition associated with (1) involved subdividing the given set in the xy -plane into rectangles. But we could equally well have used other kinds of subdivision. Let us consider this briefly without going into technical details. Imagine that the closed and bounded set A is subdivided into n subsets $S_1, \dots, S_n$, with areas $\Delta s_1, \dots, \Delta s_n$ (see Fig. 3).

³ We define the **diameter** of any closed bounded set as the maximum distance between any two of its points.

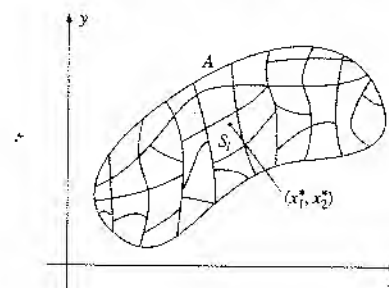


Figure 3

Choose a point (x_i^*, y_i^*) in S_i for each $i = 1, \dots, n$, and form the sum

$$f(x_1^*, y_1^*) \Delta s_1 + \dots + f(x_n^*, y_n^*) \Delta s_n = \sum_{i=1}^n f(x_i^*, y_i^*) \Delta s_i$$

Suppose we construct a whole sequence of such subdivisions with associated sums, and do it in such a way that the greatest diameter of any subset tends to 0. The limit of these sums will then equal the double integral of $f(x, y)$ over A , as defined before. This limit is independent of how we subdivide A . Moreover, it is independent of which points (x_i^*, y_i^*) are chosen in each S_i . Briefly formulated,

$$\iint_A f(x, y) dx dy = \lim_{\text{diam}(S_i) \rightarrow 0} \sum_{i=1}^n f(x_i^*, y_i^*) \Delta s_i \quad (2)$$

Riemann's definition of the double integral makes it possible to prove a number of properties that correspond to similar ones for the one-dimensional integral.

For instance, suppose f and g are continuous functions over a set A in the xy -plane, and that the double integrals of f and g over A are defined. Then

$$\iint_A [f(x, y) + g(x, y)] dx dy = \iint_A f(x, y) dx dy + \iint_A g(x, y) dx dy \quad (3)$$

$$\iint_A c f(x, y) dx dy = c \iint_A f(x, y) dx dy \quad (c \text{ constant}) \quad (4)$$

$$\iint_A f(x, y) dx dy = \iint_{A_1} f(x, y) dx dy + \iint_{A_2} f(x, y) dx dy \quad (5)$$

In (5) we assume that $A = A_1 \cup A_2$, $A_1 \cap A_2 = \emptyset$, and that $\iint_{A_1}$ and $\iint_{A_2}$ are defined.

Let f still be defined over A and suppose that there exist numbers m and M such that $m \leq f(x, y) \leq M$ for all (x, y) in A . Then there exists a number ξ in $[m, M]$ such that

$$\iint_A f(x, y) dx dy = \xi \cdot \iint_A dx dy = \xi \cdot \text{area}(A) \quad (6)$$

The number ξ is called the **average value** of f in the closed and bounded set A . If the set A is also *connected* (i.e. not the union of two disjoint closed sets) and f is continuous, then it can be shown that there exists a point $(\bar{x}, \bar{y})$ in A such that the number ξ given by (6) is equal to $f(\bar{x}, \bar{y})$. In this case (6) takes the form

$$\iint_A f(x, y) dx dy = f(\bar{x}, \bar{y}) \cdot \text{area}(A) \quad \text{for some } (\bar{x}, \bar{y}) \text{ in } A \quad (7)$$

This result is called the **mean value theorem for double integrals**.

This section has dealt with double integrals. It should be clear that the theory associated with (1) and (2) can be generalized to triple integrals and multiple integrals in general. One can also obtain formulas that correspond to (3)–(7), but for these we refer to the literature.

MS FOR SECTION 4.6

- SM 1. (a) Compute the following double integral by the method in Example 1 above:

$$\int_0^2 \left(\int_0^1 (2x - y + 1) dx \right) dy$$

- (b) Check the answer by integrating in the usual way.

7 Change of Variables

One of the most important methods of integration for single integrals is rule (4.1.4) for integration by substitution:

$$\int_a^b f(x) dx = \int_{u_1}^{u_2} f(g(u)) g'(u) du \quad (x = g(u), g(u_1) = a, g(u_2) = b) \quad (1)$$

It turns out that there is a similar rule for changing variables in multiple integrals. Let us look at this problem for double integrals first.

Change of Variables in Double Integrals

Consider the double integral $\iint_A f(x, y) dx dy$. Suppose we introduce new variables u and v together with functions h and g such that

$$x = g(u, v), \quad y = h(u, v) \quad (2)$$

With suitable restrictions on the integrands and the domains of integration, we claim that

$$\iint_A f(x, y) dx dy = \iint_{A'} f(g(u, v), h(u, v)) \left| \frac{\partial(g, h)}{\partial(u, v)} \right| du dv \quad (3)$$

Here we use the *absolute* value of the **Jacobian determinant**

$$\frac{\partial(g, h)}{\partial(u, v)} = \begin{vmatrix} \partial g / \partial u & \partial g / \partial v \\ \partial h / \partial u & \partial h / \partial v \end{vmatrix} \quad (4)$$

Also, A' is the set in the uv -plane “corresponding to” the given set A in the xy -plane. More precisely, $A' = \{(g(u, v), h(u, v)) : (u, v) \in A'\}$. Precise conditions for (3) to be true are stated in Theorem 4.7.1.

Comparing the two formulas (1) and (3), we see that introducing new variables causes two things to happen. First, the domain of integration is changed in each case. Second, a new factor appears under the integral sign. In (1) it is $g'(u)$, whereas in (3) it is the *absolute value of the Jacobian determinant* of the transformation (2).⁴

Let us see how formula (3) can be applied to a simple example.

EXAMPLE 1

Compute $I = \iint_A (x^2 + y^2 - 1) dx dy$ where A is the set in the xy -plane bounded by the lines $x + y = 1$, $x + y = 5$, $x - y = -1$, and $x - y = 1$. The set is shown in Fig. 1.

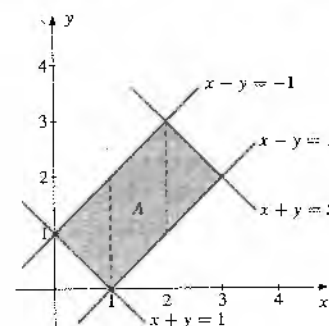


Figure 1

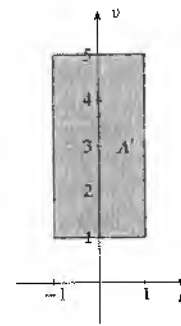


Figure 2

Solution: By subdividing the set A in a suitable way, one can compute the given integral. For example, if we use the vertical lines $x = 1$ and $x = 2$ to divide A into three parts (see Fig. 1), we can see that I is equal to

$$\int_0^1 \left(\int_{1-x}^{5-x} (x^2 + y^2 - 1) dy \right) dx + \int_1^2 \left(\int_{x-1}^{5-x} (x^2 + y^2 - 1) dy \right) dx + \int_2^4 \left(\int_{x-1}^{5-x} (x^2 + y^2 - 1) dy \right) dx$$

After a fair amount of calculation, we can find the value $I = 52/3$.

In this case, however, A is a rotated rectangle. This suggests that it might be easier to introduce $u = x - y$ and $v = x + y$ as new variables. Note that this transforms the boundary lines $x - y = -1$ and $x - y = 1$ of A into the straight lines $u = -1$ and $u = 1$, and the

⁴ Why $g'(u)$ rather than $|g'(u)|$ in (1)? Note that if $g'(u) < 0$ in (1), then $u_1 > u_2$. So if we let I denote the interval between the two endpoints u_1 and u_2 , then the right-hand side of (1) can be written as $\int_I f(g(u)) |g'(u)| du$, which is the obvious one-dimensional version of (3).

straight lines $x + y = 1$ and $x + y = 5$ into $v = 1$ and $v = 5$. Let A' be the rectangle in the uv -plane shown in Fig. 2. The transformation transforms the boundary lines of A in xy -plane into the boundary lines of A' in the uv -plane. Moreover, the interior of A is mapped in a one-to-one fashion onto the interior of A' . From $u = x - y$ and $v = x + y$ it follows that

$$x = \frac{1}{2}(u + v), \quad y = \frac{1}{2}(-u + v) \quad (i)$$

which, in a similar way, maps A' onto A . The transformation given by (i) corresponds to the transformation (2), and in this case the Jacobian is

$$\begin{vmatrix} \partial x / \partial u & \partial x / \partial v \\ \partial y / \partial u & \partial y / \partial v \end{vmatrix} = \begin{vmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{vmatrix} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

With x and y given by (i), the integrand becomes

$$x^2 + y^2 - 1 = \frac{1}{4}(u + v)^2 + \frac{1}{4}(-u + v)^2 - 1 = \frac{1}{2}u^2 + \frac{1}{2}v^2 - 1$$

Therefore, through simple calculations formula (3) yields

$$\begin{aligned} I &= \iint_A (x^2 + y^2 - 1) dx dy = \iint_{A'} \left(\frac{1}{2}u^2 + \frac{1}{2}v^2 - 1\right) \frac{1}{2} du dv \\ &= \frac{1}{2} \int_1^5 \left(\int_{-1}^1 \left(\frac{1}{2}u^2 + \frac{1}{2}v^2 - 1\right) du \right) dv = 52/3 \end{aligned}$$

Provided (3) is applicable in this case, it simplifies the computational work considerably. ■

Consider next the general double integral of $f(x, y)$ over some set A in the plane and assume that we introduce the new variables u and v as in (2).

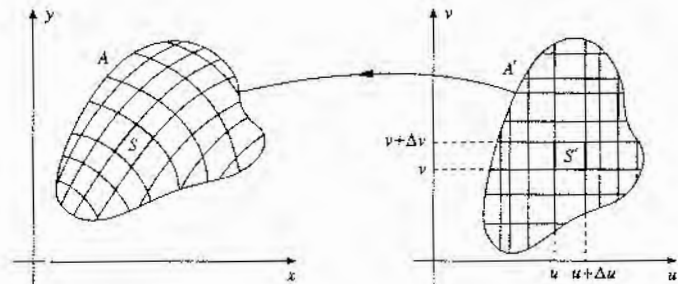


Figure 3 A curvilinear grid

We shall sketch an argument that can be extended into a proof of formula (3). Assume that g and h are C^1 functions that together map the set A' into A , and that each point (x, y) in A is the image of a unique point (u, v) in A' . The sets A and A' might be as indicated in Fig. 3.

The definition (4.6.2) of the double integral of f over A allows any sequence of subdivisions of A into subsets, provided the diameter of the largest subset converges to 0. We shall make use of this fact and employ a subdivision of A that is "induced" by the transformation (2) in the following way. Take a point in A with coordinates (x, y) . The unique point (u, v)

in A' that corresponds to (x, y) is given implicitly by (2). The numbers u and v are called the *curvilinear coordinates* of (x, y) w.r.t. the given transformation. Keep u fixed at $u = u_0$. A number of points in A will have curvilinear coordinates with this special value of u , namely those points (x, y) of A for which $x = g(u_0, v)$, $y = h(u_0, v)$. By choosing different fixed values of u , we obtain a family of curves in A . These curves cannot intersect because the correspondence between the points of A and A' is one-to-one. Similarly, by choosing different fixed values of v , we get another family of curves in A , characterized by the fact that along any particular curve, v has a fixed value. Some of the curves in the curvilinear grid obtained in this way are indicated in the set A in Fig. 3. Through the transformation (2), this curvilinear grid corresponds to the rectangular grid drawn in the set A' . If we "refine" the rectangular grid in A' , the curvilinear grid in A will also be "refined".

Consider next the rectangle S' indicated in A' in Fig. 3. Its area $\Delta S'$ is equal to $\Delta u \Delta v$, and the transformation (2) maps it to a curvilinear "rectangle" S in A . If we denote the area of S by ΔS , we obtain an approximation to the double integral by f over A as the sum

$$\sum f(x, y) \Delta S \quad (*)$$

where (x, y) is an arbitrary point in S , and we sum over all the curvilinear rectangles in A . We drop from the sum in $(*)$ those rectangles that have points in common with the boundary of A . The joint contribution to the sum in $(*)$ from all these boundary rectangles will tend to 0 as the subdivision is refined.

To proceed further we need another expression for the sum in $(*)$. First, let us find an approximate value of ΔS by using the fact that S is the image of the rectangle S' under the transformation (2). The relationship between S and S' is indicated in more detail in Fig. 4.

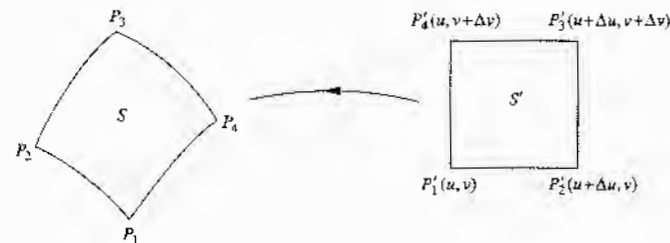


Figure 4

The point P_1 is the image of P'_1 under (2), so P_1 has coordinates $(g(u, v), h(u, v))$. The point P_2 is the image of P'_2 , so P_2 has coordinates $(g(u + \Delta u, v), h(u + \Delta u, v))$. In the same way the coordinates of P_3 and P_4 are $(g(u + \Delta u, v + \Delta v), h(u + \Delta u, v + \Delta v))$ and $(g(u, v + \Delta v), h(u, v + \Delta v))$, respectively. If Δu and Δv are small, we obtain good approximations to these coordinates by using Taylor's formula and including only first-order terms.

If the coordinates of P_i are (x_i, y_i) , $i = 1, \dots, 4$, we get

$$P_1 : x_1 = g(u, v), \quad y_1 = h(u, v)$$

$$P_2 : x_2 = g(u + \Delta u, v) \approx g(u, v) + \frac{\partial g}{\partial u} \Delta u, \quad y_2 = h(u + \Delta u, v) \approx h(u, v) + \frac{\partial h}{\partial u} \Delta u$$

$$P_3 : x_3 = g(u + \Delta u, v + \Delta v) \approx g(u, v) + \frac{\partial g}{\partial u} \Delta u + \frac{\partial g}{\partial v} \Delta v,$$

$$y_3 = h(u + \Delta u, v + \Delta v) \approx h(u, v) + \frac{\partial h}{\partial u} \Delta u + \frac{\partial h}{\partial v} \Delta v$$

$$P_4 : x_4 = g(u, v + \Delta v) \approx g(u, v) + \frac{\partial g}{\partial v} \Delta v, \quad y_4 = h(u, v + \Delta v) \approx h(u, v) + \frac{\partial h}{\partial v} \Delta v$$

For small values of Δu and Δv the curvilinear rectangle $P_1 P_2 P_3 P_4$ is approximately a parallelogram. Therefore, its area is approximately twice the area of the triangle $P_1 P_2 P_3$ with vertices (x_1, y_1) , (x_2, y_2) , (x_3, y_3) . Problem 2 asks you to show that this triangular area is half the absolute value of the following determinant:

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} = \begin{vmatrix} 1 & x_1 & y_1 \\ 0 & x_2 - x_1 & y_2 - y_1 \\ 0 & x_3 - x_1 & y_3 - y_1 \end{vmatrix} = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)$$

With the first-order approximations to the coordinates obtained above, the determinant is

$$\frac{\partial g}{\partial u} \Delta u \left(\frac{\partial h}{\partial u} \Delta u + \frac{\partial h}{\partial v} \Delta v \right) - \left(\frac{\partial g}{\partial u} \Delta u + \frac{\partial g}{\partial v} \Delta v \right) \frac{\partial h}{\partial u} \Delta u = \left(\frac{\partial g}{\partial u} \frac{\partial h}{\partial v} - \frac{\partial g}{\partial v} \frac{\partial h}{\partial u} \right) \Delta u \Delta v$$

Hence, ΔS is approximately equal to the absolute value of $\begin{vmatrix} \partial g/\partial u & \partial g/\partial v \\ \partial h/\partial u & \partial h/\partial v \end{vmatrix} \Delta u \Delta v$. The determinant here is the Jacobian of the transformation (2), so

$$\Delta S \approx \left| \frac{\partial(g, h)}{\partial(u, v)} \right| \Delta u \Delta v \quad (**)$$

It is reasonable to expect that this approximation will be better for smaller Δu and Δv .

We observed above that $\sum f(x, y) \Delta S$ is an approximation to the double integral of f over A when we sum over all the curvilinear rectangles in A that have no points in common with the boundary of A . By using (2) and (**), we therefore obtain

$$\sum f(x, y) \Delta S \approx \sum f(g(u, v), h(u, v)) \left| \frac{\partial(g, h)}{\partial(u, v)} \right| \Delta u \Delta v \quad (5)$$

There is a one-to-one correspondence between curvilinear rectangles S in A and rectangles S' in A' . It follows that the last sum is an approximation to the double integral of the function $f(g(u, v), h(u, v)) \left| \partial(g, h)/\partial(u, v) \right|$ over the set A' : if the subdivision of A' is refined in such a way that the diameter of the largest rectangle tends to 0, then passing to the limit in (5) gives (3).

Without going into the finer points of the proof, here is a precise result (see Protter and

THEOREM 4.7.1 (CHANGE OF VARIABLES IN DOUBLE INTEGRALS)

Suppose that

$$x = g(u, v), \quad y = h(u, v)$$

defines a one-to-one C^1 transformation from an open and bounded set A' in the uv -plane onto an open and bounded set A in the xy -plane, and assume that the Jacobian determinant $\partial(g, h)/\partial(u, v)$ is bounded on A' . Let f be a bounded and continuous function defined on A . Then

$$\iint_A f(x, y) dx dy = \iint_{A'} f(g(u, v), h(u, v)) \left| \frac{\partial(g, h)}{\partial(u, v)} \right| du dv$$

NOTE 1 A set in the plane is said to have *area* (or *measure*) 0 if it can be covered by a sequence of rectangles whose total area is arbitrarily small. (A set consisting of a finite number of points or a finite number of curves with finite lengths will have measure 0.) It turns out that we can always remove a subset of measure 0 from the domain of integration without affecting the value of the integral. *It is therefore sufficient if the conditions in the theorem are satisfied after suitable subsets of measure 0 are removed from A and A' .*

NOTE 2 The condition in the theorem that the transformation be one-to-one is sometimes difficult to check. Note that it is not sufficient to assume that the Jacobian determinant is different from 0 throughout A' . (See Example 2.7.3.)

Polar Coordinates

When a double integral is difficult to compute in the usual way, it is natural to search for a suitable substitution so that one can use Theorem 4.7.1. The choice of the new variables must take into account the form of the integrand as well as the domain of integration. One substitution that is often helpful is that of introducing *polar coordinates*. In that case we usually denote u and v by r and θ , and we define the transformation by

$$x = r \cos \theta, \quad y = r \sin \theta \quad (6)$$

The Jacobian is then equal to r (see Problem 2.7.8). If we assume that $r > 0$ and that θ lies in an interval of the form $[\theta_0, \theta_0 + 2\pi)$, then (6) defines a one-to-one C^1 transformation. (From Problem 2.7.8(b) we see that the transformation need not be one-to-one over an arbitrary set in the $r\theta$ -plane.) In this case (3) takes the form

$$\iint_A f(x, y) dx dy = \iint_{A'} f(r \cos \theta, r \sin \theta) r dr d\theta \quad (7)$$

Polar coordinates are particularly convenient if r or θ is constant along the boundary of the domain of integration, and/or when the integrand is particularly simple when expressed in polar coordinates. Consider the following illustrative example.

IE 2 Find $\iint_A \sqrt{x^2 + y^2} dx dy$, with $A = \{(x, y) : 4 \leq x^2 + y^2 \leq 9, \frac{1}{3}\sqrt{3}x \leq y \leq \sqrt{3}x\}$.

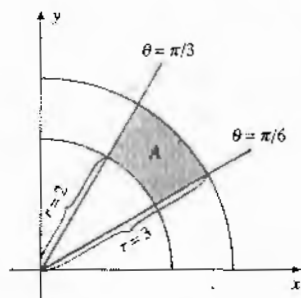


Figure 5

Solution: Figure 5 shows the set A , which is bounded by the circles $x^2 + y^2 = 4$ and $x^2 + y^2 = 9$ and the straight lines $y = \sqrt{3}x/3$ and $y = \sqrt{3}x$. The straight lines in question form angles $\pi/6$ and $\pi/3$ respectively with the x -axis. Hence, in polar coordinates A is determined by $2 \leq r \leq 3, \pi/6 \leq \theta \leq \pi/3$. The conditions in Theorem 4.7.1 are satisfied. Because $\sqrt{x^2 + y^2} = \sqrt{r^2 \cos^2 \theta + r^2 \sin^2 \theta} = r$, we get

$$\iint_A \sqrt{x^2 + y^2} dx dy = \int_{\pi/6}^{\pi/3} \left(\int_2^3 r \cdot r dr \right) d\theta = \int_{\pi/6}^{\pi/3} \left(\frac{r^3}{3} \right) d\theta = \frac{19\pi}{18}$$

To find this answer by direct integration would be unnecessarily cumbersome.

Change of Variables in Multiple Integrals

Theorem 4.7.1 can be generalized to n -dimensional integrals. We just state the result:

M 4.7.2 (CHANGE OF VARIABLES IN MULTIPLE INTEGRALS)

Suppose that $\mathbf{x} = (x_1, \dots, x_n) = \mathbf{g}(\mathbf{u}) = (g_1(\mathbf{u}), \dots, g_n(\mathbf{u}))$, where $\mathbf{u} = (u_1, \dots, u_n)$, defines a one-to-one C^1 transformation $\mathbf{g}$ from an open and bounded set A' in " $\mathbf{u}$ -space" onto an open and bounded set A in " $\mathbf{x}$ -space". Suppose that the Jacobian determinant

$$J = \frac{\partial(g_1, \dots, g_n)}{\partial(u_1, \dots, u_n)}$$

is bounded on A' . Let f be a bounded, continuous function defined on A . Then

$$\int \dots \int_A f(x_1, \dots, x_n) dx_1 \dots dx_n = \int \dots \int_{A'} f(g_1(\mathbf{u}), \dots, g_n(\mathbf{u})) |J| du_1 \dots du_n$$

Note 1 (appropriately generalized) applies equally well to the present theorem.

PROBLEMS FOR SECTION 4.7

1. Consider the double integral $I = \iint_A (x + xy) dx dy$ where A is the rectangle in the xy -plane with vertices at the points $(2,0)$, $(4,2)$, $(2,4)$ and $(0,2)$.

- Compute the integral directly.
- What integral do we obtain if we introduce new variables u and v , where $u = x - y$ and $v = x + y$? Compute its value.

SM 2. Show that the area of a triangle with vertices at (x_1, y_1) , (x_2, y_2) , and (x_3, y_3) in the xy -plane is given by half the absolute value of the determinant

$$\begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix}$$

(Hint: Draw the normals from each of the three points to the x -axis.)

SM 3. (a) Compute the following double integral by introducing polar coordinates:

$$\iint_A x^2 dx dy, \quad \text{where } A = \{(x, y) : x^2 + y^2 \leq 1/4\}$$

- What is the value of the double integral if $A = \{(x, y) : x^2 + (y - 1)^2 \leq 1/4\}$?

4. Consider the linear transformation $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by

$$x = au + bv, \quad y = cu + dv$$

- Find the Jacobian J of T . Show that T maps the unit square $[0, 1] \times [0, 1]$ in the uv -plane onto a parallelogram in the xy -plane whose area is $|J|$. (Make use of Problem 2.)
- If A' is an arbitrary bounded set in the uv -plane and the boundary of A' has measure 0, then one can prove that in general

$$\text{area}(T(A')) = |J| \cdot \text{area}(A')$$

Verify that this formula holds if T is given by (i) in Example 1 and A' is the set in Fig. 2. (See Note 1 for the definition of sets of measure 0, and Section 13.1 for the definition of the boundary of a set.)

SM 5. Let A_1 be the set $\{(x, y) : x^2 + y^2 \leq 1\}$ and let A_2 be the set in $\mathbb{R}^2$ bounded by the lines $y - 2x = -1$, $y - 2x = 1$, $y + 3x = 4$, and $y + 3x = 8$. Compute the following integrals by introducing suitable substitutions:

$$(a) \iint_{A_1} (1 - x^2 - y^2) dx dy \quad (b) \iint_{A_2} (x + y) dx dy$$

4.8 Generalized Double Integrals

So far, the treatment of multiple integrals in this chapter has dealt with integrals of bounded, continuous functions over bounded sets. We now consider briefly the problem of defining double integrals when the domain of integration is infinite and/or the integrand is unbounded.

We begin by considering a type of double integral frequently encountered in statistics,⁵

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du \quad (1)$$

The definition of this double integral is straightforward if we recall the standard way of defining integrals of one variable functions on unbounded intervals. We let $F(x, y) = \int_{-\infty}^x G(u, y) du$, where $G(u, y) = \int_{-\infty}^y f(u, v) dv$. The latter integral is, by definition, $\int_{-\infty}^y f(u, v) dv = \lim_{N \rightarrow \infty} \int_{-N}^y f(u, v) dv$, provided the limit exists. Then $F(x, y) = \int_{-\infty}^x G(u, y) du = \lim_{M \rightarrow \infty} \int_{-M}^x G(u, y) du$, provided this limit also exists.

EXAMPLE 1 Evaluate the integral (1) for $x \geq 0, y \geq 0$ if $f(u, v) = \frac{1}{4}e^{-|u|-|v|}$.

Solution: Since $e^{-|u|-|v|} = e^{-|u|}e^{-|v|}$, we get $F(x, y) = \frac{1}{4} \int_{-\infty}^x \int_{-\infty}^y e^{-|u|}e^{-|v|} dv du = \frac{1}{4} \int_{-\infty}^x e^{-|u|} du \int_{-\infty}^y e^{-|v|} dv$. Since $|u| = u$ if $u \geq 0$ and $|u| = -u$ if $u < 0$, we find that $\int_{-\infty}^x e^{-|u|} du = \int_{-\infty}^0 e^u du + \int_0^x e^{-u} du = \left[e^u \right]_{-\infty}^0 + \left[-e^{-u} \right]_0^x = 1 + (-e^{-x} + 1) = 2 - e^{-x}$. Similarly, $\int_{-\infty}^y e^{-|v|} dv = 2 - e^{-y}$, and so

$$F(x, y) = \frac{1}{4} \int_{-\infty}^x \int_{-\infty}^y e^{-|u|-|v|} dv du = \frac{1}{4} (2 - e^{-x})(2 - e^{-y})$$

Note that if $x \rightarrow \infty$ and $y \rightarrow \infty$, then $F(x, y) \rightarrow 1$.

In the last example it is natural to define $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(u, v) dv du = 1$. In general, if we integrate $f(x, y) \geq 0$ over $A = \mathbb{R}^2 = (-\infty, \infty) \times (-\infty, \infty)$, it turns out that we can define

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = \lim_{n \rightarrow \infty} \int \int_{A_n} f(x, y) dx dy \quad (2)$$

where $A_n = [-n, n] \times [-n, n]$. (Check that this definition gives the correct result in Example 1.) Problem 4 shows what can go wrong if we remove the assumption that $f(x, y) \geq 0$.

Consider more generally a bounded, continuous function f defined on an unbounded set A in the plane. We assume that $f(x, y) \geq 0$, since this simplifies matters somewhat. Our problem is to find a sensible definition of the double integral of $f(x, y)$ over A . Our point of departure is the fact that we already have defined the double integral of $f(x, y)$ over each closed, bounded subset of A . Let $A_1, A_2, \dots$ be an increasing sequence of closed, bounded subsets of A , so that

$$A_1 \subseteq A_2 \subseteq \dots \subseteq A_n \subseteq \dots \subseteq A$$

⁵ If the two random variables X and Y have a joint distribution determined by the probability density function $f(x, y)$, then $F(x, y)$ is the cumulative distribution function. For f to be a valid density

We shall say that A_n converges to A if for each closed, bounded subset A' of A there exists a number N such that for each $n \geq N$ we have $A' \subseteq A_n$. If this is the case, we define

$$\iint_A f(x, y) dx dy = \lim_{n \rightarrow \infty} \iint_{A_n} f(x, y) dx dy \quad (3)$$

if the limit exists. One can prove that if the limit exists for one such sequence $\{A_n\}$ converging to A , then it will also exist and have the same value for any other sequence of this type. One can try, in each case, to choose a convenient sequence of the required type. If the limit in (3) exists, we say that the double integral of f over A converges. If not, we say that it diverges. In order to show that the double integral of f over A diverges, it is obviously enough to find one such sequence $\{A_n\}$ as described above for which the limit in (3) does not exist.

EXAMPLE 2

Compute $\iint_A e^{-(x^2+y^2)} dx dy$ when A is the whole xy -plane ($A = \mathbb{R}^2$), (a) by using $A_n = \{(x, y) : x^2 + y^2 \leq n^2\}$; (b) by using $B_n = [-n, n] \times [-n, n]$. Use the results to prove the Poisson integral formula (4.3.3).

Solution: (a) The conditions for using (3) are satisfied and, using polar coordinates,

$$\begin{aligned} I_n &= \iint_{A_n} e^{-(x^2+y^2)} dx dy = \int_0^{2\pi} \left(\int_0^n e^{-r^2} r dr \right) d\theta \\ &= \int_0^{2\pi} \left(\left[-\frac{1}{2} e^{-r^2} \right]_0^n \right) d\theta = \frac{1}{2} (1 - e^{-n^2}) \int_0^{2\pi} d\theta = \pi (1 - e^{-n^2}) \xrightarrow{n \rightarrow \infty} \pi \end{aligned}$$

It follows that the given double integral is convergent, with value π .

(b) The integral over B_n is $J_n = \iint_{B_n} e^{-(x^2+y^2)} dx dy = \int_{-n}^n \left(\int_{-n}^n e^{-x^2} e^{-y^2} dx \right) dy$. Since the integrand is separable, $J_n = \left(\int_{-n}^n e^{-x^2} dx \right) \left(\int_{-n}^n e^{-y^2} dy \right) = \left(\int_{-n}^n e^{-x^2} dx \right)^2$. And since $A_n \subseteq B_n \subseteq A_{2n}$ for all n and $e^{-(x^2+y^2)} > 0$ everywhere, $\pi(1 - e^{-n^2}) \leq \left(\int_{-n}^n e^{-x^2} dx \right)^2 \leq \pi(1 - e^{-4n^2})$. Taking limits as $n \rightarrow \infty$, we get $\pi = \left(\int_{-\infty}^{\infty} e^{-x^2} dx \right)^2$, so $\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$. By symmetry, $\int_0^{\infty} e^{-x^2} dx = \frac{1}{2}\sqrt{\pi}$, the Poisson integral formula.

Unbounded Functions

This section concludes with an example showing how to extend the definition of double integrals to certain unbounded functions defined over bounded sets. The idea resembles the one associated with the definition (3). For more details, see Protter and Morrey (1991).

EXAMPLE 3

The function $f(x, y) = (x^2 + y^2)^{-p}$, $p > 0$, is not bounded over the set A determined by $0 < x^2 + y^2 \leq 1$, because $f(x, y) \rightarrow \infty$ as $(x, y) \rightarrow (0, 0)$. The double integral

$$\iint_A \frac{1}{(x^2 + y^2)^p} dx dy$$

is therefore so far not defined. For $n = 1, 2, \dots$, let A_n be the circular ring (or annulus) defined by $1/n^2 < x^2 + y^2 \leq 1$. Then $A_1, A_2, \dots$ form an increasing sequence of

sets $A_1 \subseteq A_2 \subseteq \cdots \subseteq A_n \subseteq \cdots$, and, as n increases, A_n will “tend to” the set $A = \{(x, y) : 0 < x^2 + y^2 \leq 1\}$, because $\bigcup_{n=1}^{\infty} A_n = A$. The double integral of f is defined over the set A_n for each $n = 1, 2, \dots$. Indeed, by introducing polar coordinates, we obtain

$$I_n = \iint_{A_n} \frac{1}{(x^2 + y^2)^p} dx dy = \int_0^{2\pi} \left(\int_{1/n}^1 \frac{1}{r^{2p}} r dr \right) d\theta = \int_0^{2\pi} \left(\int_{1/n}^1 r^{1-2p} dr \right) d\theta$$

It follows that $I_n = 2\pi \ln n$ for $p = 1$, while $I_n = \pi(1 - n^{2(p-1)})/(1-p)$ for $p \neq 1$. If $p < 1$, then $n^{2(p-1)} \rightarrow 0$ as $n \rightarrow \infty$, and $I_n \rightarrow \pi/(1-p)$. If $p \geq 1$, we see that I_n does not tend to any limit. On the basis of this observation, if $0 < p < 1$, we say that the integral is convergent, with a value $\pi/(1-p)$, whereas we say that it is divergent if $p \geq 1$. $\square$

MS FOR SECTION 4.8

- 5M 1. (a) Compute the value of the double integral $\iint_{x^2+y^2 \leq 1} \frac{1}{(x^2 + y^2)^2} dx dy$.
 (b) Discuss the convergence (for different values of p) if we replace the integrand in (a) with $(x^2 + y^2)^{-p}$, but keep the same domain of integration.
2. If $I(z) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y) dy \right) dx$ then, under appropriate conditions on $f(x, y)$, compute $I'(z)$ by using Leibniz's rule.
3. (a) Let $f(x, y) = k\sqrt{1 - x^2 - y^2}$. Find a value of k such that $\iint_{x^2+y^2 \leq 1} f(x, y) dx dy = 1$. (Then $f(x, y)$ is a joint density function for two stochastic variables X and Y .)
 (b) With the value of k from part (a), find the marginal density of X , which is defined as $f_X(x) = \int_{x^2+y^2 \leq 1} f(x, y) dy$.
- 5M 4. Let $I(b, d)$ denote the double integral in Example 4.4.3. Find the two limits

$$\lim_{b \rightarrow \infty} [\lim_{d \rightarrow \infty} I(b, d)] \quad \text{and} \quad \lim_{d \rightarrow \infty} [\lim_{b \rightarrow \infty} I(b, d)]$$

$$\text{What do your answers tell you about } \int_1^{\infty} \int_1^{\infty} (y-x)(y+x)^{-3} dx dy?$$

5. Prove that if the function F is defined by (1), then under appropriate conditions, $F''_{12}(x, y) = f(x, y)$. (Hint: Use Leibniz's rule.) Verify this equality for the function in Example 1.

HARDER PROBLEMS

$$5M 6. \text{ Compute } (a) \iint_{\mathbb{R}^2} (x^2 + y^2 + 1)^{-3/2} dx dy \quad (b) \iint_{\mathbb{R}^2} \frac{e^{-(x-y)^2}}{1 + (x+y)^2} dx dy$$

- 5M 7. Check whether the following double integrals converge and, if they do, find their values.

$$(a) \iint_{0 < x^2 + y^2 \leq 1} \frac{x^2}{(x^2 + y^2)^{3/2}} dx dy \quad (b) \iint_A \frac{-\ln(x^2 + y^2)}{\sqrt{x^2 + y^2}} dx dy$$

In (b) let $A = \{(x, y) : 0 < x^2 + y^2 \leq 1, x \geq 0, y \geq 0\}$.

5

DIFFERENTIAL EQUATIONS I:

FIRST-ORDER EQUATIONS IN ONE VARIABLE

... the task of the theory of ordinary differential equations is to reconstruct the past and predict the future of the process from a knowledge of this local law of evolution.

—V. I. Arnol'd (1973)

Economists often study the changes over time in economic variables like national income, the interest rate, the money supply, oil production, or the price of wheat. The laws of motion governing these variables are usually expressed in terms of one or more equations.

If time is regarded as continuous and the equations involve unknown functions and their derivatives, we find ourselves considering *differential equations*. In macroeconomic theory especially, but also in many other areas of economics, a certain knowledge of differential equations is essential. Another example is finance theory, where the pricing of options now requires quite advanced methods in the theory of differential equations.

The systematic study of differential equations was initiated by Newton and Leibniz in the seventeenth century, and this topic is still one of the most important in mathematics.

After the introductory Section 5.1, the short Section 5.2 merely points out how to draw a direction diagram, and how solving a differential equation is equivalent to finding a curve whose tangent at each point is given by the direction diagram.

Section 5.3 gives a systematic discussion of separable differential equations, i.e. equations of the form $\dot{x} = f(t)g(x)$.

Section 5.4 concentrates on the special properties of first-order linear differential equations, first with constant and then with variable coefficients. Several economic examples are studied.

Section 5.5 deals with exact equations and integrating factors. Exact equations appear less frequently in economics and this section is therefore in small print.

Although only a few types of differential equations have solutions given by explicit formulas, Section 5.6 shows how a clever choice of new variables can sometimes help with seemingly insoluble equations.

Section 5.7 considers what qualitative properties of a solution can be inferred, even if the equation cannot be solved analytically.

Finally, Section 5.8 is concerned with existence and uniqueness theorems for first-order equations.

5.1 Introduction

What is a differential equation? As the name suggests, it is an equation. Unlike ordinary algebraic equations, in a differential equation:

- (A) The unknown is a function, not a number.
- (B) The equation includes one or more of the derivatives of the function.

An *ordinary* differential equation is one for which the unknown is a function of only one variable. *Partial differential equations* are equations where the unknown is a function of two or more variables, and one or more of the partial derivatives of the function are included.

In this chapter we restrict attention to first-order (ordinary) differential equations—that is, equations where only the first-order derivatives of the unknown functions of one variable are included. Three typical examples are:

$$\dot{x} = ax, \quad \dot{x} + ax = b, \quad \dot{x} + ax = bx^2$$

With suitably chosen constants, these describe natural growth, growth towards a limit, and logistic growth, respectively. (Recall that we often use dot notation for the derivative, $\dot{x} = dx/dt$, especially when the independent variable is time t .) Other examples of first-order differential equations are

$$(a) \dot{x} = x + t \quad (b) \dot{K} = \alpha\sigma K + H_0 e^{\mu t} \quad (c) \dot{k} = sf(k) - \lambda k$$

In Examples 5.4.3 and 5.7.3, respectively, we shall give equations (b) and (c) interesting economic interpretations, both concerning the evolution of an economy's capital stock.

Solving equation (a), for instance, means finding all functions $x(t)$ such that, for every value of t , the derivative $\dot{x}(t)$ of $x(t)$ is equal to $x(t) + t$. In equation (b), $K(t)$ is the unknown function, whereas α , σ , H_0 , and μ are constants. In equation (c), $f(k)$ is a given function, whereas s and λ are constants. The unknown function is $k = k(t)$.

NOTE 1 We often use t to denote the independent variable. This is because most differential equations that appear in economics have time as their independent variable. The following theory is valid even if the independent variable is not time, however.

A first-order differential equation is written

$$\dot{x} = F(t, x) \quad (1)$$

where F is a given function of two variables and $x = x(t)$ is the unknown function. A **solution** of (1) in an interval I of the real line is any differentiable function φ defined on I such that $x = \varphi(t)$ satisfies (1), that is $\dot{\varphi}(t) = F(t, \varphi(t))$ for all t in I .¹ The graph of a solution is called a **solution curve** or an **integral curve**.

The equations (a), (b), and (c) are all of the form (1). For example, (a) becomes $dx/dt = F(t, x)$ with $F(t, x) = x + t$.

¹ Usually we assume that the interval I is open, but sometimes it is useful to allow closed (or half-open) intervals. If I is a closed interval, a solution is required to be continuous on I and to satisfy

EXAMPLE 1 Consider the differential equation

$$\dot{x} = x + t \quad (*)$$

- (a) Show that both $x = -t - 1$ and $x = e^t - t - 1$ are particular solutions of the equation over the entire real line.
- (b) More generally, show that $x = Ce^t - t - 1$ is a solution of (*) for all t , whatever the choice of the constant C .
- (c) Show that $x = e^t - 1$ is not a solution of (*).

Solution:

- (a) If $x = -t - 1$, then $\dot{x} = -1$ and $x + t = (-t - 1) + t = -1$. Hence, $\dot{x} = x + t$ for all t in this case. If $x = e^t - t - 1$, then $\dot{x} = e^t - 1$ and $x + t = (e^t - t - 1) + t = e^t - 1$. Again we see that (*) is satisfied for all t .
- (b) When $x = Ce^t - t - 1$, we have $\dot{x} = Ce^t - 1 = x + t$ for all t .
- (c) If $x = e^t - 1$, then $\dot{x} = e^t$ and $x + t = e^t + t - 1$. In this case, $\dot{x}$ is only equal to $x + t$ for $t = 1$, so $x = e^t - 1$ is *not* a solution of equation (*) on any interval. **|**

Example 1 illustrates the fact that a differential equation usually has infinitely many solutions. We found that $x = Ce^t - t - 1$ was a solution of $\dot{x} = x + t$ for each choice of the constant C . The answer to Problem 5.4.3 shows that no other function satisfies the equation.

The set of all solutions of a differential equation is called its **general solution**, while any specific function that satisfies the equation is called a **particular solution**.

A first-order differential equation usually has a general solution that depends on *one* constant. (Problem 5 shows why we must use the word “usually” in this statement.) If we require the solution to pass through a given point in the tx -plane, then the constant is determined uniquely, except in special cases.

EXAMPLE 2 Assuming that the general solution is $x(t) = Ce^t - t - 1$, find the solution of $\dot{x} = x + t$ that passes through the point $(t, x) = (0, 1)$.

Solution: To make the solution $x(t) = Ce^t - t - 1$ pass through $(t, x) = (0, 1)$, we must have $x(0) = 1$. Hence, $1 = Ce^0 - 0 - 1$, implying that $C = 2$. The required solution, therefore, is $x(t) = 2e^t - t - 1$. **|**

The problem in Example 2 is this: Find the unique function $x(t)$ such that

$$\dot{x}(t) = x(t) + t \quad \text{and} \quad x(0) = 1 \quad (*)$$

If $t = 0$ denotes the initial time, then $x(0) = 1$ is called an **initial condition** and we call (*) an **initial-value problem**.

Such initial-value problems arise naturally in many economic models. For instance, suppose an economic growth model involves a first-order differential equation for the accumulation of capital over time. The initial stock of capital is historically given, and therefore helps to determine the unique solution of the equation.

Qualitative Theory

When the theory of differential equations was first developed, mathematicians primarily tried to find explicit solutions for some special types of equation. It became increasingly obvious, however, that only very few equations could be solved this way. In many cases, moreover, explicit formulas for the solutions are not really needed. Instead, the main interest is in a few important properties of the solution. As a result, the theory of differential equations includes many results concerning the general behaviour of the solutions. This is the so-called *qualitative theory*. Its main results include existence and uniqueness theorems, sensitivity analysis, and investigations of the stability of equilibria. Such topics are of both theoretical interest and practical importance, and will be discussed in some detail.

Along with this qualitative theory, much work has been put into developing useful numerical methods for finding approximate solutions of differential equations. Computers are playing an increasingly important role here, but these developments are not discussed here.

PROBLEMS FOR SECTION 5.1

1. Show that $x(t) = Ce^{-t} + \frac{1}{2}e^t$ is a solution of the differential equation $\dot{x}(t) + x(t) = e^t$ for all values of the constant C .
2. Show that $x = Ct^2$ is a solution of the differential equation $t\dot{x} = 2x$ for all choices of the constant C . Find in particular the integral curve through $(1, 2)$.
3. Show that any function $x = x(t)$ that satisfies the equation $xe^{tx} = C$ is a solution of the differential equation $(1 + tx)\dot{x} = -x^2$. (Hint: Differentiate $xe^{tx} = C$ implicitly w.r.t. t .)
4. In each of the following cases, show that any function $x = x(t)$ that satisfies the equation on the left is a solution of the corresponding differential equation on the right.
 - (a) $x^2 = 2at$, $2x\dot{x} = 2t\dot{x}^2 + a$ (a is a constant)
 - (b) $\frac{1}{2}e^{t^2} + e^{-t}(x+1) + C = 0$, $x\dot{x} = te^{t^2+x}$
 - (c) $(1-t)x^2 = t^3$, $2t^3\dot{x} = x(x^2 + 3t^2)$
5. Show that $x = Ct - C^2$ is a solution of the differential equation $\dot{x}^2 = tx - x$, for all values of the constant C . Then show that it is not the general solution because $x = \frac{1}{4}t^2$ is also a solution.

HARDER PROBLEMS

6. The function $x = x(t)$ satisfies $x(0) = 0$ and the differential equation $\dot{x} = (1+x^2)t$ for all t in an open interval I around 0. Prove that $t = 0$ is a global minimum point for $x(t)$ in I , and that the function x is convex on I . (Hint: You do not have to solve the equation.)

5.2 The Direction is Given: Find the Path!

Consider again the differential equation $\dot{x} = x + t$, which was studied in Examples 5.1.1 and 5.1.2. If $x = x(t)$ is a solution, then the slope of the tangent to the graph (or integral curve) at the point (t, x) is equal to $x + t$. At the point $(t, x) = (0, 0)$ the slope is therefore equal to 0, whereas at $(1, 2)$ the slope is 3, and so on. In Fig. 1, we have drawn small straight-line segments with slopes $x + t$ through several points in the tx -plane. This gives us a so-called **direction diagram** (or **slope field**) for the differential equation $\dot{x} = x + t$. If an integral curve passes through one of these points, it will have the corresponding line segment as its tangent. This allows us to sketch curves that follow the direction of the line segments, and get a general impression of what the integral curves of $\dot{x} = x + t$ must look like.

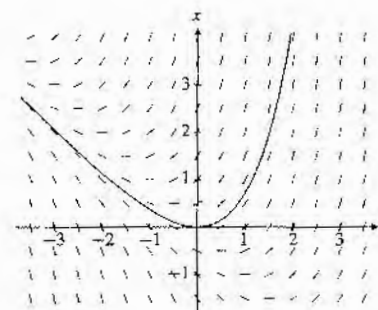


Figure 1 A direction diagram for $\dot{x} = x + t$. The integral curve through $(0, 0)$ is shown.

A direction diagram like this can be drawn for any differential equation of the form $\dot{x} = F(t, x)$. (Computer programs like *Maple* and *Mathematica* enable us to draw direction diagrams and solution curves with ease.) Whether or not it is possible to solve the equation explicitly, a direction diagram can give a rough but useful indication of how the integral curves behave. In a nutshell, the problem of solving the differential equation $\dot{x} = F(t, x)$ can be put like this: the direction is given, find the path!

PROBLEMS FOR SECTION 5.2

1. Draw a direction diagram for the differential equation $\dot{x} = x/t$ and draw some integral curves.
2. Draw a direction diagram for the differential equation $\dot{x} = -t/x$ and draw the integral curve through $(0, 2)$.

5.3 Separable Equations

Suppose that $\dot{x} = F(t, x)$, where $F(t, x)$ can be written as a product of two functions, one of which depends only on t and the other only on x . Specifically, suppose that

$$\dot{x} = f(t)g(x) \tag{1}$$

We say that this differential equation is **separable**. For instance, $\dot{x} = -2tx^2$ is obviously separable, whereas $\dot{x} = t^2 + x$ is not. (Problem 7 offers practice in deciding if a given equation is separable or not. Since separable equations are among those that can be solved in terms of integrals of known functions, it is useful to learn to distinguish between separable and nonseparable equations.)

A particular solution of (1) arises if $g(x)$ has a zero at $x = a$, so that $g(a) = 0$. In this case $x(t) \equiv a$ will be a solution of the equation, because the right- and left-hand sides are both 0 for all t . For instance, $\dot{x} = (x + 1)(x - 3)$ has the two particular solutions $x(t) \equiv -1$ and $x(t) \equiv 3$. (In addition $x = -1 + 4/(1 - Ce^{4t})$ is a solution for all values of the constant C . See Example 4 with $B = 1$, $a = -1$, and $b = 3$.)

Using differential notation, a general method for solving (1) can be expressed as follows:

GUIDE FOR SOLVING SEPARABLE DIFFERENTIAL EQUATIONS:

(A) Write equation (1) as

$$\frac{dx}{dt} = f(t)g(x) \tag{*}$$

(B) Separate the variables:

$$\frac{dx}{g(x)} = f(t) dt$$

(C) Integrate each side:

$$\int \frac{dx}{g(x)} = \int f(t) dt$$

(D) Evaluate the two integrals (if possible) to obtain a solution of (*) (possibly in implicit form). Solve for x , if possible.

(E) In addition, every zero $x = a$ of $g(x)$ gives the constant solution $x(t) \equiv a$.

To justify the method, suppose that $x = \varphi(t)$ is a function defined in an interval I such that $g(\varphi(t)) \neq 0$ throughout I . Then $x = \varphi(t)$ will solve (1) iff

$$\frac{\dot{\varphi}(t)}{g(\varphi(t))} = f(t)$$

for all t in I . But these two functions are equal in I iff

$$\int \frac{\dot{\varphi}(t)}{g(\varphi(t))} dt = \int f(t) dt$$

Suppose we substitute $x = \varphi(t)$, so that $dx = \dot{\varphi}(t) dt$ in the integral on the left-hand side. Then according to the rule of integration by substitution, the last equation is equivalent to

$$\int \frac{dx}{g(x)} = \int f(t) dt$$

Thus, $G(x) = F(t) + C$, where $G'(x) = 1/g(x)$, $F'(t) = f(t)$, and C is a constant.

NOTE 1 Suppose the function $G(x)$ is defined on an interval I where either $g(x) > 0$ everywhere, or $g(x) < 0$ everywhere. If $G(I) = \mathbb{R}$, then a solution $x(t)$ exists for all $t \in \mathbb{R}$, with values in I . But if $G(I)$ is a proper subset of $\mathbb{R}$, then $x(t) = G^{-1}(F(t) + C)$ is a solution only for a restricted range of values of C and t .

EXAMPLE 1 Solve the differential equation

$$\frac{dx}{dt} = -2tx^2$$

and find the integral curve that passes through $(t, x) = (1, -1)$.

Solution: We observe first that $x(t) \equiv 0$ is one (trivial) solution. But this does not go through $(1, -1)$, so we follow the recipe:

Separate:

$$-\frac{dx}{x^2} = 2t dt$$

Integrate:

$$-\int \frac{dx}{x^2} = \int 2t dt$$

Evaluate:

$$\frac{1}{x} = t^2 + C$$

It follows that the general solution is

$$x = \frac{1}{t^2 + C} \tag{*}$$

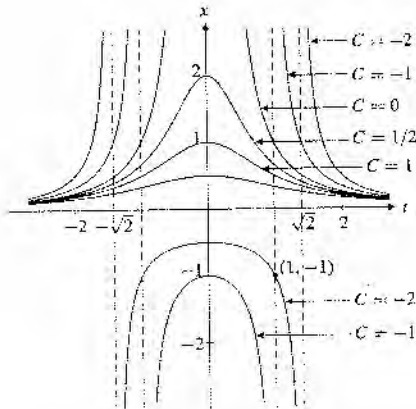


Figure 1 The solution curves $x = 1/(t^2 + C)$ for particular values of C .

To find the integral curve through $(1, -1)$, we must determine the correct value of C . Because we require $x = -1$ for $t = 1$, it follows from (*) that $-1 = 1/(1+C)$, so $C = -2$. Thus, the integral curve passing through $(1, -1)$ is $x = 1/(t^2 - 2)$.

Figure 1 shows integral curves of the form (*) for five different values of C . The constant of integration crucially affects the shape of the curve as well as its position. (Note that when $C \leq 0$, the solution is not defined over the entire real line.)

E 2 Solve the differential equation $\frac{dx}{dt} = \frac{t^3}{x^6 + 1}$.

Solution: We use the previous method, with $f(t) = t^3$ and $g(x) = 1/(x^6 + 1)$. Because $g(x)$ is never 0, there are no constant solutions. We proceed as follows:

Separate: $(x^6 + 1) dx = t^3 dt$

Integrate: $\int (x^6 + 1) dx = \int t^3 dt$

Evaluate: $\frac{1}{7}x^7 + x = \frac{1}{4}t^4 + C$

The desired functions $x = x(t)$ are those that satisfy the last equation for all t .

NOTE 2 We usually say that we have solved a differential equation even if the unknown function (as shown in Example 2) cannot be expressed explicitly. The important point is that we have found an equation involving the unknown function where the derivative of that function does not appear.

E 3 (**Economic growth**) Let $X = X(t)$ denote the national product, $K = K(t)$ the capital stock, and $L = L(t)$ the number of workers in a country at time t . Suppose that, for all $t \geq 0$,

$$(a) \quad X = AK^{1-\alpha}L^\alpha \quad (b) \quad \dot{K} = sX \quad (c) \quad L = L_0 e^{\lambda t}$$

where A, α, s, L_0 , and λ are all positive constants, with $\alpha < 1$. Derive from these equations a single differential equation to determine $K = K(t)$, and find the solution when $K(0) = K_0 > 0$. (This is a special case of the Solow–Swan model discussed in Example 5.7.3. In (a) we have a Cobb–Douglas production function, (b) says that aggregate investment is proportional to output, whereas (c) implies that the labour force grows exponentially.)

Solution: From the equations (a)–(c), we derive the single differential equation

$$\dot{K} = \frac{dK}{dt} = sAK^{1-\alpha}L^\alpha = sAL_0^\alpha e^{\alpha\lambda t} K^{1-\alpha}$$

This is clearly separable. Using the recipe yields:

$$K^{\alpha-1} dK = sAL_0^\alpha e^{\alpha\lambda t} dt, \quad \int K^{\alpha-1} dK = \int sAL_0^\alpha e^{\alpha\lambda t} dt, \quad \frac{1}{\alpha} K^\alpha = \frac{1}{\alpha\lambda} sAL_0^\alpha e^{\alpha\lambda t} + C$$

If we put $C_1 = \alpha C$, we get $K^\alpha = (sA/\lambda)L_0^\alpha e^{\alpha\lambda t} + C_1$. If $K = K_0$ for $t = 0$, we get $C_1 = K_0^\alpha - (sA/\lambda)L_0^\alpha$. Therefore the solution is

$$K = \left[K_0^\alpha + (sA/\lambda)L_0^\alpha (e^{\alpha\lambda t} - 1) \right]^{1/\alpha}$$

See Problem 9 for a closer examination of this model.

EXAMPLE 4 Solve the following differential equation when $a \neq b$:

$$\frac{dx}{dt} = B(x-a)(x-b)$$

In particular, find the solution when $B = -1/2$, $a = -1$, and $b = 2$, and draw some integral curves in this case.

Solution: Observe that both $x = a$ and $x = b$ are trivial solutions of the equation. In order to find the other solutions, separate the variables as follows. First, put all terms involving x on the left-hand side, and all terms involving t on the right-hand side. Then integrate, to get

$$\int \frac{1}{(x-a)(x-b)} dx = \int B dt$$

The next step is to transform the integrand on the left. We find that

$$\frac{1}{(x-a)(x-b)} = \frac{1}{b-a} \left(\frac{1}{x-b} - \frac{1}{x-a} \right)$$

(Verify this by expanding the right-hand side.) Hence,

$$\int \frac{1}{(x-a)(x-b)} dx = \frac{1}{b-a} \left(\int \frac{1}{x-b} dx - \int \frac{1}{x-a} dx \right)$$

Except for an additive constant, the last expression equals

$$\frac{1}{b-a} (\ln|x-b| - \ln|x-a|) = \frac{1}{b-a} \ln \frac{|x-b|}{|x-a|}$$

So, for some constant C_1 , the solution is

$$\frac{1}{b-a} \ln \frac{|x-b|}{|x-a|} = Bt + C_1 \quad \text{or} \quad \ln \frac{|x-b|}{|x-a|} = B(b-a)t + C_2$$

with $C_2 = C_1(b-a)$. So

$$\left| \frac{x-b}{x-a} \right| = e^{B(b-a)t + C_2} = e^{B(b-a)t} e^{C_2} \quad \text{or} \quad \frac{x-b}{x-a} = \pm e^{C_2} e^{B(b-a)t} = C e^{B(b-a)t} \quad (*)$$

after defining the new constant $C = \pm e^{C_2}$. Solving this last equation for x finally gives

$$\frac{dx}{dt} = B(x-a)(x-b) \iff x = \frac{b - aCe^{B(b-a)t}}{1 - Ce^{B(b-a)t}} = a + \frac{b-a}{1 - Ce^{B(b-a)t}} \quad (2)$$

For $B = -1/2$, $a = -1$, and $b = 2$, the differential equation is $\dot{x} = -\frac{1}{2}(x+1)(x-2)$. Note that $\dot{x}$ is positive for x between -1 and 2 . Hence, the integral curves rise with t in the horizontal strip between lines $x = -1$ and $x = 2$. In the same way, we can see directly from the differential equation that the integral curves are falling above and below this strip. In addition to the constant solutions $x = -1$ and $x = 2$, indicated by dashed horizontal lines in Fig. 2, we see that the general solution of the equation $\dot{x} = -\frac{1}{2}(x+1)(x-2)$ is

$$x = -1 + \frac{3}{1 - Ce^{-3t/2}}$$

Some of the associated integral curves are shown in Fig. 2.

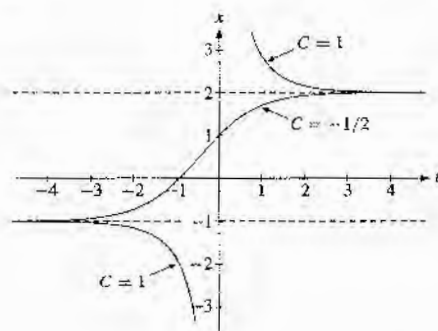


Figure 2 Some solution curves for $\dot{x} = -\frac{1}{2}(x+1)(x-2)$.

NOTE 3 In the second part of (*) we removed the absolute value sign around the fraction $(x-b)/(x-a)$, and replaced the factor e^{C_2} by $\pm e^{C_2}$, where we use $+$ if the fraction is positive, $-$ if it is negative. We claim that the sign must be the same for all t in the interval where the solution $x = x(t)$ is defined: this is because x must be continuous, and (*) was derived using the assumption that x differs from a and b everywhere. Therefore, $(x-b)/(x-a)$ is continuous and different from 0 in the whole domain of x . It follows that the fraction has the same sign everywhere—if not, the intermediate value theorem would imply that the fraction must be zero for some t . Hence the factor $\pm e^{C_2}$ has the same value for all relevant values of t , i.e. it is a constant factor (denoted by C in (*)).

LE 5 (Compound interest) Suppose that $w = w(t) > 0$ is the wealth held in an investment account at time t , and that $r(t)$ is the interest rate, with interest compounded continuously. Then

$$\dot{w} = r(t)w \quad (*)$$

which is a separable equation. Separating the variables and integrating yields

$$\int \frac{dw}{w} = \int r(t) dt$$

Therefore, $\ln w = R(t) + C_1$ where $R(t) = \int r(t) dt$. So the solution is

$$w(t) = e^{R(t)+C_1} = e^{C_1} e^{R(t)} = C e^{R(t)} \quad (**)$$

after introducing the new constant $C = e^{C_1}$. Suppose the initial value of the account is $w(0)$. Then (**) implies that $w(0) = C e^{R(0)}$, so $C = w(0) e^{-R(0)}$ and (**) becomes $w(t) = w(0) e^{R(t)-R(0)}$. But $R(t) - R(0) = \int_0^t r(s) ds$, and so

$$w(t) = w(0) e^{\int_0^t r(s) ds} = w(0) \exp \int_0^t r(s) ds$$

is the unique solution of (*) with $w(0)$ or the value of the account at time $t = 0$.

PROBLEMS FOR SECTION 5.3

1. Solve the equation $x^2 \dot{x} = t + 1$. Find the integral curve through $(t, x) = (1, 1)$.

2. Solve the following differential equations:

(a) $\dot{x} = t^3 - t$ (b) $\dot{x} = te^t - t$ (c) $e^x \dot{x} = t + 1$

SM 3. Find the general solutions of the following differential equations. Also find the integral curves through the indicated points.

(a) $t\dot{x} = x(1-t)$, $(t_0, x_0) = (1, 1/e)$ (b) $(1+t^3)\dot{x} = t^2x$, $(t_0, x_0) = (0, 2)$

(c) $x\dot{x} = t$, $(t_0, x_0) = (\sqrt{2}, 1)$ (d) $e^{2x}\dot{x} - x^2 - 2x = 1$, $(t_0, x_0) = (0, 0)$

4. Find the general solution of $\dot{x} + a(t)x = 0$. In particular, when $a(t) = a + bc^t$ (a, b , and c are positive; $c \neq 1$) show that the solution of the equation can be written in the form $x = Cp^t q^{c^t}$, where p and q are constants determined by a, b , and c , whereas C is an arbitrary constant. (This is Gompertz–Makeham's law of mortality.)

5. Explain why biological populations that develop as suggested in the figures A and B below cannot be described by differential equations of the form $\dot{N}/N = f(N)$, no matter how the function f is chosen. ($N(t)$ is the size of the population at time t .)

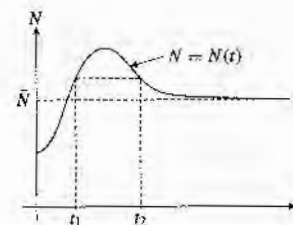


Figure A

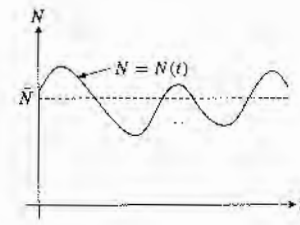


Figure B

SM 6. Find $x = x(t)$ when $\text{El}_t x = t\dot{x}/x$, the elasticity of $x(t)$ w.r.t. t , satisfies the following equations for all $t > 0$:

(a) $\text{El}_t x = a$ (b) $\text{El}_t x = at + b$ (c) $\text{El}_t x = ax + b$

7. Decide which of the following differential equations are separable:

(a) $\dot{x} = x^2 - 1$ (b) $\dot{x} = xt + t$ (c) $\dot{x} = xt + t^2$
 (d) $x\dot{x} = e^{x/t}\sqrt{1+t^2}$ (e) $\dot{x} = \sqrt{t^2+x}$ (f) $\dot{x} = F(t) + G(x)$

SM 8. The following differential equations have been studied in economics. Solve them.

(a) $\dot{K} = A\alpha_0 a^b K^{b-\alpha} e^{(\alpha v + \alpha)t}$, $b - c \neq 1$, $\alpha v + \alpha \neq 0$

(b) $\dot{x} = \frac{(\beta - \alpha x)(x - a)}{x}$, $\alpha > 0$, $\beta > 0$, $a > 0$, $\alpha a \neq \beta$

(Hint: $\frac{x}{\beta - \alpha x} = \frac{1}{\beta - \alpha a} \left(\frac{\beta}{1 - \frac{\alpha x}{a}} + \frac{\alpha}{1 - \frac{\alpha x}{a}} \right)$.)

HARDER PROBLEMS

9. (a) With reference to Example 3, show that K/L tends to $(sA/\lambda)^{1/\alpha}$ as $t \rightarrow \infty$. Compute the limit for X/L as $t \rightarrow \infty$.
- (b) Replace equation (c) in Example 3 by (c') $L = b(t + \alpha)^p$, where a , b , and p are positive constants. From (a), (b), and (c'), derive a differential equation for $K = K(t)$. Solve the equation when $K(0) = K_0$, and examine the behaviour of K/L as $t \rightarrow \infty$.
10. In connection with their study of CES (constant elasticity of substitution) production functions, Arrow, Chenery, Minhas, and Solow were led to consider the differential equation

$$\frac{dy}{dx} = \frac{y(1 - \alpha y^q)}{x} \quad (\alpha \text{ and } q \text{ are constants, } q \neq 0, x > 0, y > 0) \quad (*)$$

Use the identity $1/y + \alpha y^{q-1}/(1 - \alpha y^q) = 1/y(1 - \alpha y^q)$ to show that the general solution is

$$y = (\beta x^{-q} + \alpha)^{-1/q} \quad (**)$$

(Suppose we let $x = K/L$, $y = Y/L$, and define new constants A and a by $A = (\alpha + \beta)^{-1/q}$ and $a = \beta/(\alpha + \beta)$. Then $1 - a = \alpha/(\alpha + \beta)$ and $\alpha + \beta = A^{-q}$, so $\alpha = (1 - a)A^{-q}$ and $\beta = aA^{-q}$. Now it follows that $Y = A[aK^{-q} + (1 - a)L^{-q}]^{-1/q}$, which is a special form of the CES production function.)

5.4 First-Order Linear Equations

A **first-order linear differential equation** is one that can be written in the form

$$\dot{x} + a(t)x = b(t) \quad (1)$$

where $a(t)$ and $b(t)$ denote continuous functions of t in a certain interval, and $x = x(t)$ is the unknown function. Equation (1) is called "linear" because the left-hand side is a linear function of x and $\dot{x}$.

The following are all examples of first-order linear equations:

$$(a) \quad \dot{x} + x = t \quad (b) \quad \dot{x} + 2tx = 4t \quad (c) \quad (t^2 + 1)\dot{x} + e^t x = t \ln t$$

The first two equations are obviously of the form (1). The last one can be put into this form if we divide each term by $t^2 + 1$ to get $\dot{x} + [e^t/(t^2 + 1)]x = t \ln t/(t^2 + 1)$.

The Simplest Case

Consider the following equation with a and b as constants, where $a \neq 0$:

$$\dot{x} + ax = b \quad (?)$$

Let us multiply this equation by the positive factor e^{at} , called an **integrating factor**. We then get the equivalent equation

$$\dot{x}e^{at} + axe^{at} = be^{at} \quad (*)$$

It may not be obvious why we came up with this idea, but it turns out to be a good one because the left-hand side of (*) happens to be the derivative of the product $x e^{at}$. Thus (*) is equivalent to

$$\frac{d}{dt}(x e^{at}) = be^{at} \quad (**)$$

According to the definition of the indefinite integral, equation (**) holds for all t in an interval iff $x e^{at} = \int be^{at} dt = (b/a)e^{at} + C$ for some constant C . Multiplying this equation by e^{-at} gives the solution for x . Briefly formulated:

$$\dot{x} + ax = b \iff x = C e^{-at} + \frac{b}{a} \quad (C \text{ is a constant}) \quad (3)$$

If we let $C = 0$ in (3), we obtain the constant solution $x(t) = b/a$. We say that $x = b/a$ is an **equilibrium state**, or a **stationary state**, for the equation. Observe how this solution can be obtained from $\dot{x} + ax = b$ by letting $\dot{x} = 0$ and then solving the resulting equation for x . If the constant a is positive, then the solution $x = C e^{-at} + b/a$ converges to b/a as $t \rightarrow \infty$. In this case, the equation is said to be **stable**, because every solution of the equation converges to an equilibrium as t approaches infinity. See Section 5.7 for more on stability.

EXAMPLE 1 Find the general solution of

$$\dot{x} + 2x = 8$$

and determine whether the equation is stable.

Solution: By (3), the solution is $x = C e^{-2t} + 4$. Here the equilibrium state is $x = 4$, and the equation is stable because $a = 2 > 0$, so $x \rightarrow 4$ as $t \rightarrow \infty$. $\blacksquare$

EXAMPLE 2 (Price adjustment mechanism) Let $D(P) = a - bP$ denote the demand and $S(P) = \alpha + \beta P$ the supply of a certain commodity when the price is P . Here a , b , α , and β are positive constants. Assume that the price $P = P(t)$ varies with time, and that $\dot{P}$ is proportional to excess demand $D(P) - S(P)$. Thus,

$$\dot{P} = \lambda[D(P) - S(P)]$$

where λ is a positive constant. Inserting the expressions for $D(P)$ and $S(P)$ into this equation gives $\dot{P} = \lambda(a - bP - \alpha - \beta P)$. Rearranging, we then obtain

$$\dot{P} + \lambda(b + \beta)P = \lambda(a - \alpha)$$

According to (3), the solution is

$$P = C e^{-\lambda(b+\beta)t} + \frac{a - \alpha}{b + \beta}$$

Because $\lambda(b + \beta)$ is positive, as t tends to infinity, P converges to the equilibrium price $P^e = (a - \alpha)/(b + \beta)$ for which $D(P^e) = S(P^e)$. Thus, the equation is stable. $\blacksquare$

Variable Right-Hand Side

The method used to find the solution of (2) can immediately be applied to the following case of a variable right-hand side:

$$\dot{x} + ax = b(t)$$

Without further comment, after multiplying by the integrating factor e^{at} , we find:

$$\dot{x}e^{at} + axe^{at} = b(t)e^{at} \quad \text{or, equivalently,} \quad \frac{d}{dt}(xe^{at}) = b(t)e^{at}$$

so

$$xe^{at} = \int b(t)e^{at} dt + C$$

Multiplying the last equation by e^{-at} yields the solution for x :

$$\dot{x} + ax = b(t) \iff x = Ce^{-at} + e^{-at} \int e^{at} b(t) dt \quad (4)$$

EXAMPLE 3 (Economic growth) Consider the following model of economic growth in a developing country:

$$(a) \quad X(t) = \sigma K(t) \quad (b) \quad \dot{K}(t) = \alpha X(t) + H(t) \quad (c) \quad N(t) = N_0 e^{\rho t}$$

Here $X(t)$ is total domestic product per year, $K(t)$ is capital stock, $H(t)$ is the net inflow of foreign investment per year, and $N(t)$ is the size of the population, all measured at time t . In (a) we assume that the volume of production is simply proportional to the capital stock, with the factor of proportionality σ being called the *average productivity of capital*. In (b) we assume that the total growth of capital per year is equal to internal savings plus net foreign investment. We assume that savings are proportional to production, with the factor of proportionality α being called the *savings rate*. Finally, (c) tells us that population increases at a constant proportional rate of growth ρ .

Derive from these equations a differential equation for $K(t)$. Assume that $H(t) = H_0 e^{\mu t}$, and find the solution of the differential equation in this case, given that $K(0) = K_0$ and $\alpha\rho \neq \mu$. Find an expression for $x(t) = X(t)/N(t)$, which is domestic product per capita.

Solution: From (a) and (b), it follows that $K(t)$ must satisfy the linear equation

$$\dot{K}(t) - \alpha\sigma K(t) = H(t)$$

Put $H(t) = H_0 e^{\mu t}$ and use (4) to obtain

$$\begin{aligned} K(t) &= Ce^{\alpha\sigma t} + e^{\alpha\sigma t} \int e^{-\alpha\sigma t} H_0 e^{\mu t} dt = Ce^{\alpha\sigma t} + e^{\alpha\sigma t} H_0 \int e^{(\mu - \alpha\sigma)t} dt \\ &= Ce^{\alpha\sigma t} + e^{\alpha\sigma t} \frac{H_0}{\mu - \alpha\sigma} e^{(\mu - \alpha\sigma)t} = Ce^{\alpha\sigma t} + \frac{H_0}{\mu - \alpha\sigma} e^{\mu t} \end{aligned}$$

For $t = 0$, we obtain $K(0) = K_0 = C + H_0/(\mu - \alpha\sigma)$, so $C = K_0 - H_0/(\mu - \alpha\sigma)$. Thus, the solution is

$$K(t) = \left(K_0 - \frac{H_0}{\mu - \alpha\sigma} \right) e^{\alpha\sigma t} + \frac{H_0}{\mu - \alpha\sigma} e^{\mu t} \quad (*)$$

Per capita production is equal to $x(t) = X(t)/N(t) = \sigma K(t)/N_0 e^{\rho t}$. If we use the expression for $K(t)$ in (*), an easy calculation shows that

$$x(t) = x(0)e^{(\alpha\sigma - \rho)t} + \left(\frac{\sigma}{\alpha\sigma - \mu} \right) \frac{H_0}{N_0} e^{(\alpha\sigma - \rho)t} [1 - e^{(\mu - \alpha\sigma)t}] \quad (**)$$

Problem 10 asks you to study this model more closely. I

The General Case

We proceed to find the solution of the general linear equation (1). The trick used to solve $\dot{x} + ax = b(t)$ must be modified. We first multiply equation (1) by a suitably chosen integrating factor $e^{A(t)}$, to obtain

$$\dot{x}e^{A(t)} + a(t)xe^{A(t)} = b(t)e^{A(t)} \quad (5)$$

Now we need to find an $A(t)$ such that the left-hand side of this equation equals the derivative of $xe^{A(t)}$. But the derivative of $xe^{A(t)}$ is equal to $\dot{x}e^{A(t)} + x\dot{A}(t)e^{A(t)}$. We therefore make $A(t)$ satisfy $\dot{A}(t) = a(t)$ by choosing $A(t) = \int a(t) dt$; this makes (5) equivalent to the equation

$$\frac{d}{dt}(xe^{A(t)}) = b(t)e^{A(t)}$$

Thus $xe^{A(t)}$ is an indefinite integral of $b(t)e^{A(t)}$, so there exists a constant C such that $xe^{A(t)} = \int b(t)e^{A(t)} dt + C$. Multiplying by $e^{-A(t)}$ we obtain

$$x = Ce^{-A(t)} + e^{-A(t)} \int b(t)e^{A(t)} dt, \quad \text{where} \quad A(t) = \int a(t) dt$$

To summarize, we have shown that:

$$\dot{x} + a(t)x = b(t) \iff x = e^{-\int a(t) dt} \left(C + \int e^{\int a(t) dt} b(t) dt \right) \quad (6)$$

EXAMPLE 4 Find the general solution of $\dot{x} + 2tx = 4t$ and the integral curve through $(t, x) = (0, -2)$.

Solution: The formula in (6) can be used with $a(t) = 2t$ and $b(t) = 4t$. Then $\int a(t) dt = \int 2t dt = t^2 + C_1$. We choose $C_1 = 0$ so that $\int a(t) dt = t^2$ (choosing another value for C_1 instead gives the same general solution). Then (6) gives

$$x = e^{-t^2} \left(C + \int e^{t^2} 4t dt \right) = Ce^{-t^2} + e^{-t^2} 2e^{t^2} = Ce^{-t^2} + 2$$

If $x = -2$ for $t = 0$, then $-2 = C e^0 + 2$, and so $C = -4$. The integral curve through $(0, -2)$ has the equation $x = 2 - 4e^{-t^2}$. I

The Solution when $x(t_0) = x_0$ is Given

Assume that the value of $x(t)$ is known for $t = t_0$. Then the constant C in (6) is determined. We derive here the formula for the corresponding solution of the equation, which is sometimes useful.

Define $F(t)$ as an indefinite integral of $b(t)e^{A(t)}$, where $A(t) = \int a(t) dt$ and so $A(t) - A(s) = \int_s^t a(\xi) d\xi$. The solution in (6) then becomes

$$x(t) = Ce^{-A(t)} + e^{-A(t)} F(t)$$

Now let $t = t_0$ and solve for C to get $C = x(t_0)e^{A(t_0)} - F(t_0)$. Hence,

$$x(t) = x(t_0)e^{-[A(t)-A(t_0)]} + e^{-A(t)} [F(t) - F(t_0)]$$

By definition of $F(t)$, we have $F(t) - F(t_0) = \int_{t_0}^t b(s)e^{A(s)} ds$. So

$$e^{-A(t)} [F(t) - F(t_0)] = e^{-A(t)} \int_{t_0}^t b(s)e^{A(s)} ds = \int_{t_0}^t b(s)e^{-[A(t)-A(s)]} ds$$

(We can include $e^{-A(t)}$ in the integrand, because we are integrating w.r.t. s .) Finally, therefore, we have the following result:

$$\dot{x} + a(t)x = b(t), \quad x(t_0) = x_0 \iff x = x_0 e^{-\int_{t_0}^t a(\xi) d\xi} + \int_{t_0}^t b(s)e^{-\int_s^t a(\xi) d\xi} ds \quad (7)$$

Wealth Accumulation

As in Example 5.3.5, suppose that the amount of savings in an account at time t is $w = w(t)$. Suppose now that there are deposits and withdrawals at the rates $y(t)$ and $c(t)$, respectively. If there is continuous compounding of interest at the rate $r(t)$, then wealth at time t follows the differential equation

$$\dot{w} = r(t)w + y(t) - c(t) \quad (8)$$

This is clearly a first-order linear differential equation. According to (7), the solution is

$$w(t) = w(0)e^{\int_0^t r(s) ds} + \int_0^t [y(\tau) - c(\tau)]e^{\int_\tau^t r(s) ds} d\tau \quad (9)$$

Since $\int_\tau^t r(s) ds = \int_0^t r(s) ds - \int_0^\tau r(s) ds$, and since $\int_0^\tau r(s) ds$ is independent of τ , equation (9) can be written as:

$$w(t)e^{-\int_0^t r(s) ds} = w(0) + \int_0^t [y(\tau) - c(\tau)]e^{-\int_0^\tau r(s) ds} d\tau \quad (10)$$

Note that the discount factor to be applied to wealth at time τ is $e^{-\int_0^\tau r(s) ds}$. So equation (10) states that the present discounted value (PDV) of assets at time t is the sum of the initial assets $w(0)$ and the total PDV of all deposits, minus the total PDV of all withdrawals.

If there are no deposits or withdrawals from the account, then $y(t) = c(t) = 0$ and so (8) reduces to the separable equation $\dot{w} = r(t)w$. The general solution is $w = Ae^{\int r(t) dt}$.

PROBLEMS FOR SECTION 5.4

1. Find the general solution of $\dot{x} + \frac{1}{2}x = \frac{1}{4}$. Determine the equilibrium state of the equation, and examine whether it is stable. Also draw some typical integral curves.

2. Find the general solutions of the following linear differential equations:

$$(a) \dot{x} + x = 10 \quad (b) \dot{x} - 3x = 27 \quad (c) 4\dot{x} + 5x = 100$$

3. Find the general solution of $\dot{x} = x + t$. (See Example 5.1.1.)

4. Find the general solutions of the following differential equations, and in each case, find the integral curve through $(t, x) = (0, 1)$:

$$(a) \dot{x} - 3x = 5 \quad (b) 3\dot{x} + 2x + 16 = 0 \quad (c) \dot{x} + 2x = t^2$$

5. The differential equation in (3) is separable. Solve it by using the method in Section 5.3, and show that you obtain the same solution as that given in (3).

SM 6. Find the general solutions of the following differential equations:

$$(a) t\dot{x} + 2x + t = 0 \quad (t \neq 0) \quad (b) \dot{x} - \frac{1}{t}x = t \quad (t > 0)$$

$$(c) \dot{x} - \frac{t}{t^2 - 1}x = t \quad (t > 1) \quad (d) \dot{x} - \frac{2}{t}x + \frac{2a^2}{t^2} = 0 \quad (t > 0)$$

7. For the differential equation $\dot{x} = 2tx + t(1 + t^2)$, show that the solution $x(t)$ that passes through $(t, x) = (0, 0)$ has a local minimum at $t = 0$.

8. Prove that if $x(T) = x_T$, the solution of (1) can be expressed as

$$x(t) = x_T e^{\int_t^T a(\xi) d\xi} - \int_t^T b(s)e^{\int_s^T a(\xi) d\xi} ds$$

HARDER PROBLEMS

SM 9. Let $N = N(t)$ denote the size of a certain population, $X = X(t)$ the total product, and $x(t) = X(t)/N(t)$ the product per capita at time t . T. Haavelmo (1954) studied the model:

$$(i) \quad \dot{N}/N = \alpha - \beta N/X \quad (ii) \quad \dot{X} = AN^a$$

where α , β , and a are positive constants, with $a \neq 1$. Show that this leads to a differential equation of the form (3) for $x = x(t)$. Solve this equation and then find expressions for $N = N(t)$ and $X = X(t)$. Examine the limits for $x(t)$, $N(t)$, and $X(t)$ as $t \rightarrow \infty$ in the case $0 < a < 1$.

SM 10. (a) Consider the model in Example 3. With $H_0 = 0$ find the condition for the production per capita to increase with time. A common estimate for σ in developing countries is 0.3. If the population increases at the rate 3% per year ($\rho = 0.03$), how high must the savings rate α be for $x(t)$ to increase with time?

(b) Show that $x(t)$ given by (**) is greater than $x(0)e^{(\alpha\sigma - \rho)t}$ for all $t > 0$. (Look at the two cases $\alpha\sigma - \mu > 0$ and $\alpha\sigma - \mu < 0$ separately.) Why was this to be expected?

(c) Assume that $\alpha\sigma < \rho$. Find a necessary and sufficient condition to obtain sustained growth in production per capita. Give an economic interpretation.

5 Exact Equations and Integrating Factors

We started this chapter by studying the differential equation $\dot{x} = F(t, x)$. Only in very special cases can we find explicit analytical solutions—for example, when the equation is separable or when it is linear in x . In this section we study first-order equations of the form

$$f(t, x) + g(t, x)\dot{x} = 0 \quad (1)$$

where f and g are C^1 functions. (Of course, $\dot{x} = F(t, x)$ can be written in this form: $F(t, x) + (-1)\dot{x} = 0$.) Suppose we happen to find a function $h(t, x)$ such that

$$h'_t(t, x) = f(t, x) \quad \text{and} \quad h'_x(t, x) = g(t, x) \quad (2)$$

Note that $\frac{d}{dt}h(t, x) = h'_t(t, x) + h'_x(t, x)\dot{x}$. So if (2) is satisfied, equation (1) is equivalent to $\frac{d}{dt}h(t, x) = 0$, which is satisfied if and only if $h(t, x) = C$ for some constant C . The solutions of (1) are therefore those functions $x = x(t)$ that satisfy

$$h(t, x) = C, \quad \text{for some constant } C \quad (3)$$

E 1 The differential equation

$$1 + tx^2 + t^2x\dot{x} = 0 \quad (*)$$

is neither separable nor linear. But we might just notice that the function $h(t, x) = t + \frac{1}{2}t^2x^2$ has partial derivatives $h'_t(t, x) = 1 + tx^2$ and $h'_x(t, x) = t^2x$. Then we see that the solution of (*) is any differentiable function x defined implicitly by the equation $t + \frac{1}{2}t^2x^2 = C$, for some constant C . $\blacksquare$

The key step in finding the solution of equation (1) having the form (3) is to determine an appropriate function h . Note first a necessary condition for the existence of such a function h . In fact, if $f(t, x) = h'_t(t, x)$ and $g(t, x) = h'_x(t, x)$, then $f'_x(t, x) = h''_{tx}(t, x)$ and $g'_t(t, x) = h''_{xt}(t, x)$. Hence, by Young's theorem on the equality of second-order cross derivatives,

$$f'_x(t, x) = g'_t(t, x) \quad (4)$$

We shall show in a moment that (4) is also sufficient for the existence of a function h satisfying (2). Equation (1) is called **exact** if (4) is satisfied.

In Example 1, $f(t, x) = 1 + tx^2$ and $g(t, x) = t^2x$, and we see that (4) is satisfied because $f'_x(t, x) = g'_t(t, x) = 2tx$, so the equation is exact. Note also that if we write the separable equation (5.3.1) as $f(t) - \dot{x}/g(x) = 0$, then condition (4) holds trivially.

Next, consider equation (1) and suppose that condition (4) is satisfied. Motivated by the first equation in (2), define the function h by

$$h(t, x) = \int_{t_0}^t f(\tau, x) d\tau + \alpha(x) \quad (5)$$

where we need to choose $\alpha(x)$ appropriately. Differentiating (5) w.r.t. x and using (4) yields

$$h'_x(t, x) = \int_{t_0}^t f'_x(\tau, x) d\tau + \alpha'(x) = \int_{t_0}^t g'_t(\tau, x) d\tau + \alpha'(x) \quad (6)$$

Now $\int_{t_0}^t g'_t(\tau, x) d\tau = g(t, x) - g(t_0, x)$. To make (6) easy to solve, let us put

$$\alpha'(x) = g(t_0, x) \quad \text{with} \quad \alpha(t_0) = 0, \quad \text{so that} \quad \alpha(x) = \int_{t_0}^x g(t_0, F) dF \quad (7)$$

Together (5) and (7) imply that the left-hand side of equation (3) takes the form

$$h(t, x) = \int_{t_0}^t f(\tau, x) d\tau + \int_{t_0}^x g(t_0, \xi) d\xi \quad (8)$$

EXAMPLE 2

In Example 1, $f(t, x) = 1 + tx^2$ and $g(t, x) = t^2x$, so formula (8) yields

$$\begin{aligned} h(t, x) &= \int_{t_0}^t (1 + \tau x^2) d\tau + \int_{t_0}^x t_0^2 \xi d\xi = \left[\tau + \frac{1}{2}\tau^2 x^2 \right]_{t_0}^t + \frac{1}{2} \left[t_0^2 \xi^2 \right]_{t_0}^x \\ &= t + \frac{1}{2}t^2 x^2 - t_0 - \frac{1}{2}t_0^2 x^2 + \frac{1}{2}t_0^2 x^2 - \frac{1}{2}t_0^2 x_0^2 = t + \frac{1}{2}t^2 x^2 - t_0 - \frac{1}{2}t_0^2 x_0^2 \end{aligned}$$

Except for the constant $t_0 - \frac{1}{2}t_0^2 x_0^2$, equation (3) gives the same answer as in Example 1. $\blacksquare$

Consider equation (1) again. If the equation is not already exact, one might wonder if it can be made so by multiplying it by a suitable function $\beta(t, x)$. In fact, the equation $\beta(t, x)f(t, x) + \beta(t, x)g(t, x)\dot{x} = 0$ is exact provided that $\frac{\partial}{\partial x}[\beta(t, x)f(t, x)] = \frac{\partial}{\partial t}[\beta(t, x)g(t, x)]$, or, equivalently,

$$\beta'_t(t, x)f(t, x) + \beta(t, x)f'_t(t, x) = \beta'_t(t, x)g(t, x) + \beta(t, x)g'_t(t, x) \quad (9)$$

A function $\beta(t, x)$ satisfying (9) is called an **integrating factor** for the differential equation (1). In general it is hard to find such an integrating factor, even when one exists. But in two special cases it is relatively easy.

Case I: Suppose $(f'_x - g'_t)/g$ is a function of t alone. Then we can let $\beta(t, x) = \beta(t)$. In this case (9) reduces to $\beta'(t)f'_x = \beta(t)g'_t + \beta'(t)g$. So

$$\beta'(t) = \beta(t) \frac{f'_x - g'_t}{g}, \quad \text{and hence} \quad \beta(t) = \exp \left(\int \frac{f'_x - g'_t}{g} dt \right) \quad (10)$$

Case II: Suppose $(g'_t - f'_x)/f$ is a function of x alone. Then we can let $\beta(t, x) = \beta(x)$. In this case (9) reduces to $f\beta'(x) + \beta(x)f'_x = \beta(x)g'_t$. So

$$\beta'(x) = \beta(x) \frac{g'_t - f'_x}{f}, \quad \text{and hence} \quad \beta(x) = \exp \left(\int \frac{g'_t - f'_x}{f} dx \right) \quad (11)$$

EXAMPLE 3

Solve the differential equation $1 - (t + 2x)\dot{x} = 0$, $t > 0$, $x > 0$.

Solution: (The equation is clearly equivalent to $\dot{x} = 1/(t + 2x)$, which is neither separable nor linear.) With $f(t, x) = 1$ and $g(t, x) = -t - 2x$ we get $(g'_t - f'_x)/f = -1$, which does not depend on t , so Case II applies. By (11), $\beta(x) = \exp(\int (-1) dx) = e^{-x}$ is an integrating factor. Hence, $e^{-x} - e^{-x}(t + 2x)\dot{x} = 0$ is exact, and (8) takes the form

$$\begin{aligned} h(t, x) &= \int_{t_0}^t e^{-x} d\tau - \int_{x_0}^x e^{-t}(t_0 + 2\xi) d\xi = \left[\tau e^{-x} \right]_{t_0}^t - \left[e^{-t} (t_0\xi + \xi^2) \right]_{x_0}^x \\ &= te^{-x} + 2xe^{-x} - e^{-t_0}(t_0 + 2x_0) + 2(e^{-x} - e^{-x_0}) \end{aligned}$$

using integration by parts. The solution is then any differentiable function $x = x(t)$ that satisfies $te^{-x} + 2xe^{-x} - e^{-t_0}(t_0 + 2x_0) + 2(e^{-x} - e^{-x_0}) = C$, for some constant C . $\blacksquare$

MS FOR SECTION 5.5

1. Solve the differential equation $2t + 3x^2\dot{x} = 0$, first as a separable equation, and second by considering it as an exact equation.

- SM 2. Solve the differential equation $1 + (2 + t/x)\dot{x} = 0$, $t > 0$, $x > 0$.

5.6 Transformation of Variables

Only very special types of differential equation have solutions given by explicit formulas. However, transforming the variables sometimes converts a seemingly insoluble differential equation into one of a familiar type that we already know how to solve.

One example is **Bernoulli's equation** which has the form

$$\dot{x} + a(t)x = b(t)x^r \quad (1)$$

where the exponent r is a fixed real number, and where $a(t)$ and $b(t)$ are given continuous functions.

If $r = 0$, the equation is linear, and if $r = 1$, it is separable, since $\dot{x} = (b(t) - a(t))x$. Suppose that $r \neq 1$, and let us look for a solution with $x(t) > 0$ for all t , so that the power x^r is always well defined. If we divide (1) by x^r , we obtain

$$x^{-r}\dot{x} + a(t)x^{1-r} = b(t) \quad (*)$$

Now introduce the transformation

$$z = x^{1-r} \quad (2)$$

of the variable x . Then $\dot{z} = (1-r)x^{-r}\dot{x}$. Substituting this into (*) gives

$$\frac{1}{1-r}\dot{z} + a(t)z = b(t) \quad (3)$$

which is a linear differential equation for $z = z(t)$. Once $z(t)$ has been found, we can use (2) to determine $x(t) = z(t)^{1/(1-r)}$, which then becomes the solution of (1).

- E 1 Solve the differential equation $\dot{x} = -tx + t^3x^3$.

Solution: This is a Bernoulli equation with $r = 3$. As suggested by (2), we introduce the transformation $z = x^{1-3} = x^{-2}$. After rearranging, equation (3) then takes the form

$$\dot{z} - 2tz = -2t^3$$

This is a linear differential equation, and we can use formula (5.4.6) with $a(t) = -2t$. Because $\int a(t) dt = \int -2t dt = -t^2$, we get

$$z = Ce^{t^2} - 2e^{t^2} \int t^3 e^{-t^2} dt \quad (*)$$

If we substitute $u = -t^2$ in the last integral, then $du = -2t dt$ and we have

$$\int t^3 e^{-t^2} dt = \frac{1}{2} \int u e^u du = \frac{1}{2} u e^u - \frac{1}{2} e^u = -\frac{1}{2} t^2 e^{-t^2} - \frac{1}{2} e^{-t^2}$$

where we have used integration by parts. Now (*) yields

$$z = Ce^{t^2} - 2e^{t^2} \left(-\frac{1}{2} t^2 e^{-t^2} - \frac{1}{2} e^{-t^2} \right) = Ce^{t^2} + t^2 + 1$$

It follows that the original equation has the solutions

$$x = \pm z^{-1/2} = \pm (Ce^{t^2} + t^2 + 1)^{-1/2}$$

Here are two other examples of successful substitutions.

EXAMPLE 2 The differential equation

$$y' - 1 + 2x(y - x)^2 = 0 \quad (*)$$

evidently has $y = x$ as one solution. Define $y = x + 1/z$, where z is a function of x , and show that this substitution leads to a separable differential equation for z . Solve this equation and then find the solution of (*) that passes through the point $(x, y) = (0, -1/2)$. (Note that in this example x is the free variable and y is the unknown function, and we write y' for dy/dx .)

Solution: If $z \neq 0$, differentiating $y = x + 1/z$ w.r.t. x gives $y' = 1 - z'/z^2$. When we insert this into (*) and reorganize, we obtain an equation that reduces to $z' = 2x$, with general solution $z = x^2 + C$. We are looking for the solution with $y = -1/2$ when $x = 0$. This gives $z = -2$ when $x = 0$, so $C = -2$. Hence, the required solution is $y = x + 1/(x^2 - 2)$ (defined for $-\sqrt{2} < x < \sqrt{2}$).

EXAMPLE 3 Show that the substitution $z = x + t^2$ transforms the differential equation

$$\dot{x} = \frac{2t}{x + t^2} \quad (*)$$

into a separable differential equation for z , and use this to find the general solution.

Solution: The suggested substitution implies that $\dot{x} = \dot{z} - 2t$. Inserting this into (*) gives $\dot{z} - 2t = 2t/z$, hence

$$\dot{z} = 2t \left(1 + \frac{1}{z} \right) = 2t \frac{z + 1}{z}$$

This equation is separable, and we use the recipe from Section 5.3. Note the constant solution $z = -1$. The other solutions are found in the usual way by separating the variables:

$$\int \frac{z}{z + 1} dz = \int 2t dt$$

Since $z/(z + 1) = (z + 1 - 1)/(z + 1) = 1 - 1/(z + 1)$, we obtain $z - \ln |z + 1| = t^2 + C_1$. If we substitute $z = x + t^2$ and rearrange, we get $x - \ln |x + t^2 + 1| = C_2$.

$|x + t^2 + 1| = e^{-C_1} e^x$, i.e. $x + t^2 + 1 = \pm e^{-C_1} e^x$. Hence, if we define $C = \pm e^{-C_1}$, these solutions are given implicitly by

$$x = C e^x - t^2 - 1$$

The constant solution $z = -1$ gives $x = -t^2 - 1$, which corresponds to $C = 0$.

MS FOR SECTION 5.6

SM 1. Solve the following Bernoulli equations assuming $t > 0, x > 0$:

$$(a) \quad t\dot{x} + 2x = tx^2 \quad (b) \quad \dot{x} = 4x + 2e^t \sqrt{x} \quad (c) \quad t\dot{x} + x = x^2 \ln t$$

2. Solve the differential equation $(1 + tx)\dot{x} = -x^2, t > 0$. (Hint: Try the substitution $x = w/t$.)

SM 3. An economic growth model leads to the Bernoulli equation

$$\dot{K} = \alpha A n_0^a e^{(a\alpha + \varepsilon)t} K^b - \alpha \delta K \quad (A, n_0, a, b, \alpha, \delta, \text{ and } \varepsilon \text{ are positive constants})$$

Find the general solution of the equation when $a\alpha + \varepsilon + \alpha\delta(1 - b) \neq 0$ and $b \neq 1$.

SM 4. An economic growth model by T. Haavelmo (1954) leads to the differential equation

$$\dot{K} = \gamma_1 b K^\alpha + \gamma_2 K$$

where γ_1, γ_2, b , and α are positive constants, $\alpha \neq 1$ and $K = K(t)$ is the unknown function. The equation is separable, but solve it as a Bernoulli equation.

5. (a) Consider the equation

$$t\dot{x} = x - f(t)x^2, \quad t > 0 \quad (*)$$

where f is a given continuous function, and $x = x(t)$ is the unknown function. Show that the substitution $x = tz$ transforms $(*)$ into a separable equation in $z = z(t)$.

(b) Let $f(t) = t^3/(t^4 + 2)$ and find the solution curve for $(*)$ through the point $(1, 1)$.

6. (a) Differential equations of the form $\dot{x} = g(x/t)$, where the right-hand side depends only on the ratio x/t , are called **projective**. Prove that if we substitute $z = x/t$, a projective equation becomes a separable equation with z as the unknown function.

(b) Solve the equation $3tx^2\dot{x} = x^3 + t^3, t > 0, x > 0$.

SM 7. Find the general solution of the projective equation $\dot{x} = 1 + x/t - (x/t)^2$.

HARDER PROBLEMS

8. In general, differential equations of the form

$$\dot{x} = P(t) + Q(t)x + R(t)x^2 \quad (\text{Riccati's equation})$$

can only be solved numerically. But if we know a particular solution $u = u(t)$, the substitution $x = u + 1/z$ will transform the equation into a linear differential equation for z as a function of t . Prove this, and apply it to $t\dot{x} = x - (x/t)^2$.

5.7 Qualitative Theory and Stability

It is convenient when economic models involve differential equations that can be solved explicitly in terms of elementary functions. Usually this will make it easy to study the nature of the solution. Most kinds of differential equation do not have this nice property, however, so the nature of their solution has to be investigated in some other way.

The theory we have discussed so far is insufficient for another reason. Any economic model is based on a number of assumptions. It is often desirable to make these assumptions as weak as possible without losing the essential aspects of the problem. If a differential equation appears in the model, it therefore typically contains unspecified parameters.

As a result, when a differential equation is used to describe some particular economic phenomenon, the typical situation is as follows:

- (A) It is impossible to obtain an explicit solution of the equation.
- (B) The equation contains unspecified parameters (or even unspecified functions).

Even so, there is often much that can be said about the nature of any solution to the differential equation. In this section, we discuss, in particular, the stability of any solution.

Autonomous Equations, Phase Diagrams, and Stability

Many differential equations in economics can be expressed in the form

$$\dot{x} = F(x) \quad (1)$$

This is a special case of the equation $\dot{x} = F(t, x)$, in which t does not explicitly appear on the right-hand side. For this reason, the equation in (1) is called **autonomous**.

To examine the properties of the solutions to (1), it is useful to study its **phase diagram**. This is obtained by putting $y = \dot{x}$ and drawing the curve $y = F(x)$ in the xy -plane (or $x\dot{x}$ -plane). An example is indicated in Fig. 1.

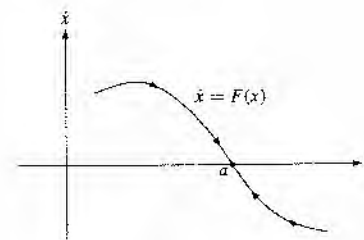


Figure 1

Any solution $x = x(t)$ of (1) has an associated $\dot{x} = \dot{x}(t)$. For every t , the pair $(x(t), \dot{x}(t))$ is a point on the curve in the phase diagram. What can be said generally about this point when t increases? If we consider a point on the curve lying above the x -axis, then $F(x(t)) > 0$ and

therefore $\dot{x}(t) = F(x(t)) > 0$, so that $x(t)$ increases with t . It follows from this observation that the point $(x(t), \dot{x}(t))$ moves from left to right in the diagram if we are above the x -axis. On the other hand, if we are at a point on the graph below the x -axis, then $\dot{x}(t) < 0$, and $x(t)$ decreases with t , so we move from right to left. These movements are indicated by arrows in Fig. 1.

One of the most important properties of a differential equation is whether it has any *equilibrium* or *stationary states*. These correspond to solutions of the equation that do not change over time. In many economic applications, it is also very important to know whether an equilibrium state is *stable*. This can often be determined even if we cannot find explicit solutions of the equation. In physics, the rest position of a pendulum (hanging downward and motionless) is stable; if it is slightly disturbed while in this position, it will swing back and forth until it gradually approaches the equilibrium state of rest. To use a common analogy, we do not expect to encounter an unstable equilibrium in the real world for the same reason that a pin will not balance on its point.

In general, we say that a point a represents an **equilibrium state** or a **stationary state** for equation (1) if $F(a) = 0$. In this case, $x(t) \equiv a$ is a solution of the equation. If $x(t_0) = a$ for some value t_0 of t , then $x(t)$ is equal to a for all t .

The example of Fig. 1 has one equilibrium state, a . It is called **globally asymptotically stable**, because if $x(t)$ is a solution to $\dot{x} = F(x)$ with $x(t_0) = x_0$, then $x(t)$ will always converge to the point on the x -axis with $x = a$ for any start point (t_0, x_0) .

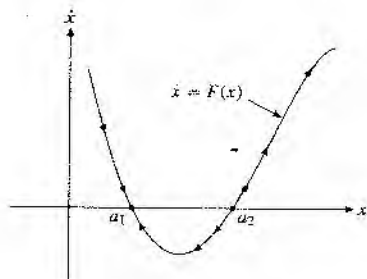


Figure 2 a_1 is a locally stable equilibrium state for $\dot{x} = F(x)$, whereas a_2 is unstable.

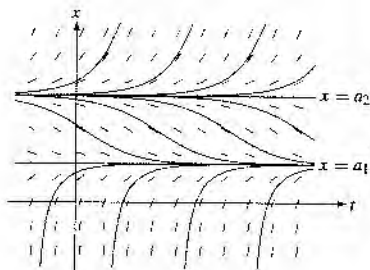


Figure 3 A corresponding directional diagram and some solution curves for $\dot{x} = F(x)$.

In Fig. 2 there are two equilibrium states, a_1 and a_2 . If we start in either of these states, then we will remain there. However, there is an important difference between the two. If $x(t)$ starts close to a_1 , but not at a_1 , then $x(t)$ will approach a_1 as t increases. On the other hand, if $x(t)$ starts close to, but not at a_2 , then $x(t)$ will move away from a_2 as t increases. We say that a_1 is a **locally asymptotically stable equilibrium state**, whereas a_2 is **unstable**. Note how this behaviour of the integral curves is confirmed by Fig. 3.

Look at Fig. 2 again. Note that at the stable point a_1 , the graph of $\dot{x} = F(x)$ has a negative slope, whereas the slope is positive at a_2 . Suppose that a is an equilibrium state for $\dot{x} = F(x)$, so that $F(a) = 0$. If $F'(a) < 0$, then $F(x)$ is positive to the left of $x = a$, and

negative to the right. Around $x = a$ the graph of $\dot{x} = F(x)$ and the directional diagram both look similar to Figs. 2 and 3 near $x = a_1$. So a is stable. On the other hand, if $F'(a) > 0$, they both look similar to Figs. 2 and 3 near $x = a_2$. Hence a is unstable. We have the following result:

- (a) $F(a) = 0$ and $F'(a) < 0 \Rightarrow a$ is a locally asymptotically stable equilibrium.
 (b) $F(a) = 0$ and $F'(a) > 0 \Rightarrow a$ is an unstable equilibrium.
- (2)

If $F(a) = 0$ and $F'(a) = 0$, then (2) is inconclusive. You should now give two different examples showing that a can be locally stable or locally unstable in this case.

EXAMPLE 1 In Section 5.4 we studied the equation

$$\dot{x} + ax = b \quad (a \neq 0)$$

It is a special case of (1), with $F(x) = b - ax$. There is a unique equilibrium state, at $x = b/a$, where $F'(x) = -a$. According to (2), $x = b/a$ is locally asymptotically stable if $a > 0$, but unstable if $a < 0$. Compare this result with the remarks following equation (5.4.3). I

EXAMPLE 2 (**Price adjustment mechanism**) We generalize Example 5.4.2 and assume that the price $P = P(t)$ satisfies the nonlinear differential equation

$$\dot{P} = F(P) = H(D(P) - S(P)) \quad (*)$$

As before, $\dot{P}$ is a function of the excess demand $D(P) - S(P)$. We assume that the function H satisfies $H(0) = 0$ and $H' > 0$, so that H is strictly increasing. If demand is greater than supply when the price is P , then $D(P) - S(P) > 0$, so $\dot{P} > 0$, and the price increases. On the other hand, the price decreases when $D(P) - S(P) < 0$. Equation (*) therefore represents what can be called a **price adjustment mechanism**.

Assume P^e is an equilibrium price at which $\dot{P} = F(P^e) = 0$ because demand $D(P^e)$ equals supply $S(P^e)$. Note that $F'(P) = H'(D(P) - S(P))(D'(P) - S'(P))$. Because $H' > 0$, we see that $F'(P)$ has the same sign as $D'(P) - S'(P)$. By (2), we see that the equilibrium price P^e is *stable* if $D'(P^e) - S'(P^e) < 0$. This condition is usually satisfied because we expect that $D' < 0$ and $S' > 0$. I

EXAMPLE 3 (**The Solow–Swan growth model**) This “neoclassical” growth model involves a constant returns to scale production function $Y = F(K, L)$ expressing national output Y as a function of the capital stock K and of the labour force L , both assumed to be positive. It is assumed that L grows at a constant proportional rate $\lambda > 0$. Also, a constant fraction $s \in (0, 1)$ of output Y is devoted to net investment $\dot{K}$. That is, $\dot{K} = sY$.

The model is usually analysed by dividing all variables by L , thus obtaining new variables

$$y = \frac{Y}{L} \quad (\text{the output/labour ratio}) \quad \text{and} \quad k = \frac{K}{L} \quad (\text{the capital/labour ratio})$$

Because there are constant returns to scale, the function F is homogeneous of degree 1. Thus $y = Y/L = F(K, L)/L = F(K/L, 1) = f(k)$, where $f(k)$ is defined as $F(k, 1)$. Also $\dot{k}/k = \dot{K}/K - \dot{L}/L = sY/K - \lambda = sy/k - \lambda$. Multiplying by k leads to the separable differential equation

$$\dot{k} = sf(k) - \lambda k \quad (3)$$

Without specifying F or f , equation (3) has no explicit solution.

Nevertheless, suppose we make the usual assumptions that $F(0, L) = 0$ and that the marginal product of capital $F'_K(K, L)$ is positive and diminishing for all L . Putting $L = 1$, it follows that $f(0) = 0$ and that $f'(k) > 0$, $f''(k) < 0$ for all $k > 0$. Provided that $sf'(0) > \lambda$ and $sf'(k) < \lambda$ for large k , the phase diagram for equation (3) will look like Fig. 4.

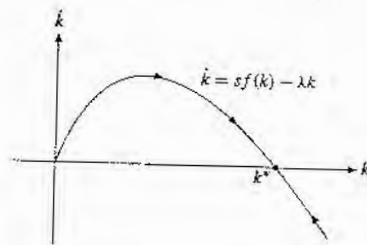


Figure 4 Phase diagram for (3), with appropriate conditions on f .

Then there is a unique equilibrium state with $k^* > 0$. It is given by

$$sf(k^*) = \lambda k^* \quad (4)$$

By studying Fig. 4 we see that k^* is stable. No matter what the initial capital/labour ratio $k(0)$ may be, $k(t) \rightarrow k^*$ as $t \rightarrow \infty$.

NOTE 1 Let us briefly discuss sufficient conditions for the existence and uniqueness of an equilibrium in the Solow-Swan model. Apart from the assumptions on F or f we have already made, we postulate the **Inada conditions**,² according to which $f'(k) \rightarrow \infty$ as $k \rightarrow 0$ and also $f'(k) \rightarrow 0$ as $k \rightarrow \infty$.

To see why these conditions are sufficient, define $G(k) = sf(k) - \lambda k$. Then $G'(k) = sf'(k) - \lambda$, and equation (3) changes to $\dot{k} = G(k)$. The assumptions on f imply that $G(0) = 0$, $G'(k) \rightarrow \infty$ as $k \rightarrow 0$, $G'(k) \rightarrow -\lambda < 0$ as $k \rightarrow \infty$, and $G''(k) = sf''(k) < 0$ for all $k > 0$. So G has a unique stationary point $\hat{k} > 0$ at which $G'(\hat{k}) = 0$. Obviously, $G(\hat{k}) > 0$. But $G'(k) < -\frac{1}{2}\lambda < 0$ for all large enough k . It follows that $G(k) \rightarrow -\infty$ as $k \rightarrow \infty$, so there is a unique point $k^* > 0$ with $G(k^*) = 0$. In addition, $G'(k^*) < 0$. According to (2), this is a sufficient condition for the local asymptotic stability of k^* .

² Named after the Japanese economist F. T. Inada, who introduced them.

General Results on Autonomous Equations

Figure 3 shows some solution curves for the equation $\dot{x} = F(x)$ which is graphed in Fig. 2. It seems that, given any one solution curve, all the others are obtained by shifting that curve horizontally to the right or to the left. This is confirmed by the following:

If $x = \varphi(t)$ is a solution of $\dot{x} = F(x)$, so is $x = \varphi(t + c)$ for any constant c .

The argument is easy: $\dot{x} = \dot{\varphi}(t + c) = F(\varphi(t + c)) = F(x)$.

Note how Fig. 3 displays two constant (equilibrium) solutions, corresponding to the zeros a_1 and a_2 of the function F , while all the other solutions appear to be either strictly increasing or strictly decreasing in the intervals where they are defined. This behaviour of the solutions of $\dot{x} = F(x)$ turns out to be typical provided F is a C^1 function. To prove this result we use Theorem 5.8.1 in the next section, which says that when F is C^1 , there is one and only one solution curve passing through a given point (t_0, x_0) in the tx -plane.

THEOREM 5.7.1

If F is a C^1 function, every solution of the autonomous differential equation $\dot{x} = F(x)$ is either constant or strictly monotone on the interval where it is defined.

Proof: Suppose first that x is a solution such that $\dot{x}(t_0) = 0$ for some t_0 , and put $a = x(t_0)$. Then $F(a) = F(x(t_0)) = \dot{x}(t_0) = 0$, so a is a zero of F . The constant function $x_a(t) \equiv a$ is then also a solution. Because both $x(t)$ and $x_a(t)$ pass through the same point (t_0, a) in the tx -plane, it follows that $x(t) = x_a(t) = a$ for all t . Hence x is a constant function. (See Note 5.8.2.)

If x is not a constant solution, then $\dot{x}(t) \neq 0$ for all t in the domain of x . Because x is differentiable and F is continuous, the derivative $\dot{x}(t) = F(x(t))$ is a continuous function of t . It follows that $\dot{x}$ must have the same sign everywhere in its domain, otherwise the intermediate value theorem would give us a zero for $\dot{x}$. Hence $\dot{x}$ is either everywhere positive or everywhere negative, and x itself must be either everywhere strictly increasing or everywhere strictly decreasing. ■

Assume still that F is C^1 . Then two different solution curves for $\dot{x} = F(x)$ cannot have common points (Theorem 5.8.1). (This holds also in the nonautonomous case.) In the present autonomous case, all solution curves crossing any given line parallel to the t axis have the same slope at the crossing points.

Let us return to the example illustrated in Fig. 2. The straight lines $x = a_1$ and $x = a_2$ in the phase diagram of Fig. 3 are solution curves. Hence no other solution curve can cross either of these lines. Consider a solution $x = \varphi(t)$ that passes through a point (t_0, x_0) where $a_1 < x_0 < a_2$. Then $\varphi(t)$ must lie in the interval (a_1, a_2) for all t , so φ is strictly decreasing with lower bound a_1 . This implies that $\varphi(t)$ must approach a limit as t approaches infinity. It is reasonable to expect (and it follows from Theorem 5.7.2 below) that the limit $a = \lim_{t \rightarrow \infty} \varphi(t)$ must be an equilibrium state for the equation $\dot{x} = F(x)$. Since there are no equilibrium states in the open interval (a_1, a_2) , we must actually have $a = a_1$. Similarly, we see that a solution that lies below a_1 will grow towards a_1 in the limit as $t \rightarrow \infty$. A solution with $x > a_2$ will tend to infinity, unless there are other equilibrium

EM 5.7.2 (A LIMIT IS AN EQUILIBRIUM)

Suppose that $x = x(t)$ is a solution of $\dot{x} = F(x)$, where the function F is continuous. Suppose that $x(t)$ approaches a (finite) limit a as t approaches ∞ . Then a must be an equilibrium state for the equation, i.e. $F(a) = 0$.

Proof: For a contradiction, suppose $F(a) > 0$. Since F is continuous, there exists a $\delta > 0$ such that $|F(x) - F(a)| < \frac{1}{2}F(a)$ for all x in $(a - \delta, a + \delta)$. In particular, $F(x) > \frac{1}{2}F(a)$ for all x in this interval. Since $\lim_{t \rightarrow \infty} x(t) = a$, there must exist a T such that $x(t)$ lies in $(a - \delta, a + \delta)$ for all $t > T$. For $t > T$ we then have $\dot{x}(t) = F(x(t)) > \frac{1}{2}F(a)$. Hence, $x(t) - x(T) = \int_T^t \dot{x}(\tau) d\tau > \frac{1}{2}F(a)(t - T)$. But the last expression tends to ∞ as $t \rightarrow \infty$. It follows that $x(t)$ also tends to ∞ as $t \rightarrow \infty$, contrary to $x(t)$ being in the interval $(a - \delta, a + \delta)$. Therefore, we cannot have $F(a) > 0$. A similar argument shows that we cannot have $F(a) < 0$ either. Hence, $F(a) = 0$.

MS FOR SECTION 5.7

- Draw phase diagrams associated with the differential equations and determine the nature of the possible equilibrium states.
 - $\dot{x} = x - 1$
 - $\dot{x} + 2x = 24$
 - $\dot{x} = x^2 - 9$
- Determine the nature of the possible equilibrium states for:
 - $\dot{x} = x^3 + x^2 - x - 1$
 - $\dot{x} = 3x^2 + 1$
 - $\dot{x} = xe^x$
- Consider the differential equation $\dot{x} = \frac{1}{2}(x^2 - 1)$, $x(0) = x_0$.
 - Find the solution of this separable differential equation, and draw some integral curves in the tx -plane. What happens to the solution as $t \rightarrow \infty$ for different initial points x_0 ?
 - Draw the phase diagram for the equation. Find the two equilibrium states. Decide whether they are stable or unstable. Compare with the results in part (a).

HARDER PROBLEMS

- The stationary state k^* defined by (4) in Example 3 depends on s and λ . Find expressions for $\partial k^*/\partial s$ and $\partial k^*/\partial \lambda$ and determine the signs of these derivatives when $f''(k) < 0$. (Show that in this case $sf'(k^*) < \lambda$.) Give an economic interpretation of the result. Prove that $F_K'(K, L) = f'(k)$.
 - Consumption per worker c is defined by $c = (Y - \dot{K})/L$. Show that when $k = k^*$, then $c = f(k^*) - \lambda k^*$. Use this to show that if consumption per worker in the stationary state is to be maximized, it is necessary that $f'(k^*) = \lambda$, i.e. $\partial F/\partial K = \lambda$. Thus, the marginal product of capital $\partial F/\partial K$ must equal the proportional rate of growth of the labour force. (This is often called "the golden rule of accumulation".)
 - Show that in the stationary state, $\dot{K}/K$ is equal to λ .

5.8 Existence and Uniqueness

For an economic model to be consistent, the equations in that model must have a solution. This is no less true when the model involves one or more differential equations. Also, if a solution does exist that satisfies the relevant initial conditions, we want to know whether the solution is unique.

Answers to such questions are provided by existence and uniqueness theorems. For first-order equations, one has the following result (a special case of Theorem 5.8.2 below):

THEOREM 5.8.1 (EXISTENCE AND UNIQUENESS I)

Consider the first-order differential equation

$$\dot{x} = F(t, x)$$

and suppose that both $F(t, x)$ and $F'_x(t, x)$ are continuous in an open set A in the tx -plane. Let (t_0, x_0) be an arbitrary point in A . Then there exists exactly one "local" solution of the equation that passes through the point (t_0, x_0) .

The following two notes make things more precise:

NOTE 1 If the conditions in the theorem are met, and (t_0, x_0) is an arbitrary point in A , then there exist an interval (a, b) around t_0 , and a function $x(t)$ defined in (a, b) , such that $x(t)$ is a solution of the equation in (a, b) with $x(t_0) = x_0$ and $(t, x(t)) \in A$ for all t in (a, b) . Note that the theorem guarantees only the *existence* of an interval as described; the length of the interval could be very small. For this reason Theorem 5.8.1 is a *local existence* theorem; it ensures the existence of a solution only in a small neighbourhood of t_0 .

NOTE 2 As for uniqueness, one can prove that if $x(t)$ and $y(t)$ are solutions of the equation lying in A with $x(t_0) = y(t_0)$, then $x(t) = y(t)$ for all t at which both solutions are defined.

EXAMPLE 1 Suppose $F(t, x)$ is the continuous function ax , in which case $F'_x(t, x) = a$ is also continuous everywhere. Theorem 5.8.1 implies that there is a unique solution curve of the associated equation $\dot{x} = ax$ passing through each point (t_0, x_0) . In fact, the required solution is $x(t) = x(t_0)e^{a(t-t_0)}$.

EXAMPLE 2 Let $F(t, x) = f(t)g(x)$. It follows from Theorem 5.8.1 that existence and uniqueness are ensured if $f(t)$ is continuous and $g(x)$ is continuously differentiable.

As pointed out in Note 1, Theorem 5.8.1 gives no information about the length of the interval on which the solution is defined. One factor which can limit this interval is that $x = x(t)$ can grow so fast that the solution "explodes" to infinity even while t remains bounded.

EXAMPLE 3 Find the largest interval on which there is a solution of $\dot{x} = x^2$ with $x(0) = 1$.

Solution: Any nonzero solution of this separable equation is of the form $x = -1/(t + C)$. Because $x(0) = 1$ gives $C = -1$, the solution is $x = 1/(1 - t)$, defined on $(-\infty, 1)$. The solution curve is shown in Fig. 1. The graph “runs off to ∞ ” as t approaches 1 from the left, and this solution cannot be extended beyond $(-\infty, 1)$. (The function $x = 1/(1 - t)$ also satisfies $\dot{x} = x^2$ for $t > 1$, but this is not part of the solution satisfying $x(0) = 1$.)

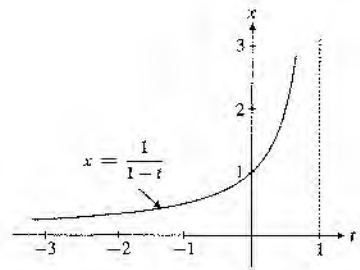


Figure 1

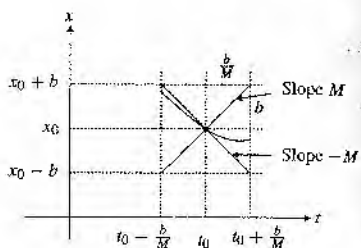


Figure 2

The following more precise result specifies an interval on which a solution is defined. Example 14.3.1 gives a well-known proof based on Picard’s method of successive approximations, explained below.

5.8.2 EXISTENCE AND UNIQUENESS

Consider the initial value problem

$$\dot{x} = F(t, x), \quad x(t_0) = x_0 \tag{1}$$

Suppose that $F(t, x)$ and $F'_x(t, x)$ are continuous over the rectangle

$$\Gamma = \{(t, x) : |t - t_0| \leq a, |x - x_0| \leq b\}$$

and let

$$M = \max_{(t,x) \in \Gamma} |F(t, x)|, \quad r = \min(a, b/M) \tag{2}$$

Then (1) has a unique solution $x(t)$ on $(t_0 - r, t_0 + r)$, and $|x(t) - x_0| \leq b$ in this interval.

NOTE 3 It may well be that the solution in Theorem 5.8.2 can be extended to a larger interval than that described in the theorem. Note that the length of this interval is $2r$, where r is the smaller of the numbers a and b/M . Of course, we cannot expect r to be larger than a . Also, the inequality $r \leq b/M$ is chosen to ensure that the solution $x(t)$ will stay inside the rectangle Γ . Because $-M \leq \dot{x} \leq M$ as long as $(t, x) \in \Gamma$, the part of any solution curve through (t_0, x_0) that is contained in Γ must lie between the two straight lines through (t_0, x_0) with slopes $\pm M$, as illustrated in Fig. 2.

EXAMPLE 4 Prove that

$$\dot{x} = 3t^2 - te^{-x^2}, \quad x(0) = 0$$

has a unique solution on the interval $(-\frac{1}{2}, \frac{1}{2})$, and that $|x(t)| \leq 1$ in this interval.

Solution: We use the notation of Theorem 5.8.2 with $a = \frac{1}{2}$ and $b = 1$. The rectangle Γ is $\Gamma = \{(t, x) : |t| \leq \frac{1}{2}, |x| \leq 1\}$. Note that $|F(t, x)| = |3t^2 - te^{-x^2}| \leq 3t^2 + |t|e^{-x^2} \leq 3/4 + 1/2 = 5/4$ for all (t, x) in Γ . The desired conclusion follows from Theorem 5.8.2, with $M = 5/4$ and $r = \min\{1/2, 1/(5/4)\} = 1/2$.

Dependence of Solutions on Parameters

Assume that the conditions of Theorem 5.8.1 or 5.8.2 are met. The unique solution will obviously depend on the initial values t_0 and x_0 . One can prove that the solution depends continuously on t_0 and x_0 , so that small changes in t_0 and x_0 cause small changes in the solution. In fact, the solution will even be differentiable as a function of (t_0, x_0) . For a precise formulation in a more general setting, see Section 7.6.

Differential equations appearing in economic models often involve a number of parameters in addition to the initial values. These parameters are often inferred imperfectly from empirical observations and so are subject to uncertainty. This gives a reason to prefer models whose solutions are not very sensitive to small perturbations of the parameters. In fact, under rather mild restrictions placed on the differential equation, one can prove that the solution depends continuously on the parameters. Again see Section 7.6.

Picard’s Method of Successive Approximations

Here is a brief indication of how to prove Theorem 5.8.2, which simultaneously suggests a method to construct an approximate solution of $\dot{x} = F(t, x)$ with $x(t_0) = x_0$.

Define the sequence of functions $\{x_n(t)\}$, $n = 0, 1, 2, \dots$, by letting $x_0(t) \equiv x_0$, and

$$x_n(t) = x_0 + \int_{t_0}^t F(s, x_{n-1}(s)) ds, \quad n = 1, 2, \dots \tag{*}$$

Assuming that F and F'_x are continuous, in Example 14.3.1 it is shown that, under the hypotheses of Theorem 5.8.2, the sequence $x_n(t)$ is well defined and converges uniformly³ to a function $x(t)$ satisfying $|x(t) - x_0| \leq b$ for all t in $(t_0 - r, t_0 + r)$. As $n \rightarrow \infty$, the left-hand side of (*) converges to $x(t)$ for each t in $(t_0 - r, t_0 + r)$, whereas it can be shown that the right-hand side converges to $x_0 + \int_{t_0}^t F(s, x(s)) ds$. So $x(t) = x_0 + \int_{t_0}^t F(s, x(s)) ds$ for all t in $(t_0 - r, t_0 + r)$. Differentiating this equation w.r.t. t yields $\dot{x}(t) = F(t, x(t))$. Moreover, $x(t_0) = x_0$, so $x(t)$ is a solution of (1).

EXAMPLE 5 Use Picard’s method to solve the initial value problem

$$\dot{x} = t + x, \quad x(0) = 0 \tag{**}$$

³ A sequence $\{x_n(t)\}$ of functions defined on an interval I is said to converge *uniformly* to a function $x(t)$ defined on I if, for each $\epsilon > 0$ there is a natural number $N(\epsilon)$ (depending on ϵ , but *not* on t) such that $|x_n(t) - x(t)| < \epsilon$ for all $n > N(\epsilon)$ and all t in I .

Solution: Here $F(t, x) = t + x$ and $x_0(t) = 0$, so we get

$$x_1(t) = 0 + \int_0^t F(s, x_0(s)) ds = \int_0^t (s + 0) ds = \frac{1}{2}t^2$$

$$x_2(t) = 0 + \int_0^t F(s, x_1(s)) ds = \int_0^t \left(s + \frac{1}{2}s^2\right) ds = \frac{1}{2}t^2 + \frac{1}{3!}t^3$$

$$x_3(t) = 0 + \int_0^t F(s, x_2(s)) ds = \int_0^t \left(s + \frac{1}{2}s^2 + \frac{1}{3!}s^3\right) ds = \frac{1}{2}t^2 + \frac{1}{3!}t^3 + \frac{1}{4!}t^4$$

By induction on n one can verify that the general expression for $x_n(t)$ is

$$x_n(t) = \frac{1}{2!}t^2 + \frac{1}{3!}t^3 + \cdots + \frac{1}{(n+1)!}t^{n+1} \quad (**)$$

But $e^t = 1 + \frac{1}{1!}t + \frac{1}{2!}t^2 + \frac{1}{3!}t^3 + \cdots$ (see e.g. EMEA). So as $n \rightarrow \infty$, we get $x_n(t) \rightarrow e^t - 1 - t$. The required solution of (*) is therefore $x(t) = \lim_{n \rightarrow \infty} x_n(t) = e^t - 1 - t$. (Check this solution by verifying directly that (*) is satisfied.)

Global Existence and Uniqueness

When analysing a dynamic model described by a differential equation, economists often simply assume that a solution exists throughout whatever interval is relevant. Is this justified?

Consider, for example, the standard growth model of Example 5.7.3 that leads to the equation $\dot{k} = sf(k) - \lambda k$ for the capital/labour ratio. Is there a solution on the whole interval $[0, \infty)$? The following example provides a sufficient condition:

E 6 Consider the initial value problem $\dot{x} = F(t, x)$, $x(t_0) = x_0$ in Theorem 5.8.2, and suppose that F and F_x are continuous everywhere. Assume too that F is uniformly bounded, i.e. there exists a number M such that $|F(t, x)| \leq M$ for all (t, x) . Then Theorem 5.8.2 applies with $r = \min(a, b/M)$. But under these assumptions r can be made arbitrarily large by choosing a and b sufficiently large. Hence, there exists a unique solution $x(t)$ defined on the whole of $(-\infty, \infty)$, i.e. we have **global** existence.

A much more powerful result (see Hartman (1982)) is the following:

5.8.3 (GLOBAL EXISTENCE AND UNIQUENESS)

Consider the initial value problem

$$\dot{x} = F(t, x), \quad x(t_0) = x_0$$

Suppose that $F(t, x)$ and $F_x(t, x)$ are continuous for all (t, x) . Suppose too that there exist continuous functions $a(t)$ and $b(t)$ such that

$$|F(t, x)| \leq a(t)|x| + b(t) \quad \text{for all } (t, x) \quad (3)$$

Given an arbitrary point (t_0, x_0) , there exists a unique solution $x(t)$ of the initial value problem, defined on $(-\infty, \infty)$. If (3) is replaced by the weaker condition

$$xF(t, x) \leq a(t)|x|^2 + b(t) \quad \text{for all } x \text{ and for all } t \geq t_0 \quad (4)$$

then the initial value problem has a unique solution defined on $[t_0, \infty)$.

EXAMPLE 7 Examine whether Theorem 5.8.3 applies to the problem $\dot{x} = -x^3$, $x(1) = 1$.

Solution: Clearly (3) is not satisfied. But $xF(t, x) = x(-x^3) = -x^4 \leq 0$, so (4) is satisfied with $a(t) = b(t) = 0$. Hence, there exists a solution on $[1, \infty)$. (In fact, this separable equation has the unique solution $x(t) = (2t - 1)^{-1/2}$. This solution is valid only for $t > 1/2$, not for all of $(-\infty, \infty)$.)

EXAMPLE 8 Consider once again the growth model of Example 5.7.3, with

$$\dot{k} = sf(k) - \lambda k, \quad k(0) = k_0 > 0 \quad (*)$$

Suppose that $f'(0) < \infty$, $f(0) = 0$, $f'(k) \rightarrow 0$ as $k \rightarrow \infty$, and $f''(k) \leq 0$ for all $k \geq 0$. This implies that $f'(k) \leq f'(0)$ for all $k \geq 0$, and the phase diagram is as in Fig. 5.7.4 with two equilibrium states, 0 and k^* . Define $F(k)$ for all k by

$$F(k) = \begin{cases} sf(k) - \lambda k, & k \geq 0 \\ sf'(0)k - \lambda k, & k < 0 \end{cases}$$

Note that for $k \geq 0$, $F(k)$ equals $\dot{k}$, as given by (*). Also, for $k \geq 0$, we have $F'(k) = sf'(k) - \lambda \leq sf'(0) - \lambda$, whereas for $k < 0$ we have $F'(k) = sf'(0) - \lambda$. Furthermore, $F'(k) \geq -\lambda$ for all k . Therefore the equation $\dot{k} = F(k)$ satisfies condition (5) with $c(t) = \max(\lambda, sf'(0) - \lambda)$. We conclude that the equation has a unique solution on $(-\infty, \infty)$.

The functions $k_1(t) \equiv 0$ and $k_2(t) \equiv k^*$ are both solutions of (*). Let $k(t)$ be any solution with $k(0) = k_0 \in (0, k^*)$. Note that uniqueness implies that no two solution curves can intersect, otherwise there would be more than one solution passing through any point of intersection. Hence, $k(t)$ will always lie in the interval $(0, k^*)$, as in the discussion that follows Theorem 5.7.1.

NOTE 4 Condition (3) will be satisfied if there exists a continuous function $c(t)$ of t such that

$$|F'_x(t, x)| \leq c(t) \quad \text{for all } (t, x) \quad (5)$$

Indeed, by the mean value theorem (see Theorem 2.1.2), $F(t, x) - F(t, 0) = F'_x(t, \theta)x$ for some θ in $[0, x]$. Hence, (5) implies that $|F(t, x)| = |F'_x(t, \theta)x + F(t, 0)| \leq |F'_x(t, \theta)||x| + |F(t, 0)| \leq a(t)|x| + b(t)$ if we let $a(t) = c(t)$ and $b(t) = |F(t, 0)|$.

NOTE 5 Condition (3) states that $|\dot{x}|$ is bounded by a linear function of $|x|$. The requirement in (4) is more complicated. As elaborated in Note 6 below, the explanation is that $x(t)$ cannot “explode” unless the product $xF(t, x) = x\dot{x} = \frac{d}{dt}(\frac{1}{2}x^2)$ is large and positive when $x(t)^2$ is large. If $x(t)$ is large and $\dot{x}(t) = F(t, x(t))$ is also large, then we are in trouble. However, if $x(t)$ is large and $\dot{x}(t) = F(t, x(t))$ is negative, then $x(t)$ decreases as t increases.

Condition (3) is actually stronger than (4). To see why, note first that $|x| \leq 1 + |x|^2$ for all x (because $u \leq 1 + u^2$ for all u). Then (3) implies that

$$\begin{aligned} xF(t, x) &\leq |x||F(t, x)| \leq a(t)|x|^2 + b(t)|x| \\ &\leq a(t)|x|^2 + b(t)(1 + |x|^2) = (a(t) + b(t))|x|^2 + b(t) \end{aligned}$$

which is inequality (4) with $a(t)$ replaced by $a(t) + b(t)$. Example 7 presented a case where (4) is satisfied, but not (3).

NOTE 6 Here is a brief explanation of why the weaker condition (4) ensures global existence for $t \geq t_0$: any solution must satisfy $(d/dt)(x(t)^2) = 2x(t)\dot{x}(t) \leq 2a(t)x(t)^2 + 2b(t)$. By (5.4.6), the linear equation $\dot{y}(t) = 2a(t)y(t) + 2b(t)$ with $y(t_0) = x(t_0)^2$ has a solution for all $t \geq t_0$. Let $z(t) = x(t)^2 - y(t)$. Then $z(t_0) = 0$ and $\dot{z}(t) \leq 2a(t)z(t)$ for $t \geq t_0$. If we let $A(t) = 2 \int_{t_0}^t a(s) ds$, then $(d/dt)[e^{-A(t)}z(t)] = e^{-A(t)}[\dot{z}(t) - 2a(t)z(t)] \leq 0$ for $t \geq t_0$. It follows that $z(t) \leq 0$, and so $x(t)^2 \leq y(t)$ for all $t \geq t_0$. Therefore, $x(t)^2$ does not explode. This makes it plausible that $x(t)$ is

NOTE 7 Suppose we drop the requirement in Theorems 5.8.1 to 5.8.3 that $F'_x(t, x)$ be continuous, but retain all the other conditions. Then a solution will still exist, but may not be unique. To ensure uniqueness, the following weaker requirement is sufficient: $F(t, x)$ is **locally Lipschitz continuous** w.r.t. x in A in the sense that, for each (t, x) in A , there exists a neighbourhood N of (t, x) in $\mathbb{R}^2$ and a constant L such that $|F(t, x') - F(t, x'')| \leq L|x' - x''|$ whenever (t, x') and (t, x'') belong to N .

MS FOR SECTION 5.8

1. Show that $x = Ct^2$ satisfies the differential equation $tx' = 2x$ for all values of the constant C . But all the corresponding solution curves pass through the point $(0, 0)$. How do you reconcile this observation with Theorem 5.8.1?
2. Use Theorem 5.8.2 to show that $\dot{x} = t^2 + e^{-x^2}$, $x(0) = 0$, has a unique solution on the interval $(-a, a)$ for every positive constant a .
3. Use Picard's method of successive approximations to solve the equation $\dot{x} = x$ with $x(0) = 1$. (Hint: Consider the Taylor expansion of e^x . See Example 5.)
4. Find the unique solution of $\dot{x} = x(1 - x)$, $x(0) = 1/2$, defined on $(-\infty, \infty)$. Show that neither of the conditions (3) or (4) in Theorem 5.8.3 is satisfied.
5. Let a and b be arbitrary constants, $a < b$, and define the function φ by

$$\varphi(t) = \begin{cases} -(t-a)^2 & \text{if } t \leq a \\ 0 & \text{if } a < t < b \\ (t-b)^2 & \text{if } t \geq b \end{cases}$$

Sketch the graph of φ when $a = -2$ and $b = 3$. Use the definition of the derivative to show that φ is differentiable at $t = a$ and $t = b$. For all choices of a and b prove that $x = \varphi(t)$ is a solution of the differential equation $\dot{x} = 2\sqrt{|x|}$ on the whole real line. Explain why this shows that the requirement in Theorem 5.8.1 that $F(t, x)$ is differentiable w.r.t. x cannot be dropped.

6

DIFFERENTIAL EQUATIONS II: SECOND-ORDER EQUATIONS AND SYSTEMS IN THE PLANE

Understanding of mathematics cannot be transmitted by painless entertainment any more than education in music can be brought by the most brilliant journalism to those who have never listened intensively. Actual contact with the content of living mathematics is necessary.
—Richard Courant (1941)

In Chapter 5 we studied only first-order differential equations. Yet many economic models are based on differential equations in which second- or higher-order derivatives appear. For example, in an important area of dynamic optimization called the *calculus of variations*, the first-order condition for optimality involves a second-order differential equation. (See Section 8.2.)

Sections 6.1–6.4 of this chapter treat the standard theory of second-order linear equations in one variable. Next, Sections 6.5 and 6.6 are devoted to systems of two simultaneous differential equations in two variables. When there are two variables, the phase plane techniques covered in Section 6.7 provide useful insights concerning the form of the solutions, and especially their long-run (asymptotic) behaviour. Section 6.8 discusses stability properties for nonlinear systems in the plane, which are important in macroeconomic theory. Saddle points, which occur in a large number of economic models, are the topic of Section 6.9.

6.1 Introduction

The typical second-order differential equation takes the form

$$\ddot{x} = F(t, x, \dot{x}) \quad (1)$$

where F is a given fixed function, $x = x(t)$ is the unknown function, and $\dot{x} = dx/dt$. Compared with Chapter 5, the new feature is the presence of the second derivative $\ddot{x} = d^2x/dt^2$. A **solution** of (1) on an interval I is a twice differentiable function that satisfies the equation.

The simplest type of second-order equation appears in the following example.

EXAMPLE 1

Find all solutions of

$$\ddot{x} = k \quad (k \text{ is a constant})$$

Solution: Because $\ddot{x} = (d/dt)\dot{x}$, direct integration implies that the equation is equivalent to $\dot{x} = \int k dt = kt + A$, for some constant A . After integrating once more, we see that the equation is satisfied iff $x = \frac{1}{2}kt^2 + At + B$. Geometrically, the solution represents for $k \neq 0$ a collection of parabolas in the tx -plane whose axes are all parallel to the x -axis.

Differential Equations where x or t is Missing

In two special cases the solution of equation (1) can be reduced to the solution of first-order equations. The two cases are

$$(a) \ddot{x} = F(t, \dot{x}) \quad (b) \ddot{x} = F(x, \dot{x}) \quad (2)$$

In case (2)(a), x is missing. We introduce the new variable $u = \dot{x}$. Then (a) becomes $\dot{u} = F(t, u)$, which is a first-order equation. If we find the general solution $u(t)$ of this first-order equation, then integrating $\dot{x}(t) = u(t)$ will yield the general solution $x(t)$ of (a).

In case (2)(b), t is not explicitly present in the equation, and the equation is called **autonomous**. Problem 6 indicates how to interchange t and x in order to transform the equation into one having the form (2)(a).

LE 2 Solve the equation $\ddot{x} = \dot{x} + t$.

Solution: Substituting $u = \dot{x}$ yields $\dot{u} = u + t$. This first-order equation has the general solution $u = Ae^t - t - 1$, where A is a constant (see Problem 5.4.3). Hence, $\dot{x} = Ae^t - t - 1$. Integrating this equation yields $x = \int (Ae^t - t - 1) dt = Ae^t - \frac{1}{2}t^2 - t + B$, where B is a second arbitrary constant.

Solving equation (1) becomes more difficult if the right-hand side includes t , the unknown function x , and its derivative $\dot{x}$. In fact, only rather special cases have explicit solutions; generally, one has to resort to numerical solutions for given initial conditions. Even so, it turns out that the *existence* of a solution of (1) can be established for almost all the equations that are likely to appear in applications. In fact, the general solution of the equation will depend on two arbitrary constants, as it did in Examples 1 and 2; that is, the solution is of the form $x = x(t; A, B)$.

In an *initial value problem*, there are specified values x_0 and $\dot{x}_0$ such that $x(t_0) = x_0$ and $\dot{x}(t_0) = \dot{x}_0$ at an "initial" time $t = t_0$. The two conditions $x(t_0, A, B) = x_0$ and $\dot{x}(t_0, A, B) = \dot{x}_0$ usually determine the constants A and B uniquely.

E 3 Solve the initial value problem $\ddot{x} = \dot{x} + t$, $x(0) = 1$, $\dot{x}(0) = 2$.

Solution: According to Example 2, the general solution of this second-order equation is $x = Ae^t - \frac{1}{2}t^2 - t + B$. Letting $x(0) = 1$ yields $1 = A + B$. Moreover, $\dot{x} = Ae^t - t - 1$, so $\dot{x}(0) = 2$ implies that $2 = A - 1$. Thus, $A = 3$ and $B = -2$, so the unique solution of the problem is $x = 3e^t - \frac{1}{2}t^2 - t - 2$.

PROBLEMS FOR SECTION 6.1

1. Find the general solutions of the following differential equations:

$$(a) \ddot{x} = t \quad (b) \ddot{x} = \sin t \quad (c) \ddot{x} = e^t + t^2$$

2. Solve the initial value problem $\ddot{x} = t^2 - t$, $x(0) = 1$, $\dot{x}(0) = 2$.

3. Solve the problem (see Example 2) $\ddot{x} = \dot{x} + t$, $x(0) = 1$, $x(1) = 2$. (In this case the constants are determined by the value of $x(t)$ at two different points of time.)

SM 4. Solve the following differential equations:

$$(a) \ddot{x} + 2\dot{x} = 8 \quad (b) \ddot{x} - 2\dot{x} = 2e^{2t} \quad (c) \ddot{x} - \dot{x} = t^2$$

5. Suppose $y > 0$ denotes wealth, and $u(y)$ is a C^2 utility function with $u'(y) > 0$ and $u''(y) < 0$ for all $y > 0$. The (positive) ratio $R_A = -u''(y)/u'(y)$ is called the **degree of absolute risk aversion**, and $R_R = yR_A$ is called the **degree of relative risk aversion**.

(a) Find an expression for $u(y)$ if $R_A = \lambda$, where λ is a constant.

(b) Find an expression for $u(y)$ if $R_R = k$, where k is a constant. Distinguish between the cases $k = 1$ and $k \neq 1$.

HARDER PROBLEMS

SM 6. (a) Consider the equation $\ddot{x} = F(x, \dot{x})$, where t is missing. A standard trick for solving the equation is to let x be the independent variable instead of t . Let t' and t'' denote the first and second derivatives of t w.r.t. x . Prove that, provided $\dot{x} \neq 0$, we have

$$\dot{x} = \frac{dx}{dt} = \frac{1}{dt/dx} \quad \text{and} \quad \ddot{x} = \frac{d\dot{x}}{dt} = -\frac{t''}{(t')^3}$$

so that the original differential equation is transformed to

$$t'' = -(t')^3 F(x, 1/t')$$

(Here t does not appear explicitly, so the method used to solve case (2)(a) will work.)

(b) Solve the equations (i) $\ddot{x} = -x\dot{x}^3$, (ii) $\ddot{x} = \dot{x}^2/x$.

7. The partial differential equation $u_{xx}(t, x) = u_t(t, x)$ (the "diffusion equation") appears in modern finance theory.

(a) Show that for every α , the function $u(t, x) = e^{i\alpha^2 t} e^{i\alpha x}$ is a solution of the equation.

(b) Suppose that the equation has a solution of the form $u(x, t) = g(y)$, where $y = x/\sqrt{t}$. Show that $g(y)$ then satisfies the equation $g''(y)/g'(y) = -\frac{1}{2}y$. Show that the solution of this equation is $g(y) = A \int e^{-\frac{1}{2}y^2} dy + B$, where A and B are constants.

6.2 Linear Differential Equations

The general second-order linear differential equation is

$$\ddot{x} + a(t)\dot{x} + b(t)x = f(t) \quad (1)$$

where $a(t)$, $b(t)$, and $f(t)$ are all continuous functions of t on some interval I . In contrast to first-order linear equations, there is no explicit solution of (1) in the general case. However, something useful can be said about the structure of the general solution.

Let us begin with the **homogeneous equation**

$$\ddot{x} + a(t)\dot{x} + b(t)x = 0 \quad (2)$$

obtained from (1) by replacing $f(t)$ by 0. We claim that if $u_1 = u_1(t)$ and $u_2 = u_2(t)$ both satisfy (2), then so does $x = Au_1 + Bu_2$ for all choices of constants A and B . In fact, since $\dot{x} = A\dot{u}_1 + B\dot{u}_2$ and $\ddot{x} = A\ddot{u}_1 + B\ddot{u}_2$, we have

$$\begin{aligned} \ddot{x} + a(t)\dot{x} + b(t)x &= A\ddot{u}_1 + B\ddot{u}_2 + a(t)(A\dot{u}_1 + B\dot{u}_2) + b(t)(Au_1 + Bu_2) \\ &= A[\ddot{u}_1 + a(t)\dot{u}_1 + b(t)u_1] + B[\ddot{u}_2 + a(t)\dot{u}_2 + b(t)u_2] \end{aligned}$$

It was assumed that both u_1 and u_2 satisfy (2), so the two expressions in square brackets are both 0. Thus, we have proved that the function $x = Au_1 + Bu_2$ satisfies (2) for all values of the constants A and B .

Suppose then that we have somehow managed to find two solutions u_1 and u_2 of (2). Does the general solution take the form $x = Au_1 + Bu_2$ for arbitrary constants A and B ? No, in order to be sure that $Au_1 + Bu_2$ is the general solution of (2), we must require u_1 and u_2 not to be constant multiples of each other—that is, they must not be proportional. (For a proof, see Section 7.1.)

Equation (1) is called a **nonhomogeneous equation**, and (2) is the homogeneous equation associated with it. Suppose we are able to find *some particular solution* $u^* = u^*(t)$ of (1). If $x(t)$ is an arbitrary solution of (1), then it is easy to see that the difference $x(t) - u^*(t)$ is a solution of the homogeneous equation (2). In fact, if $v = v(t) = x(t) - u^*(t)$, then $\dot{v} = \dot{x} - \dot{u}^*$ and $\ddot{v} = \ddot{x} - \ddot{u}^*$, so

$$\begin{aligned} \ddot{v} + a(t)\dot{v} + b(t)v &= \ddot{x} - \ddot{u}^* + a(t)(\dot{x} - \dot{u}^*) + b(t)(x - u^*) \\ &= \ddot{x} + a(t)\dot{x} + b(t)x - [\ddot{u}^* + a(t)\dot{u}^* + b(t)u^*] \\ &= f(t) - f(t) = 0 \end{aligned}$$

Thus, $x(t) - u^*(t)$ is a solution of the homogeneous equation. But then, according to the argument above, $x(t) - u^*(t) = Au_1(t) + Bu_2(t)$, where $u_1(t)$ and $u_2(t)$ are two nonproportional solutions of (2), and A and B are arbitrary constants. Conversely, if x is a function such that $x - u^*$ is a solution of the homogeneous equation, then x is a solution of the nonhomogeneous equation. All in all, we arrive at the following result:

THEOREM 6.2.1

- (a) The **general solution** of the homogeneous differential equation

$$\ddot{x} + a(t)\dot{x} + b(t)x = 0 \quad \text{is} \quad x = Au_1(t) + Bu_2(t)$$

where $u_1(t)$ and $u_2(t)$ are any two solutions that are not proportional, and A and B are arbitrary constants.

- (b) The **general solution** of the nonhomogeneous differential equation

$$\ddot{x} + a(t)\dot{x} + b(t)x = f(t) \quad \text{is} \quad x = Au_1(t) + Bu_2(t) + u^*(t)$$

where $Au_1(t) + Bu_2(t)$ is the general solution of the associated homogeneous equation (with $f(t)$ replaced by zero), and $u^*(t)$ is any **particular solution** of the nonhomogeneous equation.

EXAMPLE 1 Find the general solutions of (a) $\ddot{x} - x = 0$ and (b) $\ddot{x} - x = 5$.

Solution: (a) The problem is to find those functions that do not change when differentiated twice. You probably recall that $x = e^t$ has this property, as does $x = 2e^t$. But these two functions are proportional. So we need to find another function with the property that differentiating it twice leaves it unchanged. After some thought, you might come up with the idea of trying $x = e^{-t}$. In fact, $\dot{x} = -e^{-t}$, and so $\ddot{x} = e^{-t}$. Because e^t and e^{-t} are not proportional, the general solution is $x = Ae^t + Be^{-t}$, with A and B arbitrary constants.

(b) We need only find a particular solution of the equation. Obviously, $u(t) = -5$ will work. The general solution is therefore

$$x = Ae^t + Be^{-t} - 5 \quad (A \text{ and } B \text{ are arbitrary constants})$$

In the next section we shall give a general method for solving such equations. ■

There is no general method of discovering the two solutions of the homogeneous equation (2) that are needed for the general solution. However, in the special case when the coefficients $a(t)$ and $b(t)$ are both constants, it is always possible to find the two solutions required. The next section shows how to do this.

PROBLEMS FOR SECTION 6.2

- (a) Prove that $u_1 = e^t$ and $u_2 = te^t$ both satisfy $\ddot{x} - 2\dot{x} + x = 0$. Show that u_1 and u_2 are not proportional, and use this to find the general solution of the equation.
(b) Find the general solution of $\ddot{x} - 2\dot{x} + x = 3$.
- Show that $u_1 = \sin t$ and $u_2 = \cos t$ both are solutions of $\ddot{x} + x = 0$. What is the general solution of the equation?

3. (a) Prove that both $u_1 = e^{2t}$ and $u_2 = e^{-3t}$ are solutions of $\ddot{x} + \dot{x} - 6x = 0$. What is the general solution?

(b) Find the general solution of $\ddot{x} + \dot{x} - 6x = 6t$. (Hint: The equation has a particular solution of the form $Ct + D$.)

4. A study of the optimal exhaustion of a natural resource uses the equation

$$\ddot{x} - \frac{2-\alpha}{1-\alpha} \dot{x} + \frac{a^2}{1-\alpha} x = 0 \quad (\alpha \neq 0, \alpha \neq 1, a \neq 0)$$

Prove that $u_1 = e^{\alpha t}$ and $u_2 = e^{a^2/(1-\alpha)t}$ are both solutions. What is the general solution?

HARDER PROBLEMS

5. Let $a \neq b$ be two real numbers. Prove that the differential equation

$$(t+a)(t+b)\ddot{x} + 2(2t+a+b)\dot{x} + 2x = 0$$

has two solutions of the form $(t+k)^{-1}$ for appropriate choices of k . Find the general solution of the equation. (Hint: Let $x = (t+k)^{-1}$ and then adjust k until the function satisfies the differential equation.)

3 Constant Coefficients

Consider the *homogeneous* equation

$$\ddot{x} + a\dot{x} + bx = 0 \quad (1)$$

where a and b are arbitrary constants, and $x = x(t)$ is the unknown function. According to Theorem 6.2.1, finding the general solution of (1) requires us to discover two solutions $u_1(t)$ and $u_2(t)$ that are not proportional. Because the coefficients in (1) are constants, it seems a good idea to try possible solutions x with the property that x , $\dot{x}$, and $\ddot{x}$ are all constant multiples of each other. The exponential function $x = e^{rt}$ has this property, because $\dot{x} = re^{rt} = rx$ and $\ddot{x} = r^2e^{rt} = r^2x$. So we try adjusting the constant r in order that $x = e^{rt}$ satisfies (1). This requires us to arrange that $r^2e^{rt} + are^{rt} + be^{rt} = 0$. Cancelling the positive factor e^{rt} tells us that e^{rt} satisfies (1) iff r satisfies

$$r^2 + ar + b = 0 \quad (2)$$

This is the **characteristic equation** of the differential equation (1). It is a quadratic equation whose roots are real iff $\frac{1}{4}a^2 - b \geq 0$. Solving (2) by the quadratic formula in this case yields the two **characteristic roots**

$$r_1 = -\frac{1}{2}a + \sqrt{\frac{1}{4}a^2 - b}, \quad r_2 = -\frac{1}{2}a - \sqrt{\frac{1}{4}a^2 - b} \quad (3)$$

There are three different cases which are summed up in the following theorem:

THEOREM 6.3.1

The general solution of

$$\ddot{x} + a\dot{x} + bx = 0$$

depends on the roots of the characteristic equation $r^2 + ar + b = 0$ as follows:

(I) If $\frac{1}{4}a^2 - b > 0$, when there are two distinct real roots, then

$$x = Ae^{r_1 t} + Be^{r_2 t}, \quad \text{where } r_{1,2} = -\frac{1}{2}a \pm \sqrt{\frac{1}{4}a^2 - b}$$

(II) If $\frac{1}{4}a^2 - b = 0$, when there is a double real root, then

$$x = (A + Bt)e^{rt}, \quad \text{where } r = -\frac{1}{2}a$$

(III) If $\frac{1}{4}a^2 - b < 0$, when there are no real roots, then

$$x = e^{\alpha t} (A \cos \beta t + B \sin \beta t), \quad \alpha = -\frac{1}{2}a, \quad \beta = \sqrt{b - \frac{1}{4}a^2}$$

Proof: (I) The case $\frac{1}{4}a^2 - b > 0$ is the simplest because there are two distinct real characteristic roots r_1 and r_2 . The two functions $e^{r_1 t}$ and $e^{r_2 t}$ both satisfy (1), and are not proportional. So the general solution in this case is $Ae^{r_1 t} + Be^{r_2 t}$.

(II) If $\frac{1}{4}a^2 - b = 0$, then $r = -\frac{1}{2}a$ is a double root of (2), and $u_1 = e^{rt}$ satisfies (1). We claim that $u_2 = te^{rt}$ also satisfies (1). (See also Problem 6.) This is because $\ddot{u}_2 = e^{rt} + 2re^{rt}$ and $\ddot{u}_2 = re^{rt} + re^{rt} + tr^2e^{rt}$, which inserted into the left-hand side of (1) gives

$$\ddot{u}_2 + a\dot{u}_2 + bu_2 = e^{rt}(a + 2r) + te^{rt}(r^2 + ar + b)$$

after simplifying. But the last expression is 0 because $r = -\frac{1}{2}a$ and $r^2 + ar + b = 0$. Thus, e^{rt} and te^{rt} are indeed both solutions of equation (1). These two solutions are not proportional, so the general solution is $Ae^{rt} + Bte^{rt}$ in this case.

(III) If $\frac{1}{4}a^2 - b < 0$, the characteristic equation has no real roots. An example is the equation $\ddot{x} + x = 0$, which occurred in Problem 6.2.2; here $a = 0$ and $b = 1$, so $\frac{1}{4}a^2 - b = -1$. The general solution was $A \sin t + B \cos t$. It should, therefore, come as no surprise that when $\frac{1}{4}a^2 - b < 0$, the solution of (1) involves trigonometric functions.

Define the two functions $u_1(t) = e^{\alpha t} \cos \beta t$ and $u_2(t) = e^{\alpha t} \sin \beta t$, where α and β are defined in (III). We claim that both these functions satisfy (1). Since they are not proportional, the general solution of equation (1) in this case is as exhibited in (III).

Let us show that $u_1(t) = e^{\alpha t} \cos \beta t$ satisfies (1). We find that $\dot{u}_1(t) = \alpha e^{\alpha t} \cos \beta t - \beta e^{\alpha t} \sin \beta t$. Furthermore, $\ddot{u}_1(t) = \alpha^2 e^{\alpha t} \cos \beta t - \alpha \beta e^{\alpha t} \sin \beta t - \alpha \beta e^{\alpha t} \sin \beta t - \beta^2 e^{\alpha t} \cos \beta t$. Hence, $\ddot{u}_1 + a\dot{u}_1 + bu_1 = e^{\alpha t}[(\alpha^2 - \beta^2 + \alpha a + b) \cos \beta t - \beta(2\alpha + a) \sin \beta t]$. By using the specific values of α and β given in (III), we see that $2\alpha + a = 0$ and $\alpha^2 - \beta^2 + \alpha a + b = 0$.

$\frac{1}{4}a^2 - (b - \frac{1}{4}a^2) - \frac{1}{2}a^2 + b = 0$. This shows that $u_1(t) = e^{at} \cos \beta t$ satisfies equation (1). A similar argument shows that $u_2(t) = e^{at} \sin \beta t$ satisfies (1) as well. ■

NOTE 1 When $\frac{1}{4}a^2 - b < 0$, i.e. in case III, an alternative form of the solution is $x = Ce^{at} \cos(\beta t + D)$. (See Problem 5.)

NOTE 2 If we use complex numbers (see Section B.3), then if $a^2/4 < b$ the solutions of the characteristic equation $r^2 + ar + b = 0$ can be written as $r_{1,2} = \alpha \pm i\beta$, where $\alpha = -a/2$ and $\beta = \sqrt{b - a^2/4}$ are precisely the real numbers occurring in the solution in case III.

With $r_{1,2} = \alpha \pm i\beta$, the two complex exponential functions $e^{r_1 t} = e^{\alpha t} (\cos \beta t + i \sin \beta t)$ and $e^{r_2 t} = e^{\alpha t} (\cos \beta t - i \sin \beta t)$ both satisfy (1), as stated in case (III). But so does any linear combination of these solutions. In particular, $(e^{r_1 t} + e^{r_2 t})/2 = e^{\alpha t} \cos \beta t$ and $(e^{r_1 t} - e^{r_2 t})/2i = e^{\alpha t} \sin \beta t$ both satisfy (1), as stated in case (III).

LE 1 Find the general solutions of the following equations:

$$(a) \quad \ddot{x} - 3x = 0 \quad (b) \quad \ddot{x} - 4\dot{x} + 4x = 0 \quad (c) \quad \ddot{x} - 6\dot{x} + 13x = 0$$

Solution: (a) The characteristic equation $r^2 - 3 = 0$ has two real roots $r_1 = -\sqrt{3}$ and $r_2 = \sqrt{3}$. The general solution is

$$x = Ae^{-\sqrt{3}t} + Be^{\sqrt{3}t}$$

(b) The characteristic equation $r^2 - 4r + 4 = (r - 2)^2 = 0$ has the double root $r = 2$. Hence, the general solution is

$$x = (A + Bt)e^{2t}$$

(c) The characteristic equation $r^2 - 6r + 13 = 0$ has no real roots. According to case (III), $\alpha = -a/2 = -(-6)/2 = 3$ and $\beta = \sqrt{13 - \frac{1}{4}(-6)^2} = 2$, so the general solution is

$$x = e^{3t}(A \cos 2t + B \sin 2t)$$

The Nonhomogeneous Equation

Consider next the *nonhomogeneous* equation

$$\ddot{x} + a\dot{x} + bx = f(t) \quad (4)$$

where $f(t)$ is an arbitrary continuous function. According to Theorem 6.2.1(b), the general solution of (4) is given by

$$x = x(t) = Au_1(t) + Bu_2(t) + u^*(t) \quad (5)$$

We have explained how to find the term $Au_1(t) + Bu_2(t)$ by solving the corresponding homogeneous equation. But how do we find a particular solution $u^* = u^*(t)$ of (4)? In fact, there is a simple method of *undetermined coefficients* that works in many cases.

If $b = 0$ in (4), then the term in x is missing and the substitution $u = \dot{x}$ transforms the equation into a linear equation of the first order (see Example 6.1.2). So we may assume

(A) $f(t) = A$ (constant)

In this case we check to see if (4) has a solution that is constant, $u^* = c$. Then $\dot{u}^* = \ddot{u}^* = 0$, so the equation reduces to $bc = A$. Hence, $c = A/b$. Thus, for $b \neq 0$:

$$\ddot{x} + a\dot{x} + bx = A \quad \text{has a particular solution } u^* = A/b \quad (6)$$

(B) $f(t)$ is a polynomial

Suppose $f(t)$ is a polynomial of degree n . Then a reasonable guess is that (4) has a particular solution that is also a polynomial of degree n , of the form $u^* = A_n t^n + A_{n-1} t^{n-1} + \cdots + A_1 t + A_0$. We determine the undetermined coefficients $A_n, A_{n-1}, \dots, A_0$ by requiring u^* to satisfy (4) and equating coefficients of like powers of t .

EXAMPLE 2

Find a particular solution of $\ddot{x} - 4\dot{x} + 4x = t^2 + 2$.

Solution: The right-hand side is a polynomial of degree 2. So we let $u^* = At^2 + Bt + C$ and try adjusting A, B , and C to give a solution. We obtain $\dot{u}^* = 2At + B$, and so $\ddot{u}^* = 2A$. Inserting these expressions for $u^*, \dot{u}^*$, and $\ddot{u}^*$ into the equation yields $2A - 4(2At + B) + 4(At^2 + Bt + C) = t^2 + 2$. Collecting like terms on the left-hand side gives $4At^2 + (4B - 8A)t + (2A - 4B + 4C) = t^2 + 2$. Since this must hold for all t , we can equate coefficients of like powers of t to obtain $4A = 1$, $4B - 8A = 0$, and $2A - 4B + 4C = 2$. Solving these three equations gives $A = \frac{1}{4}$, $B = \frac{1}{2}$, and $C = \frac{7}{8}$. Hence, a particular solution is

$$u^* = \frac{1}{4}t^2 + \frac{1}{2}t + \frac{7}{8}$$

Note that the right-hand side of the given equation is $t^2 + 2$, without any t term. Yet no function of the form $Ct^2 + D$ will satisfy it; any solution must include the term $\frac{1}{2}t$. ■

EXAMPLE 3

For some differential equations in the theory of option pricing, the independent variable is the current stock price rather than time. A typical example is

$$f''(x) + af'(x) + bf(x) = \alpha x + \beta \quad (*)$$

Here $f(x)$ denotes the value of a stock option when the stock price is x . For the case of a “call option”, offering the right to buy a stock at a fixed “strike price”, the constant b is usually negative. Solve the equation in this case.

Solution: When $b < 0$, the characteristic equation has two distinct roots $r_{1,2} = -\frac{1}{2}a \pm \sqrt{\frac{1}{4}a^2 - b}$, so the homogeneous equation has the general solution

$$f(x) = Ae^{r_1 x} + Be^{r_2 x} \quad (A \text{ and } B \text{ are constants})$$

To find a particular solution $u(x)$ of (*), we try $u(x) = Px + Q$. Then $u'(x) = P$ and $u''(x) = 0$. Inserting these into (*) gives $aP + b(Px + Q) = \alpha x + \beta$, and so

$bPx + (aP + bQ) = \alpha x + \beta$. Hence $P = \alpha/b$ and $Q = (\beta b - \alpha a)/b^2$, so a particular solution is $u(x) = \alpha x/b + (\beta b - \alpha a)/b^2$. The general solution of (*) is therefore

$$f(x) = Ae^{r_1 x} + Be^{r_2 x} + \frac{\alpha}{b}x + \frac{\beta b - \alpha a}{b^2}, \quad r_{1,2} = -\frac{1}{2}a \pm \sqrt{\frac{1}{4}a^2 - b}$$

(C) $f(t) = pe^{qt}$

It seems natural to try a particular solution of the form $u^* = Ae^{qt}$. Then $\dot{u}^* = Aqe^{qt}$ and $\ddot{u}^* = Aq^2e^{qt}$. Substituting these into (4) yields $Ae^{qt}(q^2 + aq + b) = pe^{qt}$. Hence, if $q^2 + aq + b \neq 0$,

$$\ddot{x} + a\dot{x} + bx = pe^{qt} \quad \text{has the particular solution} \quad u^* = \frac{p}{q^2 + aq + b} e^{qt} \quad (7)$$

The condition $q^2 + aq + b \neq 0$ means that q is not a solution of the characteristic equation (2)—that is, that e^{qt} is not a solution of (1). If q is a simple root of $q^2 + aq + b = 0$, we look for a constant B such that Bte^{qt} is a solution. If q is a double root, then Ct^2e^{qt} is a solution for some constant C .

(D) $f(t) = p \sin rt + q \cos rt$

Again the method of undetermined coefficients works. Let $u^* = A \sin rt + B \cos rt$ and adjust the constants A and B so that the coefficients of $\sin rt$ and $\cos rt$ match. If $f(t)$ is itself a solution of the homogeneous equation, then $u^* = At \sin rt + Bt \cos rt$ will be a particular solution for suitable choices of constants A and B .

IE 4 Find a particular solution of $\ddot{x} - 4\dot{x} + 4x = 2 \cos 2t$.

Solution: Here it might seem natural to try a particular solution of the form $u = A \cos 2t$. Note, however, that the term $-4\dot{u}$ gives us a $\sin 2t$ term on the left-hand side, and no matching term occurs on the right-hand side of the equation. So we try $u^* = A \sin 2t + B \cos 2t$ instead, and adjust constants A and B appropriately. We have $\dot{u}^* = 2A \cos 2t - 2B \sin 2t$ and $\ddot{u}^* = -4A \sin 2t - 4B \cos 2t$. Inserting these expressions into the equation and rearranging, we get $8B \sin 2t - 8A \cos 2t = 2 \cos 2t$. Thus, letting $B = 0$ and $A = -1/4$, we see that $u^* = (-1/4) \sin 2t$ is a particular solution of the equation.

The technique described for obtaining particular solutions also applies if $f(t)$ is a sum, difference, or product of polynomials, exponential functions, or trigonometric functions of the type mentioned. For instance, if $f(t) = (t^2 + 1)e^{3t} + \sin 2t$, one can try $u^* = (At^2 + Bt + C)e^{3t} + D \sin 2t + E \cos 2t$. On the other hand, if $f(t)$ is an entirely different type of function such as $t \ln t$, the method of undetermined coefficients usually does not work.

Euler's Differential Equation

One type of equation that occasionally occurs in economics is Euler's differential equation,

$$(t^2 \ddot{x} + at\dot{x} + bx = 0, \quad t > 0) \quad (8)$$

One way to proceed is to look for a constant r such that $x = t^r$ satisfies the equation. If this is to work, inserting $\dot{x} = rt^{r-1}$ and $\ddot{x} = r(r-1)t^{r-2}$ into equation (8) gives

$$t^2 r(r-1)t^{r-2} + atrr^{r-1} + bt^r = 0, \quad \text{or} \quad r^2(r^2 + (a-1)r + b) = 0$$

So $x = t^r$ is a solution if

$$r^2 + (a-1)r + b = 0 \quad (9)$$

The solutions of this equation are

$$r_{1,2} = -\frac{1}{2}(a-1) \pm \frac{1}{2}\sqrt{(a-1)^2 - 4b} \quad (10)$$

If $(a-1)^2 > 4b$, equation (9) has two different real solutions r_1 and r_2 , and the general solution of (8) is

$$x = At^{r_1} + Bt^{r_2} \quad (A \text{ and } B \text{ are arbitrary constants}) \quad (11)$$

(Note that t^{r_1} and t^{r_2} are not proportional when $r_1 \neq r_2$.)

If $(a-1)^2 = 4b$, equation (9) has a double root $r = \frac{1}{2}(1-a)$. Then $u_1 = t^r$ satisfies equation (8), but how does one find another solution? It turns out that $u_2 = (\ln t)t^r$ is another solution (see Problem 7), and so the general solution is:

$$x = (A + B \ln t)t^{(1-a)/2} \quad (12)$$

If $(a-1)^2 < 4b$, the roots of (9) are complex. For instance, for the equation $t^2 \ddot{x} + t\dot{x} = 0$, equation (9) reduces to $r^2 + 1 = 0$, so $r_1 = -i$, $r_2 = i$, where i is the imaginary unit, but it is not at all clear how to interpret t^i . In fact, considerable work is needed to show that the solution in this case is

$$x(t) = t^\alpha [A \cos(\beta \ln t) + B \sin(\beta \ln t)] \quad (13)$$

where $\alpha = \frac{1}{2}(1-a)$ and $\beta = \frac{1}{2}\sqrt{4b - (a-1)^2}$.

EXAMPLE 5 Solve the equation $t^2 \ddot{x} + t\dot{x} - x = 0$.

Solution: Equation (9) reduces to $r^2 - 1 = 0$, so $r_1 = -1$ and $r_2 = 1$, so the general solution is $x = At^{-1} + Bt$.

EXAMPLE 6 In the theory of option pricing one encounters the equation

$$x^2 f''(x) + axf'(x) + bf(x) = \alpha x + \beta$$

where $f(x)$ denotes the value of a stock option when the price of the stock is x . If $(a-1)^2 > 4b$ (which is often the case in option models), the homogeneous equation has the solution

$$f(x) = Ax^{r_1} + Bx^{r_2}$$

where $r_{1,2} = -\frac{1}{2}(a-1) \pm \frac{1}{2}\sqrt{(a-1)^2 - 4b}$. We easily find $u^*(x) = \alpha x/(a+b) + \beta/b$.

NOTE 3 An alternative way of solving (8) is to introduce a new independent variable $s = \ln t$, i.e., $t = e^s$. Then $\dot{x} = dx/dt = (dx/ds)(ds/dt) = (1/t)(dx/ds)$, and differentiating $\dot{x}$ with respect to t yields

$$\ddot{x} = \frac{d}{dt} \dot{x} = \frac{d}{dt} \left(\frac{1}{t} \frac{dx}{ds} \right) = -\frac{1}{t^2} \frac{dx}{ds} + \frac{1}{t} \frac{d}{dt} \left(\frac{dx}{ds} \right) = -\frac{1}{t^2} \frac{dx}{ds} + \frac{1}{t} \frac{d^2x}{ds^2} \frac{1}{t}$$

because

$$\frac{d}{dt} \left(\frac{dx}{ds} \right) = \frac{d}{ds} \left(\frac{dx}{ds} \right) \frac{ds}{dt} = \frac{d^2x}{ds^2} \cdot \frac{1}{t}$$

Inserting these expressions into (8) yields

$$\frac{d^2x}{ds^2} + (a-1) \frac{dx}{ds} + bx = 0 \quad (14)$$

This is an ordinary second-order equation for $x(t)$ with constant coefficients.

MS FOR SECTION 6.3

Find the general solutions of the equations in Problems 1 and 2.

1. (a) $\ddot{x} - 3x = 0$ (b) $\ddot{x} + 4\dot{x} + 8x = 0$ (c) $3\ddot{x} + 8\dot{x} = 0$
 (d) $4\ddot{x} + 4\dot{x} + x = 0$ (e) $\ddot{x} + \dot{x} - 6x = 8$ (f) $\ddot{x} + 3\dot{x} + 2x = e^{5t}$

2. (a) $\ddot{x} - x = \sin t$ (b) $\ddot{x} - x = e^{-t}$ (c) $3\ddot{x} - 30\dot{x} + 75x = 2t + 1$

3. Solve the following differential equations for the specific initial conditions:

- (a) $\ddot{x} + 2\dot{x} + x = t^2$, $x(0) = 0$, $\dot{x}(0) = 1$
 (b) $\ddot{x} + 4x = 4t + 1$, $x(\pi/2) = 0$, $\dot{x}(\pi/2) = 0$

4. Find a particular solution of the differential equation

$$\ddot{L} + \gamma[\beta + \alpha(1 - \beta)]\dot{L} - \gamma\delta^*L = -\gamma\delta^*kt - \gamma\delta^*L_0 \quad (\gamma\delta^* \neq 0)$$

and then discuss when the general solution oscillates.

5. Prove that the general solution of $\ddot{x} + a\dot{x} + bx = 0$ in the case $\frac{1}{4}a^2 - b < 0$ can be written as $x = Ce^{\alpha t} \cos(\beta t + D)$, where C and D are arbitrary constants, $\alpha = -\frac{1}{2}a$, and $\beta = \frac{1}{2}\sqrt{4b - a^2}$.
6. Consider the equation $\ddot{x} + a\dot{x} + bx = 0$ when $\frac{1}{4}a^2 - b = 0$, so that the characteristic equation has a double root $r = -a/2$. Let $x(t) = u(t)e^{rt}$ and prove that this function is a solution if and only if $\ddot{u} = 0$. Conclude that the general solution is $x = (A + Bt)e^{rt}$.
7. Show that if $(a-1)^2 = 4b$, then $u_2 = (\ln t)t^{(1-a)/2}$ is a solution of (8).

8. Find the general solutions of the following equations for $t > 0$:

- (a) $t^2\ddot{x} + 5t\dot{x} + 3x = 0$ (b) $t^2\ddot{x} - 3t\dot{x} + 3x = t^2$

10. A business cycle model due to F. Dresch incorporates the equation

$$\dot{p}(t) = a \int_{-\infty}^t [D(p(\tau)) - S(p(\tau))] d\tau \quad (a > 0) \quad (*)$$

where $p(t)$ denotes a price index at time t , and $D(p)$ and $S(p)$ are aggregate demand and supply, respectively. Thus, (*) says that the rate of price increase is proportional to the accumulated total of all past excess demand. In the case when $D(p) = d_0 + d_1p$ and $S(p) = s_0 + s_1p$, where $d_1 < 0$ and $s_1 > 0$, differentiate (*) w.r.t. t in order to deduce a second-order differential equation for $p(t)$. Then find the general solution of this equation.

6.4 Stability for Linear Equations

Suppose the variables of an economic model change over time according to some differential equation (or system of differential equations). If appropriate regularity and initial conditions are imposed, there is a unique solution of the system. Also, if one or more initial conditions are changed, the solution changes. An important question is this: will small changes in the initial conditions have any effect on the long-run behaviour of the solution, or will the effect "die out" as $t \rightarrow \infty$? In the latter case the system is called **asymptotically stable**. On the other hand, if small changes in the initial conditions might lead to significant differences in the behaviour of the solution in the long run, then the system is **unstable**.

Consider in particular the second-order nonhomogeneous differential equation

$$\ddot{x} + a(t)\dot{x} + b(t)x = f(t) \quad (1)$$

Recall that the general solution of (1) is $x = Au_1(t) + Bu_2(t) + u^*(t)$, where $Au_1(t) + Bu_2(t)$ is the general solution of the associated homogeneous equation (with $f(t)$ replaced by zero), and $u^*(t)$ is a particular solution of the nonhomogeneous equation.

Equation (1) is called **globally asymptotically stable** if every solution $Au_1(t) + Bu_2(t)$ of the associated homogeneous equation tends to 0 as $t \rightarrow \infty$ for all values of A and B . Then the effect of the initial conditions "dies out" as $t \rightarrow \infty$.

If $Au_1(t) + Bu_2(t)$ tends to 0 as $t \rightarrow \infty$ for all values of A and B , then in particular $u_1(t) \rightarrow 0$ as $t \rightarrow \infty$ (choose $A = 1$, $B = 0$), and $u_2(t) \rightarrow 0$ as $t \rightarrow \infty$ (choose $A = 0$, $B = 1$). On the other hand, the condition that $u_1(t)$ and $u_2(t)$ both tend to 0 as t tends to infinity is obviously sufficient for $Au_1(t) + Bu_2(t)$ to tend to 0 as $t \rightarrow \infty$.

EXAMPLE 1 Study the stability of

- (a) $t^2\ddot{x} + 3t\dot{x} + \frac{3}{4}x = 3$ (b) $\ddot{x} + 2\dot{x} + 5x = e^t$ (c) $\ddot{x} + \dot{x} - 2x = 3t^2 + 2$

Solution:

- (a) This is an Euler equation whose general solution is $x(t) = At^{-1/2} + Bt^{-3/2} + 4$. The

- (b) The corresponding characteristic equation is $r^2 + 2r + 5 = 0$, with complex roots $r_1 = -1 + 2i$, $r_2 = -1 - 2i$, so $u_1 = e^{-t} \cos 2t$ and $u_2 = e^{-t} \sin 2t$ are linearly independent solutions of the homogeneous equation. (See (III) in Theorem 6.3.1 and Note 6.3.2.) As $t \rightarrow \infty$, both u_1 and u_2 tend to 0, since $\cos 2t$ and $\sin 2t$ are both less than or equal to 1 in absolute value and $e^{-t} \rightarrow 0$ as $t \rightarrow \infty$. The equation is therefore globally asymptotically stable.
- (c) Here $u_1 = e^t$ is one solution of the homogeneous equation. Since $u_1 = e^t$ does not tend to 0 as $t \rightarrow \infty$, the equation is *not* globally asymptotically stable. $\blacksquare$

A Useful Characterization of Stability

Recall that in the complex number $r = \alpha + i\beta$, α is the real part, and that the real part of a real number is the number itself. With these concepts, we have the following characterization of global asymptotic stability:

The equation $\ddot{x} + a\dot{x} + bx = f(t)$ is globally asymptotically stable iff both roots of the characteristic equation $r^2 + ar + b = 0$ have negative real parts. (2)

Proof: Note that $e^{(\alpha+i\beta)t} \rightarrow 0$ as $t \rightarrow \infty$ iff $\alpha < 0$. So in cases (I) and (III) of Theorem 6.3.1, the result is proved. In case (II), we note that $te^{\alpha t} \rightarrow 0$ as $t \rightarrow \infty$ iff $\alpha < 0$, so the result follows in this case as well. $\blacksquare$

This result extends easily to differential equations of order n , as we will see in Section 7.3. The following is a special result for the case of second-order equations, when $n = 2$:

$\ddot{x} + a\dot{x} + bx = f(t)$ is globally asymptotically stable $\iff a > 0$ and $b > 0$ (3)

Proof: The two roots (real or complex) r_1 and r_2 of the quadratic characteristic equation $r^2 + ar + b = 0$ have the property that $r^2 + ar + b = (r - r_1)(r - r_2) = r^2 - (r_1 + r_2)r + r_1r_2$. Hence $a = -r_1 - r_2$ and $b = r_1r_2$. In cases (I) and (II) of Theorem 6.3.1, global asymptotic stability holds iff $r_1 < 0$ and $r_2 < 0$, which is true iff $a > 0$ and $b > 0$. In case (III), when $r_{1,2} = \alpha \pm bi$, global asymptotic stability holds iff $\alpha < 0$, which is also true iff $a > 0$ and $b > 0$. $\blacksquare$

E 2 For the last two equations in Example 1, it follows immediately from (3) that (b) is stable, whereas (c) is unstable. $\blacksquare$

E 3 In a paper on growth theory, the following equation is studied:

$$\ddot{v} + \left(\mu - \frac{\lambda}{a}\right)\dot{v} + \lambda\gamma v = -\frac{\lambda}{a}\dot{b}(t)$$

where μ , λ , γ , and a are constants, and $b(t)$ is a fixed function. Assume that $a > 0$.

Solution: This is a second-order linear equation with constant coefficients. According to (3), the equation is stable iff $\mu > \lambda/a$ and $\lambda\gamma > 0$. $\blacksquare$

NOTE 1 (Asymptotically stable equilibrium states) Consider the differential equation $\ddot{x} + a\dot{x} + bx = c$ where $b \neq 0$. Then $x^* = c/b$ is an equilibrium state, since $x(t) = c/b$ is a constant solution of the equation. All solutions of the equation will tend to the equilibrium state as $t \rightarrow \infty$ iff $a > 0$ and $b > 0$. We then say that the **equilibrium state** $x^* = c/b$ is **globally asymptotically stable**.

PROBLEMS FOR SECTION 6.4

1. Determine which of the equations in Problem 6.3.1 are globally asymptotically stable, and verify (3) in this case.

2. For which values of the constant a is

$$\ddot{x} + (1 - a^2)\dot{x} + 2ax = 0$$

globally asymptotically stable?

- SM 3.** A model by T. Haavelmo leads to an equation of the type

$$\ddot{p}(t) = \gamma(a - \alpha)p(t) + k \quad (\alpha, \gamma, a, \text{ and } k \text{ are constants})$$

Solve the equation. Can the constants be chosen to make the equation globally asymptotically stable?

6.5 Simultaneous Equations in the Plane

So far we have considered finding one unknown function to satisfy a single differential equation. Many dynamic economic models, especially in macroeconomics, involve several unknown functions that satisfy a number of simultaneous differential equations.

Consider the important special case with two unknowns and two equations:

$$\begin{aligned}\dot{x} &= f(t, x, y) \\ \dot{y} &= g(t, x, y)\end{aligned} \quad (1)$$

In economic models that lead to systems of this type, $x = x(t)$ and $y = y(t)$ are state variables characterizing the economic system at a given time t . Usually, the state of the system $(x(t_0), y(t_0))$ is known at some initial time t_0 and the future development of the system is then uniquely determined. The rate of change of each variable depends not only on t and the variable itself, but on the other variable as well. In this sense, the two variables $x(t)$ and $y(t)$ “interact”. Systems of this type may exhibit very complicated behaviour.

A **solution** of (1) is a pair of differentiable functions $(x(t), y(t))$ which is defined on some interval I , and which satisfies both equations. With the assumptions imposed on f and g , if t_0 is a point in I , and x_0 and y_0 are given numbers, there will be one and only one pair of functions $(x(t), y(t))$ that satisfies (1) and has $x(t_0) = x_0, y(t_0) = y_0$.

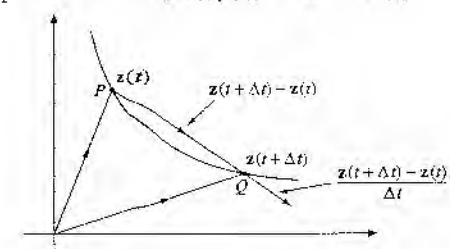


Figure 1

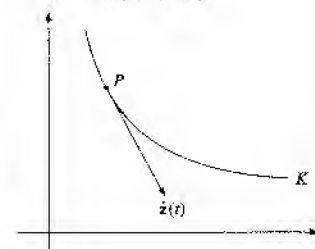


Figure 2

NOTE 1 If $(x(t), y(t))$ is a solution of (1), at time t the system is at the point $(x(t), y(t))$ in the xy -plane. As t varies, the point $(x(t), y(t))$ traces out a curve K in the xy -plane. In Fig. 1 the vector $\mathbf{z}(t) = (x(t), y(t))$ points from the origin to the point $P = (x(t), y(t))$ and $\mathbf{z}(t + \Delta t) = (x(t + \Delta t), y(t + \Delta t))$ points from the origin to the point Q . The vector $\mathbf{z}(t + \Delta t) - \mathbf{z}(t)$ points from P to Q , and $[\mathbf{z}(t + \Delta t) - \mathbf{z}(t)] / \Delta t$ points in the same direction if $\Delta t > 0$ (and in the opposite direction if $\Delta t < 0$). If t is kept fixed and Δt tends to 0, the point Q will tend to P , and the vector $[\mathbf{z}(t + \Delta t) - \mathbf{z}(t)] / \Delta t$ will tend to the **tangent vector** to the curve K at P . We see that

$$\frac{\mathbf{z}(t + \Delta t) - \mathbf{z}(t)}{\Delta t} = \left(\frac{x(t + \Delta t) - x(t)}{\Delta t}, \frac{y(t + \Delta t) - y(t)}{\Delta t} \right) \rightarrow (\dot{x}(t), \dot{y}(t)) \text{ as } \Delta t \rightarrow 0$$

Thus the vector $\dot{\mathbf{z}}(t) = (\dot{x}(t), \dot{y}(t))$, which describes how quickly $x(t)$ and $y(t)$ change when t is changed, is a tangent vector to the curve K at P , as illustrated in Fig. 2.

The general solution of (1) usually depends on two arbitrary constants A and B , and can then be written as $x = \varphi_1(t; A, B), y = \varphi_2(t; A, B)$. The two constants are determined if we specify an initial condition for each variable—for example, $x(t_0) = x_0, y(t_0) = y_0$.

How can one find the general solution of (1)? Of course, one cannot expect exact methods to work in complete generality, but explicit solutions can be found in some important cases.

One method is to reduce (1) to a second-order differential equation in only one unknown: Use the first equation in (1) to express y as a function $y = h(t, x, \dot{x})$ of t, x , and $\dot{x}$. Differentiate this last equation w.r.t. t and substitute the expressions for y and $\dot{y}$ into the second equation in (1). We then obtain a second-order differential equation to determine

EXAMPLE 1 Find the general solution of the system

$$\begin{aligned} \dot{x} &= 2x + e^t y - e^t \\ \dot{y} &= 4e^{-t} x + y \end{aligned}$$

Find also the solution that gives $x = y = 0$ for $t = 0$.

Solution: Solving the first equation for y gives $y = \dot{x}e^{-t} - 2xe^{-t} + 1$. Differentiating w.r.t. t yields $\dot{y} = \ddot{x}e^{-t} - \dot{x}e^{-t} - 2\dot{x}e^{-t} + 2xe^{-t}$. Inserting these expressions for y and $\dot{y}$ into the second equation gives $\ddot{x}e^{-t} - 3\dot{x}e^{-t} + 2xe^{-t} = 4xe^{-t} + \dot{x}e^{-t} - 2xe^{-t} + 1$, or

$$\ddot{x} - 4\dot{x} = e^t$$

Using the methods of Section 6.3, we find that the general solution for x is

$$x = A + Be^{4t} - \frac{1}{3}e^t$$

From $y = \dot{x}e^{-t} - 2xe^{-t} + 1 = (4Be^{4t} - \frac{1}{3}e^t)e^{-t} - 2(A + Be^{4t} - \frac{1}{3}e^t)e^{-t} + 1$ we get

$$y = -2Ae^{-t} + 2Be^{3t} + \frac{4}{3}$$

If $x = y = 0$ for $t = 0$, then $A + B - \frac{1}{3} = 0$ and $-2A + 2B + \frac{4}{3} = 0$. Solving these equations yields $A = \frac{1}{2}, B = -\frac{1}{6}$.

We have seen how the problem of solving (most) first-order systems of the form (1) can be transformed into the problem of solving one second-order equation in only one of the variables. On the other hand, any second-order differential equation $\ddot{x} = F(t, x, \dot{x})$ can be converted into a system of the form (1) simply by defining a new variable $y = \dot{x}$. Then $\dot{y} = \ddot{x} = F(t, x, \dot{x}) = F(t, x, y)$, and the system becomes

$$\begin{aligned} \dot{x} &= y, \\ \dot{y} &= F(t, x, y) \end{aligned} \tag{2}$$

Recursive Systems

Suppose that the two differential equations take the special form

$$\begin{aligned} \dot{x} &= f(t, x, y), \\ \dot{y} &= g(t, y) \end{aligned}$$

so that one of the two variables varies independently of the other. Then the system can be solved recursively in two steps:

- (i) First, solve $\dot{y} = g(t, y)$ as an ordinary first-order differential equation to get $y(t)$.
- (ii) second, substitute this $y = y(t)$ in the equation $\dot{x} = f(t, x, y)$ to get another first-order differential equation in $x(t)$.

Of course, a similar approach works with the pair of equations $\dot{x} = f(t, x), \dot{y} = g(t, x, y)$.

Solution Method for Autonomous Systems

When system (1) is of the form $\dot{x} = f(x, y)$, $\dot{y} = g(x, y)$, so that f and g do not depend explicitly on t , there is an alternative solution procedure: Around a point where $\dot{x} \neq 0$, we can view y as a function of x with $dy/dx = \dot{y}/\dot{x} = g(x, y)/f(x, y)$. Solve this equation to give $y = \varphi(x)$. Then $x(t)$ is found by solving $\dot{x} = f(x, \varphi(x))$. Finally, $y(t) = \varphi(x(t))$.

EX 2 Use the method described above to find the solution of the system

$$\begin{aligned}\dot{x} &= y \\ \dot{y} &= y^2/x\end{aligned}\quad (x > 0, y > 0)$$

that has $x(1) = 1$ and $y(1) = 2$.

Solution: We see that $dy/dx = \dot{y}/\dot{x} = y^2/xy = y/x$, which is a separable differential equation whose general solution is $y = Ax$. Then $\dot{x} = y = Ax$, with general solution $x = Be^{At}$. This gives $y = Ax = ABe^{At}$. If $x(1) = 1$ and $y(1) = 2$, then $1 = Be^A$ and $2 = ABe^A$. We find $A = 2$ and $B = e^{-2}$, so the solution is $x = e^{2t-2}$, $y = 2e^{2t-2}$. $\square$

Linear Systems with Constant Coefficients

Consider the linear system

$$\begin{aligned}\dot{x} &= a_{11}x + a_{12}y + b_1(t) \\ \dot{y} &= a_{21}x + a_{22}y + b_2(t)\end{aligned}\quad (3)$$

Suppose $a_{12} \neq 0$. (If $a_{12} = 0$, the first equation is a simple linear differential equation in only one unknown.) Let us derive a second-order equation by modifying the method used in Example 1. Differentiating the first equation w.r.t. t , then substituting $\dot{y}$ from the second equation in (3), we obtain $\ddot{x} = a_{11}\dot{x} + a_{12}(a_{21}x + a_{22}y + b_2(t)) + \dot{b}_1(t)$. Substituting $a_{12}y = \dot{x} - a_{11}x - b_1(t)$ from the first equation in (3), then simplifying, we have

$$\ddot{x} - (a_{11} + a_{22})\dot{x} + (a_{11}a_{22} - a_{12}a_{21})x = a_{12}b_2(t) - a_{22}b_1(t) + \dot{b}_1(t)\quad (4)$$

This is a second-order differential equation with constant coefficients. The general solution is of the form $x(t) = Au_1(t) + Bu_2(t) + u^*(t)$, where A and B are arbitrary constants. The solution for $y(t)$ is found from $a_{12}y = \dot{x} - a_{11}x - b_1(t)$, and it depends on the same two constants.

An argument similar to the above shows that y must satisfy the differential equation

$$\ddot{y} - (a_{11} + a_{22})\dot{y} + (a_{11}a_{22} - a_{12}a_{21})y = a_{21}b_1(t) - a_{11}b_2(t) + \dot{b}_2(t)\quad (5)$$

Note that (4) and (5) have the same associated homogeneous equation, so that their characteristic equations are identical.¹

¹ If we solve equations (4) and (5) separately, we end up with four constants. But these four constants cannot all be chosen independently of one another. Once we have chosen the constants for x , say, the constants for y are completely determined, because (4) gives $y = (\dot{x} - a_{11}x - b_1(t))/a_{12}$.

In fact, we obtain the following explicit formulas for the solution of system (3):

$$\begin{aligned}x(t) &= Au_1(t) + Bu_2(t) + u^*(t) \\ y(t) &= P(A, B)u_1(t) + Q(A, B)u_2(t) + \frac{1}{a_{12}}[\dot{u}^*(t) - a_{11}u^*(t) - b_1(t)]\end{aligned}\quad (6)$$

where the functions $u_1(t)$, $u_2(t)$, $P(A, B)$, and $Q(A, B)$ are defined as follows: If λ_1 and λ_2 denote the roots of the characteristic polynomial of equation (4), then:

(a) If λ_1 and λ_2 are real and different and $u_1(t) = e^{\lambda_1 t}$, $u_2(t) = e^{\lambda_2 t}$, then

$$P(A, B) = \frac{A(\lambda_1 - a_{11})}{a_{12}} \quad \text{and} \quad Q(A, B) = \frac{B(\lambda_2 - a_{11})}{a_{12}}$$

(b) If $\lambda_1 = \lambda_2$ is a real double root and $u_1(t) = e^{\lambda_1 t}$, $u_2(t) = te^{\lambda_1 t}$, then

$$P(A, B) = \frac{\lambda_1 A + B - a_{11}A}{a_{12}} \quad \text{and} \quad Q(A, B) = \frac{B(\lambda_1 - a_{11})}{a_{12}}$$

(c) If $\lambda_1 = \alpha + i\beta$ and $\lambda_2 = \alpha - i\beta$, $\beta \neq 0$, and $u_1(t) = e^{\alpha t} \cos \beta t$, $u_2(t) = e^{\alpha t} \sin \beta t$, then

$$P(A, B) = (\alpha A + \beta B - a_{11}A)/a_{12} \quad \text{and} \quad Q(A, B) = (\alpha B - \beta A - a_{11}B)/a_{12}.$$

Solutions Based on Eigenvalues

With $b_1(t) = b_2(t) = 0$, system (3) reduces to the homogeneous system

$$\begin{aligned}\dot{x} &= a_{11}x + a_{12}y \\ \dot{y} &= a_{21}x + a_{22}y\end{aligned} \iff \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}\quad (7)$$

Let us see if by an appropriate choice of numbers v_1 , v_2 , and λ we can make $(x, y) = (v_1 e^{\lambda t}, v_2 e^{\lambda t})$ a solution of (7). Inserting $\dot{x} = v_1 \lambda e^{\lambda t}$ and $\dot{y} = v_2 \lambda e^{\lambda t}$ into (7) yields

$$\begin{pmatrix} v_1 \lambda e^{\lambda t} \\ v_2 \lambda e^{\lambda t} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} v_1 e^{\lambda t} \\ v_2 e^{\lambda t} \end{pmatrix}$$

Cancelling the factor $e^{\lambda t}$ gives the equation

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \lambda \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

In the terminology of Section 1.5, $\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ is an eigenvector of the matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, with eigenvalue λ . The eigenvalues are the solutions of the equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = \lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0\quad (8)$$

The case in which A has different real eigenvalues, λ_1 and λ_2 , is the simplest. Then A has two linearly independent eigenvectors $\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ and $\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$, and the general solution of (7) is

$$\begin{pmatrix} x \\ y \end{pmatrix} = Ae^{\lambda_1 t} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + Be^{\lambda_2 t} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \quad (9)$$

where A and B are arbitrary constants.

LE 3 Solve the system $\dot{x} = 2y$, $\dot{y} = x + y$ by the eigenvalue method.

Solution: The system can be written as

$$\begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \mathbf{A} \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{with} \quad \mathbf{A} = \begin{pmatrix} 0 & 2 \\ 1 & 1 \end{pmatrix} \quad (*)$$

The characteristic polynomial of $\mathbf{A}$ is $\begin{vmatrix} 0-\lambda & 2 \\ 1 & 1-\lambda \end{vmatrix} = \lambda^2 - \lambda - 2 = (\lambda + 1)(\lambda - 2)$.

Hence the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 2$. Corresponding eigenvectors are $\begin{pmatrix} -2 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, respectively. According to (9), the general solution of (*) is therefore

$$\begin{pmatrix} x \\ y \end{pmatrix} = Ae^{-t} \begin{pmatrix} -2 \\ 1 \end{pmatrix} + Be^{2t} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -2Ae^{-t} + Be^{2t} \\ Ae^{-t} + Be^{2t} \end{pmatrix}$$

Consider a nonhomogeneous system of the form

$$\begin{aligned} \dot{x} &= a_{11}x + a_{12}y + b_1 \\ \dot{y} &= a_{21}x + a_{22}y + b_2 \end{aligned} \quad (10)$$

where b_1 and b_2 are constants. This can be transformed into a homogeneous system by introducing new variables. The method is illustrated in the next example.

LE 4 Find the solutions of the system

$$\begin{aligned} \dot{x} &= 2y + 6 \\ \dot{y} &= x + y - 3 \end{aligned} \quad (*)$$

First note that the equilibrium point (where $\dot{x} = \dot{y} = 0$) is $(6, -3)$. Introduce new variables $z = x - 6$ and $w = y + 3$ that measure the deviation of x and y from their equilibrium values. Then $\dot{z} = \dot{x}$ and $\dot{w} = \dot{y}$, so the system (*) is transformed into

$$\begin{aligned} \dot{z} &= 2(w - 3) + 6 = 2w \\ \dot{w} &= (z + 6) + (w - 3) - 3 = z + w \end{aligned}$$

According to the preceding example, the general solution is $z = -2Ae^{-t} + Be^{2t}$ and $w = Ae^{-t} + Be^{2t}$. The general solution of (*) is therefore $x = z + 6 = -2Ae^{-t} + Be^{2t} + 6$ and $y = w - 3 = Ae^{-t} + Be^{2t} - 3$.

PROBLEMS FOR SECTION 6.5

1. Find the general solutions of the following systems:

$$\begin{aligned} \text{(a)} \quad \begin{cases} \dot{x} = y \\ \dot{y} = x + t \end{cases} & \quad \text{(b)} \quad \begin{cases} \dot{x} = x + y \\ \dot{y} = x - y \end{cases} & \quad \text{(c)} \quad \begin{cases} \dot{x} = 2x - 3y \\ \dot{y} = -x + t \end{cases} \end{aligned}$$

2. Find the unique solutions of the given systems that satisfy the given initial conditions.

$$\begin{aligned} \text{(a)} \quad \begin{cases} \dot{x}(t) = a(x(t) + y(t)) \\ \dot{y}(t) = b(x(t) + y(t)) \end{cases}, & \quad x(0) = \frac{1}{2}, \quad y(0) = \frac{1}{2} \\ \text{(b)} \quad \begin{cases} \dot{x} = 2tx + y \\ \dot{y} = -2(t + x) \end{cases}, & \quad x(0) = 1, \quad y(0) = 1 \\ \text{(c)} \quad \begin{cases} \dot{x} = -2y + \sin t \\ \dot{y} = 2x + 1 - \cos t \end{cases}, & \quad x(0) = 0, \quad y(0) = 0 \end{aligned}$$

3. Find the general solution of the system

$$\begin{aligned} \dot{x} &= x + e^{2t}p \\ \dot{p} &= 2e^{-2t}x - p \end{aligned}$$

4. A model by M. J. Beckmann and H. E. Ryder includes the following system:

$$\dot{\pi}(t) = \alpha\pi(t) - \sigma(t), \quad \dot{\sigma}(t) = \pi(t) - \frac{1}{\beta}\sigma(t)$$

Find the general solution when $\alpha + 1/\beta > 2$.

5. Using the method in Example 2, find the solution curve passing through $(t, x, y) = (1, 1, \sqrt{2})$ for

$$\dot{x} = \frac{ty^2}{1+x^2}, \quad \dot{y} = \frac{txy}{1+x^2} \quad (t > 0)$$

6.6 Equilibrium Points for Linear Systems

Consider the linear system with constant coefficients

$$\begin{aligned} \dot{x} &= a_{11}x + a_{12}y + b_1 \\ \dot{y} &= a_{21}x + a_{22}y + b_2 \end{aligned} \iff \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \mathbf{A} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \quad (1)$$

The equilibrium points for this system are determined by the equations

$$\begin{aligned} a_{11}x + a_{12}y + b_1 &= 0 \\ a_{21}x + a_{22}y + b_2 &= 0 \end{aligned} \quad \text{or} \quad \begin{aligned} a_{11}x + a_{12}y &= -b_1 \\ a_{21}x + a_{22}y &= -b_2 \end{aligned} \quad (2)$$

which result from putting $\dot{x} = \dot{y} = 0$ in (1). If $|\mathbf{A}| \neq 0$, this system has a unique solution (x^*, y^*) , which is called an **equilibrium point** (or an **equilibrium state**) for the system (1). Cramer's rule tells us that the equilibrium point is

$$x^* = \frac{a_{12}b_2 - a_{22}b_1}{|\mathbf{A}|}, \quad y^* = \frac{a_{21}b_1 - a_{11}b_2}{|\mathbf{A}|} \quad (3)$$

The pair $(x(t), y(t)) = (x^*, y^*)$ with $(\dot{x}(t), \dot{y}(t)) = (0, 0)$ will then be a solution of (1). Since $\dot{x} = \dot{y} = 0$ at the equilibrium point, it follows that if the system is at (x^*, y^*) , it has always been there and will always stay there.

PLE 1 Find the equilibrium point for the system

$$\begin{aligned} \dot{x} &= -2x + y + 2 \\ \dot{y} &= -2y + 8 \end{aligned} \iff \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ 0 & -2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2 \\ 8 \end{pmatrix}$$

Find also the general solution, and examine what happens when t tends to infinity.

Solution: We easily see that $|\mathbf{A}| \neq 0$ and that $(x^*, y^*) = (3, 4)$ is the equilibrium point. The solution of the system is found by using the methods explained in the previous section. The general solution turns out to be $x(t) = Ae^{-2t} + Bte^{-2t} + 3$, $y(t) = Be^{-2t} + 4$. As t tends to infinity, $(x(t), y(t))$ tends to the equilibrium point $(3, 4)$.

In general, an equilibrium point (x^*, y^*) for (1) is called **globally asymptotically stable** if every solution tends to the equilibrium point as $t \rightarrow \infty$. Thus the equilibrium point $(3, 4)$ in Example 1 is globally asymptotically stable.

We showed in Section 6.5 (see (6.5.4) and (6.5.5)) that a solution $(x(t), y(t))$ of (1) must satisfy the two second-order equations ($\text{tr}(\mathbf{A})$ denotes the trace of $\mathbf{A}$, see Section 1.5)

$$\ddot{x} - \text{tr}(\mathbf{A})\dot{x} + |\mathbf{A}|x = a_{12}b_2 - a_{22}b_1, \quad \ddot{y} - \text{tr}(\mathbf{A})\dot{y} + |\mathbf{A}|y = a_{21}b_1 - a_{11}b_2 \quad (4)$$

If $|\mathbf{A}| \neq 0$, these equations have x^* and y^* given in (3) as their respective equilibrium points. Moreover, the characteristic equation of each of the equations in (4) is the same as the eigenvalue equation of $\mathbf{A}$. (See (6.5.8).) Using (6.4.2) and (6.4.3), we obtain the following result:

EM 6.6.1

Suppose that $|\mathbf{A}| \neq 0$. Then the equilibrium point (x^*, y^*) for the linear system

$$\begin{aligned} \dot{x} &= a_{11}x + a_{12}y + b_1 \\ \dot{y} &= a_{21}x + a_{22}y + b_2 \end{aligned} \iff \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \mathbf{A} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

is globally asymptotically stable if and only if

$$\text{tr}(\mathbf{A}) = a_{11} + a_{22} < 0 \quad \text{and} \quad |\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21} > 0$$

or equivalently, if and only if both eigenvalues of $\mathbf{A}$ have negative real part.

LE 2 Examine the stability of the equilibrium point $(0, 0)$ for the system

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -2x - y \end{aligned} \iff \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -2 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Solution: The coefficient matrix has trace -1 and determinant 2 , so according to Theorem 6.6.1 the system is globally asymptotically stable. In this case all solutions converge to the equilibrium point $(0, 0)$ as $t \rightarrow \infty$. It is easy to confirm this statement by finding the general solution of the system. The methods explained in Section 6.5 lead to a second-order equation in x ,

$$\ddot{x} + \dot{x} + 2x = 0, \quad \text{with the solution} \quad x = e^{-t/2} \left(A \cos \frac{1}{2}\sqrt{7}t + B \sin \frac{1}{2}\sqrt{7}t \right)$$

By using $y = \dot{x}$, we find a similar expression for y . We see from the formulas obtained that $x(t)$ and $y(t)$ both tend to 0 as $t \rightarrow \infty$.

Alternative Behaviour Around Equilibrium Points

In this subsection we give a brief survey of how system (1) behaves when the equilibrium point is not necessarily globally asymptotically stable. According to (6.5.6), the general solution of (1) is

$$\begin{aligned} x(t) &= Au_1(t) + Bu_2(t) + x^* \\ y(t) &= P(A, B)u_1(t) + Q(A, B)u_2(t) + y^* \end{aligned} \quad (5)$$

where $u_1(t)$ and $u_2(t)$ are described in (a)–(c) following equation (6.5.6) and (x^*, y^*) is the equilibrium point.

Disregarding the cases where one or both eigenvalues are 0 , we have the following results:

- If both eigenvalues of $\mathbf{A}$ have negative real parts, then (x^*, y^*) is globally asymptotically stable (a **sink**). All solution curves converge to the equilibrium point as $t \rightarrow \infty$.
- If both eigenvalues of $\mathbf{A}$ have positive real parts, then (x^*, y^*) is a **source**. In this case all solution curves starting away from the equilibrium point explode as t increases, i.e. $\|(x(t), y(t))\| \rightarrow \infty$ as $t \rightarrow \infty$. See Example 3(a) below.
- If the eigenvalues are real with opposite signs, with $\lambda_1 < 0$ and $\lambda_2 > 0$, then (x^*, y^*) is a so-called **saddle point**. (The eigenvalues are real and of opposite signs iff the determinant of $\mathbf{A}$ is negative. See (C) following equation (1.5.5).) In this case only solutions of the form $x(t) = Ae^{\lambda_1 t} + x^*$, $y(t) = A(\lambda_1 - a_{11})e^{\lambda_1 t}/a_{12} + y^*$ converge to the equilibrium point as $t \rightarrow \infty$. All other solution curves move away from the equilibrium point as $t \rightarrow \infty$. (See Section 6.9 for further discussion.)
- If the eigenvalues are purely imaginary ($\lambda_{1,2} = \pm i\beta$), then (x^*, y^*) is a so-called **centre**. Then all solution curves are periodic with the same period. The solution curves are ellipses or circles. See Example 3(b) below.

EXAMPLE 3

Examine the character of the equilibrium points for the following systems:

$$\begin{aligned} \text{(a)} \quad \dot{x} &= 2y \\ \dot{y} &= -x + 2y \end{aligned} \qquad \begin{aligned} \text{(b)} \quad \dot{x} &= -y \\ \dot{y} &= x \end{aligned}$$

Solution: In both cases the equilibrium point is $(0, 0)$.

- The eigenvalues are $\lambda_{1,2} = 1 \pm i$, and the general solution is $x(t) = Ae^t \cos t + Be^t \sin t$, $y(t) = \frac{1}{2}(A + B)e^t \cos t + \frac{1}{2}(-A + B)e^t \sin t$. Both $x(t)$ and $y(t)$ exhibit explosive oscillations unless $A = B = 0$.

- (b) Here the eigenvalues are $\lambda_{1,2} = \pm i$ and the general solution is $x(t) = A \cos t + B \sin t$, $y(t) = A \sin t - B \cos t$. For all t we find that $x(t)^2 + y(t)^2 = A^2 + B^2$, so the solution curves in the xy -plane are circles with centre at the equilibrium point $(0, 0)$.

AS FOR SECTION 6.6

1. Check (if possible) the stability of the following systems by using Theorem 6.6.1:

$$\begin{array}{lll} \text{(a)} \quad \begin{cases} \dot{x} = x - 8y \\ \dot{y} = 2x - 4y \end{cases} & \text{(b)} \quad \begin{cases} \dot{x} = x - 4y + 2 \\ \dot{y} = 2x - y - 5 \end{cases} & \text{(c)} \quad \begin{cases} \dot{x} = -x - 3y + 5 \\ \dot{y} = 2x - 2y + 2 \end{cases} \end{array}$$

2. For what values of the constant a are the following systems globally asymptotically stable?

$$\begin{array}{ll} \text{(a)} \quad \begin{cases} \dot{x} = ax - y \\ \dot{y} = x + ay \end{cases} & \text{(b)} \quad \begin{cases} \dot{x} = ax - (2a - 4)y \\ \dot{y} = x + 2ay \end{cases} \end{array}$$

3. Find the general solution of the system

$$\begin{cases} \dot{x} = x + 2y + 1 \\ \dot{y} = -y + 2 \end{cases}$$

- (i) by using the same method as in Example 6.5.1, (ii) by using the eigenvalue method. Is the system globally asymptotically stable?

4. (a) Solve the differential equation system

$$\begin{cases} \dot{x} = ax + 2y + \alpha \\ \dot{y} = 2x + ay + \beta \end{cases} \quad (*)$$

where α, α and β are constants, $a \neq \pm 2$.

- (b) Find the equilibrium point (x^*, y^*) , and find necessary and sufficient conditions for $(*)$ to be globally asymptotically stable.
(c) Let $a = -1$, $\alpha = -4$ and $\beta = -1$. Determine a solution curve that converges to the equilibrium point.

7 Phase Plane Analysis

The solution procedures studied in this chapter give explicit answers only for quite restricted and exceptional classes of differential equations. In this section we shall indicate how, even when explicit solutions are unavailable, geometric arguments can still shed light on the structure of the solutions of autonomous systems of differential equations in the plane.

System (6.5.1) is called **autonomous** (time independent) if f and g do not depend explicitly on t , so the equations become

$$\begin{cases} \dot{x} = f(x, y) \\ \dot{y} = g(x, y) \end{cases} \quad (\text{autonomous system}) \quad (1)$$

We assume that f and g are C^1 functions. A solution $(x(t), y(t))$ of (1) describes a curve or **path** in the xy -plane. It consists of all points $\{(x(t), y(t)) : t \in I\}$, where I is the interval of definition. If $(x(t), y(t))$ is a solution of (1), then so is $(x(t+a), y(t+a))$ for any constant a (but with a different interval I' of definition). Hence $(x(t), y(t))$ and $(x(t+a), y(t+a))$ describe the same path. (This is valid only for autonomous systems.) For the autonomous system (1), the vector $(\dot{x}(t), \dot{y}(t))$ is uniquely determined at each point $(x(t), y(t))$, so no two paths in the xy -plane can intersect.

Phase plane analysis is a technique for studying the behaviour of paths in the “phase plane” based on information obtained directly from (1).

From (1) it follows that the rates of change of $x(t)$ and $y(t)$ are given by $f(x(t), y(t))$ and $g(x(t), y(t))$, respectively. For instance, if $f(x(t), y(t)) > 0$ and $g(x(t), y(t)) < 0$ at $P = (x(t), y(t))$, then as t increases, the system moves from the point P down and to the right. The direction of motion is given by the tangent vector $(\dot{x}(t), \dot{y}(t))$ to the path at P as illustrated in Fig. 1, and the speed of motion is given by the length of the vector $(\dot{x}(t), \dot{y}(t))$.

To illustrate the dynamics of system (1), we can, in principle, draw such a vector at each point in the plane. The family of such vectors is called a **vector field**. Of course, in practice, one can draw only a small representative sample of these vectors. On the basis of the vector field one can draw paths for the system and thereby exhibit the **phase portrait** or **phase diagram** of the system.

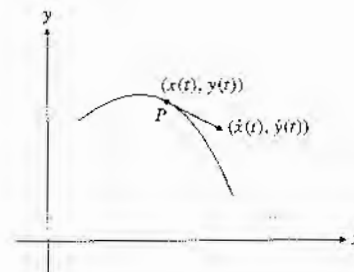


Figure 1

EXAMPLE 1 Figure 2 shows a vector field for the system studied in Example 6.6.2:

$$\dot{x} = y, \quad \dot{y} = -2x - y \quad (*)$$

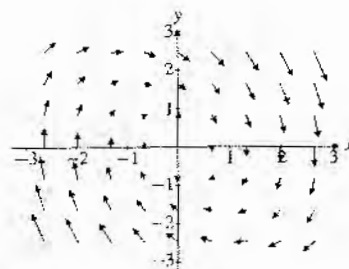


Figure 2

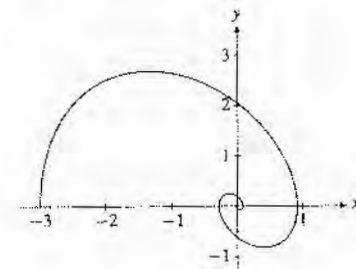


Figure 3

The lengths of the vectors have been proportionally reduced, so that the arrows will not interfere with each other. But the length of a vector still suggests the speed of motion. Note that $(\dot{x}, \dot{y}) = (0, 0)$ only at the point $(0, 0)$, which is the *equilibrium point*.

A closer study of the vector field suggests that the paths spiral towards the equilibrium point $(0, 0)$. The speed of motion decreases as one gets closer to the origin. Figure 3 shows the particular path that starts at the point $(-3, 0)$, and spirals towards $(0, 0)$. We know from Example 6.6.2 that since $(0, 0)$ is globally asymptotically stable, all paths for this linear system, wherever they start, tend to the equilibrium point $(0, 0)$ as $t \rightarrow \infty$.

In general, a point (a, b) where $f(a, b) = g(a, b) = 0$ is called an **equilibrium point** (or **stationary point**) for system (1). Because $\dot{x} = \dot{y} = 0$ at an equilibrium point E , if the system is at E , then it always will be (and always was) at E .

The equilibrium points of (1) are the points of intersection of the two curves $f(x, y) = 0$ and $g(x, y) = 0$, which are called the **nullclines** of the system.

To draw a phase diagram of (1), begin by drawing the two nullclines. At each point on the nullcline $f(x, y) = 0$, the $\dot{x}$ component is zero and the velocity vector is vertical. It points up if $\dot{y} > 0$, down if $\dot{y} < 0$.

At each point on the nullcline $g(x, y) = 0$, the $\dot{y}$ component is 0, and the velocity vector is horizontal. It points to the right if $\dot{x} > 0$, to the left if $\dot{x} < 0$.

E 2 Draw a phase diagram for system (*) in Example 1.

Solution: The nullclines and the direction of motion on paths crossing the nullclines are shown in Fig. 4. Note that the nullclines for (*) divide the phase plane into four regions or sectors, denoted by (I), (II), (III), and (IV) in Fig. 4.

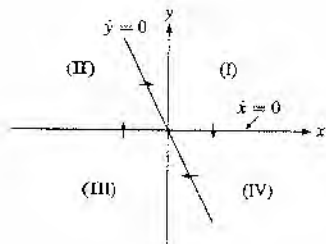


Figure 4

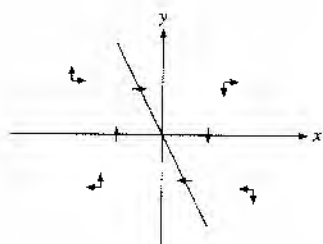


Figure 5

In sectors (I) and (II), $y > 0$, so $\dot{x} > 0$, whereas in Sectors (III) and (IV), $y < 0$ and so $\dot{x} < 0$. On the other hand, in sectors (I) and (IV), $2x + y > 0$ and so $\dot{y} < 0$, whereas in Sectors (II) and (III), $2x + y < 0$ and so $\dot{y} > 0$. (A convenient check that you have determined the direction of the arrows correctly is to pick a point in each of the four sectors and calculate $(\dot{x}, \dot{y})$ at each of these points. For example, $(2, 2)$ is in sector (I) and $(\dot{x}, \dot{y}) = (2, -6)$. The point $(-2, 2)$ is in sector (II) and $(\dot{x}, \dot{y}) = (2, 2)$. In sector (III) the point $(-2, -2)$ has $(\dot{x}, \dot{y}) = (-2, 6)$. Finally, at $(2, -2)$ in sector (IV) we have $(\dot{x}, \dot{y}) = (-2, -2)$.)

In Fig. 5, the direction of motion on a path at a point in each of the four sectors is indicated by arrows. In accordance with common practice, a separate arrow is drawn for each of the x and y directions. We usually make all the arrows have the same length. (If they were drawn with their correct lengths, they would correspond to the vectors $(\dot{x}, 0)$ and $(0, \dot{y})$. It follows that the actual direction of the path through the point would correspond to the sum of these two vectors.)

EXAMPLE 3 In a model of economic growth, capital $K = K(t)$ and consumption $C = C(t)$ satisfy the pair of differential equations

$$\begin{aligned} \dot{K} &= aK - bK^2 - C \\ \dot{C} &= w(a - 2bK)C \end{aligned} \quad (*)$$

Here a, b , and w are positive constants. Construct a phase diagram for this system, assuming that $K \geq 0$ and $C \geq 0$.

Solution: The nullcline $\dot{K} = 0$ is the parabola $C = aK - bK^2$, and the nullcline $\dot{C} = 0$ consists of the two lines $C = 0$ and $K = a/2b$. In Fig. 6 the two nullclines are drawn. There are three equilibrium points, $(0, 0)$, $(a/b, 0)$, and $(a/2b, a^2/4b)$.

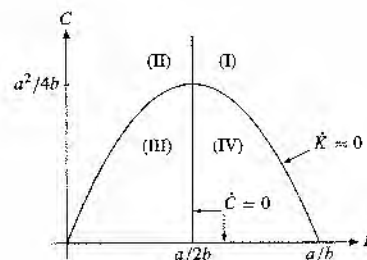


Figure 6

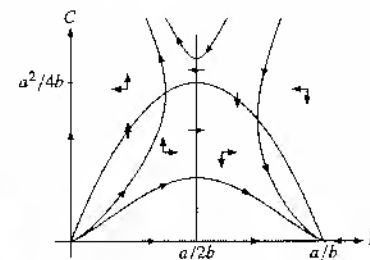


Figure 7

In sector (I), $C > aK - bK^2$ and $K > a/2b$, so $\dot{K} < 0$ and $\dot{C} < 0$. In sectors (II), (III), and (IV), we have $\dot{K} < 0$, $\dot{C} > 0$, then $\dot{K} > 0$, $\dot{C} > 0$, and $\dot{K} > 0$, $\dot{C} < 0$, respectively. The appropriate arrows are drawn in Fig. 7, which indicates some paths consistent with the arrows.

These examples show how useful information about the solution paths can be obtained by partitioning the phase plane into regions to indicate whether each of the two variables is increasing or decreasing. In particular, the partition will often suggest whether or not a certain equilibrium point is stable, in the sense that paths starting near the equilibrium point tend to that point as $t \rightarrow \infty$. However, to determine whether an equilibrium point really is stable or not, a phase diagram analysis should be supplemented with tests based on analytical methods like those set out in the subsequent sections.

MS FOR SECTION 6.7

1. Perform a phase plane analysis of the following systems and then find their explicit solutions.

$$\begin{array}{lll} \text{(a)} \quad \begin{cases} \dot{x} = y \\ \dot{y} = x \end{cases} & \text{(b)} \quad \begin{cases} \dot{x} = x + y \\ \dot{y} = x - y \end{cases} & \text{(c)} \quad \begin{cases} \dot{x} = x - 4y \\ \dot{y} = 2x - 5y \end{cases} \end{array}$$

2. Perform a phase plane analysis of the system

$$\dot{x} = x(k - ay), \quad \dot{y} = y(-h + bx), \quad x > 0, \quad y > 0$$

where a , b , h , and k are positive constants. This is the famous *Lotka–Volterra* model from mathematical biology. See Example 7.5.2.

3. In an economic model, $K = K(t)$ denotes capital, $C = C(t)$ consumption, while α , A , and r are positive constants, with $\alpha < 1$. Assume that

$$\text{(i)} \quad \dot{K} = AK^\alpha - C \quad \text{(ii)} \quad \dot{C} = C(\alpha AK^{\alpha-1} - r)$$

Perform a phase plane analysis of this system when $A = 2$, $\alpha = 0.5$, and $r = 0.05$.

- SM 4. (a) Draw a phase diagram and some typical paths for the autonomous system

$$\dot{x} = -x, \quad \dot{y} = -xy - y^2$$

- (b) Solve the system with $x(0) = -1$, $y(0) = 1$. (Hint: You need to solve a Bernoulli equation. See Section 5.6. One of the integrals cannot be evaluated.) Find the limit $\lim_{t \rightarrow \infty} (x(t), y(t))$.

5. (a) Perform a phase plane analysis of the following system, where $x > 0$ and $y > 0$.

$$\dot{x} = x(y - x - \ln x - 1), \quad \dot{y} = 1 - x \quad (*)$$

- (b) Introduce the transform $z = y - \ln x$, and show that $(*)$ becomes

$$\dot{z} = 2 - z, \quad \dot{y} = 1 - e^{y-z} \quad (**)$$

Perform a phase plane analysis of this system. (Whereas stability of the equilibrium point $(\bar{x}, \bar{y}) = (1, 2)$ is not clear from the diagram in (a), the corresponding equilibrium point $(\bar{z}, \bar{y}) = (2, 2)$ is “clearly” stable in the diagram in (b). This example is due to Conlisk and Ramanathan, *Review of Economic Studies* (1970).)

- SM 6. Consider the system

$$\dot{x} = -x, \quad \dot{y} = -x^2y$$

Here is $(0, 0)$ an equilibrium point. (In fact, $(0, b)$ is an equilibrium point for every value of b .) Find the unique solution of the system that passes through the point $(1, 1)$ for $t = 0$. Show that the corresponding path does not converge to $(0, 0)$.

6.8 Stability for Nonlinear Systems

In this section we study the stability theory for the autonomous system

$$\begin{aligned} \dot{x} &= f(x, y) \\ \dot{y} &= g(x, y) \end{aligned} \quad (1)$$

where f and g are C^1 -functions. An equilibrium point (a, b) of the system (where $f(a, b) = g(a, b) = 0$) is called **locally asymptotically stable** if any path starting near (a, b) tends to (a, b) as $t \rightarrow \infty$. An equilibrium point (a, b) is called **globally asymptotically stable** if any solution of (1) (wherever it starts) converges to (a, b) as $t \rightarrow \infty$. (A more precise definition of stability is given in Section 7.5.)

To examine whether (a, b) is locally asymptotically stable, we have to consider how solutions of the system behave in a neighbourhood of (a, b) . To this end, consider the linear approximation of the functions $f(x, y)$ and $g(x, y)$ about (a, b) . If (x, y) is sufficiently close to (a, b) , then (see Section 2.6),

$$\begin{aligned} f(x, y) &\approx f(a, b) + f'_1(a, b)(x - a) + f'_2(a, b)(y - b) \\ g(x, y) &\approx g(a, b) + g'_1(a, b)(x - a) + g'_2(a, b)(y - b) \end{aligned}$$

Because $f(a, b) = 0$ and $g(a, b) = 0$, we have $f(x, y) \approx f'_1(a, b)x + f'_2(a, b)y + b_1$ and $g(x, y) \approx g'_1(a, b)x + g'_2(a, b)y + b_2$, where $b_1 = -f'_1(a, b)a - f'_2(a, b)b$ and $b_2 = -g'_1(a, b)a - g'_2(a, b)b$. It is therefore reasonable to expect that in a neighbourhood of (a, b) , system (1) “behaves” approximately like the linear system

$$\begin{aligned} \dot{x} &= f'_1(a, b)x + f'_2(a, b)y + b_1 \\ \dot{y} &= g'_1(a, b)x + g'_2(a, b)y + b_2 \end{aligned} \quad (*)$$

Note that (a, b) is also an equilibrium point of system $(*)$, because the definitions of b_1 and b_2 imply that $f'_1(a, b)a + f'_2(a, b)b + b_1 = 0$ and $g'_1(a, b)a + g'_2(a, b)b + b_2 = 0$. According to Theorem 6.6.1, this linear system is globally asymptotically stable if and only if the eigenvalues of the matrix

$$A = \begin{pmatrix} f'_1(a, b) & f'_2(a, b) \\ g'_1(a, b) & g'_2(a, b) \end{pmatrix} \quad (2)$$

both have negative real parts, or equivalently, if and only if A has negative trace and positive determinant. Since $(*)$ “behaves” approximately like (1) near (a, b) , it is a reasonable conjecture that in this case (a, b) is a *locally asymptotically stable* equilibrium point for system (1). This conjecture is indeed correct, as the noted Russian mathematician A. M. Lyapunov demonstrated in the late 1890s. (For a more modern proof, see Coddington and Levinson (1955).)

6.8.1 (LYAPUNOV)

Suppose that f and g are C^1 functions and let (a, b) be an equilibrium point for the system

$$\dot{x} = f(x, y), \quad \dot{y} = g(x, y)$$

Let A be the Jacobian matrix $A = \begin{pmatrix} f'_1(a, b) & f'_2(a, b) \\ g'_1(a, b) & g'_2(a, b) \end{pmatrix}$.

If

$$\operatorname{tr}(A) = f'_1(a, b) + g'_2(a, b) < 0$$

and

$$|A| = f'_1(a, b)g'_2(a, b) - f'_2(a, b)g'_1(a, b) > 0$$

i.e. if both eigenvalues of A have negative real parts, then (a, b) is locally asymptotically stable.

NOTE 1 If the eigenvalues λ_1 and λ_2 of A in Theorem 6.8.1 are real with $\lambda_1 < \lambda_2 < 0$, then all the solution paths that converge to (a, b) as $t \rightarrow \infty$, become “tangent in the limit” to the line through (a, b) with the same direction as the eigenvector corresponding to λ_2 . See Theorem 6.9.1 for the case $\lambda_1 < 0 < \lambda_2$.

E 1 The system

$$\begin{aligned} \dot{x} &= f(x, y) = -3x - 2y + 8x^2 + y^3 \\ \dot{y} &= g(x, y) = 3x + y - 3x^2y^2 + y^4 \end{aligned}$$

has $(0, 0)$ as an equilibrium point. Prove that it is locally asymptotically stable.

Solution: Here $f'_1(0, 0) = -3$, $f'_2(0, 0) = -2$, $g'_1(0, 0) = 3$, and $g'_2(0, 0) = 1$, so $\operatorname{tr}(A) = -3 + 1 = -2 < 0$ and $|A| = -3 - (-6) = 3 > 0$. By Theorem 6.8.1, the equilibrium point $(0, 0)$ is locally asymptotically stable. $\square$

E 2 (Price adjustment mechanism) Consider the following extension of Example 5.7.2 to two commodities:

$$\dot{p} = H_1(D_1(p, q) - S_1(p, q)), \quad \dot{q} = H_2(D_2(p, q) - S_2(p, q)) \quad (*)$$

Here p and q denote the prices of two different commodities, $D_i(p, q)$ and $S_i(p, q)$, $i = 1, 2$ are the demand and supply for the two commodities, while H_1 and H_2 are fixed functions of one variable. Assume that $H_1(0) = H_2(0) = 0$ and that $H'_1 > 0$, $H'_2 > 0$, so that H_1 and H_2 are both strictly increasing. This implies that if there is excess demand for commodity 1, so that $D_1(p, q) - S_1(p, q) > 0$, then $\dot{p} > 0$, and thus the price of commodity 1 will increase. Similarly for commodity 2.

Suppose (p^0, q^0) is an equilibrium point for system $(*)$. By our assumptions on H_1 and H_2 , we have $D_1(p^0, q^0) = S_1(p^0, q^0)$, $D_2(p^0, q^0) = S_2(p^0, q^0)$. Thus, at prices p^0 and q^0 , demand is equal to supply for each commodity. The stability properties of the

equilibrium can be examined by appealing to Theorem 6.8.1: the equilibrium point (p^0, q^0) is asymptotically stable for system $(*)$ provided

$$H'_1(0)\left(\frac{\partial D_1}{\partial p} - \frac{\partial S_1}{\partial p}\right) + H'_2(0)\left(\frac{\partial D_2}{\partial q} - \frac{\partial S_2}{\partial q}\right) < 0 \quad (a)$$

and

$$\left(\frac{\partial D_1}{\partial p} - \frac{\partial S_1}{\partial p}\right)\left(\frac{\partial D_2}{\partial q} - \frac{\partial S_2}{\partial q}\right) > \left(\frac{\partial D_1}{\partial q} - \frac{\partial S_1}{\partial q}\right)\left(\frac{\partial D_2}{\partial p} - \frac{\partial S_2}{\partial p}\right) \quad (b)$$

All the partial derivatives are evaluated at (p^0, q^0) , and in (b) we have cancelled the positive factor $H'_1(0)H'_2(0)$. Normally, $\partial D_1/\partial p$ and $\partial D_2/\partial q$ are negative, while $\partial S_1/\partial p$ and $\partial S_2/\partial q$ are positive. (If the price of some commodity increases, then the demand goes down while supply goes up.) Because $H'_1 > 0$ and $H'_2 > 0$, we conclude that (a) is “normally” satisfied. In order to determine the sign in (b), the functions involved must be further specified. However, the left-hand side depends on “own” price effects—how p affects D_1 and S_1 , and how q affects D_2 and S_2 —whereas the right-hand side depends on “cross” price effects. $\square$

Olech's Theorem

We end this section with a brief look at a special result on global stability of an autonomous system of differential equations in the plane. (See Olech (1963).)

THEOREM 6.8.2 (OLECH)

Consider the following system, where f and g are C^1 functions in $\mathbb{R}^2$,

$$\dot{x} = f(x, y), \quad \dot{y} = g(x, y)$$

and let (a, b) be an equilibrium point. Let $A(x, y) = \begin{pmatrix} f'_1(x, y) & f'_2(x, y) \\ g'_1(x, y) & g'_2(x, y) \end{pmatrix}$, and assume that the following three conditions are all satisfied:

- (a) $\operatorname{tr}(A(x, y)) = f'_1(x, y) + g'_2(x, y) < 0$ in all of $\mathbb{R}^2$
- (b) $|A(x, y)| = f'_1(x, y)g'_2(x, y) - f'_2(x, y)g'_1(x, y) > 0$ in all of $\mathbb{R}^2$
- (c) $f'_1(x, y)g'_2(x, y) \neq 0$ in all of $\mathbb{R}^2$ or $f'_2(x, y)g'_1(x, y) \neq 0$ in all of $\mathbb{R}^2$

Then (a, b) is globally asymptotically stable.

In contrast to the Lyapunov theorem, which gives local stability, conditions (a), (b), and (c) in Olech's theorem are required to hold throughout $\mathbb{R}^2$, not only at the equilibrium point. But then these stronger “global” conditions give global stability. For an economic application, see Problem 5.

EXAMPLE 3

Use Theorem 6.8.2 to prove that $(0, 0)$ is a globally asymptotically stable equilibrium for

$$\dot{x} = f(x, y) = 1 - e^{x-y}, \quad \dot{y} = g(x, y) = -y$$

Solution: Here $f_1(x, y) = -e^{x+y}$, $f_2(x, y) = e^{x+y}$, $g_1(x, y) = 0$, and $g_2(x, y) = -1$. It follows immediately that conditions (a), (b), and (c) in Theorem 6.8.2 are all satisfied, so the equilibrium point $(0, 0)$ is globally asymptotically stable.

MS FOR SECTION 6.8

1. Show that $(6, 6)$ is a locally asymptotically stable equilibrium point for the system

$$\dot{x} = f(x, y) = y - x, \quad \dot{y} = g(x, y) = -x^2 + 8x - 2y$$

2. Determine (if possible) the local asymptotic stability of the following systems at the given stationary points by using Theorem 6.8.1:

- (a) $\dot{x} = -x + \frac{1}{2}y^2$
 $\dot{y} = 2x - 2y$ at $(0, 0)$
- (b) $\dot{x} = x - 3y + 2x^2 + y^2 - xy$
 $\dot{y} = 2x - y - e^{x-y}$ at $(1, 1)$
- (c) $\dot{x} = -x^3 - y$
 $\dot{y} = x - y^3$ at $(0, 0)$
- (d) $\dot{x} = 2x + 8 \sin y$
 $\dot{y} = 2 - e^x - 3y - \cos y$ at $(0, 0)$

3. Use Theorem 6.8.2 to show that $(0, 0)$ is a globally asymptotically stable equilibrium point for the system

$$\dot{x} = y, \quad \dot{y} = -ky - w^2x \quad (k > 0, w \neq 0)$$

4. G. Heal has studied the system

$$\dot{q} = a(p - c(q)), \quad \dot{p} = b(D(p) - q)$$

where q is the amount sold of a commodity, p is its price per unit, $c(q)$ is the average cost function, and $D(p)$ is the demand function. Here a and b are positive constants and $D'(p) < 0$. Prove that an equilibrium point (q^*, p^*) (where $p^* = c(q^*)$ and $D(p^*) = q^*$) is locally asymptotically stable provided $c''(q^*) > 0$.

5. A business cycle model by N. Kaldor uses the system

$$\dot{Y} = \alpha(I(Y, K) - S(Y, K)), \quad \dot{K} = I(Y, K) \quad (\alpha > 0) \quad (*)$$

where Y is national income, K is capital stock, $I(Y, K)$ is an investment function and $S(Y, K)$ is a savings function. Assume that $I'_Y > 0$, $I'_K < 0$, $S'_Y > 0$, $S'_K < 0$, and $I'_K - S'_K < 0$. Use Olech's theorem to prove that an equilibrium point for $(*)$ is globally asymptotically stable provided $\alpha(I'_Y - S'_Y) + I'_K < 0$ and $I'_K S'_Y < S'_K I'_Y$.

6. Suppose $K = K(t)$ denotes the capital stock of an economy and $P = P(t)$ denotes the level of pollution at time t . The development of the economy is described by the system

$$\dot{K} = K(\alpha K^{\alpha-1} - \delta), \quad \dot{P} = K^\beta - \gamma P \quad (*)$$

The constants satisfy the conditions $\alpha \in (0, 1)$, $\alpha \in (0, 1)$, $\delta > 0$, $\gamma > 0$, and $\beta > 1$. Find the equilibrium point (K^*, P^*) in the open first quadrant, and check (if possible) the stability of the point by using Theorem 6.8.1. Find an explicit expression for $K(t)$ when $K(0) = K_0 \geq 0$, and examine its limit as $t \rightarrow \infty$.

6.9 Saddle Points

Dynamic economic models often have equilibria that are not asymptotically stable. In some cases a special kind of behaviour near an equilibrium is encountered: two paths approach the equilibrium point from opposite directions as $t \rightarrow \infty$. These two paths together with the equilibrium point itself form a curve that has a tangent at the equilibrium point. All other paths that come close to the equilibrium point will move away again. The precise result is this:

THEOREM 6.9.1 (LOCAL SADDLE POINT THEOREM)

Suppose that f and g are C^1 functions and let (a, b) be an equilibrium point for

$$\dot{x} = f(x, y), \quad \dot{y} = g(x, y)$$

Let $A = \begin{pmatrix} f'_1(a, b) & f'_2(a, b) \\ g'_1(a, b) & g'_2(a, b) \end{pmatrix}$ be the Jacobian matrix, and suppose that

$$|A| = f'_1(a, b)g'_2(a, b) - f'_2(a, b)g'_1(a, b) < 0$$

or, equivalently, that the eigenvalues of A are nonzero real numbers of opposite signs.² Then there exist an open ball (i.e. an open circular disk) B with (a, b) as its centre and a curve C in B passing through (a, b) , such that:

- (1) Through every point on C there passes a solution $(x(t), y(t))$ that remains on C and converges to (a, b) as $t \rightarrow \infty$. Every solution through a point that lies in B but not on C will sooner or later leave B .
- (2) The curve C is tangent at (a, b) to the line through (a, b) with the same direction as the eigenvector corresponding to the negative eigenvalue of A .

Such an equilibrium is called a **saddle point**.³

NOTE 1 The solutions guaranteed by Theorem 6.9.1 could have their starting points very close to the equilibrium although the solutions are defined on an infinite time interval. This is why we refer to the theorem as a local result. In many economic models one needs a global version of the theorem (see e.g. Seierstad and Sydsæter (1987), Theorem 19, page 256).

NOTE 2 If the system is linear with constant coefficients,

$$\begin{aligned} \dot{x} &= a_{11}x + a_{12}y + b_1 \\ \dot{y} &= a_{21}x + a_{22}y + b_2 \end{aligned} \quad \text{with} \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} < 0$$

then there is a unique equilibrium point, which is a saddle point. In this case the paths that approach the equilibrium point will run along the straight line corresponding to the eigenvector associated with the negative eigenvalue. (See Example 1.)

² See (C) in Section 1.5.
³ The paths of the system resemble the paths taken by a drop of water falling on a horse saddle. The drop of water will converge to the centre of the saddle if it hits precisely on the ridge of the saddle, but will fall to the ground if it hits in a different place.

LE 1 Consider the following system with equilibrium $(0, 0)$:

$$\dot{x} = 2y, \quad \dot{y} = 3x - y$$

With $f(x, y) = 2y$ and $g(x, y) = 3x - y$, the matrix A in Theorem 6.9.1 is $\begin{pmatrix} 0 & 2 \\ 3 & -1 \end{pmatrix}$.

Because the determinant of A is equal to -6 , the equilibrium is a saddle point. The characteristic polynomial of A is $\lambda^2 + \lambda - 6 = (\lambda - 2)(\lambda + 3)$, so the eigenvalues are -3 and

2. An eigenvector associated with the negative eigenvalue -3 is $\begin{pmatrix} -2 \\ 3 \end{pmatrix}$. Figure 1 shows a phase diagram in which the two paths converging to the equilibrium point are indicated by dashed lines. Both lines are in the direction of the eigenvector.

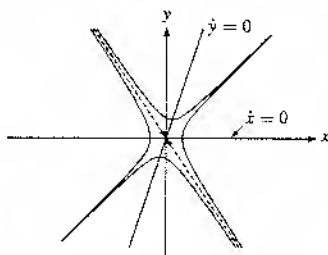


Figure 1

LE 2 Consider system $(*)$ in Example 6.7.3, with

$$\dot{K} = aK - bK^2 - C, \quad \dot{C} = w(a - 2bK)C$$

One equilibrium point is $P = (a/2b, a^2/4b)$. Here the matrix A evaluated at P is

$$A = \begin{pmatrix} a - 2bK & -1 \\ -2wbC & w(a - 2bK) \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ -wa^2/2 & 0 \end{pmatrix}$$

Thus $|A| = -wa^2/2 < 0$, so $(a/2b, a^2/4b)$ is a saddle point.

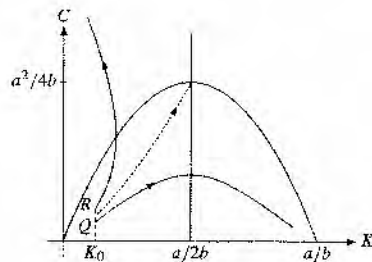


Figure 2

The evolution of the system depends critically on the values of $K(0) = K_0$ and $C(0) = C_0$. In Fig. 2 we assume that $K_0 < a/2b$.

If $C(0) = C_0$ is small, the path starts at a point like Q in Fig. 2. Then, as t increases, $K(t)$ steadily increases. On the other hand, $C(t)$ is increasing until $K(t)$ has reached the level $a/2b$, then $C(t)$ starts decreasing. If $C(0) = C_0$ is bigger, so that the path starts at a point like R , then consumption increases, whereas capital first increases and then decreases. It is not hard to imagine that, for some start point between Q and R , the path converges to the equilibrium point P . This behaviour is confirmed by Theorem 6.9.1.

Other Types of Equilibrium Point

Lyapunov's theorem (Theorem 6.8.1) and the saddle point theorem (Theorem 6.9.1) show that in some important cases the limiting behaviour of a nonlinear system near an equilibrium point is similar to the behaviour of the linearized system. In the two theorems the eigenvalues of the Jacobian matrix both had negative real parts, or were real with opposite signs, respectively.

If the eigenvalues have positive real parts, solutions that start close to the equilibrium point move away from it, and the equilibrium point is a "source".

If the eigenvalues are purely imaginary, or 0, no definite statement about the limiting character of the solution can be made. For details we refer to the literature.

PROBLEMS FOR SECTION 6.9

1. (a) Show that the equilibrium point of the following system is a saddle point:

$$\dot{x} = -\frac{1}{2}x + y, \quad \dot{y} = y - 2$$

Find also the eigenvalues of the associated matrix A in Theorem 6.9.1, and an eigenvector corresponding to the negative eigenvalue.

- (b) Draw a phase diagram in which the two paths converging to the equilibrium point are indicated. Find explicit expressions for these two paths.

- SM** 2. Find the equilibrium point and check if it is a saddle point:

$$\dot{k} = f(k) - \delta k - c, \quad \dot{c} = -c(r + \delta - f'(k))$$

Assume that δ and r are positive constants, $f(0) = 0$, $f'(k) > 0$, $f''(k) < 0$, $f'(0) > r + \delta$, and $f'(\infty) < \delta$.

- SM** 3. (a) Consider the following system of differential equations:

$$\dot{x} = x(y - x/2 - 2), \quad \dot{y} = y(1 - y/2x)$$

Find the unique equilibrium point (x_0, y_0) in $S = \{(x, y) : x > 0, y > 0\}$. Is the equilibrium point (x_0, y_0) asymptotically stable? Is it a saddle point?

- (b) Draw a phase diagram for the system and indicate the behaviour of some integral curves in the region S .

4. (a) Consider the following system of first-order differential equations:

$$\begin{aligned}\dot{x} &= y^2 - x \\ \dot{y} &= 25/4 - y^2 - (x - 1/4)^2\end{aligned}$$

Find all the equilibrium points of the system and classify them, if possible (i.e. for each of them determine if it is locally asymptotically stable, a saddle point, or neither).

(b) Draw a phase diagram for the system, and indicate some possible integral curves.

7

DIFFERENTIAL EQUATIONS III:
HIGHER-ORDER
EQUATIONS

There can be no question, however, that prolonged commitment to mathematical exercises in economics can be damaging. It leads to the atrophy of judgement and intuition.
—John Kenneth Galbraith (1971)

This chapter discusses extensions of the theory developed in the two preceding chapters. Most of Sections 7.1–7.3 present rather simple generalizations to n th-order equations of the theory of second-order equations that was discussed in Chapter 6. More specifically, Section 7.1 presents the main theory for general linear equations. Then Section 7.2 concentrates on the case of constant coefficients, after which Section 7.3 focuses on stability conditions. Next, Section 7.4 introduces systems of differential equations in n variables, and briefly discusses methods of solving them based on eigenvalues, or on an $n \times n$ matrix function called the resolvent.

Thereafter, Section 7.5 gives more formal definitions and results on stability of nonlinear systems. This section also shows how to use Lyapunov functions to decide whether an autonomous system is stable. A famous application is to the Lotka–Volterra model where a suitably constructed Lyapunov function is useful in determining the stability properties of the model.

Section 7.6 generalizes and extends the existence and uniqueness results of Section 5.8 to vector differential equations. Results on the dependence of the solutions to changes in the initial conditions are also recorded.

Finally, Section 7.7 gives a brief introduction to some types of partial differential equations, which occasionally arise in economic applications.

7.1 Linear Differential Equations

A differential equation of the n th order can usually be written in the form

$$\frac{d^n x}{dt^n} = F\left(t, x, \frac{dx}{dt}, \dots, \frac{d^{n-1}x}{dt^{n-1}}\right) = F(t, x, \dot{x}, \dots, x^{(n-1)}) \quad (1)$$

Here F is a given function of $n + 1$ variables and $x = x(t)$ is the unknown function. Sometimes we shall use the alternative notation $x^{(k)}$ for $d^k x/dt^k$.

If F is a linear function of x and of its derivatives w.r.t. t up to and including those of order $n - 1$, we usually write the equation as

$$\frac{d^n x}{dt^n} + a_1(t) \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_{n-1}(t) \frac{dx}{dt} + a_n(t)x = f(t) \quad (2)$$

where $a_1(t), \dots, a_n(t)$, and $f(t)$ are fixed continuous functions on $(-\infty, \infty)$.

The associated **homogeneous equation** is obtained when the right-hand side is 0:

$$\frac{d^n x}{dt^n} + a_1(t) \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_{n-1}(t) \frac{dx}{dt} + a_n(t)x = 0 \quad (3)$$

Suppose that $u_1(t), \dots, u_n(t)$ are n solutions of (3). Then it is easy to verify that any linear combination $C_1 u_1(t) + \cdots + C_n u_n(t)$ also satisfies (3), for all values of the constants $C_1, \dots, C_n$. (For $n = 2$ this was verified in Section 6.2.) Already for the case $n = 2$, however, we know that this is not necessarily the general solution, even if the n functions $u_i(t)$ are all different.

As in Section 1.2, say that the n functions $u_1(t), \dots, u_n(t)$ are **linearly dependent** if there exist constants $C_1, \dots, C_n$, not all 0, such that

$$C_1 u_1(t) + \cdots + C_n u_n(t) = 0 \quad \text{for all } t \quad (*)$$

Equivalently, at least one of the functions can be written as a linear combination of the others. Alternatively, if $u_1(t), \dots, u_n(t)$ are not linearly dependent, they are called **linearly independent**. Then equation (*) is satisfied for all t only if $C_1 = \cdots = C_n = 0$. If $n = 2$, linear independence is equivalent to the condition that the two functions are not proportional.

In the next section, we consider the special case when the functions $a_1(t), \dots, a_n(t)$ are all constants, independent of t . Except in this case, there is no general method for finding the general solution to (3). The following theorem exhibits the structure of the general solutions to equations (2) and (3) (for the case $n = 2$ these results were discussed in Section 6.2):

(a) The homogeneous equation (3) has the **general solution**

$$x = x(t) = C_1 u_1(t) + \cdots + C_n u_n(t) \quad (4)$$

where $u_1(t), \dots, u_n(t)$ are any n linearly independent solutions of (3) and $C_1, \dots, C_n$ are arbitrary constants.

(b) The nonhomogeneous equation (2) has the **general solution**

$$x = x(t) = C_1 u_1(t) + \cdots + C_n u_n(t) + u^*(t) \quad (5)$$

where $C_1 u_1(t) + \cdots + C_n u_n(t)$ is the general solution of the corresponding homogeneous equation (3) and $u^*(t)$ is any particular solution of the nonhomogeneous equation (2).

The proof of Theorem 7.1.1 relies on the following *existence and uniqueness* theorem (see Theorem 7.6.2 for a generalization):

THEOREM 7.1.2 (EXISTENCE AND UNIQUENESS FOR LINEAR EQUATIONS)

Suppose that $a_1(t), \dots, a_n(t)$ and $f(t)$ are all continuous functions on $(-\infty, \infty)$. Let $x_0, x_0^{(1)}, \dots, x_0^{(n-1)}$ be n given numbers and let t_0 be an arbitrary number. Then the differential equation (2) has one and only one solution $x(t)$ on $(-\infty, \infty)$ that satisfies the conditions

$$x(t_0) = x_0, \quad \frac{dx(t_0)}{dt} = x_0^{(1)}, \quad \dots, \quad \frac{d^{n-1}x(t_0)}{dt^{n-1}} = x_0^{(n-1)}$$

This result has an important corollary:

THEOREM 7.1.3 (EXISTENCE OF n LINEARLY INDEPENDENT SOLUTIONS)

Suppose that $a_1(t), \dots, a_n(t)$ are continuous functions on $(-\infty, \infty)$. Then the homogeneous differential equation (3) has n linearly independent solutions.

Proof: For $i = 1, \dots, n$, Theorem 7.1.2 yields a (unique) solution $u_i(t)$ of equation (3) such that for $k = 0, \dots, n-1$ we have $u_i^{(k)}(0) = 1$ if $k = i-1$, and $u_i^{(k)}(0) = 0$ otherwise. (Here, $u_i^{(0)} = u_i$, the "0th derivative" of u_i .) We shall prove that $u_1(t), \dots, u_n(t)$ are linearly independent.

Let $C_1, \dots, C_n$ be constants and put $w(t) = C_1 u_1(t) + \cdots + C_n u_n(t)$. Then $w^{(k)}(0) = C_1 u_1^{(k)}(0) + \cdots + C_n u_n^{(k)}(0) = C_{k+1}$ for $k = 0, \dots, n-1$. Now, suppose the constants $C_1, \dots, C_n$ are such that $w(t) = 0$ for all t . Then w and all derivatives of w must also be identically 0, and we get $C_1 = w(0) = 0, C_2 = \dot{w}(0) = 0, \dots, C_n = w^{(n-1)}(0) = 0$. Hence, the only choice of constants that makes w identically zero is $C_1 = \cdots = C_n = 0$. It follows that $u_1, \dots, u_n$ are linearly independent. ■

Proof of Theorem 7.1.1: (a) Suppose that $u_1(t), \dots, u_n(t)$ are n linearly independent solutions of (3) and let $x(t)$ be an arbitrary solution. We have to prove the existence of constants $C_1, \dots, C_n$ such that $x(t) = C_1 u_1(t) + \cdots + C_n u_n(t)$.

Assume that the solution curve for $x = x(t)$ passes through (t_0, x_0) , and let $x_0^{(2)}, x_0^{(2)}, \dots, x_0^{(n-1)}$ be given numbers. Suppose we could prove the existence of constants $C_1, \dots, C_n$ such that

$$\begin{aligned} C_1 u_1(t_0) + \cdots + C_n u_n(t_0) &= x_0 \\ C_1 \dot{u}_1(t_0) + \cdots + C_n \dot{u}_n(t_0) &= x_0^{(1)} \\ &\vdots \\ C_1 u_1^{(n-1)}(t_0) + \cdots + C_n u_n^{(n-1)}(t_0) &= x_0^{(n-1)} \end{aligned} \quad (i)$$

Then the functions $C_1 u_1(t) + \cdots + C_n u_n(t)$ and $x(t)$ would have the same value at t_0 and, moreover, they would have the same values for the first $n-1$ derivatives at t_0 . By Theorem 7.1.2 the two solutions would coincide and the proof would be complete.

By Cramer's rule, a sufficient condition for (i) to have a unique solution $C_1, \dots, C_n$ is that the Wronskian determinant

$$W(t_0) = \begin{vmatrix} u_1(t_0) & \dots & u_n(t_0) \\ \dot{u}_1(t_0) & \dots & \dot{u}_n(t_0) \\ \vdots & & \vdots \\ u_1^{(n-1)}(t_0) & \dots & u_n^{(n-1)}(t_0) \end{vmatrix} \quad (\text{ii})$$

is nonzero. We shall prove that if any n solutions $u_1(t), \dots, u_n(t)$ of (3) make the Wronskian determinant $W(t_0)$ equal to zero at any point t_0 , these solutions are linearly dependent.

Indeed, if $W(t_0) = 0$, then the columns of the Wronskian are linearly dependent according to Theorem 1.2.1. Therefore there exist numbers $\lambda_1, \dots, \lambda_n$, not all equal to 0, such that

$$\begin{aligned} \lambda_1 u_1(t_0) + \dots + \lambda_n u_n(t_0) &= 0 \\ \lambda_1 \dot{u}_1(t_0) + \dots + \lambda_n \dot{u}_n(t_0) &= 0 \\ &\vdots \\ \lambda_1 u_1^{(n-1)}(t_0) + \dots + \lambda_n u_n^{(n-1)}(t_0) &= 0 \end{aligned} \quad (\text{iii})$$

Put $\hat{x}(t) = \lambda_1 u_1(t) + \dots + \lambda_n u_n(t)$. Then $\hat{x}(t)$ solves (3) because it is a linear combination of the solutions $u_i(t)$. Moreover, the equations (iii) imply that $\hat{x}(t_0) = 0, \hat{x}'(t_0) = 0, \dots, \hat{x}^{(n-1)}(t_0) = 0$. By Theorem 7.1.2 there is only one solution of (3) that satisfies these requirements for all t . The function that is equal to 0 for all t has this property, and so $0 = \hat{x}(t) = \lambda_1 u_1(t) + \dots + \lambda_n u_n(t)$ for all t . This confirms that $u_1(t), \dots, u_n(t)$ are linearly dependent, and so completes the proof.

To prove (b), let $x(t)$ be an arbitrary solution of (2) and let $u^*(t)$ be any particular solution of (2). Then it is easy to see (by substitution) that $x(t) - u^*(t)$ satisfies the homogeneous equation (3). So there must exist constants $C_1, \dots, C_n$ and a set of linearly independent solutions $u_1(t), \dots, u_n(t)$ of (3) such that $x(t) - u^*(t) = C_1 u_1(t) + \dots + C_n u_n(t)$. The result follows immediately. ■

Variation of Parameters

We briefly describe a method for finding the solution of a nonhomogeneous linear equation once the general solution of the homogeneous equation is known. The good news is that it works (in principle) whatever is the function $f(t)$ in (2). The bad news is that it is usually quite laborious.

Suppose $u_1, \dots, u_n$ are n linearly independent solutions of the homogeneous equation (3). Let

$$x = C_1(t)u_1 + \dots + C_n(t)u_n \quad (*)$$

where the functions $C_1(t), \dots, C_n(t)$ are chosen to satisfy the $n - 1$ equations

$$\begin{aligned} \dot{C}_1(t)u_1 + \dots + \dot{C}_n(t)u_n &= 0 \\ \dot{C}_1(t)\dot{u}_1 + \dots + \dot{C}_n(t)\dot{u}_n &= 0 \\ &\vdots \\ \dot{C}_1(t)u_1^{(n-2)} + \dots + \dot{C}_n(t)u_n^{(n-2)} &= 0 \end{aligned}$$

By repeated differentiation and substitution, one verifies eventually that the function defined by (*) satisfies the nonhomogeneous equation (2) provided that

$$\dot{C}_1(t)u_1^{(n-1)} + \dots + \dot{C}_n(t)u_n^{(n-1)} = f(t)$$

Thus we have n equations which, because the functions $u_1, \dots, u_n$ are linearly independent, can be used to determine $\dot{C}_1(t), \dots, \dot{C}_n(t)$ as functions of t . These we can integrate to find $C_1(t), \dots, C_n(t)$, including an arbitrary constant for each. Then (*) gives the general solution of (2).

Solution: In this case $n = 2$. It is easy to find the two linearly independent solutions $u_1 = e^t$ and $u_2 = e^{2t}$ of the corresponding homogeneous equation. We must therefore choose $C_1(t)$ and $C_2(t)$ so that $x = C_1(t)e^t + C_2(t)e^{2t}$ is a solution. The relevant equations for determining $C_1(t)$ and $C_2(t)$ are here

$$\begin{aligned} \dot{C}_1(t)e^t + \dot{C}_2(t)e^{2t} &= 0 \\ \dot{C}_1(t)e^t + \dot{C}_2(t)2e^{2t} &= t \end{aligned}$$

The first of these equations gives $\dot{C}_1(t) = -e^t \dot{C}_2(t)$, which inserted into the second equation gives $\dot{C}_2(t)e^{2t} = t$, or $\dot{C}_2(t) = te^{-2t}$. Integrating by parts, $C_2(t) = \int te^{-2t} dt = -\frac{1}{2}te^{-2t} - \frac{1}{4}e^{-2t} + B$. Then $\dot{C}_1(t) = -e^t \dot{C}_2(t) = te^{-t}$, so $C_1(t) = -\int te^{-t} dt = te^{-t} + e^{-t} + A$. Inserting these expressions for $C_1(t)$ and $C_2(t)$ into (*), we obtain the general solution

$$x = Ae^t + Be^{2t} + \frac{1}{2}t + \frac{3}{4}$$

PROBLEMS FOR SECTION 7.1

1. Find the general solution of $\ddot{x} - 2\dot{x} - \dot{x} + 2x = 10$. (Hint: e^t, e^{-t} , and e^{2t} are solutions of the homogeneous equation.)

2. Use variation of parameters to solve $\ddot{x} - x = e^{-t}$.

3. Solve the equation $\ddot{x} + x = 1/t, t > 0$. (You may not be able to evaluate all the integrals.)

7.2 The Constant Coefficients Case

The general linear differential equation of order n with constant coefficients takes the form

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + \dots + a_{n-1} \frac{dx}{dt} + a_n x = f(t) \quad (1)$$

The associated homogeneous equation is

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + \dots + a_{n-1} \frac{dx}{dt} + a_n x = 0 \quad (2)$$

According to Theorem 7.1.1, the general solution of (1) is of the form

$$x = x(t) = C_1 u_1(t) + \dots + C_n u_n(t) + u^*(t)$$

where the functions $u_1(t), \dots, u_n(t)$ are n linearly independent solutions of (2), the numbers $C_1, \dots, C_n$ are n arbitrary constants, and $u^*(t)$ is any particular solution of (1).

Solutions of the Homogeneous Equation

Guided by the results in Chapter 6 for the case $n = 2$, we try to find solutions of (2) of the form $x = e^{rt}$ for appropriate values of r . Substituting $x = e^{rt}$ into (2) and cancelling the positive factor e^{rt} , we obtain the characteristic equation of (2) (or (1)):

$$p(r) = r^n + a_1 r^{n-1} + \dots + a_{n-1} r + a_n = 0 \quad (3)$$

where $p(r)$ is the **characteristic polynomial**. By the fundamental theorem of algebra, equation (3) has exactly n roots, real or complex, provided that each root is counted according to its multiplicity.

Suppose first that equation (3) has n distinct real roots $r_1, r_2, \dots, r_n$. Then $e^{r_1 t}, e^{r_2 t}, \dots, e^{r_n t}$ all satisfy (2), and one can prove that these n functions are linearly independent. So the general solution of (2) is

$$x(t) = C_1 e^{r_1 t} + C_2 e^{r_2 t} + \dots + C_n e^{r_n t}$$

The *general* method for finding n linearly independent solutions of (2) can be described as follows. First, find all roots of (3) and notice the multiplicity of each of them. A real root r with multiplicity 1 (i.e. a simple root) gives the solution

$$e^{rt}$$

A real root r with multiplicity p yields the p linearly independent solutions

$$e^{rt}, t e^{rt}, \dots, t^{p-1} e^{rt}$$

A pair of complex roots $r = \alpha + i\beta$, $\bar{r} = \alpha - i\beta$ with multiplicity 1 yields the two solutions

$$e^{\alpha t} \cos \beta t, \quad e^{\alpha t} \sin \beta t$$

(Complex solutions of (3) appear in complex conjugate pairs.)

A pair of complex roots $r = \alpha + i\beta$, $\bar{r} = \alpha - i\beta$, each with multiplicity q , yields the $2q$ linearly independent solutions

$$e^{\alpha t} \cos \beta t, \quad e^{\alpha t} \sin \beta t, \quad \dots, \quad t^{q-1} e^{\alpha t} \cos \beta t, \quad t^{q-1} e^{\alpha t} \sin \beta t$$

This procedure always finds n solutions of (2) that are linearly independent. It is illustrated in the following example.

E 1 Find the general solution of the equation

$$\frac{d^5 x}{dt^5} + 5 \frac{d^4 x}{dt^4} + 12 \frac{d^3 x}{dt^3} + 16 \frac{d^2 x}{dt^2} + 12 \frac{dx}{dt} + 4x = 0$$

Solution: In this constructed example, the characteristic polynomial is

$$p(r) = r^5 + 5r^4 + 12r^3 + 16r^2 + 12r + 4 = (r^2 + 2r + 2)^2(r + 1)$$

The characteristic equation, $p(r) = 0$, has the simple real root $r_1 = -1$, and the two complex roots $r_2 = -1 + i$ and $r_3 = -1 - i$, both with multiplicity 2. The general solution of the given equation is therefore

$$x = Ae^{-t} + D_1 e^{-t} \cos t + D_2 e^{-t} \sin t + D_3 t e^{-t} \cos t + D_4 t e^{-t} \sin t$$

where A and $D_1, \dots, D_4$ are arbitrary constants.

Finding a Particular Solution

In order to find the general solution of the nonhomogeneous equation (1), it remains to find a particular solution $u^* = u^*(t)$ of (1). If $f(t)$ is a linear combination of terms of the form

$$e^{at}, t^m, \cos bt, \text{ or } \sin bt \quad (*)$$

or products of such terms, then the method of undetermined coefficients developed in Section 6.3 for the case $n = 2$ will lead us to a particular solution. Consider a simple example.

EXAMPLE 2 Find the general solution of

$$\frac{d^5 x}{dt^5} + 5 \frac{d^4 x}{dt^4} + 12 \frac{d^3 x}{dt^3} + 16 \frac{d^2 x}{dt^2} + 12 \frac{dx}{dt} + 4x = t^2 + t - 1$$

Solution: The corresponding homogeneous equation was solved in Example 1. It remains to find a particular solution. The form of the right-hand side of the given equation suggests putting $u^* = At^2 + Bt + C$. We try to adjust the coefficients appropriately. We get $\ddot{u}^* = 2At + B$, $\ddot{\ddot{u}}^* = 2A$, and higher order derivatives are 0. Substituting into the given equation yields $32A + 24At + 12B + 4At^2 + 4Bt + 4C = t^2 + t - 1$, or after collecting terms,

$$4At^2 + (24A + 4B)t + (32A + 12B + 4C) = t^2 + t - 1$$

This equation is satisfied for all t when $4A = 1$, $24A + 4B = 1$, and $32A + 12B + 4C = -1$. Hence $A = \frac{1}{4}$, $B = -\frac{5}{4}$, $C = \frac{3}{2}$, and so $u^*(t) = \frac{1}{4}t^2 - \frac{5}{4}t + \frac{3}{2}$ is a particular solution. **|**

The method of undetermined coefficients depends on our ability to guess the general form of a particular solution. The method usually fails if the right-hand side is of a type different from those mentioned above. However, variation of parameters, as discussed in Section 7.1, may still work.

PROBLEMS FOR SECTION 7.2

1. Find the general solutions of the following equations:

$$(a) \quad \ddot{x} + 3\dot{x} + 3x = 3 \quad (b) \quad \frac{d^4 x}{dt^4} - 3 \frac{d^3 x}{dt^3} + \frac{d^2 x}{dt^2} + 4x = 2t - 1$$

(Hint for (b): $r^4 - 3r^3 + r^2 + 4 = (r^2 + r + 1)(r - 2)^2$.)

SM 2. Find $x = x(t)$ if $x(0) = 0$, $\dot{x}(0) = 1$, $\ddot{x}(0) = 0$ and $\ddot{x} - \ddot{x} - \dot{x} + x = 8te^{-t}$.

SM 3. In a model due to T. Haavelmo, a function $K = K(t)$ satisfies the equation

$$\ddot{K} = (\gamma_1 \kappa + \gamma_2) \dot{K} + (\gamma_1 \sigma + \gamma_3) \mu_0 e^{\mu t} \int_0^t e^{-\mu \tau} \dot{K}(\tau) d\tau$$

where $\gamma_1, \gamma_2, \gamma_3, \kappa, \sigma, \mu_0$, and μ are constants. Deduce a third-order differential equation for $K = K(t)$. Find the conditions for the characteristic equation of the third-order equation to have three different real roots. Prove that the solution in that case has the structure

$$K(t) = C_1 e^{r_1 t} + C_2 e^{r_2 t} + C_3$$

7.3 Stability of Linear Differential Equations

Global asymptotic stability for general second-order linear equations was defined in Section 6.4. As a direct generalization we say that equation (7.1.2) is **globally asymptotically stable** if the general solution $C_1 u_1(t) + \cdots + C_n u_n(t)$ of the corresponding homogeneous equation tends to 0 as $t \rightarrow \infty$, regardless of the values of the constants $C_1, \dots, C_n$. Thus the "effect of the initial conditions" dies out as t tends to ∞ .

If we put $C_j = 1$ and $C_i = 0$ for $i \neq j$, we see, in particular, that $u_j(t) \rightarrow 0$ as $t \rightarrow \infty$, and this holds for all $j = 1, \dots, n$. On the other hand, these requirements are surely sufficient for the equation to be globally asymptotically stable.

Constant Coefficients

Consider the case with constant coefficients

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_n x = f(t) \quad (1)$$

and let $u_1(t), \dots, u_n(t)$ be the n linearly independent solutions of the associated homogeneous equation obtained by the procedure described in Section 7.2. Each $u_j(t)$ corresponds to a root r_j of the characteristic equation. To simplify notation, put $r_j = \alpha + i\beta$. According to whether r_j is real ($\beta = 0$) with multiplicity 1 or multiplicity > 1 , or complex ($\beta \neq 0$) with multiplicity 1 or > 1 , the corresponding solution u_j is one of the following functions:

$$e^{\alpha t}, \quad t^k e^{\alpha t}, \quad e^{\alpha t} \cos \beta t, \quad e^{\alpha t} \sin \beta t, \quad t^k e^{\alpha t} \cos \beta t, \quad \text{or} \quad t^k e^{\alpha t} \sin \beta t$$

In each case, $u_j \rightarrow 0$ as $t \rightarrow \infty$ if and only if $\alpha < 0$. To see this, here is a detailed argument for this property in the case where $u_j = t^r e^{\alpha t} \cos \beta t$, with r as a natural number, while α and β are real numbers. As $t \rightarrow \infty$, so $t^r \rightarrow \infty$. Because $\cos \beta t$ does not tend to 0 as $t \rightarrow \infty$ for any value of β , the condition $\alpha < 0$ is *necessary* for u_j to tend to 0 as $t \rightarrow \infty$. On the other hand, if $\alpha < 0$, then $e^{\alpha t} < 1$ and thus $a = e^{-\alpha} > 1$. Hence, $t^r e^{\alpha t} = t^r / a^t \rightarrow 0$ as $t \rightarrow \infty$. (See e.g. EMEA, (7.12.3).) Because $|\cos \beta t| \leq 1$, we conclude that $u_j \rightarrow 0$ as $t \rightarrow \infty$. The condition $\alpha < 0$ is therefore necessary as well as sufficient for u_j to tend to 0 as $t \rightarrow \infty$.

7.3.1

A necessary and sufficient condition for

$$\frac{d^n x}{dt^n} + a_1 \frac{d^{n-1} x}{dt^{n-1}} + \cdots + a_n x = f(t)$$

to be globally asymptotically stable is that every root of the characteristic equation $r^n + a_1 r^{n-1} + \cdots + a_n = 0$ has a negative real part.

To check if (1) is globally asymptotically stable, therefore, it suffices to find the roots of the characteristic equation. These depend only on the coefficients $a_1, a_2, \dots, a_n$.

The case $n = 1$ is easy. The characteristic equation of $\dot{x} + a_1 x = f(t)$ is $r + a_1 = 0$, so the characteristic root is $r = -a_1$. Thus the equation is globally asymptotically stable if and only if $a_1 > 0$. For $n = 2$ it was proved in (6.4.3) that $\ddot{x} + a\dot{x} + bx = f(t)$ is globally asymptotically stable if and only if $a > 0$ and $b > 0$. On the basis of these results it is easy to find a *necessary* condition for (1) to be globally asymptotically stable:

$$\text{If (1) is globally asymptotically stable, then } a_1, \dots, a_n \text{ are all positive} \quad (2)$$

To see this, note that the characteristic polynomial $p(r) = r^n + a_1 r^{n-1} + \cdots + a_n$ can be decomposed into its first and second degree factors, i.e. factors of the form $r + c$ for real roots of the equation $p(r) = 0$ and $r^2 + ar + b$ for complex conjugate pairs of roots. If all zeros of $p(r)$ have negative real parts, then those of $r + c$ and $r^2 + ar + b$ must have negative real parts. So c , a , and b must be positive. As a product of polynomials with positive coefficients, $p(r)$ has positive coefficients only.

Except for the cases $n = 1$ and $n = 2$, the condition that $a_1, \dots, a_n$ are all positive is not sufficient for stability of (1) (see Example 2 below). We state a theorem that, in conjunction with Theorem 7.3.1, provides necessary and sufficient conditions for equation (1) to be globally asymptotically stable:¹

THEOREM 7.3.2 (HURWITZ-ROUTH)

Let

$$r^n + a_1 r^{n-1} + \cdots + a_n$$

be a polynomial of degree n with real coefficients. A necessary and sufficient condition for all roots of the polynomial to have negative real parts is that all the leading principal minors in the following $n \times n$ matrix are positive:

$$A = \begin{pmatrix} a_1 & a_3 & a_5 & \cdots & 0 & 0 \\ 1 & a_2 & a_4 & \cdots & 0 & 0 \\ 0 & a_1 & a_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & a_{n-1} & 0 \\ 0 & 0 & 0 & \cdots & a_{n-2} & a_n \end{pmatrix}$$

The k th column of the matrix A is the vector

$$(a_{2k-1}, \dots, a_{k+1}, a_k, a_{k-1}, \dots, a_{2k-n})'$$

where a_k is on the main diagonal, $a_0 = 1$, and $a_i = 0$ if $i < 0$ or $i > n$.

¹ Leading principal minors are defined in Section 1.7. For a proof, see Gantmacher (1959).

For $n = 1, 2, 3, 4$, the matrix A is given by

$$(a_1), \begin{pmatrix} a_1 & 0 \\ 1 & a_2 \end{pmatrix}, \begin{pmatrix} a_1 & a_3 & 0 \\ 1 & a_2 & 0 \\ 0 & a_1 & a_3 \end{pmatrix}, \begin{pmatrix} a_1 & a_3 & 0 & 0 \\ 1 & a_2 & a_4 & 0 \\ 0 & a_1 & a_3 & 0 \\ 0 & 1 & a_2 & a_4 \end{pmatrix} \quad (*)$$

respectively. By combining Theorems 7.3.1 and 7.3.2 we can obtain:

$$\begin{aligned} \text{(a) } \ddot{x} + a_1 x &= f(t) \text{ is globally asymptotically stable} \iff a_1 > 0 \\ \text{(b) } \ddot{x} + a_1 \dot{x} + a_2 x &= f(t) \text{ is globally asymptotically stable} \\ &\iff a_1 > 0 \text{ and } a_2 > 0 \\ \text{(c) } \ddot{x} + a_1 \ddot{x} + a_2 \dot{x} + a_3 x &= f(t) \text{ is globally asymptotically stable} \\ &\iff a_1 > 0, a_3 > 0, \text{ and } a_1 a_2 - a_3 > 0 \end{aligned} \quad (3)$$

These equivalences are in accordance with our earlier results for $n = 1$ and $n = 2$. For $n = 3$ the requirements in Theorem 7.3.2 are:

$$a_1 > 0, \begin{vmatrix} a_1 & a_3 \\ 1 & a_2 \end{vmatrix} = a_1 a_2 - a_3 > 0, \begin{vmatrix} a_1 & a_3 & 0 \\ 1 & a_2 & 0 \\ 0 & a_1 & a_3 \end{vmatrix} = (a_1 a_2 - a_3) a_3 > 0$$

which is equivalent to (c) in (3).

1 Prove that the equation $\ddot{x} + 3\dot{x} + 7x = e^{3t}$ is globally asymptotically stable.

Solution: Here $a_1 = 3 > 0$, $a_3 = 7 > 0$, and $a_1 a_2 - a_3 = 21 - 7 = 14 > 0$, so that this third-order equation is globally asymptotically stable.

2 Prove that the equation $\ddot{x} + \ddot{x} + \dot{x} + x = \sin t$ is not globally asymptotically stable.

Solution: This fourth-order equation is globally asymptotically stable if and only if all the leading principal minors of the last matrix in (*) are positive. But we find that

$$\begin{vmatrix} a_1 & a_3 \\ 1 & a_2 \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 1 & 1 \end{vmatrix} = 0$$

Hence the equation is not globally asymptotically stable.

IS FOR SECTION 7.3

1. Check global asymptotic stability for Problem 7.2.1(a).

SM 2. Use (3) to show that the equation $\ddot{x} + 4\dot{x} + 5x = 0$ is globally asymptotically stable. Confirm the result by finding the general solution of the equation. (Hint: Recall that integer roots of the characteristic equation divide the constant term 2.)

7.4 Systems of Differential Equations

A **normal system** of n first-order equations in n variables takes the form

$$\begin{aligned} \frac{dx_1}{dt} &= f_1(t, x_1, \dots, x_n) \\ &\vdots \\ \frac{dx_n}{dt} &= f_n(t, x_1, \dots, x_n) \end{aligned} \quad (1)$$

EXAMPLE 1 Transform the following second-order system into the form (1)

$$\ddot{x}_1 = F_1(t, x_1, \dot{x}_1, x_2, \dot{x}_2), \quad \ddot{x}_2 = F_2(t, x_1, \dot{x}_1, x_2, \dot{x}_2)$$

Solution: Introduce new unknowns u_1, u_2, u_3 , and u_4 defined by $u_1 = x_1$, $u_2 = \dot{x}_1$, $u_3 = x_2$, and $u_4 = \dot{x}_2$. Then the system is transformed into the first-order system

$$\dot{u}_1 = u_2, \quad \dot{u}_2 = F_1(t, u_1, u_2, u_3, u_4), \quad \dot{u}_3 = u_4, \quad \dot{u}_4 = F_2(t, u_1, u_2, u_3, u_4)$$

which is in the form (1).

EXAMPLE 2 Prove that the n th-order differential equation $\frac{d^n x}{dt^n} = F\left(t, x, \frac{dx}{dt}, \dots, \frac{d^{n-1}x}{dt^{n-1}}\right)$ can be transformed into a normal system.

Solution: Define $y_1 = x$, $y_2 = dx/dt$, $\dots$, $y_n = d^{n-1}x/dt^{n-1}$. Then the given equation takes the form

$$\dot{y}_1 = y_2, \quad \dot{y}_2 = y_3, \quad \dots, \quad \dot{y}_{n-1} = y_n, \quad \dot{y}_n = F(t, y_1, y_2, \dots, y_n)$$

This is a normal system of n first-order equations.

The examples above indicate that general systems of n th-order differential equations can usually be reduced to the normal form (1). Therefore, in studying systems of differential equations, not much is lost if we restrict attention to normal systems.

A **solution** of (1) is a set of functions $x_1 = x_1(t), \dots, x_n = x_n(t)$ that satisfy all the equations. Geometrically, such a solution describes a curve in $\mathbb{R}^n$. The parameter t usually denotes time, and as t varies, we say that the system "moves along" the curve. The vector $\dot{\mathbf{x}}(t) = (\dot{x}_1(t), \dots, \dot{x}_n(t))$ is the **velocity vector** associated with $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))$. The space with coordinates $x_1, \dots, x_n$ is called the **phase space** associated with system (1).

In the case $n = 3$, as t varies, the vector $\mathbf{x}(t) = (x_1(t), x_2(t), x_3(t))$ traces out a curve in $\mathbb{R}^3$, and

$$\frac{\mathbf{x}(t + \Delta t) - \mathbf{x}(t)}{\Delta t} = \left(\frac{x_1(t + \Delta t) - x_1(t)}{\Delta t}, \frac{x_2(t + \Delta t) - x_2(t)}{\Delta t}, \frac{x_3(t + \Delta t) - x_3(t)}{\Delta t} \right)$$

tends to the vector $\dot{\mathbf{x}}(t) = (\dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t))$ as a limit as $\Delta t \rightarrow 0$. As in the case $n = 2$ in Section 6-5 (see Fig. 6.5.1), $\dot{\mathbf{x}}(t)$ is a tangent vector to the curve.

Let $F(t, \mathbf{x}(t))$ denote the vector with components $f_i(t, \mathbf{x}(t)) = f_i(t, x_1(t), \dots, x_n(t))$, $i = 1, \dots, n$. Then (1) can be written in the concise form

$$\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x}) \quad (2)$$

In economic models that lead to systems of the type (1), the functions $x_1(t), \dots, x_n(t)$ are state variables characterizing the given economic system at time t . Usually the state of the system at some definite time t_0 is known, so $x(t_0) = (x_1(t_0), \dots, x_n(t_0))$ is given. Now, according to the existence and uniqueness theorem for (1), if f_i and $\partial f_i / \partial x_j$ are continuous for all $i = 1, \dots, n, j = 1, \dots, n$, then there is one and only one vector of functions $x_1(t), \dots, x_n(t)$ that satisfies (1) and has the prescribed values for $t = t_0$ (see Theorem 7.6.1).

In the case of (1), the general solution usually depends on n arbitrary constants,

$$x_1 = \varphi_1(t; C_1, \dots, C_n), \quad \dots, \quad x_n = \varphi_n(t; C_1, \dots, C_n)$$

For each choice of $C_1, \dots, C_n$ the solution describes a curve in $\mathbb{R}^n$.

In Example 2 we showed that an n th-order differential equation can be transformed into a system of first-order equations. Sometimes one can find the solution of system (1) by going the other way around. The method was illustrated for the case $n = 2$ in Section 6.5.

Linear Systems

In some models the functions $f_1, \dots, f_n$ appearing in (1) are linear, or it may be an acceptable approximation to treat them as linear. Then the system is

$$\begin{aligned} \dot{x}_1 &= a_{11}(t)x_1 + \dots + a_{1n}(t)x_n + b_1(t) \\ &\dots\dots\dots(3) \\ \dot{x}_m &= a_{m1}(t)x_1 + \dots + a_{mn}(t)x_n + b_m(t) \end{aligned}$$

By a method similar to that applied to the case $n = 2$ in Section 6.5, the problem of solving (3) can be transformed into the problem of solving one n th-order linear differential equation in one unknown function, say x_1 . When x_1 has been found, we can also find $x_2, \dots, x_n$.

Let $\mathbf{x}$, $\dot{\mathbf{x}}$ and $\mathbf{b}(t)$ be the three column vectors with components $x_1, \dots, x_n, \dot{x}_1, \dots, \dot{x}_n$ and $b_1(t), \dots, b_n(t)$, respectively. Let $\mathbf{A}$ denote the matrix $(a_{ij}(t))_{(n \times n)}$. Then (3) can be written in the matrix form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{b}(t) \quad (4)$$

A particularly important case occurs when all the functions $a_{ij}(t)$ are constants. Then

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}(t) \iff \dot{x}_i = a_{i1}x_1 + \dots + a_{in}x_n + b_i(t), \quad i = 1, \dots, n \quad (5)$$

In the same way that we derived (6.5.4) from (6.5.3), we deduce from (5) an n th-order

Their characteristic equations coincide with the eigenvalue equation for the matrix $\mathbf{A}$, as we saw in Section 6.5 for the case $n = 2$. On the basis of Theorem 7.3.1 we deduce the following result:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}(t) \text{ is globally asymptotically stable} \iff \begin{matrix} \text{all the eigenvalues of } \mathbf{A} \text{ have negative real parts} \end{matrix} \quad (6)$$

Global asymptotic stability of $\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{b}(t)$ means, in particular, that the general solution of $\dot{\mathbf{x}} = \mathbf{Ax}$ tends to the zero vector $\mathbf{0}$ as $t \rightarrow \infty$, regardless of the initial conditions.

Solutions Based on Eigenvalues

The system

$$\dot{x} = Ax + b \quad (7)$$

can alternatively be solved by using methods from linear algebra, as shown for $n = 2$ in Section 6.5. Suppose first that $\mathbf{h} = \mathbf{0}$. We search for numbers λ and $v_1, v_2, \dots, v_n$ such that the vector function $\mathbf{x} = \mathbf{v}e^{\lambda t} = (v_1 e^{\lambda t}, v_2 e^{\lambda t}, \dots, v_n e^{\lambda t})$ satisfies $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. With $\mathbf{x} = \mathbf{v}e^{\lambda t}$ we have $\dot{\mathbf{x}} = \lambda e^{\lambda t} \mathbf{v}$, so $\lambda e^{\lambda t} \mathbf{v} = \mathbf{A}(\mathbf{v}e^{\lambda t}) = e^{\lambda t} \mathbf{A}\mathbf{v}$. Cancelling the common factor $e^{\lambda t}$ yields

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \quad (8)$$

Hence any nonzero solution $\mathbf{v}$ is an eigenvector of the matrix $\mathbf{A}$ with eigenvalue λ .

The case where A has n different real eigenvalues, $\lambda_1, \lambda_2, \dots, \lambda_n$, is the simplest. Then A has n linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$, and the general solution of $\dot{\mathbf{x}} = A\mathbf{x}$ is

$$x(t) = C_1 e^{\lambda_1 t} v_1 + \dots + C_n e^{\lambda_n t} v_n$$

Suppose that $\mathbf{x}^0$ is an equilibrium point for (7) in the sense that $\mathbf{A}\mathbf{x}^0 + \mathbf{b} = 0$. If we define $\mathbf{w} = \mathbf{x} - \mathbf{x}^0$, then $\mathbf{w}$ measures the deviation of $\mathbf{x}$ from the equilibrium state $\mathbf{x}^0$. Then $\dot{\mathbf{w}} = \dot{\mathbf{x}}$, which inserted into (7) gives $\dot{\mathbf{w}} = \dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{A}(\mathbf{w} + \mathbf{x}^0) + \mathbf{b} = \mathbf{A}\mathbf{w} + \mathbf{A}\mathbf{x}^0 + \mathbf{b} = \mathbf{A}\mathbf{w}$. In this way the nonhomogeneous system (7) can be reduced to a homogeneous system.

NOTE 1 The solution of the scalar initial value problem $\dot{x} = ax$, $x(t_0) = x^0$, is $x = e^{a(t-t_0)}x^0$. It is tempting to conjecture that the more general initial value problem $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$, $\mathbf{x}(t_0) = \mathbf{x}^0$ has the solution $\mathbf{x} = e^{\mathbf{A}(t-t_0)}\mathbf{x}^0$. This is correct if e to the power of a matrix is properly defined. In fact, if $\mathbf{A}$ is an $n \times n$ -matrix and t is any number, we define

$$e^{At} = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2!}\mathbf{A}^2 + \frac{t^3}{3!}\mathbf{A}^3 + \dots \quad (9)$$

One can show that this series converges for all t , and that $(d/dt)e^{At} = Ae^{At}$. Since $(d/dt)(e^{At}\mathbf{c}) = Ae^{At}\mathbf{c} = A(e^{At}\mathbf{c})$, it follows that $e^{At}\mathbf{c}$ is a solution to $\dot{\mathbf{x}} = A\mathbf{x}$ for every constant vector $\mathbf{c}$, and only solutions of this kind are possible.

Moreover, one can show that $(e^{A\tau})^{-1} = e^{-A\tau}$ and that $e^{A(t+s)} = e^{At}e^{As}$. Note, however, that e^{At+Bt} is *not* equal to $e^{At}e^{Bt}$ unless $AB = BA$.

The Resolvent

We shall explain how one can (in principle) solve the general linear equation system (4). Consider first the homogeneous system

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} \quad (*)$$

For any fixed t_0 and for each $j = 1, \dots, n$, this equation has a unique vector solution $\mathbf{p}_j(t) = (p_{1j}(t), \dots, p_{nj}(t))'$, $t \in \mathbb{R}$, satisfying $\mathbf{p}_j(t_0) = \mathbf{e}_j$, where $\mathbf{e}_j$ is the j th standard unit vector in $\mathbb{R}^n$. (Existence and uniqueness follow from Theorem 7.6.2 below.) The **resolvent** of $(*)$ is the $n \times n$ matrix

$$\mathbf{P}(t, t_0) = \begin{pmatrix} p_{11}(t) & \cdots & p_{1n}(t) \\ \vdots & \ddots & \vdots \\ p_{n1}(t) & \cdots & p_{nn}(t) \end{pmatrix} \quad (10)$$

whose columns are the solutions $\mathbf{p}_j(t)$. Considering each column separately, this matrix evidently satisfies $\mathbf{P}(t, t_0) = (d/dt)\mathbf{P}(t, t_0) = \mathbf{A}(t)\mathbf{P}(t, t_0)$ and $\mathbf{P}(t_0, t_0) = \mathbf{I}_n$, where $\mathbf{I}_n$ is the unit matrix of order n . The solution of $(*)$ with $\mathbf{x}(t_0) = \mathbf{x}^0 = x_1^0 \mathbf{e}_1 + \cdots + x_n^0 \mathbf{e}_n$ is obviously $\mathbf{x}(t) = x_1^0 \mathbf{p}_1(t) + \cdots + x_n^0 \mathbf{p}_n(t) = \mathbf{P}(t, t_0)\mathbf{x}^0$.

For the nonhomogeneous equation (4), one can then derive the following:

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{b}(t), \quad \mathbf{x}(t_0) = \mathbf{x}^0 \iff \mathbf{x}(t) = \mathbf{P}(t, t_0)\mathbf{x}^0 + \int_{t_0}^t \mathbf{P}(t, s)\mathbf{b}(s) ds \quad (11)$$

To show that this $\mathbf{x}(t)$ is a solution, we observe from Leibniz's formula that

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \dot{\mathbf{P}}(t, t_0)\mathbf{x}^0 + \mathbf{P}(t, t)\mathbf{b}(t) + \int_{t_0}^t \dot{\mathbf{P}}(t, s)\mathbf{b}(s) ds \\ &= \mathbf{A}(t)\mathbf{P}(t, t_0)\mathbf{x}^0 + \int_{t_0}^t \mathbf{A}(t)\mathbf{P}(t, s)\mathbf{b}(s) ds + \mathbf{b}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t) \end{aligned}$$

The matrix $\mathbf{P}(t, s)$ denotes the resolvent at time t when the initial point of time is s . Because all the components of each vector $\mathbf{p}_j(t)$ are unique, it follows that $\mathbf{P}(t, t_0) \cdot \mathbf{P}(t_0, t) = \mathbf{I}$, so $\mathbf{P}(t, t_0)$ has an inverse. More generally (also by uniqueness), $\mathbf{P}(t, s) = \mathbf{P}(t, \tau) \cdot \mathbf{P}(\tau, s)$ for all t, s , and τ . In particular, $\mathbf{P}(t, s) = \mathbf{P}(t, t_0)(\mathbf{P}(s, t_0))^{-1}$.

If $\mathbf{A}(t)$ is the constant matrix $\mathbf{A}$, then $\mathbf{P}(t, s) = \mathbf{P}(t - s, 0) = e^{\mathbf{A}(t-s)}$ for all t and s . Hence,

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}(t), \quad \mathbf{x}(t_0) = \mathbf{x}^0 \iff \mathbf{x} = \mathbf{P}(t, t_0)\mathbf{x}^0 + \int_{t_0}^t \mathbf{P}(t - s, 0)\mathbf{b}(s) ds \quad (12)$$

Note finally that if $\mathbf{P}(t, s)$ is the resolvent of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$, then $t \mapsto \mathbf{P}(s, t)'$ (the transpose of $\mathbf{P}(s, t)$) is the resolvent of the equation

$$\dot{\mathbf{z}}(t) = -\mathbf{A}(t)'\mathbf{z}(t) \quad (13)$$

To prove that $t \mapsto \mathbf{P}(s, t)'$ satisfies (13), first differentiate $\mathbf{P}(t, s)\mathbf{P}(s, t) = \mathbf{I}_n$ w.r.t. t to obtain $((\partial/\partial t)\mathbf{P}(t, s))\mathbf{P}(s, t) + \mathbf{P}(t, s)(\partial/\partial t)\mathbf{P}(s, t) = \mathbf{0}$, which implies that $\mathbf{A}(t)\mathbf{P}(t, s)\mathbf{P}(s, t) + \mathbf{P}(t, s)(\partial/\partial t)\mathbf{P}(s, t) = \mathbf{0}$, or $(\partial/\partial t)\mathbf{P}(s, t) = -\mathbf{P}(t, s)^{-1}\mathbf{A}(t) = -\mathbf{P}(s, t)\mathbf{A}(t)$, i.e. $(\partial/\partial t)\mathbf{P}(s, t)' = -\mathbf{A}(t)'\mathbf{P}(s, t)'$.

7.5 Stability for Nonlinear Systems

This section generalizes the stability theory of autonomous systems in the plane to the more general autonomous system

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \dots, x_n) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, \dots, x_n) \end{aligned} \quad (1)$$

in n dimensions. We assume that $f_1, \dots, f_n$ are all C^1 functions.

A point $\mathbf{a} = (a_1, \dots, a_n)$ is called an **equilibrium point** for (1) if

$$f_1(a_1, \dots, a_n) = 0, \quad \dots, \quad f_n(a_1, \dots, a_n) = 0$$

Note that $x_1 = x_1(t) = a_1, \dots, x_n = x_n(t) = a_n$ is then a solution of the system. If $x_1, \dots, x_n$ are state variables for some economic (or biological or physical) system, and (1) is satisfied, then $\mathbf{a}$ is an **equilibrium state**.

An equilibrium point $\mathbf{a}$ is **stable** if all nearby solution curves remain nearby. The point $\mathbf{a}$ is **locally asymptotically stable** if each solution curve that starts near $\mathbf{a}$ not only stays near $\mathbf{a}$ but, in addition, converges to $\mathbf{a}$. We now define precisely the concepts of stable and locally asymptotically stable equilibrium points:

DEFINITION OF STABILITY

The equilibrium state $\mathbf{a} = (a_1, \dots, a_n)$ for system (1) is **stable** if for each $\varepsilon > 0$ there exists a $\delta > 0$ (that generally depends on ε) such that every solution $\mathbf{x}(t)$ of (1) with $\|\mathbf{x}(t_0) - \mathbf{a}\| < \delta$ for some t_0 is defined for all $t > t_0$ and satisfies the inequality

$$\|\mathbf{x}(t) - \mathbf{a}\| < \varepsilon \quad \text{for all } t > t_0$$

If $\mathbf{a}$ is stable and, in addition, there exists a $\delta_0 > 0$ such that

$$\|\mathbf{x}(t_0) - \mathbf{a}\| < \delta_0 \implies \lim_{t \rightarrow \infty} \|\mathbf{x}(t) - \mathbf{a}\| = 0$$

then $\mathbf{a}$ is called **locally asymptotically stable**.

An equilibrium that is not stable is called **unstable**. See the illustrations of these concepts in Fig. 1. For an example of an equilibrium state that is stable, but not asymptotically stable, see the Lotka-Volterra model in Example 2 below.

Theorem 6.8.1, on locally asymptotically stable equilibrium points for systems of autonomous equations in the plane, has a natural extension to n dimensions (see Hirsch

1. Find the general solution of the following system

$$\dot{x}_1 = -x_1 + x_2 + x_3, \quad \dot{x}_2 = x_1 - x_2 + x_3, \quad \dot{x}_3 = x_1 + x_2 + x_3$$

by each of the following three methods: (i) solve a third-order differential equation for x_1 ; (ii) find the eigenvalues and eigenvectors of the corresponding matrix $\mathbf{A}$; (iii) find the resolvent.

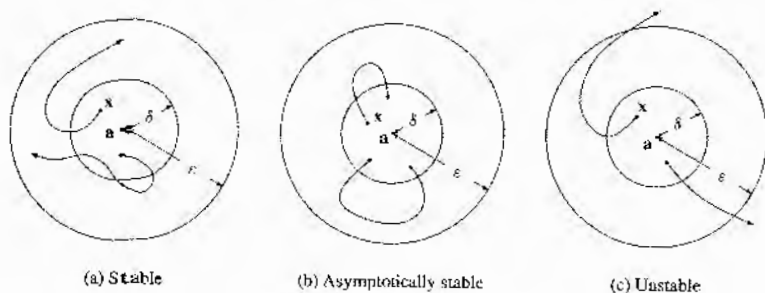


Figure 1

7.5.1 (LYAPUNOV)

Let $\mathbf{a} = (a_1, \dots, a_n)$ be an equilibrium point for system (1) and let $\mathbf{A}$ be the Jacobian matrix

$$\mathbf{A} = \begin{pmatrix} \frac{\partial f_1(\mathbf{a})}{\partial x_1} & \dots & \frac{\partial f_1(\mathbf{a})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{a})}{\partial x_1} & \dots & \frac{\partial f_n(\mathbf{a})}{\partial x_n} \end{pmatrix} \quad (2)$$

If all the eigenvalues of $\mathbf{A}$ have negative real parts, then $\mathbf{a}$ is locally asymptotically stable. If at least one eigenvalue of $\mathbf{A}$ has positive real part, then $\mathbf{a}$ is unstable.

NOTE 1 The condition that all the eigenvalues of $\mathbf{A}$ have negative real parts is sufficient but not necessary for $\mathbf{a}$ to be locally asymptotically stable. For instance, $\dot{x} = -x^3$ has the general solution $x = C(2C^2t + 1)^{-1/2}$, with $a = 0$ as a locally asymptotically stable point. But the matrix $\mathbf{A}$ has the eigenvalue $f'(0) = 0$.

NOTE 2 Let $x = a$ be an equilibrium point for the scalar equation $\dot{x} = F(x)$. The matrix $\mathbf{A}$ in (2) is then the 1×1 matrix $(F'(a))$, and the only eigenvalue of $\mathbf{A}$ is $\lambda = F'(a)$. Thus, we conclude from the theorem that $x = a$ is locally asymptotically stable provided $F'(a) < 0$, and $x = a$ is unstable provided $F'(a) > 0$. This accords with the results in (5.7.2).

Lyapunov Functions

Let $\mathbf{a} = (a_1, \dots, a_n)$ be an equilibrium point for system (1) and let $V(\mathbf{x})$ be a C^1 function defined in an open neighbourhood Ω of $\mathbf{a}$. We call $V(\mathbf{x})$ **positive definite** in Ω if

$$V(\mathbf{a}) = 0 \text{ and } V(\mathbf{x}) > 0 \text{ for all } \mathbf{x} \text{ in } \Omega \text{ with } \mathbf{x} \neq \mathbf{a}$$

Thus $V(\mathbf{x})$ is positive definite in Ω if it has a unique minimum at $\mathbf{a}$, with minimum value 0.

Let $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))$ be a solution of (1). The derivative of $V(\mathbf{x}(t))$ w.r.t. t is

$$\frac{d}{dt} V(\mathbf{x}(t)) = \sum_{i=1}^n \frac{\partial V(\mathbf{x}(t))}{\partial x_i} \frac{dx_i}{dt} = \sum_{i=1}^n \frac{\partial V(\mathbf{x}(t))}{\partial x_i} f_i(\mathbf{x}(t)) = \nabla V(\mathbf{x}(t)) \cdot \mathbf{f}(\mathbf{x}(t)) \quad (3)$$

Define $\dot{V}(\mathbf{x}) = \nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = \sum_{i=1}^n V_i'(\mathbf{x}) f_i'(\mathbf{x})$. If V is positive definite and $\dot{V}(\mathbf{x}) < 0$ for all $\mathbf{x}$ in Ω , then $V(\mathbf{x})$ is called a **Lyapunov function** for system (1). If in addition $\dot{V}(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{a}$ in Ω , then $V(\mathbf{x})$ is called a **strong (or strict) Lyapunov function** for the system.

THEOREM 7.5.2 (LYAPUNOV)

Let $\mathbf{a} = (a_1, \dots, a_n)$ be an equilibrium point for system (1). If there is a Lyapunov function for the system in an open neighbourhood Ω of $\mathbf{a}$, then $\mathbf{a}$ is a stable equilibrium point. If there is a strong Lyapunov function for the system, then $\mathbf{a}$ is locally asymptotically stable.

Proof: Choose $\varepsilon > 0$ so small that $\bar{B} = \{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\| \leq \varepsilon\} \subseteq \Omega$. Let $V(\mathbf{x})$ be a Lyapunov function. The boundary $A = \{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\| = \varepsilon\}$ of $\bar{B}$ is compact, and since V is continuous, V has a minimum value α over A . Because $V(\mathbf{x})$ is positive definite, $\alpha > 0$. By continuity of V , we can choose a $\delta > 0$ such that $V(\mathbf{x}) < \alpha$ for all $\mathbf{x}$ in the ball $C = \{\mathbf{x} : \|\mathbf{x} - \mathbf{a}\| < \delta\}$. Then $\delta \leq \varepsilon$ and $C \subseteq \bar{B}$. Also, by definition of a Lyapunov function, V cannot increase along any solution curve. So if a solution curve for (1) starts at a point $\mathbf{x}_0$ in C , then that solution curve can never meet A . Hence any solution starting in C never leaves $\bar{B}$. Thus $\mathbf{a}$ is stable.

Now suppose that V is a **strong** Lyapunov function. We want to prove that if $\mathbf{x}(t)$ is a solution starting at a point $\mathbf{x}(0) = \mathbf{x}_0$ in C , then $\mathbf{x}(t) \rightarrow \mathbf{a}$ as $t \rightarrow \infty$. Evidently, $V(\mathbf{x}(t))$ converges to some limit $V^* \geq 0$. By the mean value theorem, for each $k = 1, 2, \dots$ there exists a t_k in $(k, k+1)$ such that $\dot{V}(\mathbf{x}(t_k)) = V(\mathbf{x}(k+1)) - V(\mathbf{x}(k))$. By compactness of $\bar{B}$, a subsequence $\{\mathbf{x}(t_{k_j})\}$ converges to some point $\mathbf{x}^*$ in $\bar{B}$. Taking limits in the last equality, we get $\dot{V}(\mathbf{x}^*) = V^* - V^* = 0$. But $\mathbf{x}^* = \mathbf{a}$ is the only point where $\dot{V} = 0$. Hence, $0 = V(\mathbf{a}) = V(\mathbf{x}^*) = \lim_{j \rightarrow \infty} V(\mathbf{x}(t_{k_j})) = \lim_{j \rightarrow \infty} V(\mathbf{x}(t)) = V^*$. For a contradiction, assume that $\mathbf{x}(t)$ does not converge to $\mathbf{a}$. Then for some ε there exists a sequence $\{t^k\}$ such that $t^k \rightarrow \infty$ and $\|\mathbf{x}(t^k) - \mathbf{a}\| \geq \varepsilon$ for all k . By compactness of $\bar{B}$ the sequence $\{t^k\}$ has a subsequence $\{s_j\} = \{t^{k_j}\}$ such that $\{\mathbf{x}(s_j)\}$ converges to a point $\mathbf{x}_*$ in $\bar{B}$. Then $V(\mathbf{x}_*) = \lim_{j \rightarrow \infty} V(\mathbf{x}(s_j)) = 0$, so $\mathbf{x}_* = \mathbf{a}$, contradicting $\|\mathbf{x}(s_j) - \mathbf{a}\| \geq \varepsilon$ for all j . ■

NOTE 3 Actually, a proof is needed of the fact that a solution starting at $\mathbf{x}_0$ at $t = 0$ can be extended to a solution defined for all $t > 0$. (See e.g. Hirsch et al. (2004).)

According to our previous terminology, an equilibrium point for (1) is **globally asymptotically stable** if every solution of (1) (wherever it starts) converges to $\mathbf{a}$ as t tends to ∞ . A slight modification of the proof of Theorem 7.5.2 yields the following result:

THEOREM 7.5.3

Let $\mathbf{a} = (a_1, \dots, a_n)$ be an equilibrium point for system (1), and assume that there exists a strong Lyapunov function $V(\mathbf{x})$ for (1) that is defined in all of $\mathbb{R}^n$ and is such that $V(\mathbf{x}) \rightarrow \infty$ as $\|\mathbf{x}\| \rightarrow \infty$. Then $\mathbf{a}$ is globally asymptotically stable.

Proof: Consider a solution with $x(0) = x_0$. Let $\tilde{B} = \{x \in \mathbb{R}^n : V(x) \leq V(x_0)\}$. The set $\tilde{B}$ is bounded because $V(x) > V(x_0)$ when $\|x\|$ is sufficiently large, and $\tilde{B}$ is closed because V is continuous. Hence, $\tilde{B}$ is compact. Now proceed as in the second paragraph of the proof of Theorem 7.5.2.

E 1 Prove that $V(x, y) = x^2 + y^2$ is a Lyapunov function for the system

$$\dot{x} = -x - y, \quad \dot{y} = x - y$$

with equilibrium point $(0, 0)$. Prove that $(0, 0)$ is globally asymptotically stable.

Solution: $V(x, y)$ is clearly positive definite, and $\dot{V} = 2x\dot{x} + 2y\dot{y} = -2x^2 - 2y^2 < 0$ for all $(x, y) \neq (0, 0)$, so V is a Lyapunov function for the system. In fact, V is a *strong* Lyapunov function, defined over the entire plane, and $V(x, y) = x^2 + y^2 \rightarrow \infty$ as $\|(x, y)\| = \sqrt{x^2 + y^2} \rightarrow \infty$, so Theorem 7.5.3 tells us that $(0, 0)$ is globally asymptotically stable.

E 2 Consider the celebrated Lotka–Volterra predator–prey model

$$\dot{x} = x(k - ay), \quad \dot{y} = y(-h + bx) \quad (i)$$

with a, b, h , and k all positive constants. Here x is the population of prey (say rabbits) and y is the population of predators (say foxes). The rate of rabbit population growth is a decreasing function of the fox population, but the rate of fox population growth is an increasing function of the rabbit population.

Note that there are two equilibrium points, $(0, 0)$ and $(x_0, y_0) = (h/b, k/a)$. By studying a phase diagram it is easy to see that $(0, 0)$ is not stable. (See Problem 6.7.3.)

To check for stability of (x_0, y_0) , consider the function

$$H(x, y) = b(x - x_0 \ln x) + a(y - y_0 \ln y) \quad (ii)$$

We claim that $L(x, y) = H(x, y) - H(x_0, y_0)$ is a Lyapunov function for (i). (To understand how to arrive at this function, see Problem 4.) Note that $L'_x = b(1 - x_0/x)$ and $L'_y = a(1 - y_0/y)$, so (x_0, y_0) is a stationary point for L . Moreover, $L''_{xx} = bx_0/x^2 > 0$, $L''_{yy} = ay_0/y^2 > 0$, and $L''_{xy} = 0$, so L is strictly convex for all $x > 0, y > 0$. It follows that (x_0, y_0) is the unique minimum point for L . But $L(x_0, y_0) = 0$, so $L(x, y)$ is positive definite. Moreover, with $x_0 = h/b, y_0 = k/a$, we obtain for all $x > 0, y > 0$,

$$\begin{aligned} \dot{L} &= b\left(1 - \frac{x_0}{x}\right)\dot{x} + a\left(1 - \frac{y_0}{y}\right)\dot{y} \\ &= b\left(1 - \frac{x_0}{x}\right)x(k - ay) + a\left(1 - \frac{y_0}{y}\right)y(-h + bx) = 0 \end{aligned} \quad (iii)$$

We conclude from Theorem 7.5.2 that (x_0, y_0) is a stable equilibrium point.

In fact, (iii) implies that $L(x, y)$ is constant along solution curves for (i). One can prove that the curves $L(x, y) = \text{constant}$ are closed curves. The populations of predator and prey will therefore oscillate cyclically. For further discussion of this model see Hirsch et al. (2004), Chom. 11. (See also Fig. A6.7.3 in the answer section.)

PROBLEMS FOR SECTION 7.5

1. Prove that $(0, 0)$ is a locally asymptotically stable point for each of the following systems. Is it also globally asymptotically stable?

$$(a) \quad \dot{x} = -y - x^3, \quad \dot{y} = x - y^3. \quad (\text{Try } V(x, y) = x^2 + y^2.)$$

$$(b) \quad \dot{x} = -\frac{7}{4}x + \frac{1}{4}y, \quad \dot{y} = \frac{3}{4}x - \frac{1}{4}y. \quad (\text{Try } V(x, y) = 12x^2 + 12xy + 20y^2.)$$

- (c) Test the stability of (b) by using Theorem 7.5.1 as well.

2. Consider the following differential equation for $p > 0$:

$$\dot{p} = a(b/p - c), \quad \text{where } a, b \text{ and } c \text{ are positive constants}$$

Find the equilibrium point and prove that it is locally asymptotically stable by using Theorem 7.5.2. (Hint: $V(p) = (p - b/c)^2$.)

3. Consider the system of differential equations

$$\dot{x}_i = u'_i(x_1, \dots, x_n), \quad i = 1, \dots, n$$

where $u(x) = u(x_1, \dots, x_n)$ is C^1 in an open set Ω about $\mathbf{0} = (0, \dots, 0)$. Suppose $u(x)$ has a global maximum at $\mathbf{0}$ and that $\nabla u(x) \neq \mathbf{0}$ when $x \neq \mathbf{0}$. Prove that $\mathbf{0}$ is locally asymptotically stable. (Hint: Put $V(x) = u(\mathbf{0}) - u(x)$ and use Theorem 7.5.2.)

- SM** 4. (a) Consider a system of differential equations of the form

$$\dot{x} = f_1(x)f_2(y), \quad \dot{y} = g_1(x)g_2(y)$$

Eliminate t from the system to obtain

$$\frac{dy}{dx} = \frac{\dot{y}}{\dot{x}} = \frac{g_1(x)g_2(y)}{f_1(x)f_2(y)} \quad \text{at points where } \dot{x} \neq 0 \quad (*)$$

Deduce that $H(x, y) = \int (g_1(x)/f_1(x)) dx - \int (f_2(y)/g_2(y)) dy$ is constant along each solution curve.

- (b) Show that for the Lotka–Volterra system (i) in Example 2, $H(x, y)$ is given by (ii) in Example 2.

- SM** 5. Consider the following generalization of the Lotka–Volterra system:

$$\dot{x} = kx - axy - \epsilon x^2, \quad \dot{y} = -hy + bxy - \delta y^2$$

where a, b, h, k, δ , and ϵ are positive constants, with $bk > h\epsilon$. (In particular, the rabbit population grows logistically in the absence of foxes.) Verify that $x_0 = (ah + k\delta)/(ab + \delta\epsilon)$, $y_0 = (bk - h\epsilon)/(ab + \delta\epsilon)$ is an equilibrium point. Prove that $L(x, y)$ defined in Example 2, with (x_0, y_0) as given in the present case, is a Lyapunov function with $\dot{L} = -\epsilon b(x - x_0)^2 - a\delta(y - y_0)^2$ along a solution curve. Conclusion?

7.6 Qualitative Theory

The local and global existence and uniqueness theorems of Section 5.8 can be generalized to systems of differential equations in n variables, regarded as vector differential equations in n dimensions. The motivation and the examples that were presented for the scalar case are important aids towards understanding the general theorems.

THEOREM 7.6.1 (LOCAL EXISTENCE AND UNIQUENESS)

Consider the initial value problem

$$\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1)$$

Suppose that the elements of the vector $\mathbf{F}(t, \mathbf{x})$ and the matrix $\mathbf{F}'_{\mathbf{x}}(t, \mathbf{x})$ are continuous over the $(n+1)$ -dimensional rectangle $\Gamma = \{(t, \mathbf{x}) : |t - t_0| \leq a, \|\mathbf{x} - \mathbf{x}_0\| \leq b\}$, and let

$$M = \max_{(t, \mathbf{x}) \in \Gamma} \|\mathbf{F}(t, \mathbf{x})\|, \quad r = \min(a, b/M) \quad (2)$$

Then (1) has a unique solution $\mathbf{x}(t)$ on $(t_0 - r, t_0 + r)$ and $\|\mathbf{x}(t) - \mathbf{x}_0\| \leq b$ in this interval.

THEOREM 7.6.2 (GLOBAL EXISTENCE AND UNIQUENESS)

Consider the initial value problem (1). Suppose that the elements of the vector $\mathbf{F}(t, \mathbf{x})$ and the matrix $\mathbf{F}'_{\mathbf{x}}(t, \mathbf{x})$ are continuous functions for all $(t, \mathbf{x})$, and suppose that there exist continuous scalar functions $a(t)$ and $b(t)$ such that

$$\|\mathbf{F}(t, \mathbf{x})\| \leq a(t)\|\mathbf{x}\| + b(t) \quad \text{for all } (t, \mathbf{x}) \quad (3)$$

Given an arbitrary point $(t_0, \mathbf{x}_0)$, there exists a unique solution $\mathbf{x}(t)$ of (1), defined on $(-\infty, \infty)$. If (3) is replaced by the requirement

$$\mathbf{x} \cdot \mathbf{F}(t, \mathbf{x}) \leq a(t)\|\mathbf{x}\|^2 + b(t) \quad \text{for all } \mathbf{x} \text{ and all } t \geq t_0 \quad (4)$$

then the initial value problem (1) has a unique solution defined on $[t_0, \infty)$.

The notes to Theorems 5.8.2 and 5.8.3 for the scalar case are relevant also for Theorems 7.6.1 and 7.6.2. (For proofs of these theorems and of Theorem 7.6.3 below, see Hartman (1982).)

NOTE 1 Condition (3) is satisfied if, for all $(t, \mathbf{x})$,

$$\sup_{\|\mathbf{y}\| \leq 1} \|\mathbf{F}'_{\mathbf{x}}(t, \mathbf{x})\mathbf{y}\| \leq c(t) \quad \text{for some continuous function } c(t) \quad (5)$$

Dependence on Initial Conditions

How does the solution of a differential equation change when the initial conditions change? A precise result is formulated in the next theorem. We need to spell out some crucial assumptions:

(B) For each $(t, \mathbf{x})$ in A there exist a number $r > 0$ and an interval (a, b) that contains t , with $(a, b) \times B(\mathbf{x}; r) \subseteq A$, and a constant L such that for all $\mathbf{x}', \mathbf{x}''$ in $B(\mathbf{x}; r)$ and for all t in (a, b) ,

$$\|\mathbf{F}(t, \mathbf{x}') - \mathbf{F}(t, \mathbf{x}'')\| \leq L\|\mathbf{x}' - \mathbf{x}''\| \quad (\mathbf{F} \text{ is locally Lipschitz continuous w.r.t. } \mathbf{x}) \quad (6)$$

THEOREM 7.6.3 (CONTINUOUS DEPENDENCE ON INITIAL CONDITIONS)

Consider the vector differential equation $\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x})$, where $\mathbf{F}(t, \mathbf{x})$ satisfies both (A) and the Lipschitz condition (B). Suppose that $\tilde{\mathbf{x}}(t)$ is a solution of the equation on an interval $[a, b]$ with $(t, \tilde{\mathbf{x}}(t))$ in A , and let $\tilde{t}_0 \in (a, b)$, $\tilde{\mathbf{x}}^0 = \tilde{\mathbf{x}}(\tilde{t}_0)$. Then there exists a neighbourhood $N = (\tilde{t}_0 - \alpha, \tilde{t}_0 + \alpha) \times B(\tilde{\mathbf{x}}^0; r)$ with $r > 0$ and $\alpha > 0$, such that for every $(t_0, \mathbf{x}^0)$ in N there exists a unique solution through $(t_0, \mathbf{x}^0)$ defined on $[a, b]$ whose graph lies in A . If this solution is denoted by $\mathbf{x}(t; t_0, \mathbf{x}^0)$, then for every t in $[a, b]$ the function $(t_0, \mathbf{x}^0) \mapsto \mathbf{x}(t; t_0, \mathbf{x}^0)$ is continuous in N . If $\mathbf{F}$ and $\mathbf{F}'_{\mathbf{x}}$ are continuous in A , then $\mathbf{x}(t; t_0, \mathbf{x}^0)$ is a C^1 function of $(t_0, \mathbf{x}^0)$ in N , and

$$\frac{\partial \mathbf{x}(t; t_0, \tilde{\mathbf{x}}^0)}{\partial \mathbf{x}^0} = \mathbf{P}(t, \tilde{t}_0) \quad (7)$$

$$\frac{\partial \mathbf{x}(t; t_0, \tilde{\mathbf{x}}^0)}{\partial t_0} = -\mathbf{P}(t, \tilde{t}_0) \cdot \mathbf{F}(\tilde{t}_0, \tilde{\mathbf{x}}^0) \quad (8)$$

where the $n \times n$ matrix $\mathbf{P}(t, \tilde{t}_0)$ is the resolvent of the linear differential equation

$$\dot{\mathbf{z}} = \mathbf{F}'_{\mathbf{x}}(t, \tilde{\mathbf{x}}(t))\mathbf{z} \quad (9)$$

Let us test Theorem 7.6.3 on a simple example.

EXAMPLE 1

Consider the system $\dot{x}_1 = 2x_2$, $\dot{x}_2 = x_1 + x_2$, which can be written as $\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x})$ if we put

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \mathbf{F}(t, \mathbf{x}) = \begin{pmatrix} 2x_2 \\ x_1 + x_2 \end{pmatrix}$$

Conditions (A) and (B) in Theorem 7.6.3 are satisfied everywhere for this linear system.

The general solution of $\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x})$ with $x_1(t_0) = x_1^0$ and $x_2(t_0) = x_2^0$ is

$$x_1 = \frac{2}{3}(x_1^0 - x_2^0)e^{t_0}e^{-t} + \frac{1}{3}(x_1^0 + 2x_2^0)e^{-2t_0}e^{2t}, \quad x_2 = -\frac{1}{3}(x_1^0 - x_2^0)e^{t_0}e^{-t} + \frac{1}{3}(x_1^0 + 2x_2^0)e^{-2t_0}e^{2t} \quad (i)$$

We see that x_1 and x_2 are C^1 functions of (t_0, x_1^0, x_2^0) .

In the present case, the left-hand side of (7) is the following 2×2 matrix (evaluated at (x_1^0, x_2^0, t_0) rather than at $(\tilde{x}_1^0, \tilde{x}_2^0, \tilde{t}_0)$),

$$\begin{pmatrix} \partial x_1 / \partial x_1^0 & \partial x_1 / \partial x_2^0 \\ \partial x_2 / \partial x_1^0 & \partial x_2 / \partial x_2^0 \end{pmatrix} = \begin{pmatrix} \frac{2}{3}e^{t_0}e^{-t} + \frac{1}{3}e^{-2t_0}e^{2t} & -\frac{2}{3}e^{t_0}e^{-t} + \frac{2}{3}e^{-2t_0}e^{2t} \\ -\frac{1}{3}e^{t_0}e^{-t} + \frac{1}{3}e^{-2t_0}e^{2t} & \frac{1}{3}e^{t_0}e^{-t} + \frac{2}{3}e^{-2t_0}e^{2t} \end{pmatrix} \quad (ii)$$

Notice that, because $\mathbf{F}(t, \mathbf{x})$ is linear in $\mathbf{x}$, the differential equation in (9) is identical to $\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x})$. Thus we see that the right-hand side of (7) is the resolvent of $\dot{\mathbf{x}} = \mathbf{F}(t, \mathbf{x})$. According to (7.4.10), the two columns of $\mathbf{P}(t, t_0)$ are obtained from (i) by putting $x_1^0 = x_1(t_0) = 1$, $x_2^0 = x_2(t_0) = 0$, and $x_1^0 = x_1(t_0) = 0$, $x_2^0 = x_2(t_0) = 1$, respectively. Hence,

$$\mathbf{P}(t, t_0) = \begin{pmatrix} \frac{2}{3}e^{t_0}e^{-t} + \frac{1}{3}e^{-2t_0}e^{2t} & -\frac{2}{3}e^{t_0}e^{-t} + \frac{2}{3}e^{-2t_0}e^{2t} \\ -\frac{1}{3}e^{t_0}e^{-t} + \frac{1}{3}e^{-2t_0}e^{2t} & \frac{1}{3}e^{t_0}e^{-t} + \frac{2}{3}e^{-2t_0}e^{2t} \end{pmatrix} \quad (iii)$$

Because the matrices in (ii) and (iii) are identical, (7) is confirmed.

Using (i), the left-hand side of (8) evaluated at (t_0, x_1^0, x_2^0) is

$$\begin{pmatrix} \partial x_1 / \partial t_0 \\ \partial x_2 / \partial t_0 \end{pmatrix} = \begin{pmatrix} \frac{2}{3}(x_1^0 - x_2^0)e^{4t_0}e^{-t} - \frac{2}{3}(x_1^0 + 2x_2^0)e^{-2t_0}e^{2t} \\ -\frac{1}{3}(x_1^0 - x_2^0)e^{4t_0}e^{-t} - \frac{2}{3}(x_1^0 + 2x_2^0)e^{-2t_0}e^{2t} \end{pmatrix} \quad (\text{iv})$$

The right-hand side of (8) is $-\mathbf{P}(t, t_0) \cdot \mathbf{F}(t_0, \mathbf{x}^0) = -\mathbf{P}(t, t_0) \cdot \begin{pmatrix} 2x_2^0 \\ x_1^0 + x_2^0 \end{pmatrix}$. Using (iii) we see that this matrix product is equal to the column vector in (iv). Thus (8) is confirmed.

7.7 A Glimpse at Partial Differential Equations

In an ordinary differential equation, the unknown function depends on a single variable, and the equation involves the ordinary derivative of that function. In a partial differential equation, however, the unknown function depends on two or more variables, and the equation involves the partial derivatives of that function. When such an equation involves only derivatives of the first order, it is said to be of first order. For example, the general partial differential equation of first order in two variables has the form

$$F(x, y, z, \partial z / \partial x, \partial z / \partial y) = 0 \quad (1)$$

where $z = z(x, y)$ is the unknown C^1 function. Here are two simple examples.

EXAMPLE 1 Find the most general C^1 function $z = z(x, y)$ satisfying $\partial z / \partial y = 3x^2y - y^2$.

Solution: First, we keep x constant, and integrate $3x^2y - y^2$ w.r.t. y . The most general function of x and y whose derivative w.r.t. y equals $3x^2y - y^2$, is $3x^2 \frac{1}{2}y^2 - \frac{1}{3}y^3 + C$. But when x is variable, note that C could be an arbitrary function of x . Thus, the general solution is $z = \frac{3}{2}x^2y^2 - \frac{1}{3}y^3 + \varphi(x)$, where $\varphi(x)$ is any C^1 function of x .

EXAMPLE 2 Find the solutions of $z'_x + az'_y = 0$.

Solution: Provided that $z'_y \neq 0$, the equation can be rewritten as $z'_x / z'_y = -a$. Economists should recognize this as saying that the marginal rate of substitution between x and y is equal to the constant $-a$. This suggests that the level curves of $z(x, y)$ should be straight lines of slope a , and so that the general solution is $z = g(ax - y)$ for an arbitrary differentiable function g . Indeed, using the chain rule, $z'_x = ag'(ax - y)$ and $z'_y = -g'(ax - y)$, so $z'_x + az'_y = 0$, so any such function does satisfy the equation. The technique set out below shows that this is the general solution.

Quasi-linear Equations of First Order

Consider the general quasi-linear partial differential equation of first order,

$$P(x, y, z) \frac{\partial z}{\partial x} + Q(x, y, z) \frac{\partial z}{\partial y} = R(x, y, z) \quad (2)$$

where $P = P(x, y, z)$, $Q = Q(x, y, z)$, and $R = R(x, y, z)$ are all defined in an open set Ω in 3-space, and we assume that $P \neq 0$ in Ω .

The problem is to find all functions $z = z(x, y)$ that satisfy (2). The graph of such a function is called a solution surface or an **integral surface** for (2).

In order to find the general solution of (2), Lagrange proposed the following recipe:

RECIPE FOR SOLVING (2)

(A) Solve the following pair of ordinary differential equations, called the *characteristic equations*:

$$\frac{dy}{dx} = \frac{Q}{P}, \quad \frac{dz}{dx} = \frac{R}{P} \quad (3)$$

where x is the independent variable. The solutions of this simultaneous system can be written in the form $y = \varphi_1(x, C_1, C_2)$, $z = \varphi_2(x, C_1, C_2)$, where C_1 and C_2 are constants of integration. Solving these equations for C_1 and C_2 yields

$$u(x, y, z) = C_1, \quad v(x, y, z) = C_2 \quad (4)$$

(B) Then the general solution $z = \varphi(x, y)$ of (2) is given implicitly by the equation

$$\Phi(u(x, y, z), v(x, y, z)) = 0 \quad (5)$$

where Φ is any differentiable function of two variables, provided that z occurs in equation (5) (otherwise the equation cannot define z as a function of x and y).

If $P = 0$ somewhere, one can use $dx/dy = P/Q$ and $dz/dy = R/Q$ as characteristic equations instead. The recipe usually gives all solutions of (2), but problems arise when $P(x, y, z) = 0$ and $Q(x, y, z) = 0$ simultaneously.

EXAMPLE 3

(a) Find the general solution of $\frac{\partial z}{\partial x} + \frac{y}{x} \frac{\partial z}{\partial y} = \frac{y}{z}$ ($x \neq 0, z \neq 0$).

(b) Find the only solution satisfying the boundary condition $z(3, y) = y$.

Solution: (a) The equations in (A) are, with $P = 1$, $Q = y/x$, and $R = y/z$,

$$\frac{dy}{dx} = \frac{y}{x}, \quad \frac{dz}{dx} = \frac{y}{z}$$

The first equation is separable, with solution $y = C_1x$. Inserting this expression into the second equation gives $dz/dx = C_1x/z$, so $z dz = C_1x dx$, and hence $\frac{1}{2}z^2 = \frac{1}{2}C_1x^2 + \frac{1}{2}C_2$.

Solving $y = C_1x$ and $\frac{1}{2}z^2 = \frac{1}{2}C_1x^2 + \frac{1}{2}C_2$ for C_1 and C_2 yields

$$C_1 = y/x, \quad C_2 = z^2 - xy$$

The general solution $z(x, y)$ of the given equation is then defined implicitly by the equation

$$\Phi(y/x, z^2 - xy) = 0$$

where Φ is an arbitrary differentiable function.

If $\Phi'_2 \neq 0$, we can express $z^2 - xy$ as a C^1 function φ of y/x , so that $z^2 - xy = \varphi(y/x)$, or

$$z^2 = xy + \varphi(y/x) \quad (*)$$

Any function $z = z(x, y)$ that satisfies the latter equation for some C^1 function φ is a solution of the given equation.

(b) The boundary condition requires, geometrically, that the integral surface intersects the plane $x = 3$ along the curve $z = y$. When $z(3, y) = y$, then $z(3, y)^2 = y^2$ and $(*)$ for $x = 3$ yields $y^2 = 3y + \varphi(y/3)$, or $\varphi(y/3) = y^2 - 3y$, i.e. $\varphi(u) = (3u)^2 - 3(3u) = 9u^2 - 9u$, where $u = y/3$. Hence, a solution $z(x, y)$ that satisfies the boundary condition must satisfy the equation $z(x, y)^2 = xy + 9(y/x)^2 - 9y/x$.

E 4 Suppose $z = z(x, y)$ has constant elasticity a w.r.t. x , i.e.

$$\frac{x}{z} \frac{\partial z}{\partial x} = a, \quad (x \neq 0, z \neq 0) \quad (i)$$

What does the method above tell us about $z(x, y)$?

Solution: If we compare (i) with (2), we see that $P = x/z$, $Q = 0$, $R = a$. So the equations in (A) take the form

$$\frac{dy}{dx} = 0, \quad \frac{dz}{dx} = a \frac{z}{x}$$

Hence $y = C_1$, $z = C_2x^a$, or $C_1 = y$, $C_2 = zx^{-a}$. The solution of (i) is therefore $\Phi(y, zx^{-a}) = 0$. If $\Phi'_2 \neq 0$, we get $zx^{-a} = \varphi(y)$, i.e.

$$z = \varphi(y)x^a \quad (ii)$$

where φ is an arbitrary differentiable function. (Because the elasticity of x^a w.r.t. x is a and $\varphi(y)$ is independent of x , it is clear that when z is given by (ii), it has elasticity a w.r.t. x .)

These examples indicate that the set of solutions of a given partial differential equation is often enormous. For ordinary differential equations the general solution depends on one or more arbitrary constants, according to whether the equation is of first or higher order. However, for any such equation, the different solutions are all functions of the same type. For partial differential equations like (1), typically the general solution depends on an arbitrary

Why the Recipe Works

Let us give a brief explanation of why the recipe works. The functions P , Q , and R in (2) can be regarded as the components of a vector in $\mathbb{R}^3$. As (x, y, z) varies, the function

$$(x, y, z) \mapsto (P, Q, R) = (P(x, y, z), Q(x, y, z), R(x, y, z)) \quad (*)$$

is called the **vector field of the differential equation** (2). For simplicity, we shall assume that P , Q , and R are C^1 functions defined throughout $\mathbb{R}^3$, and that $(P, Q, R) \neq (0, 0, 0)$ everywhere. The (direction of the) vector (P, Q, R) at a point in $\mathbb{R}^3$ is called the **characteristic direction** at that point.

If $z = \varphi(x, y)$ is the equation of a surface in $\mathbb{R}^3$, then the vector $(\partial z / \partial x, \partial z / \partial y, -1)$ is orthogonal to its tangent plane (see Note 2.1.1). The scalar product of this vector and (P, Q, R) is

$$\left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}, -1 \right) \cdot (P, Q, R) = P \frac{\partial z}{\partial x} + Q \frac{\partial z}{\partial y} - R \quad (**)$$

Hence, $z = \varphi(x, y)$ is a solution of equation (2) if and only if the characteristic direction (P, Q, R) is tangent to the graph of φ at every point.

If $t \mapsto (x(t), y(t), z(t))$ is a parametric representation of a curve γ in $\mathbb{R}^3$, then the tangent vector to γ is $(\dot{x}(t), \dot{y}(t), \dot{z}(t))$. (See Section 7.4.) The curve γ is called a **characteristic** for equation (2) if the tangent to γ at every point has the same direction as the vector (P, Q, R) , i.e. if there is a nonzero function α such that

$$(\dot{x}(t), \dot{y}(t), \dot{z}(t)) = \alpha(t)(P(x(t), y(t), z(t)), Q(x(t), y(t), z(t)), R(x(t), y(t), z(t)))$$

for all t . After a suitable change in the parameter t , we can assume $\alpha(t) \equiv 1$, so that $(\dot{x}, \dot{y}, \dot{z}) = (P, Q, R)$.

It follows from the existence and uniqueness theorem (Theorem 7.6.1) that every point of $\mathbb{R}^3$ has exactly one characteristic for (2) passing through it. Furthermore, if $z = \varphi(x, y)$ is a solution of (2) and $\mathbf{x}^0 = (x_0, y_0, z_0)$ is a point on the graph of φ , then the entire characteristic through $\mathbf{x}^0$ lies in the graph of φ . To see why, let $\gamma : t \mapsto (x(t), y(t), z(t))$ be a parametric representation of the characteristic through $\mathbf{x}^0$. We can assume that $\gamma(0) = \mathbf{x}^0$. Let $V(t) = z(t) - \varphi(x(t), y(t))$. Then $V(0) = z_0 - \varphi(x_0, y_0) = 0$, and the derivative of V is

$$\begin{aligned} \dot{V}(t) &= \dot{z}(t) - \varphi'_1(x(t), y(t))\dot{x}(t) - \varphi'_2(x(t), y(t))\dot{y}(t) \\ &= R(x, y, z) - \varphi'_1(x, y)P(x, y, z) - \varphi'_2(x, y)Q(x, y, z) = 0 \end{aligned}$$

where we have suppressed the parameter t in the second line. But then V must be a constant function with $V(t) = V(0) = 0$ for all t . So $\gamma(t)$ belongs to the graph of φ for all t .

Hence, the graph of a solution of (2) is a union of characteristics for the equation. For example, if $\varphi(x, y)$ solves (2) and we find all characteristics through the points $(x_0, y, \varphi(x_0, y))$ as y varies, then the graph of φ consists of the union of these as y varies over $\mathbb{R}$. On the other hand, if $\psi(x, y)$ is a function whose graph is a union of characteristics for (2), then at every point on the graph, the characteristic direction is tangent to the graph, and therefore ψ is a solution of (2).

To find the characteristics of (2), we need to solve the differential equations

$$\dot{x} = P(x, y, z), \quad \dot{y} = Q(x, y, z), \quad \dot{z} = R(x, y, z) \quad (6)$$

If $P \neq 0$, we can eliminate t and use x as a free variable instead. Then $dy/dx = \dot{y}/\dot{x} = Q/P$ and $dz/dx = \dot{z}/\dot{x} = R/P$, precisely the equations (3) in the recipe. For every pair of constants C_1 and C_2 , each of the two equations in (4) is the equation of a surface, and the intersection of these surfaces is a characteristic of equation (2). (Solving (4) for y and z gives back $y = \varphi_1$, $z = \varphi_2$.) At each point $\mathbf{x}^0 = (x_0, y_0, z_0)$ of this characteristic the vector $(P(\mathbf{x}^0), Q(\mathbf{x}^0), R(\mathbf{x}^0))$ is tangent to the curve, and therefore tangent to each of the surfaces $u(x, y, z) = C_1$ and $v(x, y, z) = C_2$. This implies that

$$(P(\mathbf{x}^0), Q(\mathbf{x}^0), R(\mathbf{x}^0)) \cdot \nabla u(\mathbf{x}^0) = (P(\mathbf{x}^0), Q(\mathbf{x}^0), R(\mathbf{x}^0)) \cdot \nabla v(\mathbf{x}^0) = 0 \quad (***)$$

If we keep C_2 fixed, say, and vary the constant C_1 , we would expect the resulting family of characteristic curves to make up a surface, which would then be a solution surface for (2). More generally, we could vary C_1 and C_2 simultaneously in a proper manner. One way to do so is to demand that they satisfy an equation $\Phi(C_1, C_2) = 0$, where Φ is a C^1 function with a nonzero gradient. This leads to the equation (5).

Suppose $\Phi(u, v)$ is such a function, and let $F(x, y, z) = \Phi(u(x, y, z), v(x, y, z))$. Then (5) is equivalent to the equation $F(x, y, z) = 0$. Straightforward calculations give

$$\nabla F(x, y, z) = \Phi'_1(u, v) \nabla u(x, y, z) + \Phi'_2(u, v) \nabla v(x, y, z)$$

Hence, using (***) gives

$$(P, Q, R) \cdot \nabla F(x, y, z) = 0$$

which means that the surface $F(x, y, z) = 0$ is the graph of a solution of (2), provided $F'_3(x, y, z) \neq 0$. This shows that the recipe does indeed lead to solutions of the partial differential equation (2).

It can be shown that the recipe yields all solutions of (2), provided that P and Q are never simultaneously equal to 0.

A More General Case

Consider the more general problem of finding all functions $z = z(x_1, \dots, x_n)$ of n variables satisfying the general quasi-linear partial differential equation

$$P_1 \frac{\partial z}{\partial x_1} + P_2 \frac{\partial z}{\partial x_2} + \dots + P_n \frac{\partial z}{\partial x_n} = Q \quad (7)$$

Here $P_1, \dots, P_n$, and Q are functions of $x_1, \dots, x_n$ and z .

It turns out that the method used above for solving (2) can be generalized. We solve (7) in the following way: Assume that $P_1 \neq 0$ and find the general solution of the system

$$\frac{dx_2}{dx_1} = \frac{P_2}{P_1}, \quad \dots, \quad \frac{dx_n}{dx_1} = \frac{P_n}{P_1}, \quad \frac{dz}{dx_1} = \frac{Q}{P_1} \quad (8)$$

in the form

$$x_2 = \psi_2(x_1; C_1, \dots, C_n)$$

$$\dots\dots\dots$$

$$x_n = \psi_n(x_1; C_1, \dots, C_n)$$

$$z = \psi_{n+1}(x_1; C_1, \dots, C_n)$$

Solving (if possible) for $C_1, \dots, C_n$, we obtain

$$u_1(x_1, x_2, \dots, x_n, z) = C_1, \dots, u_n(x_1, x_2, \dots, x_n, z) = C_n$$

If Φ is an arbitrary differentiable function of n variables, and at least one of the functions $u_1, \dots, u_n$ involves z , then the general solution of (7) is given implicitly by

$$\Phi(u_1(x_1, \dots, x_n, z), \dots, u_n(x_1, \dots, x_n, z)) = 0 \quad (9)$$

i.e., a solution $z = z(x_1, \dots, x_n)$ of this equation is a solution of (7).

5. Solve the equation

$$x_1 \frac{\partial z}{\partial x_1} + x_2 \frac{\partial z}{\partial x_2} + x_3 \frac{\partial z}{\partial x_3} = x_1 x_2 x_3 \quad (x_1 \neq 0) \quad (i)$$

Solution: Here

$$\frac{dx_2}{dx_1} = \frac{x_2}{x_1}, \quad \frac{dx_3}{dx_1} = \frac{x_3}{x_1}, \quad \frac{dz}{dx_1} = \frac{x_1 x_2 x_3}{x_1} = x_2 x_3$$

The first two equations give us

$$(ii) \quad x_2 = C_1 x_1 \quad \text{and} \quad (iii) \quad x_3 = C_2 x_1$$

Inserting these into the third equation in (i) yields $dz = C_1 C_2 x_1^2 dx_1$. Hence, $z + C_3 = \frac{1}{3} C_1 C_2 x_1^3$, where C_3 is an arbitrary constant. Putting $C_4 = 3C_3$, we obtain

$$3z + C_4 = C_1 C_2 x_1^3 \quad (iv)$$

Solving (ii), (iii), and (iv) for C_1, C_2 , and C_4 , we obtain

$$C_1 = x_2/x_1, \quad C_2 = x_3/x_1, \quad C_4 = x_1 x_2 x_3 - 3z \quad (v)$$

The general solution of the given equation is therefore $\Phi(x_2/x_1, x_3/x_1, x_1 x_2 x_3 - 3z) = 0$, or, if $\Phi'_3 \neq 0$,

$$z = \frac{1}{3} x_1 x_2 x_3 - \varphi\left(\frac{x_2}{x_1}, \frac{x_3}{x_1}\right)$$

where φ is an arbitrary C^1 function of two variables. I

For further discussions about these methods, see e.g. Zauderer (1989).

PROBLEMS FOR SECTION 7.7

SM 1. Find the general solutions of

$$(a) \quad \frac{\partial z}{\partial x} = x^2 + xy^2 - e^x y \quad (b) \quad \frac{\partial z}{\partial x} + 2 \frac{\partial z}{\partial y} = 3 \quad (c) \quad x^2 \frac{\partial z}{\partial x} + y^2 \frac{\partial z}{\partial y} = z^2$$

2. Find the general solution of the partial differential equation

$$y^2 \frac{\partial z}{\partial x} + y \frac{\partial z}{\partial y} = \frac{1}{2} z, \quad y > 0, \quad z > 0$$

GM 3. (a) Find the general solution of the partial differential equation

$$x \frac{\partial z}{\partial x} - y \frac{\partial z}{\partial y} = x, \quad x > 0, \quad y > 0, \quad z > 0 \quad (*)$$

(b) Find a solution $z = f(x, y)$ of (*) such that $f(x, 1) = x^2$ for all x .

4. Find all functions $z = z(x, y)$ that satisfy $\text{El}_x z - \text{El}_y z = x$. (Here $\text{El}_x z$ denotes the partial derivative of z with respect to x , and likewise for $\text{El}_y z$.)

5. In utility theory we encounter the following problem: Find all functions $U = U(x_1, x_2)$ with the property that the ratio between the marginal utilities w.r.t. x_1 and x_2 depends on (say) x_1 only. Thus we must solve the equation

$$\frac{\partial U}{\partial x_1} = f(x_1) \frac{\partial U}{\partial x_2} = 0$$

where f is a given function. Solve this problem.

6. Euler's theorem on homogeneous functions states that $z = z(x, y)$ satisfies the equation

$$x \frac{\partial z}{\partial x} + y \frac{\partial z}{\partial y} = nz$$

if and only if $z(x, y)$ is homogeneous of degree n , i.e. $z(\lambda x, \lambda y) = \lambda^n z(x, y)$ for all scalars $\lambda > 0$. (See e.g. EMEA.) Make use of the method described above to confirm the "only if" part of this result.

- SM 7. In a problem in the theory of production, McElroy studies the equation

$$v_1 \frac{\partial x}{\partial v_1} + v_2 \frac{\partial x}{\partial v_2} = x\varepsilon(x)$$

where $\varepsilon(x)$ is a given positive function and v_1, v_2 , and x are positive. Prove that the solution of the equation can be written in the form $x = F(g(v_1, v_2))$ where g is homogeneous of degree 1, while F is an increasing function. (Thus x is a homothetic function of v_1, v_2 , see e.g. EMEA, Section 12.7.)

8. A model by W. Leontief requires finding the most general function $z = F(x_1, x_2, x_3)$ satisfying the differential equation

$$\frac{\partial z}{\partial x_1} = f(x_1, x_2) \frac{\partial z}{\partial x_2} \quad (f \text{ a given function})$$

Prove that the general solution is $z = G(\varphi(x_1, x_2), x_3)$, where G is an arbitrary differentiable function of two variables, and $\varphi(x_1, x_2)$ is a certain differentiable function of x_1 and x_2 .

8

CALCULUS OF VARIATIONS

We are usually convinced more easily by reasons we have found ourselves than by those which have occurred to others.

—Pascal (1670)

This chapter gives a brief introduction to the classical calculus of variations. The next two chapters deal with optimal control theory, which is a modern generalization of the classical theory that offers a unified method for treating very general dynamic optimization problems in continuous time. Control theory is now used by a large number of economists, even when they face a problem that can be solved with the calculus of variations. Economics students interested in dynamic optimization problems should therefore make a serious effort to learn the basic facts of optimal control theory. It is considerably easier to understand the modern methods, however, if one knows something about the prior calculus of variations theory.

The calculus of variations actually has a rather long history. Some of its main results were established by Euler and Lagrange as early as in the 18th century. Since then the subject has formed an important part of applied mathematics. In economics, some of its first applications were by Ramsey (1928) to an optimal savings problem (see Example 8.1.1), and by Hotelling (1931) to a problem of finding the optimal extraction of a natural resource (see Examples 9.1.2 and 9.8.1).

Section 8.1 presents a version of the Ramsey model, whose purpose was to give a simplified answer to a problem of crucial importance: how much should a nation save? It also states the simplest general problem in the calculus of variations.

Section 8.2 presents the Euler equation, while Section 8.3 gives its rather easy and very instructive proof. Even easier is the proof of the sufficiency of the Euler equation when appropriate concavity conditions are imposed.

Section 8.4 uses the Euler equation to characterize the solution to the Ramsey problem.

Section 8.5 is concerned with different types of terminal conditions that are often present in economic problems. To determine the optimal solution, a transversality condition at the end of the time period is required.

The Simplest Problem

We begin by introducing a problem from optimal growth theory that is closely related to Ramsey's pioneering discussion of optimal saving. It forms the basis of much recent work in macroeconomic theory, as discussed in the textbooks by Blanchard and Fischer (1989) and Barro and Sala-i-Martin (1995).

(How much should a nation save?) Consider an economy evolving over time where $K = K(t)$ denotes the capital stock, $C = C(t)$ consumption, and $Y = Y(t)$ net national product at time t . Suppose that

$$Y = f(K), \quad \text{where } f'(K) > 0 \text{ and } f''(K) \leq 0 \quad (\text{i})$$

Thus net national product is a strictly increasing, concave function of the capital stock alone. For each t assume that

$$f(K(t)) = C(t) + \dot{K}(t) \quad (\text{ii})$$

which means that output, $Y(t) = f(K(t))$, is divided between consumption, $C(t)$, and investment, $\dot{K}(t)$. Moreover, let $K(0) = K_0$ be the historically given capital stock existing "today" at $t = 0$, and suppose that there is a fixed planning period $[0, T]$. Now, for each choice of investment function $\dot{K}(t)$ on the interval $[0, T]$, capital is fully determined by $K(t) = K_0 + \int_0^t \dot{K}(\tau) d\tau$, and (ii) in turn determines $C(t)$. The question Ramsey and his successors have addressed is how much investment would be desirable. Higher consumption today is in itself preferable, but equation (ii) tells us that it leads to a lower rate of investment. This in turn results in a lower capital stock in the future, thus reducing the possibilities for future consumption. One must somehow find a way to reconcile the conflict between higher consumption now and more investment in the future.

To this end, assume that the society has a utility function U , where $U(C)$ is the utility (flow) the country enjoys when the total consumption is C . Suppose too that

$$U'(C) > 0, \quad U''(C) < 0$$

so that U is strictly increasing and strictly concave. (This assumption implies that when society has a high level of consumption, it enjoys a lower increase in satisfaction from a given increase in consumption than when there is a low level of consumption.) Now, as is common in this literature, introduce a discount rate r to reflect the idea that the present may matter more than the future. That is, for each $t \geq 0$ we multiply $U(C(t))$ by the discount factor e^{-rt} . However, Ramsey himself criticized such "impatience", so he put $r = 0$, in effect. Anyway, assume that the goal of investment policy is to choose $K(t)$ for t in $[0, T]$ in order to make the total discounted utility over the period $[0, T]$ as large as possible. Another way of formulating the problem is: Find the path of capital $K = K(t)$, with $K(0) = K_0$, that maximizes

$$\int_0^T U(C(t))e^{-rt} dt = \int_0^T U(f(K(t)) - \dot{K}(t))e^{-rt} dt \quad (\text{1})$$

Usually, some "terminal condition" on $K(T)$ is imposed—for example, that $K(T) = K_T$, where K_T is given. One possibility is $K_T = 0$, with no capital left for times after T . This problem is studied in Example 8.4.1.

Example 1 is a special case of the simplest general problem in the calculus of variations. This takes the form

$$\max \int_{t_0}^{t_1} F(t, x, \dot{x}) dt \quad \text{subject to} \quad x(t_0) = x_0, \quad x(t_1) = x_1 \quad (\text{2})$$

Here F is a given "well-behaved" function of three variables, whereas t_0 , t_1 , x_0 , and x_1 are given numbers. More precisely: Among all well-behaved functions $x(t)$ that satisfy $x(t_0) = x_0$ and $x(t_1) = x_1$, find one making the integral in (2) as large as possible.

Geometrically, the problem is illustrated in Fig. 1. The two end points $A = (t_0, x_0)$ and $B = (t_1, x_1)$ are given in the tx -plane. For each smooth curve that joins the points A and B , the integral in (2) has a definite value. Find the curve that makes the integral as large as possible.

So far the integral in (2) has been maximized. Because minimizing the integral of $F(t, x, \dot{x})$ leads to the same function $x = x(t)$ as maximizing the integral of $-F(t, x, \dot{x})$, there is an obvious relationship between the maximization and minimization problems.

The first known application of the calculus of variations was to the "brachistochrone problem".¹ Given two points A and B in a vertical plane, the time required for a particle to slide along a curve from A to B under the sole influence of gravity will depend on the shape of the curve. The problem is to find the curve along which the particle goes from A to B as quickly as possible. (See Fig. 2.)

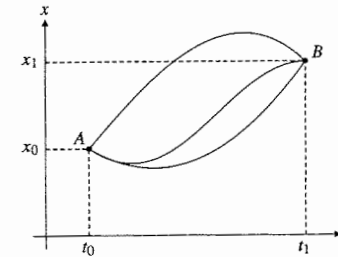


Figure 1

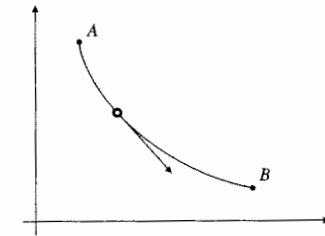


Figure 2

One's first reaction to the problem might be that it is easy, because the straight line joining A to B must be the solution. This is not correct. (Actually, the straight line between A and B solves another variational problem: Find the shortest curve joining A and B . See Problem 8.2.6.) In 1696, the Swiss mathematician John Bernoulli proved that the solution is a part of a curve called a *cycloid*. This starts out steeply so that the particle can accelerate rapidly, acquiring momentum in both the vertical and horizontal directions, before the curve flattens out as it approaches B . Using elementary physics one can show that the brachistochrone problem reduces to the problem of minimizing an integral of the type appearing in (2).

¹ The word "brachistochrone" is derived from two Greek roots meaning "shortest" and "time".

FOR SECTION 8.1

1. The graphs of the following two functions

$$(i) x = (e^2 - 1)t \quad (ii) x = x(t) = e^{1+t} - e^{1-t}$$

both pass through the points $(0, 0)$ and $(1, e^2 - 1)$ in the tx -plane. Compute the value of $J(x) = \int_0^1 (x^2 + \dot{x}^2) dt$ in each case. Which function gives $J(x)$ the lower value?

The Euler Equation

The simplest variational problem is this:²

$$\max \int_{t_0}^{t_1} F(t, x, \dot{x}) dt \quad \text{subject to} \quad x(t_0) = x_0, \quad x(t_1) = x_1 \quad (1)$$

Already in 1744 the Swiss mathematician L. Euler proved that a function $x(t)$ can only solve problem (1) if $x(t)$ satisfies the differential equation

$$\frac{\partial F}{\partial x} - \frac{d}{dt} \left(\frac{\partial F}{\partial \dot{x}} \right) = 0 \quad (2)$$

called the **Euler equation**. Here $\partial F / \partial x$ denotes the derivative $F'_2(t, x, \dot{x})$ of $F(t, x, \dot{x})$ w.r.t. the second variable, whereas $\partial F / \partial \dot{x}$ denotes $F'_3(t, x, \dot{x})$.

Replacing F with $-F$ does not change (2), so the Euler equation also represents a necessary condition for solving the corresponding minimization problem.

Note that in equation (2), the term $(d/dt)(\partial F / \partial \dot{x})$ denotes the *total derivative* of $\partial F / \partial \dot{x}$ w.r.t. t , allowing for the dependence of $\partial F / \partial \dot{x}$ on all three variables which depend on t . Assuming that $x = x(t)$ is C^2 , one finds that³

$$\frac{d}{dt} \left(\frac{\partial F(t, x, \dot{x})}{\partial \dot{x}} \right) = \frac{\partial^2 F}{\partial t \partial \dot{x}} + \frac{\partial^2 F}{\partial x \partial \dot{x}} \cdot \dot{x} + \frac{\partial^2 F}{\partial \dot{x} \partial \dot{x}} \cdot \ddot{x} \quad (*)$$

Inserting this into (2) and rearranging, the Euler equation becomes

$$\frac{\partial^2 F}{\partial \dot{x} \partial \dot{x}} \cdot \ddot{x} + \frac{\partial^2 F}{\partial x \partial \dot{x}} \cdot \dot{x} + \frac{\partial^2 F}{\partial t \partial \dot{x}} - \frac{\partial F}{\partial x} = 0 \quad (3)$$

Alternatively, equation (3) can be written as $F''_{33}\ddot{x} + F''_{32}\dot{x} + F''_{31} - F'_2 = 0$, so we see that the Euler equation is a differential equation of the second order (if $F''_{33} \neq 0$).

The Euler equation gives a *necessary condition* for optimality. It is not, in general, sufficient. By analogy with static optimization problems, it is natural to expect that appropriate concavity (convexity) requirements on $F(t, x, \dot{x})$ will ensure optimality. In the next section we shall prove that for the maximization problem (1), concavity of $F(t, x, \dot{x})$ w.r.t. $(x, \dot{x})$ is sufficient, while the convexity of $F(t, x, \dot{x})$ w.r.t. $(x, \dot{x})$ is sufficient for the corresponding minimization problem.

² In the next section we specify the regularity conditions to be imposed on F and $x(t)$.

³ According to the chain rule we have $\frac{d}{dt} G(u(t), v(t), w(t)) = G'_1 \dot{u} + G'_2 \dot{v} + G'_3 \dot{w}$. Equation (*) follows by letting $G = \partial F / \partial \dot{x}$ and $u = t, v = x, w = \dot{x}$.

EXAMPLE 1

Solve the problem: $\max \int_0^2 (4 - 3x^2 - 16\dot{x} - 4\dot{x}^2)e^{-t} dt, \quad x(0) = -\frac{8}{3}, \quad x(2) = \frac{1}{3}.$

Solution: Here $F(t, x, \dot{x}) = (4 - 3x^2 - 16\dot{x} - 4\dot{x}^2)e^{-t}$, so $\partial F / \partial x = -6xe^{-t}$ and $\partial F / \partial \dot{x} = (-16 - 8\dot{x})e^{-t}$. Using Equation (2) requires finding $(d/dt)[(-16 - 8\dot{x})e^{-t}]$. The product rule for differentiation gives

$$\frac{d}{dt}[(-16 - 8\dot{x})e^{-t}] = 16e^{-t} - 8\ddot{x}e^{-t} + 8\dot{x}e^{-t}$$

so the Euler equation reduces to $-6xe^{-t} - 16e^{-t} + 8\ddot{x}e^{-t} - 8\dot{x}e^{-t} = 0$. Cancelling the nonzero common factor $8e^{-t}$ yields

$$\ddot{x} - \dot{x} - \frac{3}{4}x = 2 \quad (*)$$

This is a linear differential equation of the second order with constant coefficients (see Section 6.3). The characteristic equation is $r^2 - r - \frac{3}{4} = 0$, with roots $r_1 = -1/2$ and $r_2 = 3/2$. The nonhomogeneous equation (*) has a particular solution $-8/3$, so the general solution is $x = x(t) = Ae^{-\frac{1}{2}t} + Be^{\frac{3}{2}t} - 8/3$ where A and B are arbitrary constants. The boundary conditions $x(0) = -8/3$ and $x(2) = 1/3$ imply that $0 = A + B$ and $Ae^{-1} + Be^3 = 3$. It follows that $A = -3/(e^3 - e^{-1})$ and $B = -A$, so

$$x = x(t) = -\frac{3}{e^3 - e^{-1}}e^{-\frac{1}{2}t} + \frac{3}{e^3 - e^{-1}}e^{\frac{3}{2}t} - \frac{8}{3}$$

This is the only solution of the Euler equation that satisfies the given boundary conditions. The function $F(t, x, \dot{x}) = (4 - 3x^2 - 16\dot{x} - 4\dot{x}^2)e^{-t}$ is concave in $(x, \dot{x})$, as a sum of concave functions. We conclude that we have found the solution of the problem. ■

EXAMPLE 2

Consider the simple macroeconomic problem of trying to steer the state $y(t)$ of the economy over the course of a planning period $[0, T]$ toward the desired level $\hat{y}$, independent of t , by means of the control $u(t)$, where $\dot{y}(t) = u(t)$. Because using the control is costly, the objective is to minimize the integral $\int_0^T [(y(t) - \hat{y})^2 + c(u(t))^2] dt$ with $y(T) = \hat{y}$, where c is a positive constant.

It is more convenient to define $x(t)$ as the difference $y(t) - \hat{y}$ between the original state variable and the target level $\hat{y}$, so that the target value of x is 0. Then $u(t) = \dot{x}(t)$. This leads to the following variational problem:

$$\min \int_0^T (x^2 + c\dot{x}^2) dt, \quad x(0) = x_0, \quad x(T) = 0$$

where x_0 is the initial deviation from the target level. Find the optimal solution $x^*(t)$.

Solution: In this case the integrand is $F(t, x, \dot{x}) = x^2 + c\dot{x}^2$, so $\partial F / \partial x = 2x$ and $\partial F / \partial \dot{x} = 2c\dot{x}$. The Euler equation is $2x - (d/dt)(2c\dot{x}) = 0$, or

$$\ddot{x} - (1/c)x = 0$$

The general solution is

$$x = Ae^{rt} + Be^{-rt}, \quad \text{where } r = 1/\sqrt{c}$$

The initial condition $x(0) = x_0$ and the terminal condition $x(T) = 0$ yield the two equations $A + B = x_0$ and $Ae^{rT} + Be^{-rT} = 0$. These two equations in A and B have the solution $A = -x_0 e^{-rT} / (e^{rT} - e^{-rT})$ and $B = x_0 e^{rT} / (e^{rT} - e^{-rT})$. It follows that the only possible solution is

$$x^*(t) = \frac{x_0}{e^{rT} - e^{-rT}} [e^{r(T-t)} - e^{-r(T-t)}]$$

The function $F = x^2 + cx^2$ is convex in $(x, \dot{x})$ (as a sum of convex functions), so the solution to the problem has been found. $\square$

Important Special Cases

If the integrand $F(t, x, \dot{x})$ in (1) does not depend on x , then the Euler equation reduces to $(d/dt)F_{\dot{x}}(t, x, \dot{x}) = 0$, so $F_{\dot{x}}(t, x, \dot{x}) = C$ for some constant C . This is a first-order differential equation.

In many variational problems in economics the integrand $F(t, x, \dot{x})$ in (1) is of the form $F(x, \dot{x})$, so that t does not enter explicitly. It is then possible to reduce the Euler equation to a first-order equation. The trick is to calculate the total derivative of the expression $F(x, \dot{x}) - \dot{x} \partial F(x, \dot{x}) / \partial \dot{x}$:

$$\frac{d}{dt} \left[F(x, \dot{x}) - \dot{x} \frac{\partial F(x, \dot{x})}{\partial \dot{x}} \right] = \dot{x} \frac{\partial F}{\partial x} + \ddot{x} \frac{\partial F}{\partial \dot{x}} - \ddot{x} \frac{\partial F}{\partial \dot{x}} - \dot{x} \frac{d}{dt} \frac{\partial F}{\partial \dot{x}} = \dot{x} \left[\frac{\partial F}{\partial x} - \frac{d}{dt} \left(\frac{\partial F}{\partial \dot{x}} \right) \right] \quad (*)$$

It follows that if the Euler equation is satisfied for all t in $[t_0, t_1]$, then the expression in $(*)$ is 0. That is, the derivative of $F - \dot{x} \partial F / \partial \dot{x}$ must be 0 for all t . In this case the Euler equation implies that

$$F - \dot{x} \frac{\partial F}{\partial \dot{x}} = C \quad (C \text{ constant}) \quad (4)$$

for some constant C . This is a first-order differential equation which, in general, is easier to handle than the (second-order) Euler equation. Because of the possibility that $\dot{x} = 0$, the Euler equation is not quite equivalent to (4). Every solution of the Euler equation is clearly a solution to (4) for some constant C . On the other hand, for each value of C , any solution of (4) that is not constant on any interval is a solution of the Euler equation. But if $\dot{x} = 0$ on some interval, then x may not solve the Euler equation.

It follows from (3) that the Euler equation associated with $F(x, \dot{x})$ has a constant solution $x = k$ on some interval if and only if $F_x(k, 0) = 0$. Hence, if $F_x(x, 0) \neq 0$ for all x , then the Euler equation is equivalent to equation (4). Equation (4) is then called a **first integral** of the Euler equation.

EXAMPLE 3 Consider Example 8.1.1 with no discounting, so that $r = 0$. The objective (1) then becomes $\int_0^T U(C(t)) dt = \int_0^T U(f(K(t)) - \dot{K}(t)) dt$. In this case the integrand is $F = U(C) = U(f(K) - \dot{K})$, so $F_{\dot{K}} = -U'(C)$. Hence equation (4) reduces to

$$U(C) + \dot{K} U'(C) = c \quad (c \text{ is a constant}) \quad (i)$$

One usually assumes that $f' > 0$ and $U' > 0$, so $F_K = U'(C)f'(K) > 0$, and (i) is equivalent to the Euler equation. $\square$

PROBLEMS FOR SECTION 8.2

1. Solve the problem $\max \int_0^1 (4xt - \dot{x}^2) dt$, $x(0) = 2$, $x(1) = 2/3$.

2. Solve the problem $\min \int_0^1 (t\dot{x} + x^2) dt$, $x(0) = 1$, $x(1) = 0$.

3. Find the Euler equation associated with $J(x) = \int_0^1 F(t, x, \dot{x}) dt$ when

(a) $F(t, x, \dot{x}) = x^2 + \dot{x}^2 + 2xe^t$ (b) $F(t, x, \dot{x}) = -e^{t-ax}$
(c) $F(t, x, \dot{x}) = [(x - \dot{x})^2 + x^2]e^{-at}$ (d) $F(t, x, \dot{x}) = 2tx + 3x\dot{x} + t\dot{x}^2$

4. Solve the problem $\min \int_0^1 (x^2 + 2tx\dot{x} + \dot{x}^2) dt$, $x(0) = 1$, $x(1) = 2$.

5. Solve the problem $\min \int_0^1 (x^2 + tx + t\dot{x} + \dot{x}^2) dt$, $x(0) = 0$, $x(1) = 1$.

6. The part of the graph of the function $x = x(t)$ that joins the points (t_0, x_0) and (t_1, x_1) has length given by $L = \int_{t_0}^{t_1} \sqrt{1 + \dot{x}^2} dt$. Find the $x(t)$ that minimizes L . Comment on the answer.

7. Solve the problem $\min \int_1^2 (x^2 + tx\dot{x} + t^2\dot{x}^2) dt$, $x(1) = 0$, $x(2) = 1$.

8. H.Y. Wan considers the problem of finding a function $x = x(t)$ that maximizes

$$\int_0^T [N(\dot{x}(t)) + \dot{x}(t)f(x(t))]e^{-rt} dt$$

where N and f are given C^1 functions, r and T are positive constants, $x(0) = x_0$, and $x(T) = x_T$. Deduce the Euler equation, $\frac{d}{dt}N'(\dot{x}) = r[N'(\dot{x}) + f(x)]$.

8.3 Why the Euler Equation is Necessary

In the previous section we showed how to use the Euler equation to find solution candidates to variational problems. In this section we formulate and prove a precise result in which the regularity conditions on F and x are specified.

The Euler equation plays a similar role in the calculus of variations as the familiar first-order conditions in static optimization. The main result is summed up in the following theorem. The ensuing proof is very instructive and should be studied carefully by students who want an insight into dynamic optimization.⁴

⁴ We assume in the proof that the admissible functions are C^2 . This ensures that the time derivative $(d/dt)F_{\dot{x}}(t, x, \dot{x})$ is well defined. A more elaborate argument (see Gelfand and Fomin (1963)), allows a related result to be proved assuming only that the admissible functions are C^1 .

THEOREM 8.3.1 NECESSARY CONDITIONS/SUFFICIENT CONDITIONS

Suppose that F is a C^2 function of three variables. Suppose that $x^*(t)$ maximizes or minimizes

$$J(x) = \int_{t_0}^{t_1} F(t, x, \dot{x}) dt \quad (1)$$

among all admissible functions $x(t)$, i.e. all C^1 functions $x(t)$ defined on $[t_0, t_1]$ that satisfy the boundary conditions

$$x(t_0) = x_0, \quad x(t_1) = x_1. \quad (x_0 \text{ and } x_1 \text{ given numbers})$$

Then $x^*(t)$ is a solution of the Euler equation

$$\frac{\partial F(t, x, \dot{x})}{\partial x} - \frac{d}{dt} \left(\frac{\partial F(t, x, \dot{x})}{\partial \dot{x}} \right) = 0 \quad (2)$$

If $F(t, x, \dot{x})$ is concave (convex) in $(x, \dot{x})$, an admissible $x^*(t)$ that satisfies the Euler equation solves the maximization (minimization) problem.

Proof: To prove that the Euler equation is *necessary*, we need only compare the optimal solution with members of a special class of functions. Suppose $x^* = x^*(t)$ is an optimal solution to the maximization problem, and let $\mu(t)$ be any C^2 function that satisfies $\mu(t_0) = \mu(t_1) = 0$. For each real number α , define a perturbed function $x(t)$ by $x(t) = x^*(t) + \alpha\mu(t)$. (See Fig. 1.)

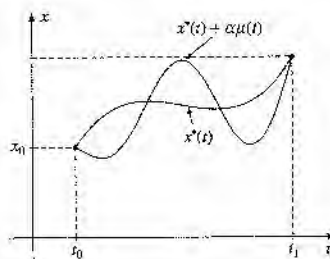


Figure 1

Note that if α is small, the function $x(t)$ is "near" the function $x^*(t)$. Clearly, $x(t)$ is admissible, because it is C^2 with $x(t_0) = x^*(t_0) + \alpha\mu(t_0) = x_0 + \alpha \cdot 0 = x_0$ and $x(t_1) = x^*(t_1) + \alpha\mu(t_1) = x_1 + \alpha \cdot 0 = x_1$. If the function $\mu(t)$ is kept fixed, then $J(x^* + \alpha\mu)$ is a function $I(\alpha)$ of only the single scalar α , given by

$$I(\alpha) = \int_{t_0}^{t_1} F(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t)) dt \quad (3)$$

Obviously $I(0) = J(x^*)$. Also, because of the hypothesis that $x^*(t)$ is optimal, $J(x^*) \geq J(x^* + \alpha\mu)$ and so $I(0) \geq I(\alpha)$ for all α . Hence, the function I has a maximum at $\alpha = 0$. Because I is a differentiable function and $\alpha = 0$ is an interior point in the domain of I , one must have $I'(0) = 0$. (Obviously, this equation must hold for the minimization problem as well.) This condition allows one to deduce the Euler equation.

Indeed, looking at (3), note that to calculate $I'(\alpha)$ requires differentiating under the integral sign. We apply formula (4.2.1) to obtain

$$I'(\alpha) = \int_{t_0}^{t_1} \frac{\partial}{\partial \alpha} F(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t)) dt$$

According to the chain rule, $\frac{\partial}{\partial \alpha} F(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t)) = F'_2 \cdot \mu(t) + F'_3 \cdot \dot{\mu}(t)$, where F'_2 and F'_3 are evaluated at $(t, x^*(t) + \alpha\mu(t), \dot{x}^*(t) + \alpha\dot{\mu}(t))$. For $\alpha = 0$ this is $(t, x^*(t), \dot{x}^*(t))$, so

$$I'(0) = \int_{t_0}^{t_1} [F'_2(t, x^*(t), \dot{x}^*(t)) \cdot \mu(t) + F'_3(t, x^*(t), \dot{x}^*(t)) \cdot \dot{\mu}(t)] dt$$

or, in more compact notation,

$$I'(0) = \int_{t_0}^{t_1} \left[\frac{\partial F^*}{\partial x} \mu(t) + \frac{\partial F^*}{\partial \dot{x}} \dot{\mu}(t) \right] dt \quad (i)$$

where the asterisks indicate that the derivatives are evaluated at $(t, x^*, \dot{x}^*)$.

To proceed further, integrate the second term of the integrand by parts to obtain

$$\int_{t_0}^{t_1} \frac{\partial F^*}{\partial \dot{x}} \dot{\mu}(t) dt = \left[\left(\frac{\partial F^*}{\partial \dot{x}} \right) \mu(t) \right]_{t_0}^{t_1} - \int_{t_0}^{t_1} \frac{d}{dt} \left(\frac{\partial F^*}{\partial \dot{x}} \right) \mu(t) dt$$

Inserting this result into (i) and rearranging the terms gives

$$I'(0) = \int_{t_0}^{t_1} \left[\frac{\partial F^*}{\partial x} - \frac{d}{dt} \left(\frac{\partial F^*}{\partial \dot{x}} \right) \right] \mu(t) dt + \left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_1} \mu(t_1) - \left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_0} \mu(t_0) \quad (4)$$

However, the function μ satisfies $\mu(t_0) = \mu(t_1) = 0$, so the last two terms of (4) are zero. Hence, the first-order condition $I'(0) = 0$ reduces to

$$\int_{t_0}^{t_1} \left[\frac{\partial F^*}{\partial x} - \frac{d}{dt} \left(\frac{\partial F^*}{\partial \dot{x}} \right) \right] \mu(t) dt = 0 \quad (ii)$$

In the argument leading to this result, $\mu(t)$ was a fixed function. But (ii) must be valid for *all* functions $\mu(t)$ that are C^2 on $[t_0, t_1]$ and that are 0 at t_0 and at t_1 . It stands to reason then that the bracketed expression in (ii) must be 0 for all t in $[t_0, t_1]$. (See Theorem 8.3.2 below.) It follows that $x^*(t)$ satisfies the Euler equation.

To prove that solving the Euler equation gives a *sufficient* condition, suppose that $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$. Suppose too that $x^* = x^*(t)$ satisfies the Euler equation as well as the boundary conditions $x^*(t_0) = x_0$ and $x^*(t_1) = x_1$. Let $x = x(t)$ be an arbitrary admissible function in the problem. Because $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$, Theorem 2.4.1 implies that

$$F(t, x, \dot{x}) - F(t, x^*, \dot{x}^*) \leq \frac{\partial F(t, x^*, \dot{x}^*)}{\partial x} (x - x^*) + \frac{\partial F(t, x^*, \dot{x}^*)}{\partial \dot{x}} (\dot{x} - \dot{x}^*) \quad (iii)$$

Using the Euler equation, reversing the inequality (iii) yields (with simplified notation)

$$\begin{aligned} F^* - F &\geq \frac{\partial F^*}{\partial x} (x^* - x) + \frac{\partial F^*}{\partial \dot{x}} (\dot{x}^* - \dot{x}) \\ &= \left[\frac{d}{dt} \left(\frac{\partial F^*}{\partial \dot{x}} \right) \right] (x^* - x) + \frac{\partial F^*}{\partial \dot{x}} (\dot{x}^* - \dot{x}) = \frac{d}{dt} \left[\frac{\partial F^*}{\partial \dot{x}} (x^* - x) \right] \end{aligned} \quad (iv)$$

Because (iv) is valid for all t in $[t_0, t_1]$, integrating yields

$$\int_{t_0}^{t_1} (F^* - F) dt \geq \int_{t_0}^{t_1} \frac{d}{dt} \left[\frac{\partial F^*}{\partial \dot{x}} (x^* - x) \right] dt = \left[\frac{\partial F^*}{\partial \dot{x}} (x^* - x) \right]_{t_0}^{t_1} \quad (5)$$

However, the functions $x^*(t)$ and $x(t)$ both satisfy the boundary conditions: $x^*(t_0) = x(t_0) (= x_0)$ and $x^*(t_1) = x(t_1) (= x_1)$. So the last expression in (5) is equal to 0. It follows that

$$\int_{t_0}^{t_1} [F(t, x^*, \dot{x}^*) - F(t, x, \dot{x})] dt \geq 0$$

for every admissible function $x = x(t)$. This confirms that $x^*(t)$ solves the maximization problem. The corresponding result for the minimization problem is easily derived. ■

For the interested reader we now show how equation (ii) in the proof above implies the Euler equation. First we need:

THEOREM 8.3.2 (THE FUNDAMENTAL LEMMA)

Suppose that f is a continuous function on $[t_0, t_1]$, and that $\int_{t_0}^{t_1} f(t)\mu(t) dt = 0$ for every function $\mu = \mu(t)$ that is C^2 in this interval and satisfies $\mu(t_0) = \mu(t_1) = 0$. Then $f(t) = 0$ for all t in $[t_0, t_1]$.

Proof: Suppose there exists a number s in (t_0, t_1) such that $f(s) > 0$. We construct a function μ which is C^2 in (t_0, t_1) and satisfies $\mu(t_0) = \mu(t_1) = 0$, as well as $\int_{t_0}^{t_1} f(t)\mu(t) dt > 0$. Indeed, because f is continuous, there must be an interval (α, β) with s in $(\alpha, \beta) \subseteq (t_0, t_1)$ such that $f(t) > 0$ for all t in (α, β) . For all t in $[t_0, t_1]$ define (see Fig. 2)

$$\mu(t) = \begin{cases} 0 & \text{if } t \notin (\alpha, \beta) \\ \varphi(t) & \text{if } t \in (\alpha, \beta) \end{cases}$$

where $\varphi(t) = (t - \alpha)^3(\beta - t)^3$.

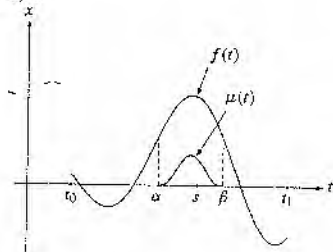


Figure 2

Because $(\alpha, \beta) \subseteq (t_0, t_1)$, we have $\mu(t_0) = \mu(t_1) = 0$. Moreover, $\varphi(t) = \varphi'(t) = \varphi''(t) = 0$ at $t = \alpha$ and $t = \beta$. Then $\mu(t) = \mu'(t) = \mu''(t) = 0$ at $t = \alpha$ and $t = \beta$, and it becomes obvious that $\mu(t)$ is C^2 everywhere. Now $f(t) \cdot \mu(t) > 0$ in (α, β) , while $f(t) \cdot \mu(t) = 0$ outside (α, β) . Therefore $\int_{t_0}^{t_1} f(t)\mu(t) dt = \int_{\alpha}^{\beta} f(t)\mu(t) dt > 0$. So the hypotheses of the theorem imply that $f(t) = 0$ for all t in $[t_0, t_1]$. Similarly, one can show that $f(t) \leq 0$ for all t in $[t_0, t_1]$. Therefore, $f(t) = 0$ for all t in (t_0, t_1) . By continuity it follows that $f(t) = 0$ for all t in $[t_0, t_1]$. ■

To apply this theorem to (ii) above, put $f(t) = (\partial F^* / \partial x) - (d/dt)(\partial F^* / \partial \dot{x})$. With our assumptions on F and $x = x(t)$, the function f is continuous on $[t_0, t_1]$, and so $f(t) = 0$ for all t in $[t_0, t_1]$, which reduces to the Euler equation.

NOTE 1 Suppose that in the standard variational problem the requirement

$$h(t, x, \dot{x}) > 0 \quad (*)$$

is imposed as an extra condition, where h is a given C^1 function of three variables. In order to be admissible a function $x(t)$ must then satisfy $h(t, x(t), \dot{x}(t)) > 0$ for all t in $[t_0, t_1]$. A simple case is one in which it is assumed that $x > 0$. If $x^*(t)$ solves the variational problem (8.2.1) with $(*)$ imposed, then $x^*(t)$ has to satisfy the Euler equation. In fact, because of the continuity assumption, $h(t, x(t), \dot{x}(t))$ must be positive for all $x(t)$ sufficiently close to $x^*(t)$, and the result follows because the Euler equation was derived by comparing the value of the objective functional only with the values for functions close to $x^*(t)$. Note that in this case the function F in (1) need only be defined and C^2 for triples $(t, x, \dot{x})$ that satisfy $h(t, x, \dot{x}) > 0$.

Suppose that for all t in $[t_0, t_1]$ the function $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$ for all $(x, \dot{x})$ satisfying $h(t, x, \dot{x}) > 0$, and that $h(t, x, \dot{x})$ is quasi-concave in $(x, \dot{x})$. Then an admissible $x^*(t)$ satisfying the Euler equation is optimal.

NOTE 2 A variational problem does not necessarily have a solution. For example, in the Ramsey model of Example 8.1.1, it is obvious that with realistic production functions the required terminal capital stock at the end of the planning period can be set so high that there is no admissible solution. Problem 3 includes another example where an optimal solution does not exist.

Note that the existence of a solution to a variational problem can often be established even if the concavity or convexity conditions in Theorem 8.3.1 are not satisfied. (By analogy, many non-concave functions of one variable have a maximum.) We refer to Section 10.4.

PROBLEMS FOR SECTION 8.3

1. Show that there is no solution to the problem

$$\max \int_0^1 (x^2 + \dot{x}^2) dt, \quad x(0) = 0, \quad x(1) = 0$$

(Hint: Let $x(t) = a(t - t^2)$, compute the integral, and let $a \rightarrow \infty$.)

2. (a) Write down the Euler equation associated with the problem

$$\max \int_0^T U(\bar{c} - \dot{x}e^{rt}) dt, \quad x(0) = x_0, \quad x(T) = 0$$

where $x = x(t)$ is the unknown function, T , $\bar{c}$, r , and x_0 are positive constants, and U is a given C^1 function of one variable.

- (b) Put $U(\bar{c}) = -e^{-w/v}$, where v is a positive constant. Write down and solve the Euler equation in this case, then explain why you have solved the problem.

HARDER PROBLEMS

3. Show that the problem $\min \int_a^1 t\dot{x}^2 dt$, $x(a) = 0$, $x(1) = 1$ has a solution if $a \in (0, 1)$, but not if $a = 0$.

4. (a) R. M. Goodwin considers the problem of finding the function $y(t)$ that maximizes the integral

$$\int_0^1 \ln[y - \sigma \dot{y} - \bar{z}l(t)] dt$$

Here σ and $\bar{z}$ are positive constants and $l(t)$ is a given positive function. Find the Euler equation.

- (b) Suppose that $l(t) = l_0 e^{\alpha t}$ and then find the solution of the equation when $\alpha\sigma \neq 1$.

8.4 Optimal Savings

This section considers in more detail the finite horizon optimal savings problem of Example 8.1.1.

EXAMPLE 1 Find the Euler equation for the Ramsey problem in Example 8.1.1:

$$\max \int_0^T U(f(K(t)) - \dot{K}(t))e^{-rt} dt, \quad K(0) = K_0, \quad K(T) = K_T$$

Deduce an expression for the corresponding relative rate of change of consumption, $\dot{C}/C$, where $C = f(K) - \dot{K}$. Also, show that the concavity condition in Theorem 8.3.1 is satisfied if $f'(K) > 0$, $f''(K) \leq 0$, $U'(C) > 0$, and $U''(C) < 0$.

Solution: Let $F(t, K, \dot{K}) = U(C)e^{-rt}$ with $C = f(K) - \dot{K}$. Then we find that $\partial F/\partial K = U'(C)f'(K)e^{-rt}$ and $\partial F/\partial \dot{K} = -U'(C)e^{-rt}$, so that the Euler equation reduces to $U'(C)f'(K)e^{-rt} - \frac{d}{dt}(-U'(C)e^{-rt}) = 0$. Both $U'(C)$ and e^{-rt} depend on t , so by the product rule for differentiation, $\frac{d}{dt}(U'(C)e^{-rt}) = U''(C)\dot{C}e^{-rt} - rU'(C)e^{-rt}$. Multiplying by e^{rt} and rearranging, it follows that

$$U'(C)(f'(K) - r) + U''(C)\dot{C} = 0$$

and so we obtain

$$\frac{\dot{C}}{C} = \frac{U'(C)}{CU''(C)}(r - f'(K)) = \frac{r - f'(K)}{\tilde{\omega}}$$

where $\tilde{\omega} = \text{El}_C U'(C) = CU''(C)/U'(C)$ is the **elasticity of marginal utility** with respect to consumption. Note that $\tilde{\omega} < 0$ because it is assumed that $U'(C) > 0$ and $U''(C) < 0$. (An estimate sometimes used for $\tilde{\omega}$ is -0.6 .) It follows that at any time t ,

$$\frac{\dot{C}}{C} > 0 \iff f'(K(t)) > r$$

Hence, consumption increases if and only if the marginal productivity of capital exceeds the discount rate.

On the other hand, if $f'(K) < r$, there is so much impatience to consume that consumption starts off high, then declines over time.

If we use the fact that $\dot{C} = f'(K)\dot{K} - \dot{K}$ in equation (*), and divide it by $U''(C)$, we get

$$\dot{K} - f'(K)\dot{K} + \frac{U'(C)}{U''(C)}(r - f'(K)) = 0 \quad (1)$$

Because f is concave ($f''(K) \leq 0$), it follows that $f'(K) - \dot{K}$ is also concave in $(K, \dot{K})$, as a sum of concave functions. The function U is increasing and concave, so $U(f(K) - \dot{K})e^{-rt}$ is also concave in $(K, \dot{K})$ (see Theorem 2.3.5(a)). Any solution of (1) that satisfies the boundary conditions must therefore be a solution of the problem.

Equation (1) is a complicated second-order differential equation. Explicit solutions are obtainable only in special cases. But note how interesting economic conclusions have been obtained anyway.

EXAMPLE 2 Solve Example 1 when $f(K) = bK$ and $U(C) = C^{1-v}/(1-v)$, where $b > 0$, $v > 0$, $v \neq 1$, and $b \neq (b-r)/v$.

Solution: Equation (1) yields

$$\dot{K} - \left(b - \frac{r-b}{v}\right)\dot{K} + \frac{b-r}{v}bK = 0$$

For $b \neq (b-r)/v$, this second-order differential equation has the general solution

$$K(t) = Ae^{bt} + Be^{(b-r)t/v} \quad (*)$$

The constants A and B are determined by the equations $K_0 = A + B$, $K_T = Ae^{bT} + Be^{(b-r)T/v}$. Because $f(K)$ is concave and U is increasing and concave, the function $K(t)$ given by (*), with the constants determined by these two equations, solves the problem.

PROBLEMS FOR SECTION 8.4

1. Find the Euler equation and solve the variational problem

$$\max \int_0^T e^{-t/4} \ln(2K - \dot{K}) dt, \quad K(0) = K_0, \quad K(T) = K_T$$

2. (a) Solve the problem

$$\max \int_0^T e^{-t/10} \left(\frac{1}{100} tx - \dot{x}^2 \right) dt, \quad x(0) = 0, \quad x(T) = S$$

- (b) Let $T = 10$ and $S = 20$ and find the solution in this case.

3. We generalize Example 1. Let $Y(t) = f(K(t), t)$ and replace $U(C)e^{-rt}$ by $U(C, t)$. Assume also that capital depreciates at the proportional rate δ , so that $C = f(K, t) - \dot{K} - \delta K$. The problem then becomes

$$\max \int_0^T U(f(K, t) - \dot{K} - \delta K, t) dt, \quad K(0) = K_0, \quad K(T) = K_T$$

What is the Euler equation in this case? Find an expression for $\dot{C}/C$.

4. A monopolist's production of a commodity per unit of time is $x = x(t)$. Suppose $b(x)$ is the associated cost function. At time t , let $D(p(t), \dot{p}(t))$ be the demand for the commodity per unit of time when the price is $p(t)$. If production at any time is adjusted to meet demand, the monopolist's total profit in the time interval $[0, T]$ is given by

$$\int_0^T [pD(p, \dot{p}) - b(D(p, \dot{p}))] dt$$

Suppose that $p(0)$ and $p(T)$ are given. The monopolist's natural problem is to find a price function $p(t)$ that maximizes the total profit.

- (a) Find the Euler equation associated with this problem.
 (b) Let $b(x) = \alpha x^2 + \beta x + \gamma$ and $x = D(p, \dot{p}) = Ap + B\dot{p} + C$, where α, β, γ, B , and C are positive constants, while A is negative. Solve the Euler equation in this case.

8.5 More General Terminal Conditions

So far in our variational problems, the initial and terminal values of the unknown function have all been fixed. In economic applications the initial point is usually fixed because it represents a historically given initial situation. On the other hand, in many models the terminal value of the unknown function can be free, or subject to more general restrictions. This section considers two of the most common terminal conditions that appear in economic models.

The problems we study are briefly formulated as

$$\max \int_{t_0}^{t_1} F(t, x, \dot{x}) dt, \quad x(t_0) = x_0, \quad \text{(a) } x(t_1) \text{ free or (b) } x(t_1) \geq x_1 \quad (1)$$

If the terminal condition is (a), any C^1 function is admissible if its graph joins the fixed point (t_0, x_0) to any point on the vertical line $t = t_1$, as illustrated in Fig. 1. We simply don't care where on the line $t = t_1$ the graph ends.

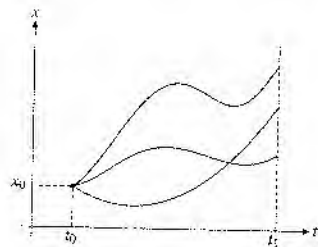


Figure 1 $x(t_1)$ free

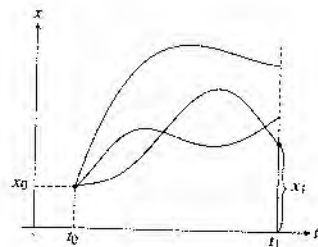


Figure 2 $x(t_1) \geq x_1$

Under terminal condition (b), any C^1 function is admissible if its graph joins the fixed point (t_0, x_0) to any point on or above the level x_1 on the vertical line $t = t_1$, as illustrated in Fig. 2.

Inequality terminal conditions like this are often encountered in economics. For instance, in the optimal savings model of the previous section it makes sense to replace the terminal condition $K(T) = K_T$ by $K(T) \geq K_T$.

The inequality sign in (b) sometimes needs to be reversed. For example, if $x(t)$ denotes the total stock of a pollutant in a lake, then $x(t_1) \leq x_1$ means that at the end of the planning period pollution should not exceed a prescribed level x_1 .

An important observation concerning (1) is that an optimal solution to either of the two problems must satisfy the Euler equation: Suppose $x^*(t)$ solves either problem, and let $\bar{x} = x^*(t_1)$. Then, in particular, $x^*(t)$ solves the corresponding variational problem with fixed terminal point $(t_1, \bar{x})$. According to Theorem 8.3.1, the function $x^*(t)$ must then satisfy the Euler equation. The condition $x^*(t_0) = x_0$ places one restriction on the constants in the general solution of the Euler equation. A so-called *transversality condition* is needed to determine both constants. The relevant transversality condition is spelled out in the next theorem.

THEOREM 8.5.1 (TRANSVERSALITY CONDITIONS)

If $x^*(t)$ solves problem (1) with either (a) or (b) as the terminal condition, then $x^*(t)$ must satisfy the Euler equation (8.2.2). With the terminal condition (a), the transversality condition is

$$\left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_1} = 0 \quad (2)$$

With the terminal condition (b), the transversality condition is

$$\left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_1} \leq 0, \quad \text{with} \quad \left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_1} = 0 \quad \text{if} \quad x^*(t_1) > x_1 \quad (3)$$

If $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$, then an admissible $x^*(t)$ that satisfies both the Euler equation and the appropriate transversality condition will solve problem (1).

Proof: We already know that the Euler equation must be satisfied. To derive the transversality conditions, we define the function J by $J(x) = \int_{t_0}^{t_1} F(t, x, \dot{x}) dt$, and we compare its value at $x^*(t)$ with its value at the perturbed function $x(t) = x^*(t) + \alpha \mu(t)$. In both cases we require that $\mu(t_0) = 0$.

Suppose the terminal condition is (a). The value of $x(t_1)$ is unconstrained, so the perturbed function $x(t)$ is admissible whatever the value of $\mu(t_1)$. Defining $I(\alpha)$ by (8.3.3), once again $I'(0)$ is given by (8.3.4). But the Euler equation is satisfied and $\mu(t_0) = 0$, so the condition that $I'(0) = 0$ reduces to

$$\left(\frac{\partial F^*}{\partial \dot{x}} \right)_{t=t_1} \mu(t_1) = 0$$

Because $\mu(t_1)$ can be chosen different from 0, the conclusion is that (2) must hold.

Now suppose that $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$. The argument leading to (8.3.5) holds as before. Evaluating the last expression in (8.3.5) at the upper limit yields 0 because of the transversality condition (2). Evaluating it at the lower limit also yields 0, because $x^*(t_0) = x(t_0) = x_0$. Again we conclude that $x^*(t)$ solves the maximization problem.

Suppose the terminal condition is (b). For $x = x^* + \alpha\mu$ to be admissible, μ must be chosen so that $\mu(t_0) = 0$ and $x^*(t_1) + \alpha\mu(t_1) \geq x_1$. There are two cases to consider:

(I) $x^*(t_1) > x_1$. In this case the optimal candidate “overshoots” the target. Choose $|\mu(t_1)|$ and $|\alpha|$ small enough so that $|\mu(t_1)| \cdot |\alpha| < x^*(t_1) - x_1$. Define $I(\alpha)$ as before by (8.3.3). Then $I(\alpha)$ must have a local maximum for $\alpha = 0$, so that $I'(0) = 0$, where $I'(0)$ is given in (8.3.4). Because the Euler equation is satisfied for $x^* = x^*(t)$ and $\mu(t_0) = 0$, we find as in the proof of (2) that $(\partial F/\partial \dot{x})_{t=t_1} \mu(t_1) = 0$. Choosing $\mu(t_1)$ different from 0 yields $(\partial F/\partial \dot{x})_{t=t_1} = 0$.

(II) $x^*(t_1) = x_1$. The requirement $x^*(t_1) + \alpha\mu(t_1) \geq x_1$ gives $\alpha\mu(t_1) \geq 0$ in this case. Choose $\mu(t)$ such that $\mu(t_0) = 0$ and $\mu(t_1) > 0$. Then $x^*(t) + \alpha\mu(t)$ is admissible for all $\alpha \geq 0$, and therefore $I(\alpha) \leq I(0)$ for all $\alpha \geq 0$. This implies that $I'(0) \leq 0$, and so $I'(0) = (\partial F/\partial \dot{x})_{t=t_1} \mu(t_1) \leq 0$. Because $\mu(t_1) > 0$, this yields $(\partial F/\partial \dot{x})_{t=t_1} \leq 0$.

Taken together, the conclusions in (I) and (II) reduce to (3), because $\mu(t_1) > 0$.

If $F(t, x, \dot{x})$ is concave in $(x, \dot{x})$, the argument leading up to (8.3.5) is again valid, and the last expression in (8.3.5) is now equal to

$$\left(\frac{\partial F^*}{\partial \dot{x}}\right)_{t=t_1} [x^*(t_1) - x(t_1)] \quad (*)$$

If $x^*(t_1) > x_1$, then $(\partial F^*/\partial \dot{x})_{t=t_1} = 0$ and the expression in (*) is equal to 0. If $x^*(t_1) = x_1$, then $x^*(t_1) - x(t_1) = x_1 - x(t_1) \leq 0$, since $x(t_1) \geq x_1$. Because $(\partial F^*/\partial \dot{x})_{t=t_1} \leq 0$ according to (3), the product in (*) is ≥ 0 . Thus the expression in (8.3.5) is always ≥ 0 , so the conclusion follows. ■

NOTE 1 Condition (3) is a little tricky. It says that $\partial F/\partial \dot{x}_{t=t_1}$ is always less than or equal to 0, but equal to 0 if $x^*(t_1)$ overshoots, in the sense that it is greater than x_1 .

NOTE 2 If we minimize the integral in (1), the theorem is still valid if (3) is replaced by

$$\left(\frac{\partial F}{\partial \dot{x}}\right)_{t=t_1} \geq 0, \quad \text{with} \quad \left(\frac{\partial F}{\partial \dot{x}}\right)_{t=t_1} = 0 \quad \text{if} \quad x^*(t_1) > x_1 \quad (4)$$

and we require $F(t, x, \dot{x})$ to be convex.

NOTE 3 If the inequality sign in (1)(b) is reversed, so is the inequality sign $\leq$ in (3).

EXAMPLE 1 Find the solutions of the following problems:

$$\max \int_0^1 (1 - x^2 - \dot{x}^2) dt, \quad x(0) = 1, \quad \text{with (a) } x(1) \text{ free or (b) } x(1) \geq 2$$

Solution: The Euler equation is easily seen to be $\ddot{x} - x = 0$, with general solution $x(t) = Ae^t + Be^{-t}$. The condition $x(0) = 1$ gives $1 = A + B$, so an optimal solution of either problem must be of the form $x^*(t) = Ae^t + (1 - A)e^{-t}$, and thus $\dot{x}^*(t) = Ae^t - (1 - A)e^{-t}$. Furthermore, $\partial F/\partial \dot{x} = -2\dot{x}$.

With (a) as the terminal condition, (2) requires $\dot{x}^*(1) = 0$, so $Ae^1 - (1 - A)e^{-1} = 0$, and hence $A = 1/(e^2 + 1)$. The only possible solution is therefore

$$x^*(t) = \frac{1}{e^2 + 1} (e^t + e^2 e^{-t})$$

Because $F(t, x, \dot{x}) = 1 - x^2 - \dot{x}^2$ is concave in $(x, \dot{x})$, the solution has been found.

With (b) as the terminal condition, we require $x^*(1) = Ae + (1 - A)e^{-1} \geq 2$, so $A \geq (2e - 1)/(e^2 - 1)$. Suppose $x^*(1) > 2$. Then condition (3) would give $\partial F^*/\partial \dot{x} = -2\dot{x}^* = 0$ at $t = 1$. Hence $\dot{x}^*(1) = 0$, and so $A = 1/(e^2 + 1)$. But this would violate the inequality $A \geq (2e - 1)/(e^2 - 1)$. We conclude that $x^*(1) = 2$, with $A = (2e - 1)/(e^2 - 1)$. Then $\dot{x}^*(1) = Ae + (1 - A)e^{-1} = 2(e^2 - e + 1)/(e^2 - 1) > 0$ for this value of A , implying that $\partial F^*/\partial \dot{x} = -2\dot{x}^*(1)$ is ≤ 0 . The only possible solution is therefore

$$x^*(t) = \frac{1}{e^2 - 1} ((2e - 1)e^t + (e^2 - 2e)e^{-t})$$

Because $F(t, x, \dot{x}) = 1 - x^2 - \dot{x}^2$ is concave in $(x, \dot{x})$, the solution has been found. ■

EXAMPLE 2 Consider the macroeconomic problem of Example 8.2.2, but now assume that $x(T)$ is unrestricted. Find the optimal solution $x^*(t)$, and discuss what happens to the terminal state $x^*(T)$ as the horizon $T \rightarrow \infty$ and also as $c \rightarrow 0$.

Solution: Again the Euler equation is $\ddot{x} - x/c = 0$, and the general solution satisfying the initial condition is $x = Ae^{rt} + (x_0 - A)e^{-rt}$ where $r = 1/\sqrt{c}$. The transversality condition (2) reduces to $\dot{x}(T) = 0$. Because $\dot{x}(t) = rAe^{rt} - r(x_0 - A)e^{-rt}$, this implies that $rAe^{rT} - r(x_0 - A)e^{-rT} = 0$. It follows that the only possible solution to the problem is

$$x^*(t) = \frac{x_0}{e^{rT} + e^{-rT}} [e^{r(T-t)} + e^{-r(T-t)}]$$

Note that then $x^*(T) = 2x_0/(e^{rT} + e^{-rT}) \rightarrow 0$ as $T \rightarrow \infty$. Also, as $c \rightarrow 0$, so $r \rightarrow \infty$ and therefore $x^*(T) \rightarrow 0$. In fact, because

$$\frac{e^{r(T-t)} + e^{-r(T-t)}}{e^{rT} + e^{-rT}} = \frac{e^{-rt} + e^{-r(2T-t)}}{1 + e^{-2rT}} \rightarrow 0 \quad \text{as} \quad r \rightarrow \infty$$

it follows that $x^*(t) \rightarrow 0$ for each $t > 0$ even with T fixed. This is not surprising. As c becomes small, the costs become negligible, so $x^*(t)$ gets adjusted to 0 almost immediately.

Because $F = x^2 + \dot{x}^2$ is convex in $(x, \dot{x})$, the optimal solution has been found. ■

EXAMPLE 3 Let $A(t)$ denote a pensioner's wealth at time t , and let w be the (constant) pension income per unit of time. Suppose that the person can borrow and save at the same constant rate of interest r . Consumption per unit of time at time t is then given by $C(t) = rA(t) + w - \dot{A}(t)$. Suppose the pensioner plans consumption from now, $t = 0$, until the expected time of death T , so as to maximize

$$\int_0^T U(C(t))e^{-\rho t} dt = \int_0^T U(rA(t) + w - \dot{A}(t))e^{-\rho t} dt$$

where U is a utility function with $U' > 0$, $U'' < 0$, and ρ is a discount rate. Suppose that present wealth is A_0 , and the minimum desired legacy is A_T , so that an admissible wealth function must satisfy $A(0) = A_0$ and $A(T) \geq A_T$. Characterize the possible solutions. (This model has been studied by Atkinson (1971).)

Solution: The objective function is $F(t, A, \dot{A}) = U(rA + w - \dot{A})e^{-\rho t}$. The Euler equation is easily shown to be

$$\ddot{A} - r\dot{A} + (\rho - r)U'/U'' = 0 \quad (*)$$

Because $U' > 0$, one has $\partial F/\partial \dot{A} = -U'(C)e^{-\rho t} < 0$ everywhere. Therefore (3) implies that $A^*(T) = A_T$. Hence, any optimal solution $A^*(t)$ of the problem must satisfy (*) with $A^*(0) = A_0$ and $A^*(T) = A_T$. Because of the requirement imposed on U , the function $F(t, A, \dot{A})$ is concave in $(A, \dot{A})$ (for the same reason as in Example 8.4.1). Note that we have not proved that (*) really has a solution that satisfies the boundary conditions. See Problem 4 for a special case.

PROBLEMS FOR SECTION 8.5

1. Solve the problem

$$\min \int_0^1 (r\dot{x} + x^2) dt, \quad x(0) = 1, \quad \text{(i) with } x(1) \text{ free, (ii) with } x(1) \geq 1$$

- SM 2. (a) Solve the variational problem

$$\max \int_0^1 (10 - \dot{x}^2 - 2x\dot{x} - 5x^2)e^{-t} dt, \quad x(0) = 0, \quad x(1) = 1$$

- (b) What is the optimal solution if (i) $x(1)$ is free? (ii) $x(1) \geq 2$?

3. J. K. Sengupta has considered the problem

$$\min \int_0^T (\alpha_1 \tilde{Y}^2 + \alpha_2 \tilde{G}^2) dt, \quad \dot{\tilde{Y}} = r_1 \tilde{Y} - r_2 \tilde{G}, \quad \tilde{Y}(0) = Y_0, \quad \tilde{Y}(T) \text{ free}$$

where $\alpha_1, \alpha_2, r_1, r_2, T$, and Y_0 are given positive constants. Formulate this as a variational problem with $\tilde{Y} = \tilde{Y}(t)$ as the unknown function. Find the corresponding Euler equation, and solve the problem.

- SM 4. Solve the problem in Example 3 when $U(C) = a - e^{-bC}$, with $a > 0$ and $b > 0$.

- SM 5. (a) A community wants to plant trees to cover a 1500 hectare piece of land over a period of 5 years. Let $x(t)$ be the number of hectares that have been planted by time t , and let $u(t)$ be the rate of planting, $\dot{x}(t) = u(t)$. Let the cost per unit of time of planting be given by the function $C(t, u)$. The total discounted cost of planting at the rate $u(t)$ in the period from $t = 0$ to $t = 5$, when the rate of interest is r , is then $\int_0^5 C(t, u)e^{-rt} dt$. Write down the necessary conditions for the problem

$$\min \int_0^5 C(t, \dot{x})e^{-rt} dt, \quad x(0) = 0, \quad x(5) \geq 1500$$

- (b) Solve the problem when $r = 0$, and $C(t, u) = g(u)$, with $g(0) = 0$, $g'(u) > 0$, and $g''(u) > 0$.

9

CONTROL THEORY: BASIC TECHNIQUES

A person who insists on understanding every tiny step before going to the next is liable to concentrate so much on looking at his feet that he fails to realize he is walking in the wrong direction.

—I. Stewart (1975)

Optimal control theory is a modern extension of the classical calculus of variations. Whereas the Euler equation dates back to 1744, the main result in optimal control theory, the **maximum principle**, was developed as recently as the 1950s by a group of Russian mathematicians (Pontryagin et al. (1962)). This principle gives necessary conditions for optimality in a wide range of dynamic optimization problems. It includes all the necessary conditions that emerge from the classical theory, but can also be applied to a significantly wider range of problems.

Since 1960, thousands of papers in economics have used optimal control theory. Its applications include, for instance, economic growth, inventory control, taxation, extraction of natural resources, irrigation, and the theory of regulation under asymmetric information.

This chapter contains some important results for the one state variable case that are used widely in the economics literature. ("What every young economist should know about optimal control theory.")

After the introductory Section 9.1, Section 9.2 deals with the simple case in which there are no restrictions on the control variable or on the terminal state. Although such problems can usually be solved by using the calculus of variations, their relative simplicity make them ideal as a starting point for introducing some of the main concepts and ideas in control theory.

Some additional concepts like the control region, piecewise continuous controls, and required regularity conditions are spelled out in Section 9.3.

Section 9.4 goes on to consider different alternative terminal conditions on the state variable. The brief Section 9.5 shows how to formulate a calculus of variations problem as an optimal control problem. The Euler equation is easily derived as an implication of the maximum principle.

Section 9.6 is concerned with sensitivity results: what happens to the optimal value function when the parameters change? In a growth theory setting, these parameters are: the beginning and end of the planning period; the initial capital stock; and the amount of capital to leave at the end of the planning period. It turns out that the adjoint variable can be given interesting economic interpretations, somewhat similar to those for Lagrange multipliers.

A rigorous proof of the maximum principle is beyond the level of this book. However, the Mangasarian sufficiency theorem is proved quite easily in Section 9.7. An important generalization (the Arrow sufficiency theorem) is also explained.

In some economic problems the final time (say the end of the planning period) is a variable to be chosen optimally along with the optimal control. Such problems are discussed in Section 9.8. Note 9.8.1 points out that the conditions in the Mangasarian or Arrow theorems are not sufficient for optimality in variable time problems.

In optimal economic growth models there is often a discount factor in the objective function. Then a slight reformulation of the maximum principle, called the current value formulation, is frequently used. This approach is explained in Section 9.9.

Some economic models include a "scrap value" in the objective function. Section 9.10 explains how to adjust the maximum principle to take care of this case.

Most control models in the economics literature assume an infinite time horizon. Section 9.11 is an attempt to give some correct results and examples in this area where there is some confusion, even in leading textbooks.

Even control problems for which explicit solutions are unobtainable can sometimes be analysed by the phase diagram technique explained in Section 9.12.

9.1 The Basic Problem

Consider a system whose state at time t is characterized by a number $x(t)$, the **state variable**. The process that causes $x(t)$ to change can be controlled, at least partially, by a **control function** $u(t)$. We assume that the rate of change of $x(t)$ depends on t , $x(t)$, and $u(t)$. The state at some initial point t_0 is typically known, $x(t_0) = x_0$. Hence, the evolution of $x(t)$ is described by a controlled differential equation

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0 \quad (1)$$

Suppose we choose some control function $u(t)$ defined for $t \geq t_0$. Inserting this function into (1) gives a first-order differential equation for $x(t)$. Because the initial point is fixed, a unique solution of (1) is usually obtained.

By choosing different control functions $u(t)$, the system can be steered along different paths, not all of which are equally desirable. As usual in economic analysis, assume that it is possible to measure the benefits associated with each path. More specifically, assume that the benefits can be measured by means of the integral

$$J = \int_{t_0}^{t_1} f(t, x(t), u(t)) dt \quad (2)$$

where f is a given function. Here, J is called the **objective** or the **criterion**. Certain restrictions are often placed on the final state $x(t_1)$. Moreover, the time t_1 at which the process stops is not necessarily fixed. The fundamental problem that we study is:

Among all pairs $(x(t), u(t))$ that obey the differential equation in (1) with $x(t_0) = x_0$ and that satisfy the constraints imposed on $x(t_1)$, find one that maximizes (2).

EXAMPLE 1. (Economic growth) Consider the control problem

$$\max \int_0^T (1-s)f(k) dt, \quad \dot{k} = sf(k), \quad k(0) = k_0, \quad k(T) \geq k_T, \quad 0 \leq s \leq 1$$

Here $k = k(t)$ is the real capital stock of a country and $f(k)$ is its production function. Moreover, $s = s(t)$, the control variable, is the fraction of production set aside for investment, and it is natural to require that $s \in [0, 1]$. The quantity $(1-s)f(k)$ is the flow of consumption per unit of time. We wish to maximize the integral of this quantity over $[0, T]$, i.e. to maximize total consumption over the period $[0, T]$. The constant k_0 is the initial capital stock, and the condition $k(T) \geq k_T$ means that we wish to leave a capital stock of at least k_T to those who live after time T . (Example 9.6.3(b) studies a special case of this model.)

EXAMPLE 2. (Oil extraction) Let $x(t)$ denote the amount of oil in a reservoir at time t . Assume that at $t = 0$ the field contains K barrels of oil, so that $x(0) = K$. If $u(t)$ is the rate of extraction, then integrating each side of (*) yields $x(t) - x(0) = -\int_0^t u(\tau) d\tau$, or $x(t) = K - \int_0^t u(\tau) d\tau$ for each $t \geq 0$. That is, the amount of oil left at time t is equal to the initial amount K , minus the total amount that has been extracted during the time span $[0, t]$, namely $\int_0^t u(\tau) d\tau$. Differentiating gives

$$\dot{x}(t) = -u(t), \quad x(0) = K \quad (*)$$

Suppose that the market price of oil at time t is known to be $q(t)$, so that the sales revenue per unit of time at t is $q(t)u(t)$. Assume further that the cost C per unit of time depends on t , x and u , so that $C = C(t, x, u)$. The instantaneous profit per unit of time at time t is then

$$\pi(t, x(t), u(t)) = q(t)u(t) - C(t, x(t), u(t))$$

If the discount rate is r , the total discounted profit over the interval $[0, T]$ is

$$\int_0^T [q(t)u(t) - C(t, x(t), u(t))] e^{-rt} dt \quad (**)$$

It is natural to assume that $u(t) \geq 0$, and that $x(T) \geq 0$.

Problem I: Find the rate of extraction $u(t) \geq 0$ that maximizes (**) subject to (*) and $x(T) \geq 0$ over a fixed extraction period $[0, T]$.

Problem II: Find the rate of extraction $u(t) \geq 0$ and also the optimal terminal time T that maximizes (**) subject to (*) and $x(T) \geq 0$.

These are two instances of *optimal control problems*. Problem I has a fixed terminal time T , whereas Problem II is referred to as a free terminal time problem. See Example 9.8.1.

9.2 A Simple Case

We begin by studying a control problem with no restrictions on the control variable and no restrictions on the terminal state—that is, no restrictions are imposed on the value of $x(t)$ at $t = t_1$. Given the fixed times t_0 and t_1 , our problem is

$$\text{maximize } \int_{t_0}^{t_1} f(t, x(t), u(t)) dt, \quad u(t) \in (-\infty, \infty) \quad (1)$$

subject to

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0, \quad x_0 \text{ fixed}, \quad x(t_1) \text{ free} \quad (2)$$

Given any control function $u(t)$ defined on $[t_0, t_1]$, the associated solution of the differential equation in (2) with $x(t_0) = x_0$ will usually be uniquely determined on the whole of $[t_0, t_1]$. A pair $(x(t), u(t))$ that satisfies (2) is called an **admissible pair**. Among all admissible pairs we search for an **optimal pair**, i.e. a pair of functions that maximizes the integral in (1).

Notice that the problem is to maximize the objective w.r.t. u subject to the constraint (2). Because this constraint is a differential equation on the interval $[t_0, t_1]$, it can be regarded as an infinite number of equality constraints, one for each time t in $[t_0, t_1]$.

Economists usually incorporate equality constraints in their optimization problems by forming a Lagrangian function, with a Lagrange multiplier corresponding to each constraint. Analogously, we associate a number $p(t)$, called the **co-state variable**, with the constraint (2) for each t in $[t_0, t_1]$. The resulting function $p = p(t)$ is called the **adjoint function** associated with the differential equation. Corresponding to the Lagrangian function in the present problem is the **Hamiltonian** H .¹ For each time t in $[t_0, t_1]$ and each possible triple (x, u, p) , of the state, control, and adjoint variables, the Hamiltonian is defined by

$$H(t, x, u, p) = f(t, x, u) + pg(t, x, u) \quad (3)$$

A set of necessary conditions for optimality is given in the following theorem (some regularity conditions required are discussed in the next section):

THEOREM 9.2.1 (THE MAXIMUM PRINCIPLE)

Suppose that $(x^*(t), u^*(t))$ is an optimal pair for problem (1)–(2). Then there exists a continuous function $p(t)$ such that, for each t in $[t_0, t_1]$,

$$u = u^*(t) \text{ maximizes } H(t, x^*(t), u, p(t)) \text{ for } u \text{ in } (-\infty, \infty) \quad (4)$$

$$\dot{p}(t) = -H'_x(t, x^*(t), u^*(t), p(t)), \quad p(t_1) = 0 \quad (5)$$

NOTE 1 The requirement that $p(t_1) = 0$ in (5) is called a **transversality condition**. So condition (5) tells us that in the case where $x(t_1)$ is free, the adjoint variable vanishes at t_1 .

The conditions in Theorem 9.2.1 are necessary, but not sufficient for optimality. The following theorem, which is a special case of Theorem 9.4.2 below, gives sufficient conditions:

¹ The correspondence is rather loose. Something closer to the Lagrangian would be the function $f(t, x, u) + p(g(t, x, u) - \dot{x})$.

THEOREM 9.2.2 (MANGASARIAN)

If the requirement

$$H(t, x, u, p(t)) \text{ is concave in } (x, u) \text{ for each } t \text{ in } [t_0, t_1] \quad (6)$$

is added to the requirements in Theorem 9.2.1, then we obtain *sufficient* conditions. That is, if we find a triple $(x^*(t), u^*(t), p(t))$ that satisfies (2), (4), (5), and (6), then $(x^*(t), u^*(t))$ is optimal.

NOTE 2 Changing $u(t)$ on a small interval causes $f(t, x, u)$ to change immediately. Moreover, at the end of this interval $x(t)$ has changed and this change is transmitted throughout the remaining time interval. In order to steer the process optimally, the choice of $u(t)$ at each instant of time must anticipate the future changes in $x(t)$. In short, we have to plan ahead. In a certain sense, the adjoint function $p(t)$ takes care of this need for forward planning. Equation (5) implies that $p(t) = \int_t^{t_1} H'_x(s, x^*(s), u^*(s), p^*(s)) ds$.

NOTE 3 If the problem is to minimize the objective in (1), then we can rewrite the problem as one of maximizing the negative of the original objective function. Alternatively, we could reformulate the maximum principle for the minimization problem: an optimal control will minimize the Hamiltonian, and convexity of $H(t, x, u, p(t))$ w.r.t. (x, u) is the relevant sufficient condition.

Since the control region is $(-\infty, \infty)$, a *necessary* condition for (4) is that

$$H'_u(t, x^*(t), u^*(t), p(t)) = 0 \quad (7)$$

If $H(t, x(t), u, p(t))$ is concave in u , condition (7) is also *sufficient* for the maximum condition (4) to hold, because we recall that an interior stationary point for a concave function is (globally) optimal.

It is helpful to see how these conditions allow some simple examples to be solved.

EXAMPLE 1

Solve the problem

$$\max \int_0^T [1 - tx(t) - u(t)^2] dt, \quad \dot{x}(t) = u(t), \quad x(0) = x_0, \quad x(T) \text{ free}, \quad u \in \mathbb{R}$$

where x_0 and T are given positive constants.

Solution. The Hamiltonian is $H(t, x, u, p) = 1 - tx - u^2 + pu$, which is concave in u , so the control $u = u^*(t)$ maximizes $H(t, x^*(t), u, p(t))$ w.r.t. u if and only if it satisfies $H'_u = -2u + p(t) = 0$. Thus $u^*(t) = \frac{1}{2}p(t)$. Because $H'_x = -t$, the conditions in (5) reduce to $\dot{p}(t) = t$ and $p(T) = 0$. Integrating gives $p(t) = \frac{1}{2}t^2 + C$ with $\frac{1}{2}T^2 + C = 0$, so

$$p(t) = -\frac{1}{2}(T^2 - t^2) \quad \text{and then} \quad u^*(t) = -\frac{1}{4}(T^2 - t^2)$$

Because $\dot{x}^*(t) = u^*(t) = -\frac{1}{4}(T^2 - t^2)$, integrating $\dot{x}^*(t) = u^*(t)$ with $x^*(0) = x_0$ gives

$$x^*(t) = x_0 - \frac{1}{4}T^2t + \frac{1}{12}t^3$$

Thus, there is only one pair $(x^*(t), u^*(t))$ that, together with $p(t)$, satisfies both necessary conditions (4) and (5). We have therefore found the only possible pair that can solve the problem. Because $H(t, x, u, p) = 1 - tx - u^2 + pu$ is concave in (x, u) for each fixed t (it is a sum of concave functions), $(x^*(t), u^*(t))$ is indeed optimal. $\square$

EXAMPLE 2 Solve the problem

$$\max_{u(t) \in (-\infty, \infty)} \int_0^T (-x^2 - \frac{1}{2}u^2)e^{-2t} dt, \quad \dot{x} = x + u, \quad x(0) = 1, \quad x(T) \text{ free}$$

Solution: The Hamiltonian is $H(t, x, u, p) = (-x^2 - \frac{1}{2}u^2)e^{-2t} + p(x+u)$. The maximum principle states that if an admissible pair $(x^*(t), u^*(t))$ solves the problem, then there exists a function p defined on $[0, T]$ such that:

(i) For every t in $[0, T]$, $u = u^*(t)$ maximizes $H(t, x^*(t), u, p(t))$ for u in $(-\infty, \infty)$.

(ii) $\dot{p}(t) = -H'_x(t, x^*(t), u^*(t), p(t)) = 2x^*(t)e^{-2t} - p(t)$ and $p(T) = 0$.

Since $H'_u = -ue^{-2t} + p$, it follows from (i) that $u^*(t) = e^{2t}p(t)$. The equation $\dot{x}^* = x^* + u^*$ then yields $\dot{x}^*(t) = x^*(t) + e^{2t}p(t)$. Thus x^* and p must satisfy the system

$$\begin{aligned}\dot{x} &= x + e^{2t}p \\ \dot{p} &= 2e^{-2t}x - p\end{aligned}$$

of two simultaneous equations in the plane, which appeared as Problem 6.5.3. The general solution is $x = Ae^{(1+\sqrt{2})t} + Be^{(1-\sqrt{2})t}$ and $p = A\sqrt{2}e^{(\sqrt{2}-1)t} - B\sqrt{2}e^{(-\sqrt{2}-1)t}$.

It remains to determine the constants A and B so that $x^*(0) = 1$ and $p(T) = 0$. This yields $A + B = 1$ and $A\sqrt{2}e^{(\sqrt{2}-1)T} - B\sqrt{2}e^{(-\sqrt{2}-1)T} = 0$. The solution is $A = (1 + e^{2\sqrt{2}T})^{-1}$ and $B = e^{2\sqrt{2}T}(1 + e^{2\sqrt{2}T})^{-1}$. The corresponding functions x^* , u^* , and p satisfy the conditions in the maximum principle, and H is concave w.r.t. (x, u) , so it follows from Theorem 9.2.2 that this is the optimal solution. $\square$

EXAMPLE 3 (A macroeconomic control problem) Consider once again the macroeconomic model of Example 8.2.2. If we drop the terminal constraint at the end of the planning period, we face the following control problem:

$$\min \int_0^T [x(t)^2 + cu(t)^2] dt, \quad \dot{x}(t) = u(t), \quad x(0) = x_0, \quad x(T) \text{ free}$$

where $u(t) \in \mathbb{R}$ and $c > 0$, x_0 , and T are given. Use the maximum principle to solve the problem.

Solution: We maximize $-\int_0^T [x(t)^2 + cu(t)^2] dt$. The Hamiltonian is $H(t, x, u, p) = -x^2 - cu^2 + pu$, so $H'_x = -2x$ and $H'_u = -2cu + p$. A necessary condition for $u = u^*(t)$ to maximize the Hamiltonian is that $H'_u = 0$ at $u = u^*(t)$, or that $-2cu^*(t) + p(t) = 0$. Therefore, $u^*(t) = p(t)/2c$. The differential equation for $p(t)$ is

$$\dot{p}(t) = -H'_x(t, x^*(t), u^*(t), p(t)) = 2x^*(t) \quad (*)$$

From $\dot{x}^*(t) = u^*(t)$ and $u^*(t) = p(t)/2c$, we have

$$\dot{x}^*(t) = p(t)/2c \quad (**)$$

The two first-order differential equations (*) and (**) can be used to determine the functions p and x^* . Differentiate (*) w.r.t. t and then use (**) to obtain $\ddot{p}(t) = 2\dot{x}^*(t) = p(t)/c$, whose general solution is

$$p(t) = Ae^{rt} + Be^{-rt}, \quad \text{where } r = 1/\sqrt{c}$$

Imposing the boundary conditions $p(T) = 0$ and $\dot{p}(0) = 2x^*(0) = 2x_0$ implies that $Ae^{rT} + Be^{-rT} = 0$ and $r(A - B) = 2x_0$. These two equations determine A and B , and they yield $A = 2x_0e^{-rT}/[r(e^{rT} + e^{-rT})]$ and $B = -2x_0e^{rT}/[r(e^{rT} + e^{-rT})]$. Therefore,

$$p(t) = \frac{2x_0}{r} \frac{e^{-r(T-t)} - e^{r(T-t)}}{e^{rT} + e^{-rT}} \quad \text{and} \quad x^*(t) = \frac{1}{2} \dot{p}(t) = x_0 \frac{e^{r(T-t)} + e^{-r(T-t)}}{e^{rT} + e^{-rT}}$$

The Hamiltonian $H = -x^2 - cu^2 + pu$ is concave in (x, u) , so by Mangasarian's theorem, this is the solution to the problem. (The same result was obtained in Example 8.5.2.) $\square$

PROBLEMS FOR SECTION 9.2

Solve the control problems 1–5:

$$1. \max_{u(t) \in (-\infty, \infty)} \int_0^2 [e^t x(t) - u(t)^2] dt, \quad \dot{x}(t) = -u(t), \quad x(0) = 0, \quad x(2) \text{ free}$$

$$2. \max_{u(t) \in (-\infty, \infty)} \int_0^1 [1 - u(t)^2] dt, \quad \dot{x}(t) = x(t) + u(t), \quad x(0) = 1, \quad x(1) \text{ free}$$

$$\textcircled{SM} 3. \min_{u(t) \in (-\infty, \infty)} \int_0^1 [x(t) + u(t)^2] dt, \quad \dot{x}(t) = -u(t), \quad x(0) = 0, \quad x(1) \text{ free}$$

$$\textcircled{SM} 4. \max_{u(t) \in (-\infty, \infty)} \int_0^{10} [1 - 4x(t) - 2u(t)^2] dt, \quad \dot{x}(t) = u(t), \quad x(0) = 0, \quad x(10) \text{ free}$$

$$\text{CSM } 5. \max_{u(t) \in (-\infty, \infty)} \int_0^T (x - u^2) dt, \quad \dot{x} = x + u, \quad x(0) = 0, \quad x(T) \text{ free}$$

CSM 6. (a) Write down conditions (7) and (5) for the problem

$$\max_{u(t) \in (-\infty, \infty)} \int_0^T [qf(K) - c(I)] dt, \quad \dot{K} = I - \delta K, \quad K(0) = K_0, \quad K(T) \text{ free}$$

Here is an economic interpretation: $K = K(t)$ denotes the capital stock of a firm, $f(K)$ is the production function, q is the price per unit of output, $I = I(t)$ is investment, $c(I)$ is the cost of investment, δ is the rate of depreciation of capital, K_0 is the initial capital stock, and T is the fixed planning horizon.

(b) Let $f(K) = K - 0.03K^2$, $q = 1$, $c(I) = I^2$, $\delta = 0.1$, $K_0 = 10$, and $T = 10$. Derive a second-order differential equation for K , and explain how to find the solution.

9.3 Regularity Conditions

In most applications of control theory to economics, the control functions are explicitly or implicitly restricted in various ways. For instance, $u(t) \geq 0$ was a natural restriction in the oil extraction problem of Section 9.1; it means that one cannot pump oil back into the reservoir.

In general, assume that $u(t)$ takes values in a fixed subset U of the reals, called the **control region**. In the oil extraction problem, then, $U = [0, \infty)$. Actually, an important aspect of control theory is that the control region may be a closed set, so that $u(t)$ can take values at the boundary of U . (In the classical calculus of variation, by contrast, one usually considered open control regions, although developments around 1930–1940 paved the way for the modern theory.)

What regularity conditions is it natural to impose on the control function $u(t)$? Among the many papers in the economics literature that use control theory, the majority assume implicitly or explicitly that the control functions are continuous. Consequently, many of our examples and problems will deal with continuous controls. Yet in some applications, continuity is too restrictive. For example, the control variable $u(t)$ could be the fraction of investment allocated to one plant, with the remaining fraction $1 - u(t)$ is allocated to another. Then it is natural to allow control functions that suddenly switch all the investment from one plant to the other. Because they alternate between extremes, such functions are often called **bang-bang** controls. A simple example of such a control is

$$u(t) = \begin{cases} 1 & \text{for } t \text{ in } [t_0, t'] \\ 0 & \text{for } t \text{ in } (t', t_1] \end{cases}$$

which involves a single shift at time t' . In this case $u(t)$ is *piecewise continuous*, with a jump discontinuity at $t = t'$.

By definition, a function of one variable has a **finite jump** at a point of discontinuity if it has (finite) one-sided limits from both above and below at that point. A function is **piecewise**

continuous if it has at most a finite number of discontinuities in each finite interval, with finite jumps at each point of discontinuity. The value of a control $u(t)$ at each isolated point of discontinuity will affect neither the integral objective nor the state, but let us agree to choose the value of $u(t)$ at a point of discontinuity t' as the left-hand limit of $u(t)$ at t' . Then $u(t)$ will be **left-continuous**, as illustrated in Fig. 1. Moreover, if the control problem concerns the time interval $[t_0, t_1]$, we shall assume that $u(t)$ is continuous at both end points of this interval.

What is meant by a “solution” of $\dot{x} = g(t, x, u)$ when $u = u(t)$ has discontinuities? A **solution** is a **continuous** function $x(t)$ that has a derivative that satisfies the equation, except at points where $u(t)$ is discontinuous. The graph of $x(t)$ will, in general, have “kinks” at the points of discontinuity of $u(t)$, and it will usually not be differentiable at these kinks. It is, however, still continuous at the kinks.

For the oil extraction problem in Example 9.1.2, Fig. 1 shows one possible control function, whereas Fig. 2 shows the corresponding development of the state variable. The rate of extraction is initially a constant u_0 on the interval $[0, t']$, then a different constant u_1 (with $u_1 < u_0$) on $(t', t'']$. Finally, on (t'', T) , the rate of extraction $u(t)$ gradually declines from a level lower than u_1 until the field is exhausted at time T . Observe that the graph of $x(t)$ is connected, but has kinks at t' and t'' .

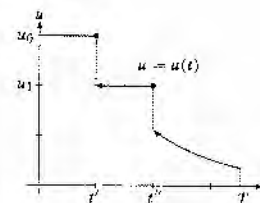


Figure 1

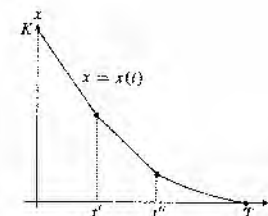


Figure 2

So far no restrictions have been placed on the functions $g(t, x, u)$ and $f(t, x, u)$. For the analysis presented in this chapter, it suffices to assume that f , g , and their first-order partial derivatives w.r.t. x and u are continuous in (t, x, u) . These continuity properties will be implicitly assumed from now on.

Necessary Conditions, Sufficient Conditions, and Existence

In static optimization theory there are three main types of result that can be used to find possible global solutions: theorems giving necessary conditions for optimality (typically, first-order conditions); theorems giving sufficient conditions (typically, first-order conditions supplemented by appropriate concavity/convexity requirements); and existence theorems (typically, the extreme value theorem).

In control theory the situation is similar. The maximum principle, in different versions, gives *necessary* conditions for optimality, i.e. conditions that an optimal control *must* satisfy. These conditions do not guarantee that the maximization problem has a solution.

The second type of theorem consists of *sufficiency results*, of the kind originally developed by Mangasarian. Theorems of this type impose certain concavity/convexity requirements on the functions involved. If a control function $u^*(t)$ (with corresponding state variable $x^*(t)$ and adjoint variable $p(t)$) satisfies the stated sufficient conditions, then $(x^*(t), u^*(t))$ solves the maximization problem. But these sufficient conditions are rather demanding, and in many problems there are optimal solutions even though the sufficient conditions are not satisfied.

Existence theorems give conditions which ensure that an optimal solution of the problem really exists. The conditions needed for existence are less stringent than the sufficient conditions. Existence theorems are used (in principle) in the following way: One finds, by using the necessary conditions, all the "candidates" for a solution of the problem. If the existence of an optimal solution is assured, then an optimal solution can be found by simply examining which of the candidates gives the largest value of the objective function. (This direct comparison of different candidates is unnecessary if we use sufficient conditions.)

9.4 The Standard Problem

Section 9.2 studied a control problem with no restriction on the control function at any time, and also no restriction on the state variable at the terminal time; $x(t_1)$ was free. These features are unrealistic in many economic models, as has already been pointed out.

This section considers the "standard end-constrained problem"

$$\max \int_{t_0}^{t_1} f(t, x(t), u(t)) dt, \quad u \in U \subseteq \mathbb{R} \quad (1)$$

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0 \quad (2)$$

with one of the following terminal conditions imposed

$$(a) \ x(t_1) = x_1, \quad (b) \ x(t_1) \geq x_1, \quad \text{or} \quad (c) \ x(t_1) \text{ free} \quad (3)$$

Again, t_0, t_1, x_0 , and x_1 are fixed numbers and U is the fixed control region. A pair $(x(t), u(t))$ that satisfies (2) and (3) with $u(t) \in U$ is called an **admissible pair**. Among all admissible pairs we seek an **optimal pair**, i.e. a pair of functions that maximizes the integral in (1).

In order to formulate correct necessary conditions, we define the Hamiltonian as

$$H(t, x, u, p) = p_0 f(t, x, u) + p g(t, x, u) \quad (4)$$

The new feature is the constant number p_0 in front of $f(t, x, u)$. If $p_0 \neq 0$, we can divide by p_0 to get a new Hamiltonian in which $p_0 = 1$, in effect. But if $p_0 = 0$, this normalization is impossible.

The following result is proved in Fleming and Rishel (1975):

THEOREM 9.4.1 (THE MAXIMUM PRINCIPLE: STANDARD END CONSTRAINTS)

Suppose that $(x^*(t), u^*(t))$ is an optimal pair for the standard end-constrained problem (1)–(3). Then there exist a continuous function $p(t)$ and a number p_0 , which is either 0 or 1, such that for all t in $[t_0, t_1]$ we have $(p_0, p(t)) \neq (0, 0)$ and, moreover:

(A) The control $u^*(t)$ maximizes $H(t, x^*(t), u, p(t))$ w.r.t. $u \in U$, i.e.

$$H(t, x^*(t), u, p(t)) \leq H(t, x^*(t), u^*(t), p(t)) \text{ for all } u \text{ in } U \quad (5)$$

(B) $\dot{p}(t) = -H'_x(t, x^*(t), u^*(t), p(t))$ (6)

(C) Corresponding to each of the terminal conditions in (3) there is a **transversality condition** on $p(t_1)$:

- (a') $p(t_1)$ no condition
 - (b') $p(t_1) \geq 0$, with $p(t_1) = 0$ if $x^*(t_1) > x_1$
 - (c') $p(t_1) = 0$
- (7)

NOTE 1 In some "bizarre" problems the conditions in the theorem are only satisfied with $p_0 = 0$. (See Problem 9.) Note that when $p_0 = 0$ the conditions in the maximum principle do not change at all if f is replaced by any arbitrary function. In fact, when $p_0 = 0$, then (5) takes the form $p(t)g(t, x^*(t), u, p(t)) \leq p(t)g(t, x^*(t), u^*(t), p(t))$ for all u in U .

In the examples and problems to follow we shall assume without proof that $p_0 = 1$. An exception occurs in Example 4 where we show the type of argument needed to prove that $p_0 = 1$. (Almost all papers in the economics literature that use control theory assume that the problem is "normal" in the sense that $p_0 = 1$.)

If $x(t_1)$ is free, then according to (7)(c'), $p(t_1) = 0$. Since $(p_0, p(t_1))$ cannot be $(0, 0)$, we conclude that in this case $p_0 = 1$ and Theorem 9.2.1 is correct as stated.

NOTE 2 If the inequality sign in (3)(b) is reversed, so are the inequality signs in (7)(b').

NOTE 3 The derivative $\dot{p}(t)$ in (6) does not necessarily exist at the discontinuity points of $u^*(t)$, and (6) need hold only wherever $u^*(t)$ is continuous. If U is a convex set and the function H is strictly concave in u , one can show that an optimal control $u^*(t)$ must be continuous.

The conditions in the maximum principle are necessary, but generally not sufficient for optimality. The following theorem gives sufficient conditions (a proof of the result is given in Section 9.7, see Theorem 9.7.1):

THEOREM 9.4.2 (MANGASARIAN)

Suppose that $(x^*(t), u^*(t))$ is an admissible pair with a corresponding adjoint function $p(t)$ such that the conditions (A)–(C) in Theorem 9.4.1 are satisfied with $p_0 = 1$. Suppose further that the control region U is convex and that $H(t, x, u, p(t))$ is concave in (x, u) for every t in $[t_0, t_1]$. Then $(x^*(t), u^*(t))$ is an optimal pair.

When applying Theorems 9.4.1 and 9.4.2, we can often use the following approach:

- For each triple (t, x, p) , maximize $H(t, x, u, p)$ w.r.t. $u \in U$. In many cases, this maximization yields a unique maximum point $u = \hat{u}(t, x, p)$.
- Insert this function into the differential equations (2) and (6) to obtain

$$\dot{x}(t) = g(t, x(t), \hat{u}(t, x(t), p(t))) \quad \text{and} \quad \dot{p}(t) = -H'_x(t, x(t), \hat{u}(t, x(t), p(t)), p(t))$$

This gives two differential equations to determine the functions $x(t)$ and $p(t)$.

- The two constants in the general solution $(x(t), p(t))$ of these differential equations are determined by combining the initial condition $x(t_0) = x_0$ with the terminal conditions and the transversality conditions (7). The state variable obtained in this way is denoted by $x^*(t)$, and the corresponding control variable by $u^*(t) = \hat{u}(t, x^*(t), p(t))$. The pair $(x^*(t), u^*(t))$ is then a candidate for optimality.

This sketch suggests that the maximum principle may contain enough information to give only one or perhaps a few solution candidates.

EXAMPLE 1 Solve the problem

$$\max \int_0^1 x(t) dt, \quad \dot{x}(t) = x(t) + u(t), \quad x(0) = 0, \quad x(1) \text{ free}, \quad u \in [-1, 1]$$

Solution: Looking at the objective function, we see that it pays to have $x(t)$ as large as possible all the time, and from the differential equation it follows that this is obtained by having u as large as possible all the time, i.e. $u(t) = 1$ for all t . So this must be the optimal control. Let us confirm this by using the maximum principle.

The Hamiltonian function with $p_0 = 1$ is $H(t, x, u, p) = x + px + pu$, which is linear and hence concave in (x, u) , so Theorem 9.4.2 applies. The differential equation (6), together with the appropriate version (c') of the transversality condition (7), gives

$$\dot{p} = -1 - p, \quad p(1) = 0$$

This differential equation is especially simple because it is linear with constant coefficients. According to (5.4.3), the general solution is $p(t) = Ae^{-t} - 1$, where A is determined by $0 = p(1) = Ae^{-1} - 1$, which gives $A = e$. Hence, $p(t) = e^{1-t} - 1$, and we see that $p(t) > 0$ for all t in $[0, 1]$. Since the optimal control should maximize $H(t, x^*(t), u, p(t))$, we see from the expression for H that we must have $u^*(t) = 1$ for all t in $[0, 1]$. The corresponding path $x^*(t)$ for the state variable x satisfies the equation $\dot{x}^*(t) = x^*(t) + 1$, with general solution $x^*(t) = Be^t - 1$. Since $x^*(0) = 0$, we obtain $B = 1$, and so

$$x^*(t) = e^t - 1$$

We see now that $u^*(t)$, $x^*(t)$, and $p(t)$ satisfy all the requirements in Theorem 9.4.2. We conclude that we have found the solution to the problem. $\square$

EXAMPLE 2 (Optimal consumption) Consider a consumer who expects to live from the present time, when $t = 0$, until time T . Let $c(t)$ denote his consumption expenditure at time t and $y(t)$ his predicted income. Let $w(t)$ denote his wealth at time t . Then

$$\dot{w}(t) = r(t)w(t) + y(t) - c(t) \quad (*)$$

where $r(t)$ is the instantaneous rate of interest at time t . Suppose the consumer wants to maximize the "lifetime intertemporal utility function"

$$\int_0^T e^{-\alpha t} u(c(t)) dt$$

where $\alpha > 0$, and $u'(c) > 0$, $u''(c) < 0$ for all $c > 0$. The dynamic constraint is $(*)$ above. In addition, $w(0) = w_0$ is given, and there is the terminal constraint $w(T) \geq 0$ preventing the consumer from ending in debt.

This is an optimal control problem with $w(t)$ as the state variable and $c(t)$ as the control variable. We assume that $c(t) > 0$, so that the control region is $(0, \infty)$. We will first characterize the optimal consumption path in general, then find an explicit solution in some instructive special cases.

The Hamiltonian for this problem is $H(t, w, c, p) = e^{-\alpha t} u(c) + p[r(t)w + y - c]$, with $p_0 = 1$. Let $c^* = c^*(t)$ be an optimal solution. Then $H'_c = 0$ at c^* , i.e.

$$e^{-\alpha t} u'(c^*(t)) = p(t) \quad (i)$$

Hence, the adjoint variable is equal to the discounted value of marginal utility. Also,

$$\dot{p}(t) = -H'_w = -p(t)r(t) \quad (ii)$$

so that the adjoint variable decreases at a proportional rate equal to the rate of interest. Notice that (ii) is a separable differential equation whose solution is (see Example 5.3.5)

$$p(t) = p(0) \exp \left[- \int_0^t r(s) ds \right] \quad (iii)$$

Special case 1 (Constant consumption): Suppose that $r(t) = r$, independent of time, and $\alpha = r$. Then (iii) reduces to $p(t) = p(0)e^{-rt}$, and (i) becomes $e^{-rt} u'(c^*(t)) = p(0)e^{-rt}$, or $u'(c^*(t)) = p(0)$. It follows that $c^*(t)$ is a constant, $c^*(t) = \bar{c}$, independent of time. Now $(*)$ becomes $\dot{w} = rw + y(t) - \bar{c}$, whose solution is

$$w^*(t) = e^{rt} \left[w_0 + \int_0^t e^{-rs} y(s) ds - \frac{\bar{c}}{r} (1 - e^{-rt}) \right] \quad (iv)$$

Because of (7)(b'), the terminal constraint $w^*(T) \geq 0$ implies that

$$p(T) \geq 0, \quad \text{with } p(T) = 0 \text{ if } w^*(T) > 0$$

It follows that if $w^*(T) > 0$, then $p(T) = 0$, which contradicts (i). Thus $w^*(T) = 0$, so it is optimal for the consumer to leave no legacy after time T . The condition $w^*(T) = 0$ determines the optimal level of $\bar{c}$, which is²

$$\bar{c} = \frac{r}{1 - e^{-rT}} \left[w_0 + \int_0^T e^{-rs} y(s) ds \right]$$

Special case 2 (Isoelastic utility): Suppose that the utility function u takes the special form

$$u(c) = \frac{(c - \underline{c})^{1-\varepsilon}}{1-\varepsilon} \quad (\varepsilon > 0; \varepsilon \neq 1) \quad \text{or} \quad u(c) = \ln(c - \underline{c}) \quad (v)$$

Then $u'(c) = (c - \underline{c})^{-\varepsilon}$ in both cases, with $\varepsilon = 1$ when $u(c) = \ln(c - \underline{c})$. Note that when $\underline{c} = 0$, the *elasticity of marginal utility* is $\text{El}_c u'(c) = cu''(c)/u'(c) = -\varepsilon$.

When $\underline{c} > 0$, the level $\underline{c}$ of consumption can be regarded as minimum subsistence, below which consumption should never be allowed to fall, if possible. With utility given by (v), equation (i) can be solved explicitly for $c^*(t)$. In fact

$$c^*(t) = \underline{c} + [e^{\alpha t} p(t)]^{-1/\varepsilon} \quad (vi)$$

In order to keep the algebra manageable, we restrict attention once again to the case when $r(t) = r$, independent of time, but now $r \neq \alpha$ is allowed. Still, $p(t) = p(0)e^{-rt}$ and so (vi) implies that

$$c^*(t) = \underline{c} + [e^{(\alpha-r)t} p(0)]^{-1/\varepsilon} = \underline{c} + Ae^{\gamma t}$$

where $A = p(0)^{-1/\varepsilon}$ and $\gamma = (r - \alpha)/\varepsilon$. Then (*) becomes

$$\dot{w} = rw + \gamma(t) - \underline{c} - Ae^{\gamma t}$$

Multiplying this first-order equation by the integrating factor e^{-rt} leads to

$$\frac{d}{dt}(e^{-rt}w) = e^{-rt}(\dot{w} - rw) = e^{-rt}(\gamma(t) - \underline{c} - Ae^{\gamma t})$$

Integrating each side from 0 to t gives

$$e^{-rt}w(t) - w_0 = \int_0^t e^{-rs} \gamma(s) ds - \frac{\underline{c}}{r}(1 - e^{-rt}) - \frac{A}{r - \gamma}[1 - e^{-(r-\gamma)t}]$$

In particular,

$$w(T) = e^{rT}w_0 + \int_0^T e^{r(T-s)} \gamma(s) ds - \frac{\underline{c}}{r}(e^{rT} - 1) - \frac{A}{r - \gamma}(e^{rT} - e^{\gamma T})$$

Again $p(T) > 0$ and thus $w^*(T) = 0$, so the optimal path involves choosing $p(0)$ such that $p(0) = A^{-\varepsilon}$, where

$$A = \frac{r - \gamma}{e^{rT} - e^{\gamma T}} \left[e^{rT}w_0 + \int_0^T e^{r(T-s)} \gamma(s) ds - \frac{\underline{c}}{r}(e^{rT} - 1) \right]$$

² This is the same answer as that derived in Example 2.4.2, equation (iii).

There are two significantly different cases involved here. The first is when $r > \alpha$ and so $\gamma > 0$. Then consumption grows over time starting from the level $\underline{c} + A$. But if $r < \alpha$ and so $\gamma < 0$, then optimal consumption shrinks over time. This makes sense because $r < \alpha$ is the case when the agent discounts future utility at a rate α that exceeds the rate of interest.

The previous case with constant consumption is when $\gamma = 0$. The same solution emerges in the limit as $\varepsilon \rightarrow \infty$, which represents the case when the consumer is extremely averse to fluctuations in consumption.

It is also worth examining what happens to the solution as the horizon T recedes to infinity. The normal case is when $r > \gamma$, which is always true when $\varepsilon \geq 1$; it is also true when $0 < \varepsilon < 1$ and $\alpha > (1 - \varepsilon)r$. In this normal case, we see that

$$A = \frac{r - \gamma}{1 - e^{-(r-\gamma)T}} \left[w_0 + \int_0^T e^{-rt} \gamma(t) dt - \frac{\underline{c}}{r}(1 - e^{-rT}) \right] \rightarrow (r - \gamma) \left[w_0 + \int_0^\infty e^{-rt} \gamma(t) dt - \frac{\underline{c}}{r} \right]$$

as $T \rightarrow \infty$. Using the methods introduced in Section 9.11, this limit can be shown to give the solution to the infinite horizon problem.

In the abnormal case when $r \leq \gamma$, the limiting value of A is zero. Then the limiting value of $c^*(t)$ is the minimum allowable level $\underline{c}$ for all t . This is the worst admissible consumption path; in fact, there is no infinite horizon optimum in this case. $\blacksquare$

EXAMPLE 3 Solve the following problem (where the optimal control is bang-bang):

$$\max \int_0^1 (2x - x^2) dt, \quad \dot{x} = u, \quad x(0) = 0, \quad x(1) = 0, \quad u \in [-1, 1]$$

Solution: The Hamiltonian is $H = 2x - x^2 + pu$, so an optimal control $u^*(t)$ must maximize $2x^*(t) - (x^*(t))^2 + p(t)u$ subject to $u \in [-1, 1]$. Only the term $p(t)u$ depends on u , so

$$u^*(t) = \begin{cases} 1 & \text{if } p(t) > 0 \\ -1 & \text{if } p(t) < 0 \end{cases} \quad (*)$$

The differential equation for $p(t)$ is

$$\dot{p}(t) = -H_x'(t, x^*(t), u^*(t), p(t)) = 2x^*(t) - 2 = 2(x^*(t) - 1) \quad (**)$$

Note that $\dot{x}^*(t) = u^*(t) \leq 1$. Because $x^*(0) = 0$, it follows that $x^*(t) < 1$ for all t in $[0, 1]$. Then (**) implies that $p(t)$ is strictly decreasing in $[0, 1]$.

Suppose there could be a solution with $p(1) \geq 0$. Because $p(t)$ is strictly decreasing in $[0, 1]$, one would have $p(t) > 0$ in $[0, 1]$, and then (*) would imply that $u^*(t) = 1$ for all t . In this case, $\dot{x}^*(t) = 1$ for all t in $[0, 1]$. With $x^*(0) = 0$ we get $x^*(t) = t$ and thus $x^*(1) = 1$, which is incompatible with the terminal condition $x^*(1) = 0$.

Thus, $p(t)$ must satisfy $p(1) < 0$. Suppose $p(t) < 0$ for all t in $(0, 1]$. Then from (*), $u^*(t) = -1$ for all such t , so $x^*(t) = -t$ with $x^*(1) = -1$, again violating the terminal condition. Hence, for some t^* in $(0, 1)$, the function $p(t)$ switches from being positive (or possibly zero) to being negative. A possible path for $p(t)$ is shown in Fig. 1.

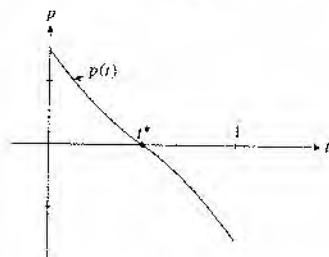


Figure 1

Recalling our convention that u^* should be left-continuous, we have $u^*(t) = 1$ in $[0, t^*]$ and $u^*(t) = -1$ in $(t^*, 1]$. On $[0, t^*)$, therefore, $\dot{x}^*(t) = 1$, and with $x^*(0) = 0$ this yields $x^*(t) = t$. Since $x^*(t)$ is required to be continuous at t^* , we have $x^*(t^*) = x^*(t^{*-}) = t^*$. In $(t^*, 1]$, we have $\dot{x}^*(t) = -1$, so $x^*(t) = -t + C$ for some constant C . Because $x^*(t)$ is continuous at t^* , $x^*(t^{*+}) = x^*(t^*) = t^*$, so $C = 2t^*$. Hence, $x^*(t) = -t + 2t^*$. Then $x^*(1) = 0$ implies that $t^* = 1/2$. We conclude that the optimal solution is

$$u^*(t) = \begin{cases} 1 & \text{in } [0, 1/2] \\ -1 & \text{in } (1/2, 1] \end{cases} \quad x^*(t) = \begin{cases} t & \text{in } [0, 1/2] \\ 1-t & \text{in } (1/2, 1] \end{cases}$$

To find $p(t)$, note that $\dot{p}(t) = 2x^*(t) - 2 = 2t - 2$ in $[0, 1/2]$. Because $p(1/2) = 0$, we have $p(t) = t^2 - 2t + 3/4$. In the interval $(1/2, 1]$, $(**)$ implies that $\dot{p}(t) = -2t$ and, because $p(t)$ is continuous with $p(1/2) = 0$, the adjoint function is $p(t) = -t^2 + 1/4$. For this function $p(t)$ the maximum condition $(*)$ is satisfied. Since the Hamiltonian is concave in (x, u) , we have found the optimal solution. $\square$

The last example shows a typical kind of argument needed to prove that $p_0 \neq 0$.

EXAMPLE 3

Consider Example 3 again. Including the multiplier p_0 , the Hamiltonian function (4) is $H = p_0(2x - x^2) + pu$, and the differential equation (6) for p is $\dot{p} = -H_x = -p_0(2 - 2x^*(t))$. Suppose $p_0 = 0$. Then $\dot{p} = 0$ and so p is a constant, $\bar{p}$. Because $(p_0, p(t)) = (p_0, \bar{p}) \neq (0, 0)$, that constant $\bar{p}$ is not 0. Now, an optimal control must maximize $pu = \bar{p}u$ subject to $u \in [-1, 1]$. If $\bar{p} > 0$, then obviously $u^*(t) = 1$ for all t in $[0, 1]$. This means that $\dot{x}^*(t) = 1$, with $x^*(0) = 0$, so $x^*(t) = t$. This violates the terminal condition $x^*(1) = 0$. If $\bar{p} < 0$, then obviously $u^*(t) = -1$, and $\dot{x}^*(t) = -1$ for all t in $[0, 1]$, with $x^*(0) = 0$. Therefore $x^*(t) = -t$, which again violates the terminal condition. We conclude that $p_0 = 0$ is impossible, so $p_0 = 1$. $\square$

PROBLEMS FOR SECTION 9.4

1. What is the obvious solution to the problem

$$\max \int_0^T x(t) dt, \quad \dot{x}(t) = u(t), \quad x(0) = 0, \quad x(T) \text{ free}, \quad u(t) \in [0, 1]$$

where T is a fixed positive constant? Compute the associated value, $V(T)$, of the objective function. Find the solution also by using Theorem 9.4.2.

2. Solve the problem: $\max \int_0^1 (1 - x^2 - u^2) dt$, $\dot{x} = u$, $x(0) = 0$, $x(1) \geq 1$, $u \in \mathbb{R}$.

3. Consider the problem in Example 9.2.1.

- (a) Replace $u \in \mathbb{R}$ by $u \in [0, 1]$ and find the optimal solution.
(b) Replace $u \in \mathbb{R}$ by $u \in [-1, 1]$ and find the optimal solution, provided $T > 2$.

4. Solve the following problems and compute the corresponding values V of the objective function.

(a) $\max_{u \in [0, 1]} \int_0^{10} x dt$, $\dot{x} = u$, $x(0) = 0$, $x(10) = 2$

(b) $\max_{u \in [0, 1]} \int_0^T x dt$, $\dot{x} = u$, $x(0) = x_0$, $x(T) = x_1$ (with $x_0 < x_1 < x_0 + T$)

5. (a) Given the fixed positive number T , write down the conditions in Theorem 9.4.1 for the problem

$$\max \int_0^T -(u^2 + x^2) dt, \quad \dot{x} = au, \quad x(0) = 1, \quad x(T) \text{ free}, \quad u(t) \in [0, 1]$$

and find the solution when $a \geq 0$.

- (b) Find the solution if $a < 0$. (Hint: Try $u^*(t) \in (0, 1)$ for all t .)

6. Solve the following special case of Problem 1 in Example 9.1.2:

$$\max \int_0^5 [10u - (u^2 + 2)]e^{-0.1t} dt, \quad \dot{x} = -u, \quad x(0) = 10, \quad x(5) \geq 0, \quad u \geq 0$$

7. (From Kamien and Schwartz (1991).) A firm has an order of B units of a commodity to be delivered at time T . Let $x(t)$ be the stock at time t . We assume that the cost per unit of time of storing $x(t)$ units is $ax(t)$. The increase in $x(t)$, which equals production per unit of time, is $u(t) = \dot{x}(t)$. Assume that the total cost of production per unit of time is equal to $b(u(t))^2$. Here a and b are positive constants. So the firm's cost minimization problem is

$$\min \int_0^T [ax(t) + bu(t)^2] dt, \quad \dot{x}(t) = u(t), \quad x(0) = 0, \quad x(T) = B, \quad u(t) \geq 0$$

- (a) Write down the necessary conditions implied by Theorem 9.4.1.
(b) Find the only possible solution to the problem and explain why it really is a solution. (Hint: Distinguish between the cases $B \geq aT^2/4b$ and $B < aT^2/4b$.)

8. Find the only possible solution to the problem

$$\max \int_0^2 (x^2 - 2u) dt, \quad \dot{x} = u, \quad x(0) = 1, \quad x(2) \text{ free}, \quad u \in [0, 1]$$

(Hint: Show that $p(t)$ is strictly decreasing.)

9. Consider the problem $\max \int_0^1 -u dt$, $\dot{x} = u^2$, $x(0) = x(1) = 0$, $u \in \mathbb{R}$.

- (a) Explain why $u^*(t) = x^*(t) = 0$ solves the problem.
(b) Show that the conditions in the maximum principle are satisfied only for $p_0 = 0$.

9.5 The Maximum Principle and the Calculus of Variations

The introduction to this chapter claimed that optimal control theory extends the classical calculus of variations. Consider what the maximum principle has to say about the standard variational problem

$$\max \int_{t_0}^{t_1} F(t, x(t), \dot{x}(t)) dt, \quad x(t_0) = x_0, \quad \begin{cases} \text{(a) } x(t_1) = x_1 \\ \text{(b) } x(t_1) \geq x_1 \\ \text{(c) } x(t_1) \text{ free} \end{cases} \quad (1)$$

where one of the alternative conditions (a), (b), and (c) is imposed. To transform this to a control problem, simply use $u(t) = \dot{x}(t)$ as a control variable. Because there are no restrictions on $\dot{x}(t)$ in the variational problem, nor are there any on the control function $u(t)$. Hence, $U = \mathbb{R}$.

The control problem has the particularly simple differential equation $\dot{x}(t) = u(t)$. The Hamiltonian is $H(t, x, u, p) = p_0 F(t, x, u) + pu$. The maximum principle states that if $u^*(t)$ solves the problem, then H as a function of u must be maximized at $u = u^*(t)$. Because $U = \mathbb{R}$, a necessary condition for this maximum is

$$H'_u(t, x^*(t), u^*(t), p(t)) = p_0 F'_u(t, x^*(t), u^*(t)) + p(t) = 0 \quad (*)$$

Since $(p_0, p(t)) \neq (0, 0)$, equation (*) implies that $p_0 \neq 0$, so $p_0 = 1$. The differential equation for $p(t)$ is

$$\dot{p}(t) = -H'_x(t, x^*(t), u^*(t), p(t)) = -F'_x(t, x^*(t), u^*(t)) \quad (**)$$

Differentiating (*) with respect to t yields

$$\frac{d}{dt} (F'_u(t, x^*(t), u^*(t))) + \dot{p}(t) = 0 \quad (***)$$

Since $u^* = \dot{x}^*$, it follows from (**) and (***) that

$$F'_x(t, x^*, \dot{x}^*) - \frac{d}{dt} (F'_x(t, x^*, \dot{x}^*)) = 0 \quad (2)$$

which is the Euler equation. Moreover, (*) implies that

$$p(t) = -F'_x(t, x^*, \dot{x}^*) \quad (3)$$

Using (3) it is easy to check that the transversality conditions in (9.4.7) are precisely those set out in Section 8.5. Note also that concavity of the Hamiltonian with respect to (x, u) is equivalent to concavity of $F(t, x, \dot{x})$ with respect to $(x, \dot{x})$.

Thus the maximum principle confirms all the main results found in Chapter 8. Actually, it contains more information about the solution of the optimization problem. For instance, according to the maximum principle, for every t in $[t_0, t_1]$ the Hamiltonian attains its maximum at $u^*(t)$. Assuming that F is a C^2 function, not only is $H'_u = 0$, but also $H''_{uu} \leq 0$, implying that $F''_{\dot{x}\dot{x}} \leq 0$. This is the so-called **Legendre condition** in the calculus of variations. (Also, continuity of $p(t)$ and (3) together give the **Weierstrass-Erdmann corner**

condition, requiring F'_x to be continuous as a function of t . This is a well-known result in the classical theory.)

PROBLEMS FOR SECTION 9.5

1. Find the only possible solution to the following problem by using both the calculus of variations and control theory:

$$\max \int_0^1 (2xe^{-t} - 2x\dot{x} - \dot{x}^2) dt, \quad x(0) = 0, \quad x(1) = 1$$

2. Solve the following problem by using both the calculus of variations and control theory:

$$\max \int_0^2 (3 - x^2 - 2\dot{x}^2) dt, \quad x(0) = 1, \quad x(2) \geq 4$$

3. Solve the following problem by using both the calculus of variations and control theory:

$$\max \int_0^1 (-2\dot{x} - \dot{x}^2)e^{-t/10} dt, \quad x(0) = 1, \quad x(1) = 0$$

4. At time $t = 0$ an oil field is known to contain $\bar{x}$ barrels of oil. It is desired to extract all of the oil during a given time interval $[0, T]$. If $x(t)$ is the amount of oil left at time t , then $-\dot{x}$ is the extraction rate (which is ≥ 0 when $x(t)$ is decreasing). Assume that the world market price per barrel of oil is given and equal to ae^{at} . The extraction costs per unit of time are assumed to be $\dot{x}(t)^2 e^{\beta t}$. The profit per unit of time is then $\pi = -\dot{x}(t)ae^{at} - \dot{x}(t)^2 e^{\beta t}$. Here a , α , and β are constants, $a > 0$. This leads to the variational problem

$$\max \int_0^T [-\dot{x}(t)ae^{at} - \dot{x}(t)^2 e^{\beta t}] e^{-rt} dt, \quad x(0) = \bar{x}, \quad x(T) = 0, \quad (*)$$

where r is a positive constant. Find the Euler equation for problem (*), and show that at the optimum $3\pi/\partial \dot{x} = ce^{rt}$ for some constant c . Derive the same result by using control theory.

5. S. Ström considers the problem

$$\max_x \int_0^T \{U(x(t)) - b(x(t)) - gz(t)\} dt, \quad \dot{z}(t) = ax(t), \quad z(0) = z_0, \quad z(T) \text{ free}$$

Here $U(x)$ is the utility enjoyed by society consuming x , whereas $b(x)$ is total cost and $z(t)$ is the stock of pollution at time t . Assume that U and b satisfy $U' > 0$, $U'' < 0$, $b' > 0$, and $b'' > 0$. The control variable is $x(t)$, whereas $z(t)$ is the state variable. The constants a and g are positive.

- (a) Write down the conditions of the maximum principle. Show that the adjoint function is given by $p(t) = g(T - t)$, $t \in [0, T]$, and prove that if $x^*(t) > 0$ solves the problem, then

$$U'(x^*(t)) = b'(x^*(t)) + ag(T - t) \quad (*)$$

- (b) Prove that a solution of (*) with $x^*(t) > 0$ must solve the problem. Show that $x^*(t)$ is strictly increasing. (Hint: Differentiate (*) with respect to t .)

9.6 Adjoint Variables as Shadow Prices

Like the Lagrange multipliers used to solve static constrained optimization problems in Chapter 3, the adjoint function $p(t)$ in the maximum principle can be given an interesting price interpretation.

Consider the standard end-constrained problem (9.4.1)–(9.4.3). Suppose that it has a unique optimal solution $(x^*(t), u^*(t))$ with a unique corresponding adjoint function $p(t)$. The corresponding value of the objective function will depend on x_0, x_1, t_0 , and t_1 , so it is denoted by

$$V(x_0, x_1, t_0, t_1) = \int_{t_0}^{t_1} f(t, x^*(t), u^*(t)) dt \quad (1)$$

We call V the (optimal) **value function**. (When $x(t_1)$ is free, x_1 is not an argument of V .)

Suppose x_0 is changed slightly. In general, both $u^*(t)$ and $x^*(t)$ will change over the whole interval $[t_0, t_1]$. For typical problems in control theory, there is no guarantee that V is differentiable at a particular point. But at any point where it is differentiable,

$$\frac{\partial V(x_0, x_1, t_0, t_1)}{\partial x_0} = p(t_0) \quad (2)$$

The number $p(t_0)$ therefore measures the marginal change in the optimal value function as x_0 increases.

EXAMPLE 1 In Example 9.2.1 the objective function was $\int_0^T [1 - tx(t) - u(t)^2] dt$, and the solution was $u^*(t) = -\frac{1}{4}(T^2 - t^2)$, $x^*(t) = x_0 - \frac{1}{4}T^2t + \frac{1}{12}t^3$, with $p(t) = -\frac{1}{2}(T^2 - t^2)$. So the value function is

$$V(x_0, T) = \int_0^T [1 - tx^*(t) - (u^*(t))^2] dt = \int_0^T [1 - x_0t + \frac{1}{4}T^2t^2 - \frac{1}{12}t^4 - \frac{1}{16}(T^2 - t^2)^2] dt$$

This last integral could be evaluated exactly, but fortunately we do not need to. Instead, simply differentiating V w.r.t. x_0 under the integral sign using formula (4.2.1) gives

$$\frac{\partial V(x_0, T)}{\partial x_0} = \int_0^T (-t) dt = -\frac{1}{2}T^2$$

On the other hand, $p(0) = -\frac{1}{2}T^2$, so (2) is confirmed. 1

Formula (2) interprets $p(t)$ at time $t = t_0$. What about $p(t)$ at an arbitrary $t \in (t_0, t_1)$? We want an interpretation that relates to the value function for the problem defined over the whole interval $[t_0, t_1]$, not only the subinterval $[t, t_1]$. Consider again problem (9.4.1)–(9.4.3), but assume that all admissible paths $x(t)$ of the state variable are forced to have a jump equal to v at $t \in (t_0, t_1)$, so that $x(t^+) - x(t^-) = v$. Suppose all admissible $x(t)$ are continuous elsewhere. The optimal value function V for this problem will depend on v . Suppose that $(x^*(t), u^*(t))$ is the optimal solution of the problem for $v = 0$. Then, under

certain conditions, it can be shown that V as a function of v is defined in a neighbourhood of $v = 0$, that V is differentiable w.r.t. v at $v = 0$, and that

$$\left(\frac{\partial V}{\partial v} \right)_{v=0} = p(t) \quad (3)$$

The adjoint variable $p(t)$ is the first-order approximate change in the value function (1) due to an enforced unit jump increase in $x(t)$.³

A General Economic Interpretation

Consider a firm that seeks to maximize its profit over a planning period $[t_0, t_1]$. The state of the firm at time t is described by its capital stock $x(t)$. At each time t the firm can influence its immediate profit, as well as the change in its future capital stock. Suppose the firm can choose its control variable $u(t)$ within certain limits, so that $u(t) \in U = [u_0, u_1]$. Let the profit flow at time t be $f(t, x(t), u(t))$ per unit of time, so that the total profit in the time period $[t_0, t_1]$ is

$$\int_{t_0}^{t_1} f(t, x(t), u(t)) dt$$

The rate of change in the capital stock depends on the present capital stock as well as on the value chosen for $u(t)$ at time t . Thus,

$$\dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0$$

where x_0 is the given capital stock at time $t = t_0$. The control variable $u(t)$ not only influences the immediate profit but also, via the differential equation, influences the rate of change of the capital stock and thereby the future capital stock, which again changes the total profit.

Suppose we have found the optimal solution to this problem, with corresponding adjoint function $p(t)$. According to (3), $p(t)$ is a “shadow price” of the capital stock variable, since $p(t)$ measures the marginal profit of capital. The Hamiltonian is $H = f(t, x, u) + p(t)g(t, x, u)$. Consider a small time interval $[t, t + \Delta t]$. Over this time interval, $\Delta x \approx g(t, x, u) \Delta t$ and so

$$H \Delta t = f(t, x, u) \Delta t + p(t)g(t, x, u) \Delta t \approx f(t, x, u) \Delta t + p(t) \Delta x$$

Hence, $H \Delta t$ is the sum of the instantaneous profit $f(t, x, u) \Delta t$ earned in the time interval $[t, t + \Delta t]$ and the contribution $p(t) \Delta x$ to the total profit produced by the extra capital Δx at the end of this time period. The maximum principle requires choosing at each time the value of u that maximizes H , and hence $H \Delta t$.

³ Economists have realized for a long time that the adjoint can be interpreted as a shadow price. Dorfman (1969) has an illuminating discussion on the economic interpretations, extending the material in the next subsection. For precise results and references, see e.g. Seierstad and Sydsæter (1987).

Other Sensitivity Results

Consider once again the standard end-constrained problem (9.4.1)–(9.4.3) and its optimal value function (1). It turns out that, provided V is differentiable, the effects on V of small changes in x_1 , t_0 , and t_1 can also be expressed very simply. Define

$$H^*(t) = H(t, x^*(t), u^*(t), p(t)) \quad (4)$$

Then

$$\frac{\partial V}{\partial x_0} = p(t_0), \quad \frac{\partial V}{\partial x_1} = -p(t_1), \quad \frac{\partial V}{\partial t_0} = -H^*(t_0), \quad \frac{\partial V}{\partial t_1} = H^*(t_1) \quad (5)$$

The first of these equations was discussed above. As for the second, it is like the first, except that requiring the state x_1 to be larger at time t_1 has an effect that is the opposite of allowing x_0 to be larger at time t_0 . For example, in the capital accumulation interpretation in the previous subsection, increasing the initial capital stock x_0 by one unit increases the total profit by approximately $p(t_0)$. On the other hand, increasing the capital which must be left at the end of the planning period t_1 decreases the total profit earned by approximately $p(t_1)$. The third equality is similar to the fourth except for the change of sign. In the capital accumulation interpretation, increasing t_1 makes the planning period longer and the total profit increases (if the instantaneous profit is positive). On the other hand, increasing t_0 makes the planning period shorter, so the total profit decreases. The last equality is illustrated in the next example.

NOTE 1 Consider the standard end-constrained problem with $x(t_1)$ free. If $(x^*(t), u^*(t))$ is an optimal pair with corresponding adjoint function $p(t)$, then according to condition (9.4.7)(c'), $p(t_1) = 0$. This makes sense because of the second formula in (5) and the economic interpretation above: if there is no reason to care about the capital stock at the end of the planning period, its shadow price should be equal to 0.

EXAMPLE 2 Verify the last equality in (5) for the problem in Example 1.

Solution: Differentiating the value function $V(x_0, T)$ from Example 1 w.r.t. T , using the Leibniz rule (4.2.3) yields

$$\frac{\partial V}{\partial T} = 1 - x_0 T + \frac{1}{4} T^4 - \frac{1}{12} T^4 + \int_0^T \left[\frac{1}{2} t^2 T - \frac{1}{8} (T^2 - t^2) 2T \right] dt$$

Integrating and simplifying gives

$$\frac{\partial V}{\partial T} = 1 - x_0 T + \frac{1}{6} T^4$$

Now, $H^*(T) = 1 - T x^*(T) - (u^*(T))^2 + p(T) u^*(T) = 1 - x_0 T + \frac{1}{6} T^4$, because $u^*(T) = 0$ and $x^*(T) = x_0 - \frac{1}{6} T^3$. Thus the last result in (5) is confirmed. $\square$

EXAMPLE 3 (Economic growth) Consider the following problem in economic growth theory due to Shell (1967):

$$\max \int_0^T (1 - s(t)) e^{\rho t} f(k(t)) e^{-\delta t} dt$$

$$\text{s.t. } \dot{k}(t) = s(t) e^{\rho t} f(k(t)) - \lambda k(t), \quad k(0) = k_0, \quad k(T) \geq k_T > k_0, \quad 0 \leq s(t) \leq 1$$

Here $k(t)$ is the capital stock (a state variable), $s(t)$ is the savings rate (a control variable), and $f(k)$ is a production function. Suppose that $f(k) > 0$ whenever $k \geq k_0 e^{-\lambda T}$, that $f'(k) > 0$, and that $\rho, \delta, \lambda, T, k_0$, and k_T are all positive constants.

- Suppose $(k^*(t), s^*(t))$ solves the problem. Write down the conditions in the maximum principle in this case. What are the possible values of $s^*(t)$?
- Put $\rho = 0$, $f(k) = ak$, $a > 0$, $\delta = 0$ and $\lambda = 0$. Suppose that $T > 1/a$ and that $k_0 e^{aT} > k_T$. Find the only possible solution to the problem, distinguishing between two different cases.
- Compute the value function in case (b) and then verify the relevant equalities in (5).

Solution: (a) The Hamiltonian is $H = (1 - s) e^{\rho t} f(k) e^{-\delta t} + p(s e^{\rho t} f(k) - \lambda k)$. If $(k^*(t), s^*(t))$ solves the problem, then in particular, $s^*(t)$ must solve

$$\max_s (1 - s) e^{\rho t} f(k^*(t)) e^{-\delta t} + p(t) [s e^{\rho t} f(k^*(t)) - \lambda k^*(t)] \quad \text{subject to } s \in [0, 1]$$

Disregarding the terms that do not depend on s , $s^*(t)$ must maximize the expression $e^{\rho t} f(k^*(t)) (-e^{-\delta t} + p(t)) s$ for $s \in [0, 1]$. Hence, we must choose

$$s^*(t) = \begin{cases} 1 & \text{if } p(t) > e^{-\delta t} \\ 0 & \text{if } p(t) < e^{-\delta t} \end{cases} \quad (\text{i})$$

A possible optimal control can therefore only take the values 1 and 0 (except if $p(t) = e^{-\delta t}$). Except where $s^*(t)$ is discontinuous,

$$\dot{p}(t) = -(1 - s^*(t)) e^{\rho t} f'(k^*(t)) e^{-\delta t} - p(t) s^*(t) e^{\rho t} f'(k^*(t)) + \lambda p(t) \quad (\text{ii})$$

The transversality condition (9.4.7)(b') gives

$$p(T) \geq 0 \quad \text{with} \quad p(T) = 0 \quad \text{if} \quad k^*(T) > k_T \quad (\text{iii})$$

For a more extensive discussion of the model, see Shell (1967).

(b) Briefly formulated, the problem reduces to

$$\max \int_0^T (1 - s) a k dt, \quad \dot{k} = a s k, \quad k(0) = k_0, \quad k(T) \geq k_T > k_0$$

with $s \in [0, 1]$, $a > 0$, $T > 1/a$, and $k_0 e^{aT} > k_T$.

The Hamiltonian is $H = (1 - s) a k + p a s k$. The differential equation (ii) is now

$$\dot{p}(t) = -a + s^*(t) a (1 - p(t)) \quad (\text{iv})$$

whereas (i) implies that

$$s^*(t) = \begin{cases} 1 & \text{if } p(t) > 1 \\ 0 & \text{if } p(t) < 1 \end{cases} \quad (\text{v})$$

From (iv) and (v) it follows that

$$\dot{p}(t) = -a < 0 \text{ if } p(t) < 1, \text{ whereas } \dot{p}(t) = -ap(t) \text{ if } p(t) > 1 \quad (\text{vi})$$

In both cases $\dot{p}(t) < 0$, so $p(t)$ is strictly decreasing.

Suppose $p(0) \leq 1$, which implies that $p(t) < 1$ throughout $(0, T]$. Then by (v), $s^*(t) \equiv 0$, and so $k^*(t) \equiv k_0$, which contradicts $k^*(T) \geq k_T > k_0$. Hence, $p(0) > 1$. Then there are two possible paths for $p(t)$, which are shown in Fig. 1. In the first case $p(T) = 0$; in the second case, $p(T) > 0$.

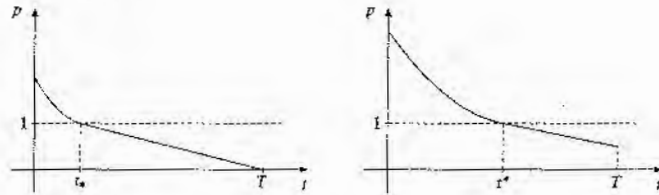


Figure 1 Two possible paths for $p(t)$.

Case I: $p(T) = 0$. Since $p(t)$ is continuous and strictly decreasing with $p(0) > 1$ and $p(T) = 0$, there is a unique t_* in $(0, T)$ such that $p(t_*) = 1$, with $p(t) > 1$ in $[0, t_*)$ and $p(t) < 1$ in $(t_*, T]$. Then $s^*(t) = 1$ in $[0, t_*)$ and $s^*(t) = 0$ in $(t_*, T]$. By (vi), $\dot{p}(t) = -ap(t)$ in $[0, t_*)$ and $\dot{p}(t) = -a$ in $(t_*, T]$. On $(t_*, T]$, we have $p(t) = -a(t - T)$, because we have assumed $p(T) = 0$. But $p(t_*) = 1$, so $1 = -a(t_* - T)$, implying that $t_* = T - 1/a$. Furthermore, $p(t) = e^{a(T-t)-1}$ on $[0, T - 1/a]$. This gives the following solution candidate:

$$\text{For } t \in [0, T - 1/a], \quad s^*(t) = 1, \quad k^*(t) = k_0 e^{a(T-t)-1}, \quad \text{and } p(t) = e^{a(T-t)-1} \quad (\text{vii})$$

$$\text{For } t \in (T - 1/a, T], \quad s^*(t) = 0, \quad k^*(t) = k_0 e^{a(T-t)-1}, \quad \text{and } p(t) = -a(t - T) \quad (\text{viii})$$

It remains to check that $k^*(T) \geq k_T$. This reduces to $k_0 e^{a(T-1/a)-1} \geq k_T$, i.e. $k_0 e^{aT-1} \geq k_T$. The latter inequality holds if and only if

$$t_* = T - \frac{1}{a} \geq \frac{1}{a} \ln \left(\frac{k_T}{k_0} \right) \quad (\text{ix})$$

Case II: $p(T) > 0$. In this case, by (iii), $k^*(T) = k_T$. If it were true that $p(T) \geq 1$, then one would have $p(t) > 1$ and so $s^*(t) = 1$ for all t in $[0, T]$, implying that $k^*(T) = k_0 e^{aT} > k_T$, a contradiction. So there exists a unique t^* in $[0, T)$ such that $p(t^*) = 1$. Similar arguments to those for case I suggest the following as an optimal solution:

$$\text{For } t \in [0, t^*], \quad s^*(t) = 1, \quad k^*(t) = k_0 e^{at}, \quad \text{and } p(t) = e^{a(t-t^*)} \quad (\text{x})$$

$$\text{For } t \in (t^*, T], \quad s^*(t) = 0, \quad k^*(t) = k_0 e^{at^*}, \quad \text{and } p(t) = 1 - a(t - t^*) \quad (\text{xi})$$

From $k^*(T) = k_T$ it follows that $e^{at^*} = k_T/k_0$, so

$$t^* = \frac{1}{a} \ln \left(\frac{k_T}{k_0} \right) \quad (\text{xii})$$

We note that $t^* < T$ is equivalent to $k_0 e^{aT} > k_T$, as assumed. All of this was derived under the assumption that $p(T) > 0$, i.e. $1 - a(T - t^*) > 0$, which gives

$$T - \frac{1}{a} < \frac{1}{a} \ln \left(\frac{k_T}{k_0} \right) = t^* \quad (\text{xiii})$$

Putting the two cases together, there is only one solution candidate, with

$$s^*(t) = 1 \text{ if } t \in [0, \bar{t}], \quad s^*(t) = 0 \text{ if } t \in (\bar{t}, T] \quad (\text{xiv})$$

where $\bar{t} = \max \{ T - 1/a, (1/a) \ln(k_T/k_0) \}$.

In Example 9.7.3 we shall prove that this is the optimum.

(c) For case I in (b) we have

$$V(k_0, k_T, T) = \int_{T-1/a}^T a k_0 e^{a(T-t)-1} dt = a k_0 e^{a(T-1/a)-1} [T - (T - 1/a)] = k_0 e^{aT-1}$$

so $\partial V / \partial k_0 = e^{aT-1} = p(0)$, using (vii). Also $\partial V / \partial k_T = 0 = -p(T)$. Finally, $H^*(T) = (1 - s^*(T)) a k^*(T) + p(T) a s^*(T) k^*(T) = a k^*(T) = a k_0 e^{aT-1} = \partial V / \partial T$.

For case II,

$$V(k_0, k_T, T) = \int_{t^*}^T a k_0 e^{at} dt = a k_0 e^{at^*} (T - t^*) = a k_T (T - \frac{1}{a} \ln k_T + \frac{1}{a} \ln k_0)$$

Hence $\partial V / \partial k_0 = k_T / k_0$, and we see that $p(0) = e^{at^*} = k_T / k_0$ also. Moreover, $\partial V / \partial k_T = a(T - \frac{1}{a} \ln k_T + \frac{1}{a} \ln k_0) - 1 = a(T - t^*) - 1$, and $-p(T) = a(T - t^*) - 1$ also. Finally, $\partial V / \partial T = a k_T$ and $H^*(T) = a k^*(T) = a k_0 e^{at^*} = a k_0 (k_T / k_0) = a k_T$. $\square$

PROBLEMS FOR SECTION 9.6

SM 1. (a) Solve the control problem

$$\max \int_0^T (x - \frac{1}{2} u^2) dt, \quad \dot{x} = u, \quad x(0) = x_0, \quad x(T) \text{ free}, \quad u(t) \in \mathbb{R}$$

(b) Compute the optimal value function $V(x_0, T)$, and verify the relevant equalities in (5).

2. Verify that $V'(T) = H^*(T)$ for Problem 9.4.1.

3. Verify (5) for Problem 9.4.4(b).

HARDER PROBLEMS

4. (a) Given the positive constant T , find the only possible solution to the problem:

$$\max \int_0^T (2x^2 e^{-2t} - ue^t) dt, \quad \dot{x} = ue^t, \quad x(0) = 1, \quad x(T) \text{ free}, \quad u \in [0, 1]$$

- (b) Compute the value function $V(T)$ and verify that $V'(T) = H^*(T)$.

5. Consider the problem $\max \int_0^1 ux \, dt$, $\dot{x} = 0$, $x(0) = x_0$, $x(1)$ free, $u \in [0, 1]$.

- (a) Prove that if $x_0 < 0$, then the optimal control is $u^* = 0$, and if $x_0 > 0$, then the optimal control is $u^* = 1$.
 (b) Show that the value function $V(x_0)$ is not differentiable at $x_0 = 0$.

9.7 Sufficient Conditions

The maximum principle provides necessary conditions for optimality. Only solution candidates fulfilling these necessary conditions can possibly solve the problem. However, the maximum principle by itself cannot tell us whether a given candidate is optimal or not, nor does it tell us whether or not an optimal solution exists.

The following result, originally due to Mangasarian (1966), has been referred to before (Theorem 9.4.2). In fact, it is quite easy to prove.

THEOREM 9.7.1 (MANGASARIAN)

Consider the standard end-constrained problem (9.4.1)–(9.4.3) with U an interval of the real line. Suppose the admissible pair $(x^*(t), u^*(t))$ satisfies all the conditions (9.4.5)–(9.4.7) of the maximum principle, with the associated adjoint function $p(t)$, and with $p_0 = 1$. Then, if

$$H(t, x, u, p(t)) \text{ is concave w.r.t. } (x, u) \text{ for all } t \text{ in } [t_0, t_1] \quad (1)$$

the pair $(x^*(t), u^*(t))$ solves the problem.

If $H(t, x, u, p(t))$ is strictly concave w.r.t. (x, u) , then the pair $(x^*(t), u^*(t))$ is the unique solution to the problem.

NOTE 1 Suppose that U is an open interval (u_0, u_1) . Then the concavity of $H(t, x, u, p(t))$ in u implies that the maximization condition (9.4.5) is equivalent to the first-order condition $\partial H^*/\partial u = \partial H(t, x^*(t), u^*(t), p(t))/\partial u = 0$. (See Theorem 3.1.2.) The concavity of $H(t, x, u, p(t))$ in (x, u) is satisfied, for example, if f and pg are concave in (x, u) , or if f is concave and g is linear in (x, u) .

Suppose that $U = [u_0, u_1]$. If $u^*(t) \in (u_0, u_1)$, then $\partial H^*/\partial u = 0$. If the lower limit $u^*(t) = u_0$ maximizes the Hamiltonian, then $\partial H^*/\partial u \leq 0$, because otherwise if $\partial H^*/\partial u > 0$, then the Hamiltonian would attain greater values just to the right of u_0 . If the upper limit $u^*(t) = u_1$ maximizes the Hamiltonian, then we see in a similar way that $\partial H^*/\partial u \geq 0$. Because the Hamiltonian is concave in u , it follows that if $U = [u_0, u_1]$, then the maximum condition (9.4.5) is equivalent to the conditions:

$$\frac{\partial H^*}{\partial u} \begin{cases} \leq 0 & \text{if } u^*(t) = u_0 \\ = 0 & \text{if } u^*(t) \in (u_0, u_1) \\ \geq 0 & \text{if } u^*(t) = u_1 \end{cases} \quad (2)$$

These conditions are illustrated in Fig. 1.

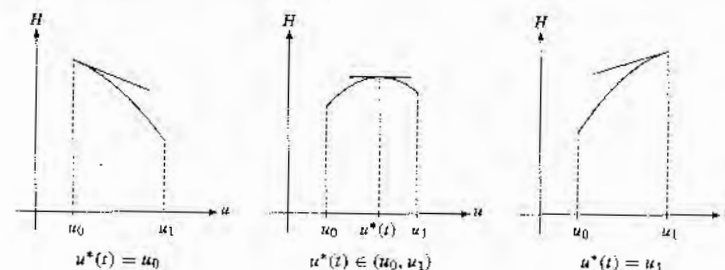


Figure 1

If the Hamiltonian is concave in u , the maximization condition in (9.4.5) can be replaced by the inequality

$$\frac{\partial H^*}{\partial u} (u^*(t) - u) \geq 0 \quad \text{for all } u \text{ in } [u_0, u_1] \quad (3)$$

If $u^*(t) \in (u_0, u_1)$, condition (3) reduces to $\partial H^*/\partial u = 0$. If $u^*(t) = u_0$, then $u^*(t) - u = u_0 - u < 0$ for all u in $(u_0, u_1]$, so (3) is equivalent to $\partial H^*/\partial u \leq 0$. On the other hand, if $u^*(t) = u_1$, then $u^*(t) - u = u_1 - u > 0$ for all u in $[u_0, u_1)$, so (3) is equivalent to $\partial H^*/\partial u \geq 0$.

Proof of Theorem 9.7.1: Suppose that $(x, u) = (x(t), u(t))$ is an arbitrary alternative admissible pair. We must show that

$$D_u = \int_{t_0}^{t_1} f(t, x^*(t), u^*(t)) dt - \int_{t_0}^{t_1} f(t, x(t), u(t)) dt \geq 0$$

First, simplify notation by writing H^* instead of $H(t, x^*(t), u^*(t), p(t))$ and H instead of $H(t, x(t), u(t), p(t))$, etc. Then, using the definition of the Hamiltonian and the fact that $\dot{x}^*(t) = g(t, x^*(t), u^*(t))$ and $\dot{x}(t) = g(t, x(t), u(t))$, we have $f^* = H^* - p\dot{x}^*$ and $f = H - p\dot{x}$. Therefore,

$$D_u = \int_{t_0}^{t_1} (H^* - H) dt + \int_{t_0}^{t_1} p(\dot{x} - \dot{x}^*) dt \quad (*)$$

Because H is concave in (x, u) , Theorem 2.4.1 implies that

$$H - H^* \leq \frac{\partial H^*}{\partial x}(x - x^*) + \frac{\partial H^*}{\partial u}(u - u^*) \quad (**)$$

Now, $\dot{p} = -\partial H^*/\partial x$, so $(*)$ and $(**)$ together imply that

$$D_u \geq \int_{t_0}^{t_1} [\dot{p}(x - x^*) + p(\dot{x} - \dot{x}^*)] dt + \int_{t_0}^{t_1} \frac{\partial H^*}{\partial u}(u^* - u) dt$$

Because of (3), the second integral is ≥ 0 . Moreover, according to the rule for differentiating a product, $\dot{p}(x - x^*) + p(\dot{x} - \dot{x}^*) = (d/dt)(p(x - x^*))$. Hence,

$$D_u \geq \int_{t_0}^{t_1} \frac{d}{dt} [p(x - x^*)] dt = \left[p(t)[x(t) - x^*(t)] \right]_{t_0}^{t_1} = p(t_1)(x(t_1) - x^*(t_1)) \quad (***)$$

where the last equality holds because the contribution from the lower limit of integration is $p(t_0)(x(t_0) - x^*(t_0)) = p(t_0)(x_0 - x_0) = 0$.

Now one can use the terminal condition (9.4.3) and the transversality condition (9.4.7) to show that the last term in $(***)$ is always ≥ 0 . Indeed, if (9.4.3)(a) holds, then $x(t_1) - x^*(t_1) = x_1 - x_1 = 0$. But if (9.4.3)(b) holds, then $p(t_1) \geq 0$ and so if $x^*(t_1) = x_1$, then $p(t_1)(x(t_1) - x^*(t_1)) = p(t_1)[x(t_1) - x_1] \geq 0$ because $x(t_1) \geq x_1$. Alternatively, if $x^*(t_1) > x_1$, then $p(t_1) = 0$, and the term is 0. Finally, if (9.4.3)(c) holds, then $p(t_1) = 0$, and the term is 0. In all cases, therefore, one has $D_u \geq 0$.

If H is strictly concave in (x, u) , then the inequality $(**)$ is strict for $(x, u) \neq (x^*, u^*)$, and so $D_u > 0$ unless $x(t) = x^*(t)$ and $u(t) = u^*(t)$ for all t . Hence (x^*, u^*) is the unique solution to the problem.

Most of the control problems presented so far can be solved by using Mangasarian's sufficient conditions. However, in many important economic models the Hamiltonian is not concave. Arrow has suggested a weakening of this concavity condition. Define

$$\hat{H}(t, x, p) = \max_{u \in U} H(t, x, u, p) \quad (4)$$

assuming that the maximum value is attained. The function $\hat{H}(t, x, p)$ is called the **maximized Hamiltonian**. Then one can show:

THEOREM 9.7.2 (ARROW'S SUFFICIENT CONDITIONS)

Suppose that $(x^*(t), u^*(t))$ is an admissible pair in the standard end-constrained problem (9.4.1)–(9.4.3) that satisfies all the requirements in the maximum principle, with $p(t)$ as the adjoint function, and with $p_0 = 1$. Suppose further that

$$\hat{H}(t, x, p(t)) \text{ is concave in } x \text{ for every } t \in [t_0, t_1] \quad (5)$$

Then $(x^*(t), u^*(t))$ solves the problem.

A proof and further discussion of this result are postponed to Section 10.1.

NOTE 2 Here is an important generalization of the theorem: Suppose the problem imposes the constraint that $x(t)$ must belong to a convex set $A(t)$ for all t . Suppose also that $x^*(t)$ is an interior point of $A(t)$ for every t . Then Theorem 9.7.2 is still valid, and $x \mapsto \hat{H}(t, x, p(t))$ need only be concave for x in $A(t)$.

EXAMPLE 1 Consider the problem

$$\max \int_0^2 (u^2 - x) dt, \quad \dot{x} = u, \quad x(0) = 0, \quad x(2) \text{ free}, \quad 0 \leq u \leq 1$$

- (a) Find the only possible solution candidate by using the maximum principle.
(b) Use Theorem 9.7.2 to prove that the pair found in (a) is optimal.

Solution: (a) The Hamiltonian with $p_0 = 1$ is $H(t, x, u, p) = u^2 - x + pu$. Because $H'_x = -1$, the differential equation for $p = p(t)$ becomes $\dot{p} = -H'_x = 1$. The solution of this equation with $p(2) = 0$ is $p(t) = t - 2$. According to the maximum condition (9.4.5), for each t in $[0, 2]$, an optimal control $u^*(t)$ must maximize H subject to $u \in [0, 1]$. Hence, $u^*(t)$ must maximize $g(u) = u^2 + (t - 2)u$, with t fixed, subject to $u \in [0, 1]$. Note that $g(u)$ is a strictly convex function, so its maximum cannot occur at an interior point of $[0, 1]$. At the end points, $g(0) = 0$ and $g(1) = t - 1$. Thus the maximum of g depends on the value of t . Clearly, if $t < 1$ the maximum of g occurs at $u = 0$, and if $t > 1$, the maximum occurs at $u = 1$. Thus the only possible optimal control that is continuous on the left at $t = 1$ is the bang-bang control

$$u^*(t) = \begin{cases} 0 & \text{if } t \in [0, 1] \\ 1 & \text{if } t \in (1, 2] \end{cases}$$

In the interval $[0, 1]$ one has $\dot{x}^*(t) = u^*(t) = 0$, and $x^*(0) = 0$, so $x^*(t) = 0$. In the interval $(1, 2]$ one has $\dot{x}^*(t) = u^*(t) = 1$, and $x^*(1) = 0$, so $x^*(t) = t - 1$. We have found the only possible pair that can solve the problem.

(b) The Hamiltonian with $p(t) = t - 2$ is $H(t, x, u, p) = u^2 - x + (t - 2)u$, which is strictly convex in u . The maximized Hamiltonian is seen to be

$$\hat{H}(t, x, p(t)) = \max_{u \in [0, 1]} u^2 - x + (t - 2)u = \begin{cases} -x & \text{if } t \in [0, 1] \\ -x + t - 1 & \text{if } t \in (1, 2] \end{cases}$$

For each t in $[0, 2]$, the maximized Hamiltonian is linear in x , hence concave. The optimality of $u^*(t)$ follows from Theorem 9.7.2.

The following example illustrates an important aspect of Theorem 9.7.2: It is enough to show that the maximized Hamiltonian is concave as a function of x with $p(t)$ as the adjoint function derived from the maximum principle.

EXAMPLE 2 Use Theorem 9.7.2 to prove that an optimal control for the problem

$$\max \int_0^1 3u \, dt, \quad \dot{x} = u^3, \quad x(0) = 0, \quad x(1) \leq 0, \quad u \in [-2, \infty)$$

is $u^*(t) = 1$ in $[0, 8/9]$ and $u^*(t) = -2$ in $(8/9, 1]$, with $p(t) = -1$.

Solution: The Hamiltonian with $p(t) \equiv -1$ is $H(t, x, u, p) = 3u - u^3$, which is not concave in (x, u) . But both $u^*(t) \equiv 1$ and $u^*(t) \equiv -2$ maximize $3u - u^3$ subject to $u \in [-2, \infty)$. (See Fig. 2.) So the maximized Hamiltonian is $\hat{H} = \max_{u \in [-2, \infty)} (3u - u^3) \equiv 2$, which is concave. Because $p(1) = -1$, the result in Note 9.4.2 implies that $x^*(1) = 0$. The function $x^*(t)$ must satisfy the equation $\dot{x}^*(t) = u^*(t)^3$ for each t , and also have $x^*(0) = 0$ and $x^*(1) = 0$. One possibility is $x^*(t) \equiv t$ in $[0, 8/9]$, with $u^*(t) \equiv 1$, and $x^*(t) \equiv 8 - 8t$ in $(8/9, 1]$, with $u^*(t) \equiv -2$. Because all the conditions in Theorem 9.7.2 are satisfied, this is a solution. But the solution is not unique. One could also have, for example, $x^*(t) \equiv -8t$ in $[0, 1/9]$ with $u^*(t) \equiv -2$, and $x^*(t) \equiv t - 1$ in $(1/9, 1]$ with $u^*(t) \equiv 1$.

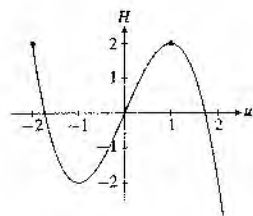


Figure 2 For Example 2.

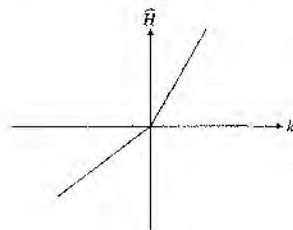


Figure 3 For Example 3.

Our final example makes use of Note 2.

EXAMPLE 3 Consider the capital accumulation model of Example 9.6.3(b). Prove that the proposed solution candidate is optimal.

Solution: The Hamiltonian is $H(t, k, s, p) = (1-s)ak + pask = ak[1 + (p-1)s]$. This function is not concave in (k, s) for $p \neq 1$, because then $H''_{kk}H''_{ss} - (H''_{ks})^2 = -a^2(p-1)^2 < 0$. The function $\hat{H}$ defined in (4) is

$$\hat{H}(t, k, p(t)) = ak \max_{s \in [0, 1]} [1 + (p(t) - 1)s]$$

Given $\bar{t}$ defined as in the solution to Example 9.6.3(b), we found that for $t \in [0, \bar{t})$, the adjoint variable is $p(t) > 1$. It follows that $\hat{H}(t, k, p(t)) = ap(t)k$ for $k \geq 0$, while $\hat{H}(t, k, p(t)) = ak$ for $k \leq 0$. For $t \in (\bar{t}, T)$, however, the adjoint variable is $p(t) < 1$, and it follows that $\hat{H}(t, k, p(t)) = ak$ for $k \geq 0$, while $\hat{H}(t, k, p(t)) = ap(t)k$ for $k \leq 0$. It is tempting to suggest that, because $\hat{H}$ is linear in each case, $\hat{H}$ must be concave in k . But the graph in Fig. 3 shows that $\hat{H}$ is convex, and not concave.

Define $A(t) = \{k : k \geq 0\}$. Certainly, the optimal $k^*(t)$ is positive for all t , so we can impose the constraint that $k(t) \in A(t)$ without affecting the solution. Moreover, $k^*(t)$ is an interior point of $A(t)$ for every t , so we can apply Note 2 provided that $\hat{H}(t, k, p(t))$ is concave as a function of k on the domain $A = [0, \infty)$. But for $k \geq 0$ we have

$$\hat{H}(t, k, p(t)) = \begin{cases} akp(t) & \text{if } p(t) > 1 \\ ak & \text{if } p(t) \leq 1 \end{cases}$$

which is linear in k , and so is concave. The candidate suggested in Example 9.6.3(b) is therefore optimal.

NOTE 3 To give a complete solution of an optimal control problem using the Mangasarian (or Arrow) sufficiency results, it is necessary to prove that there is a pair $(x^*(t), u^*(t))$ satisfying all the requirements. In problems where it is impossible to find explicit solutions for $x^*(t)$ and $u^*(t)$, this means that we must prove that there exist admissible solutions of the differential equations which are valid for the whole interval $[t_0, t_1]$. (This is almost never checked in the economics literature.)

NOTE 4 (What to do if even the Arrow condition fails) If the maximized Hamiltonian is not concave, then the Mangasarian condition also fails. Nevertheless, even if $\hat{H}$ is not concave in x , it is still possible that $x^*(t)$ may maximize $\hat{H} + \dot{p}x$ for each t in $[t_0, t_1]$. This is sufficient for optimality.

When even this weaker sufficient condition fails, we can still be sure that, if there is a unique solution candidate and a solution really does exist, then that candidate must be it. For static optimization problems we relied on the extreme value theorem to demonstrate that, under certain conditions, there does exist an optimal solution. In Section 10.4 we discuss analogous existence theorems for control problems.

PROBLEMS FOR SECTION 9.7

SM 1. (a) Solve the control problem

$$\max \int_0^1 (100 - x - \frac{1}{2}u^2) dt, \quad \dot{x} = u, \quad x(0) = x_0, \quad x(1) = x_1, \quad u \in (-\infty, \infty)$$

(b) Verify that $\partial V / \partial x_0 = p(0)$ and $\partial V / \partial x_1 = -p(1)$, where V is the optimal value function.

SM 2. (a) Find the only possible solution to

$$\max \int_0^{10} (1-s)\sqrt{k} dt, \quad \dot{k} = s\sqrt{k}, \quad k(0) = 1, \quad k(10) \text{ free}, \quad s \in [0, 1]$$

(b) Use Theorem 9.7.2 to prove that the solution candidate in (a) is optimal.

SM 3. (a) Solve the problem (where T, α , and β are positive constants, $\alpha \neq 2\beta$)

$$\max \int_0^T e^{-\beta t} \sqrt{u} dt \quad \text{when} \quad \dot{x}(t) = \alpha x(t) - u(t), \quad x(0) = 1, \quad x(T) = 0, \quad u(t) \geq 0$$

(b) What happens if the terminal condition $x(T) = 0$ is changed to $x(T) \geq 0$?

4. Let f be a C^1 -function defined on a set A in $\mathbb{R}^n$, and let S be a convex set in the interior of A . Show that if $\bar{x}^0$ maximizes $f(x)$ in S , then $\nabla f(\bar{x}^0) \cdot (\bar{x}^0 - x) \geq 0$ for all x in S . (Hint: Define the function $g(t) = f(t\bar{x} + (1-t)\bar{x}^0)$ for t in $[0, 1]$. Then $g(0) \geq g(t)$ for all t in $[0, 1]$.)

9.8 Variable Final Time

In the optimal control problems studied so far the time interval has been fixed. Yet for some control problems in economics, the final time is also a variable to be chosen optimally, along with the function $u(t)$, $t \in [t_0, t_1]$. One instance is the optimal extraction problem of Example 9.1.2, where it is natural to choose for how long to extract the resource, as well as how fast. Another example is the minimal time problem in which the objective is to steer a system from its initial state to a desired state as quickly as possible.

The **variable final time problem** considered here can be briefly formulated as follows (note that the choice variables u and t_1 are indicated below the max sign):

$$\max_{u, t_1} \int_{t_0}^{t_1} f(t, x, u) dt, \quad \dot{x}(t) = g(t, x, u), \quad x(t_0) = x_0, \quad \begin{cases} \text{(a) } x(t_1) = x_1 \\ \text{(b) } x(t_1) \geq x_1 \\ \text{(c) } x(t_1) \text{ free} \end{cases} \quad (1)$$

(Either (a), or (b), or (c) is imposed.) The only difference from the standard end-constrained problem is that t_1 can now be chosen. Thus, the problem is to maximize the integral in (1) over all admissible control functions $u(t)$ that, over the time interval $[t_0, t_1]$, bring the system from x_0 to a point satisfying the terminal conditions. In contrast to the previous problems, the admissible control functions may be defined on different time intervals.

Suppose $(x^*(t), u^*(t))$ is an optimal solution defined on $[t_0, t_1^*]$. Then the conditions (9.4.5)–(9.4.7) in the maximum principle are still valid on the interval $[t_0, t_1^*]$, because the pair $(x^*(t), u^*(t))$ must be optimal for the corresponding fixed time problem with $t_1 = t_1^*$. In fact, here is a modified maximum principle:

THEOREM 9.8.1 (THE MAXIMUM PRINCIPLE WITH VARIABLE FINAL TIME)

Let $(x^*(t), u^*(t))$ be an admissible pair defined on $[t_0, t_1^*]$ which solves problem (1) with t_1 free ($t_1 \in (t_0, \infty)$). Then all the conditions in the maximum principle (Theorem 9.4.1) are satisfied on $[t_0, t_1^*]$, and, in addition,

$$H(t_1^*, x^*(t_1^*), u^*(t_1^*), p(t_1^*)) = 0 \quad (2)$$

Compared with a fixed final time problem there is one additional unknown t_1^* . Fortunately, (2) is one extra condition.

One method for solving variable final time problems is first to solve the problem with t_1 fixed for every $t_1 > t_0$. Next, consider t_1 as an unknown to be determined by condition (2).

According to (9.6.5), $\partial V / \partial t_1 = H(t_1^*, x^*(t_1^*), u^*(t_1^*), p(t_1^*))$ if V is differentiable. Thus, condition (2) is precisely as expected. For a formal proof, see Hestenes (1966).

NOTE 1 (A common misunderstanding) Concavity of the Hamiltonian in (x, u) is not sufficient for optimality when t_1 is free. For sufficiency results when the final time is variable, see Seierstad and Sydsæter (1987), Sections 2.9 and 6.7.

EXAMPLE 1 Consider Problem II in Example 9.1.2 for the special case when the cost function $C = C(t, u)$ is independent of x and convex in u , with $C_{uu}'' > 0$. Thus, the problem is

$$\max_{u, T} \int_0^T [q(t)u(t) - C(t, u(t))] e^{-rt} dt \quad \text{s.t. } \dot{x}(t) = -u(t), \quad x(0) = K, \quad x(T) \geq 0, \quad u(t) \geq 0$$

What does the maximum principle imply for this problem?

Solution: Suppose $(x^*(t), u^*(t))$, defined on $[0, T^*]$, solves this problem. The Hamiltonian with $p_0 = 1$ is $H(t, x, u, p) = [q(t)u - C(t, u)]e^{-rt} + p(-u)$, and the maximum principle states that there exists a continuous function $p(t)$ such that

$$u^*(t) \text{ maximizes } [q(t)u - C(t, u)]e^{-rt} - p(t)u \text{ subject to } u \geq 0 \quad (i)$$

$$\dot{p}(t) = -\frac{\partial H}{\partial x} = 0, \quad p(T^*) \geq 0, \quad \text{with } p(T^*) = 0 \text{ if } x^*(T^*) > 0 \quad (ii)$$

$$[q(T^*)u^*(T^*) - C(T^*, u^*(T^*))]e^{-rT^*} = p(T^*)u^*(T^*) \quad (iii)$$

Because $p(t)$ is continuous, (ii) implies that $p(t) = \bar{p} \geq 0$, where $\bar{p}$ is a constant.

Put $g(u) = [q(t)u - C(t, u)]e^{-rt} - \bar{p}u$. Because $C(t, u)$ is convex in u and the other terms are linear in u , the function $g(u)$ is concave. According to (i), $u^*(t)$ maximizes $g(u)$ subject to $u \geq 0$. If $u^*(t) = 0$, then $g'(u^*(t)) = g'(0) \leq 0$. If $u^*(t) > 0$, then $g'(u^*(t)) = 0$. Therefore (i) implies that

$$[q(t) - C'_u(t, u^*(t))]e^{-rt} - \bar{p} \leq 0 \quad (= 0 \text{ if } u^*(t) > 0) \quad (iv)$$

Because g is concave, this condition is also sufficient for (i) to hold.

At any time t where $u^*(t) > 0$, equation (iv) implies that

$$q(t) - C'_u(t, u^*(t)) = \bar{p}e^{rt} \quad (v)$$

The left-hand side is the marginal profit from extraction, $\partial \pi / \partial u$. Therefore, whenever it is optimal to have positive extraction, we have the following rule due to Hotelling (1931):

HOTELLING'S RULE

Positive optimal extraction requires the marginal profit to increase exponentially at a rate equal to the constant discount factor r . (3)

Putting $t = T^*$ in (v), and using (iii), we deduce that if $u^*(T^*) > 0$, then

$$C'_u(T^*, u^*(T^*)) = \frac{C(T^*, u^*(T^*))}{u^*(T^*)} \quad (vi)$$

Terminate extraction at a time when the marginal cost of extraction is equal to average cost!

If the problem has a solution with $u^*(t) > 0$, then (v) and (vi) both hold. If $C(T^*, 0) > 0$, then $u^*(T^*) > 0$, because $u^*(T^*) = 0$ contradicts (iii).

We have not proved that there exists an optimal solution. (For a more thorough discussion of this problem, see Seierstad and Sydsæter (1987), Section 2.9, Example 11.)

PROBLEMS FOR SECTION 9.8

SM 1. Find the only possible solution to the following variable final time problems:

$$(a) \max_{u, T} \int_0^T (x - t^3 - \frac{1}{2}u^2) dt, \quad \dot{x} = u, \quad x(0) = 0, \quad x(T) \text{ free}, \quad u \in \mathbb{R}$$

$$(b) \max_{u, T} \int_0^T (-9 - \frac{1}{2}u^2) dt, \quad \dot{x} = u, \quad x(0) = 0, \quad x(T) = 16, \quad u \in \mathbb{R}$$

SM 2. Solve problem 9.4.7 with T free.

3. Consider the optimal extraction problem over a fixed extraction period,

$$\max_{u(t) \geq 0} \int_0^T [ae^{\alpha t} u(t) - e^{\beta t} u(t)^2 - c] e^{-rt} dt, \quad \dot{x}(t) = -u(t), \quad x(0) = K, \quad x(T) = 0$$

Here $x(t)$ and $u(t)$ have the same interpretation as in Example 1, with $q(t) = ae^{\alpha t}$ as the world market price, and $e^{\beta t} u(t)^2 + c$ as the cost of extraction, with $c > 0$.

(a) One can prove that if $u^*(t)$ is optimal, then $u^*(t) > 0$ for all t . (You are not required to show this.) The adjoint function is a constant $\bar{p}$. Find $u^*(t)$ expressed in terms of $\bar{p}$. Then find $x^*(t)$ and $\bar{p}$ for the case $\alpha = \beta = 0$, $r > 0$.

(b) Let $T > 0$ be subject to choice (keeping the assumptions $\alpha = \beta = 0$, $r > 0$). Prove that the necessary conditions lead to an equation for determining the optimal T^* that has a unique positive solution. Assume that $\max_u (au - u^2 - c) > 0$, i.e. $a^2 > 4c$.

9.9 Current Value Formulations

Many control problems in economics have the following structure:

$$\max_{u \in U \subseteq \mathbb{R}} \int_0^{t_1} f(t, x, u) e^{-rt} dt, \quad \dot{x} = g(t, x, u), \quad x(t_0) = x_0, \quad \begin{cases} (a) x(t_1) = x_1 \\ (b) x(t_1) \geq x_1 \\ (c) x(t_1) \text{ free} \end{cases} \quad (1)$$

(Either (a), or (b), or (c) is imposed.) The new feature is the explicit appearance of the discount factor e^{-rt} . For such problems it is often convenient to formulate the maximum principle in a slightly different form.

The ordinary Hamiltonian is $H = p_0 f(t, x, u) e^{-rt} + pg(t, x, u)$. Multiply it by e^{rt} to obtain the **current value Hamiltonian** $H^c = He^{rt} = p_0 f(t, x, u) + e^{rt} pg(t, x, u)$. Introducing $\lambda = e^{rt} p$ as the **current value shadow price** for the problem, one can write H^c in the form (where we put $p_0 = \lambda_0$)

$$H^c(t, x, u, \lambda) = \lambda_0 f(t, x, u) + \lambda g(t, x, u) \quad (2)$$

Note that if $\lambda = e^{rt} p$, then $\dot{\lambda} = re^{rt} p + e^{rt} \dot{p} = r\lambda + e^{rt} \dot{p}$ and so $\dot{p} = e^{-rt}(\dot{\lambda} - r\lambda)$. Also, $H^c = He^{rt}$ implies that $\partial H^c / \partial x = e^{rt}(\partial H / \partial x)$. So $\dot{p} = -\partial H / \partial x$ takes the form $\dot{\lambda} - r\lambda = -\partial H^c / \partial x$. In fact, the following result is obtained:

THEOREM 9.9.1 (THE MAXIMUM PRINCIPLE: CURRENT VALUE FORMULATION)

Suppose that the admissible pair $(x^*(t), u^*(t))$ solves problem (1) and let H^c be the current value Hamiltonian (2). Then there exists a continuous function $\lambda(t)$ and a number λ_0 , either 0 or 1, such that for all t in $[t_0, t_1]$ we have $(\lambda_0, \lambda(t)) \neq (0, 0)$, and:

$$(A) u = u^*(t) \text{ maximizes } H^c(t, x^*(t), u, \lambda(t)) \text{ for } u \in U \quad (3)$$

$$(B) \dot{\lambda}(t) - r\lambda(t) = -\frac{\partial H^c(t, x^*(t), u^*(t), \lambda(t))}{\partial x} \quad (4)$$

(C) Finally, the transversality conditions are:

$$(a') \lambda(t_1) \text{ no condition}$$

$$(b') \lambda(t_1) \geq 0, \text{ with } \lambda(t_1) = 0 \text{ if } x^*(t_1) > x_1 \quad (5)$$

$$(c') \lambda(t_1) = 0$$

The Mangasarian and Arrow sufficiency results from Section 9.7 have immediate extensions to problem (1). The conditions in Theorem 9.9.1 are sufficient for optimality if $\lambda_0 = 1$ and

$$H^c(t, x, u, \lambda(t)) \text{ is concave in } (x, u) \quad (\text{Mangasarian}) \quad (6)$$

or (more generally)

$$\hat{H}^c(t, x, \lambda(t)) = \max_{u \in U} H^c(t, x, u, \lambda(t)) \text{ is concave in } x \quad (\text{Arrow}) \quad (7)$$

or (more generally still)

$$x = x^*(t) \text{ maximizes } \hat{H}^c + (\dot{\lambda}(t) - r\lambda(t))x$$

EXAMPLE 1 Solve the following problem using the current value formulation:

$$\max_{u \geq 0} \int_0^{20} (4K - u^2) e^{-0.25t} dt, \quad \dot{K} = -0.25K + u, \quad K(0) = K_0, \quad K(20) \text{ is free}$$

An economic interpretation is that $K(t)$ is the value of a firm's capital stock, which depreciates at the constant proportional rate 0.25 per unit of time, whereas $u(t)$ is gross investment, which costs $u(t)^2$ because the marginal cost of investment increases. Finally, profits are discounted at the constant proportional rate 0.25 per unit of time.

Solution: The current value Hamiltonian is $H^c = 4K - u^2 + \lambda(-0.25K + u)$ (with $\lambda_0 = 1$), and so $\partial H^c/\partial u = -2u + \lambda$ and $\partial H^c/\partial K = 4 - 0.25\lambda$. Assuming that $u^*(t) > 0$ (we try this assumption in the following), $\partial(H^c)^*/\partial u = 0$, so $u^*(t) = 0.5\lambda(t)$. The adjoint function λ satisfies

$$\dot{\lambda} - 0.25\lambda = -\partial(H^c)^*/\partial K = -4 + 0.25\lambda, \quad \lambda(20) = 0$$

It follows that

$$\lambda(t) = 8(1 - e^{0.5t-10}) \quad \text{and} \quad u^*(t) = 0.5\lambda = 4(1 - e^{0.5t-10})$$

Note that $u^*(t) > 0$ in $[0, 20)$. The time path of $K^*(t)$ is found from $\dot{K}^* = -0.25K^* + u^* = -0.25K^* + 4(1 - e^{0.5t-10})$. Solving this linear differential equation with $K^*(0) = K_0$, we get

$$K^*(t) = (K_0 - 16 + \frac{16}{3}e^{-10})e^{-0.25t} + 16 - \frac{16}{3}e^{0.5t-10}$$

Here $H^c = (4K - u^2) + \lambda(-0.25K + u)$ is concave in (K, u) , so we have found the optimal solution.

Note that the pair $(K^*(t), \lambda(t))$ must satisfy the system

$$\begin{aligned} \dot{\lambda} &= 0.5\lambda - 4, & \lambda(20) &= 0 \\ \dot{K} &= -0.25K + 0.5\lambda, & K(0) &= K_0 \end{aligned}$$

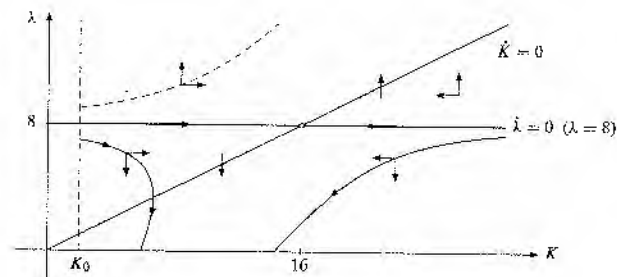


Figure 1 Phase diagram for Example 1.

Figure 1 shows a phase diagram for this system. When $K_0 < 16$ as in the figure, the left curve drawn with a solid line is consistent with the indicated arrows. Initially the capital stock increases, and investment is reduced. Then, after the curve hits the line $\dot{K} = 0$, the capital stock decreases and investment is reduced until it eventually is 0. The dotted curve is also consistent with the arrows, but there is no way the curve can satisfy $\lambda(20) = 0$ —the required investment is too high to lead to an optimal solution. (When λ is large, so is $u = 0.5\lambda$, and the integrand $4K - u^2$ becomes large negative.)

The diagrammatic analysis related to Fig. 1 in the last example is in a way superfluous since the solution has already been completely specified. But it is very useful in some problems where explicit solutions are unobtainable. See Section 9.12.

PROBLEMS FOR SECTION 9.9

1. Find the solution to Example 9.2.2 using the current value formulation.
- SM 2. Find the solution of Problem 9.4.6 using the current value formulation.
- SM 3. Find the solution of Problem 9.5.3 using the current value formulation.

9.10 Scrap Values

In some economic optimization problems it is natural to include within the optimality criterion an additional function representing the value or utility associated with the terminal state. This gives the typical problem

$$\max_{u(t) \in U} \left\{ \int_0^T f(t, x(t), u(t)) dt + S(x(T)) \right\}, \quad \dot{x}(t) = g(t, x(t), u(t)), \quad x(0) = x_0 \quad (1)$$

The function $S(x)$ is called a **scrap value function**, and we shall assume that it is C^1 .

Suppose that $(x^*(t), u^*(t))$ solves this problem (with no additional condition on $x(T)$). Then, in particular, that pair is a solution to the corresponding problem with fixed terminal point $(T, x^*(T))$. For all admissible pairs in this new problem, the scrap value function $S(x^*(T))$ is constant. But then $(x^*(t), u^*(t))$ must satisfy all the conditions in the maximum principle, except the transversality conditions. The correct transversality condition for problem (1)

$$p(T) = S'(x^*(T)) \quad (2)$$

This is quite natural if we use the general economic interpretation explained in Section 9.6. In fact, if $x(t)$ denotes the capital stock of a firm, then according to (2), the shadow price of capital at the end of the planning period is equal to the marginal scrap value of the terminal stock.

NOTE 1 If $S(x) \equiv 0$, then (2) reduces to $p(T) = 0$, which is precisely as expected in a problem with no restrictions on $x(T)$.

One way to show that (2) is the correct transversality condition involves transforming problem (1) into one studied before. Indeed, suppose that $(x(t), u(t))$ is an admissible pair for the problem (1). Then $\frac{d}{dt} S(x(t)) = S'(x(t))\dot{x}(t) = S'(x(t))g(t, x(t), u(t))$. So, by integration,

$$S(x(T)) - S(x(0)) = \int_0^T S'(x(t))g(t, x(t), u(t)) dt$$

Here $S(x(t_0)) = S(x_0)$ is a constant, so if the objective function in (1) is replaced by

$$\int_{t_0}^t [f(t, x(t), u(t)) + S'(x(t))g(t, x(t), u(t))] dt \quad (3)$$

then the new problem is of a type studied previously with no scrap value, still with $x(t_1)$ free. Let the Hamiltonian for this new problem be $H_1 = f + S'(x)g + qg = f + (q + S'(x))g$, with adjoint variable q . An optimal pair $(x^*(t), u^*(t))$ for this problem must have the following properties:

(a) $u = u^*(t)$ maximizes $H_1(t, x^*(t), u, q(t))$ for $u \in U$

(b) $\dot{q}(t) = -\partial H_1^* / \partial x$, $q(t_1) = 0$

Define $p(t) = q(t) + S'(x^*(t))$. Problem 7 asks you to prove that, if $H = f + pg$ is the ordinary Hamiltonian associated with problem (1), then $u^*(t)$ maximizes $H(x^*(t), u, p(t))$ for $u \in U$ and $\dot{p}(t) = -\partial H^* / \partial x$, with $p(t_1) = 0$.

Appropriate concavity conditions again ensure optimality, as shown in the next theorem:

THEOREM 9.10.1 (SUFFICIENT CONDITIONS WITH SCRAP VALUE)

Suppose $(x^*(t), u^*(t))$ is an admissible pair for the scrap value problem (1) and suppose there exists a continuous function $p(t)$ such that for all t in $[t_0, t_1]$,

(A) $u = u^*(t)$ maximizes $H(t, x^*(t), u, p(t))$ w.r.t. $u \in U$

(B) $\dot{p}(t) = -H_x^*(t, x^*(t), u^*(t), p(t))$, $p(t_1) = S'(x^*(t_1))$

(C) $H(t, x, u, p(t))$ is concave in (x, u) and $S(x)$ is concave

Then $(x^*(t), u^*(t))$ solves the problem.

Proof: Suppose that $(x(t), u(t))$ is an arbitrary admissible pair. We must show that

$$D_u = \int_{t_0}^{t_1} f(t, x^*(t), u^*(t)) dt + S(x^*(t_1)) - \int_{t_0}^{t_1} f(t, x(t), u(t)) dt - S(x(t_1)) \geq 0$$

Because $S(x)$ is C^1 and concave, $S(x(t_1)) - S(x^*(t_1)) \leq S'(x^*(t_1))(x(t_1) - x^*(t_1))$. Combining this with the inequality $\int_{t_0}^{t_1} (f^* - f) dt \geq p(t_1)(x(t_1) - x^*(t_1))$ that was derived as formula (***) in the proof of the Theorem 9.7.1, we get

$$D_u \geq [p(t_1) - S'(x^*(t_1))](x(t_1) - x^*(t_1)) = 0$$

where the last equality follows from (B). So $D_u \geq 0$.

NOTE 2 The theorem still holds if the concavity of H in (x, u) is replaced by the Arrow condition requiring $\widehat{H}(t, x, p(t))$ to exist and be concave in x . Or, more generally still, if $x^*(t)$ maximizes $\widehat{H}(t, x, p(t)) + \dot{p}(t)x$.

EXAMPLE 1 Solve the problem

$$\max_{u \in (-\infty, \infty)} \left\{ \int_0^1 -\frac{1}{2}u^2 dt + \sqrt{x(1)} \right\}, \quad \dot{x} = x + u, \quad x(0) = 0, \quad x(1) \text{ free}$$

Solution: We have $H = -\frac{1}{2}u^2 + p(x + u)$ and $S(x) = \sqrt{x} = x^{1/2}$. Hence $H'_u = -u + p$ and $H'_x = p$. Since $u \in (-\infty, \infty)$, maximization w.r.t. u requires that $H'_u = 0$, implying that $u = p$. So we have the differential equations $\dot{x} = x + u = x + p$, $\dot{p} = -H'_x = -p$. The latter equation has the solution $p(t) = Ae^{-t}$. Then $\dot{x} = x + p = x + Ae^{-t}$, and this linear differential equation has the solution $x = Be^t - \frac{1}{2}Ae^{-t}$, where the constant B is determined by $x(0) = B - \frac{1}{2}A = 0$. Hence, $B = \frac{1}{2}A$, so that $x(t) = \frac{1}{2}A(e^t - e^{-t})$. The constant A is determined by the transversality condition $p(1) = Ae^{-1} = S'(x(1)) = \frac{1}{2}(x(1))^{-1/2} = \frac{1}{2}[\frac{1}{2}A(e^1 - e^{-1})]^{-1/2}$. Solving for A we find $A = e[2(e^2 - 1)]^{-1/3}$. Thus we have the following candidate for an optimal solution:

$$u(t) = p(t) = Ae^{-t}, \quad x(t) = \frac{1}{2}A(e^t - e^{-t}), \quad A = e[2(e^2 - 1)]^{-1/3}$$

Because the Hamiltonian is concave in (x, u) , and the scrap value function is concave in x , this is the solution. $\square$

Current Value Formulation

Many control problems in economics have the following structure:

$$\max_{u \in U \subseteq \mathbb{R}} \left\{ \int_{t_0}^{t_1} f(t, x, u)e^{-rt} dt + S(x(t_1))e^{-rt_1} \right\}, \quad \dot{x} = g(t, x, u), \quad x(t_0) = x_0 \quad (4)$$

$$(a) \ x(t_1) = x_1, \quad (b) \ x(t_1) \geq x_1, \quad \text{or} \quad (c) \ x(t_1) \text{ free} \quad (5)$$

(Either (a), or (b), or (c) is imposed.) The new features as compared with problem (1) are the discount factor (or interest rate) r , and the reintroduction of the alternative terminal conditions in the standard problem. (If $x(t_1)$ is fixed as in S(a), the scrap value function is a constant.)

The current value Hamiltonian for the problem is

$$H^c(t, x, u, \lambda) = \lambda_0 f(t, x, u) + \lambda g(t, x, u) \quad (6)$$

and the correct necessary conditions are as follows:

THEOREM 9.10.2 (CURRENT VALUE MAXIMUM PRINCIPLE: SCRAP VALUE)

Suppose that the admissible pair $(x^*(t), u^*(t))$ solves problem (4)–(5). Then there exist a continuous function $\lambda(t)$ and a number λ_0 , either 0 or 1, such that for all t in $[t_0, t_1]$ we have $(\lambda_0, \lambda(t)) \neq (0, 0)$, and:

- (A) $u = u^*(t)$ maximizes $H^c(t, x^*(t), u, \lambda(t))$ for $u \in U$
 (B) $\dot{\lambda}(t) - r\lambda(t) = -\frac{\partial H^c(t, x^*(t), u^*(t), \lambda(t))}{\partial x}$ whenever $u^*(t)$ is continuous
 (C) Finally, the transversality conditions are:
 (a') $\lambda(t_1)$ no condition
 (b') $\lambda(t_1) \geq \lambda_0 S'(x^*(t_1))$ (with = if $x^*(t_1) > x_1$)
 (c') $\lambda(t_1) = \lambda_0 S'(x^*(t_1))$

The following sufficiency result is a straightforward extension of Theorem 9.10.1:

THEOREM 9.10.3 (SUFFICIENT CONDITIONS)

The conditions in Theorem 9.10.2 with $\lambda_0 = 1$ are sufficient if U is convex, $H^c(t, x, u, \lambda(t))$ is concave in (x, u) , and $S(x)$ is concave in x .

EXAMPLE 2 Consider the following problem:

$$\max \left\{ \int_0^T (x - u^2)e^{-0.1t} dt + ax(T)e^{-0.1T} \right\}$$

$$\dot{x} = -0.4x + u, \quad x(0) = 1, \quad x(T) \text{ is free}, \quad u \in \mathbb{R}$$

where a is a positive constant. Solve the problem.

Solution: The current value Hamiltonian, with $\lambda_0 = 1$, is $H^c(t, x, u, \lambda) = x - u^2 + \lambda(-0.4x + u)$, which is concave in (x, u) . Moreover, $S(x) = ax$ is linear, and hence concave in x . The conditions in the maximum principle are therefore sufficient. Because H^c is concave in u and $u \in \mathbb{R}$, the maximum of the Hamiltonian occurs when

$$\frac{\partial H^c(t, x^*(t), u^*(t), \lambda(t))}{\partial u} = -2u^*(t) + \lambda(t) = 0 \quad (\text{i})$$

Next, the differential equation for λ is

$$\dot{\lambda}(t) - 0.1\lambda(t) = -\partial H^c / \partial x = -1 + 0.4\lambda(t) \quad (\text{ii})$$

Because $x(T)$ is free and $S(x) = ax$, condition (C)(c') in Theorem 9.10.2 yields

$$\lambda(T) = a \quad (\text{iii})$$

By integrating the linear differential equation (ii), using (iii), we obtain

$$\lambda(t) = (a - 2)e^{-0.5(T-t)} + 2$$

From (i), $u^*(t) = \frac{1}{2}\lambda(t)$. Because $x^*(t)$ must satisfy the linear differential equation $\dot{x}^* = -0.4x^* + u^* = -0.4x^* + \frac{1}{2}(a - 2)e^{-0.5(T-t)} + 1$, with $x^*(0) = 1$, one has

$$x^*(t) = \frac{5}{2} + \frac{5}{9}(a - 2)e^{-0.5(T-t)} - \left(\frac{3}{2} + \frac{5}{9}(a - 2)e^{-0.5T}\right)e^{-0.4t}$$

All the conditions in Theorem 9.10.3 are satisfied, so this is the solution. $\blacksquare$

EXAMPLE 3

(Feeding a fish optimally) Let $x(t)$ be the weight of a fish at time t and let $P(t, x)$ be the price per kilogram of a fish whose weight is x at time t . Furthermore, let $u(t)$ denote the amount of fish food per unit of time measured as a proportion of the weight of a fish, and let $c > 0$ be the constant cost of a kilogram of fish food. If the interest rate is r , then the present value of the profit from feeding the fish and then catching it at the fixed time T is

$$x(T)P(T, x(T))e^{-rT} - \int_0^T cx(t)u(t)e^{-rt} dt \quad (\text{i})$$

Suppose that

$$\dot{x}(t) = x(t)g(t, u(t)), \quad x(0) = x_0 > 0 \quad (\text{ii})$$

so that the proportional rate of growth in the weight of the fish is a known function $g(t, u(t))$. The natural problem is to find the feeding function $u^*(t)$ and the corresponding weight function $x^*(t)$ that maximize (i) subject to the constraint (ii) and $u(t) \geq 0$.

- (a) Write down necessary conditions for $(x^*(t), u^*(t))$, with corresponding adjoint function $\lambda(t)$, to solve the problem. Deduce an equation that $u^*(t)$ must satisfy if $u^*(t) > 0$.
 (b) Suppose $P(t, x) = a_0 + a_1x$, and $g(t, u) = a - be^{st}/u$, where all the constants are positive, with $s > r$. Characterize the only possible solution.

Solution: (a) The current value Hamiltonian is $H^c(t, x, u, \lambda) = -cxu + \lambda xg(t, u)$, and the scrap value function is $S(x) = xP(t, x)$. Thus $\partial H^c / \partial x = -cu + \lambda g(t, u)$, $\partial H^c / \partial u = x(-c + \lambda g_u(t, u))$, and $S'(x) = P(t, x) + xP'_x(t, x)$.

According to the maximum principle, there exists a continuous function $\lambda(t)$ such that

$$u^*(t) \text{ maximizes } x^*(t)(-cu + \lambda(t)g(t, u)) \text{ for } u \geq 0 \quad (\text{iii})$$

and

$$\dot{\lambda}(t) - r\lambda(t) = -\frac{\partial (H^c)^*}{\partial x} = cu^*(t) - \lambda(t)g(t, u^*(t)) \quad (\text{iv})$$

Furthermore, condition (C)(c') in Theorem 9.10.2 takes the form

$$\lambda(T) = P(T, x^*(T)) + x^*(T)P'_x(T, x^*(T)) \quad (\text{v})$$

From (iii) it follows that if $u^*(t) > 0$, then $\partial(H^0)^*/\partial u = 0$. If $x^*(t)$ is not 0 then

$$\lambda(t)g'_u(t, u^*(t)) = c \quad (\text{vi})$$

(b) We have $g'_u(t, u) = be^{rt}/u^2$, so (vi) yields $\lambda(t)be^{rt}/(u^*(t))^2 = c$. Then $\lambda(t) > 0$, and with $u^*(t) > 0$, we obtain $u^*(t) = \sqrt{b/c} e^{\frac{1}{2}rt} (\lambda(t))^{1/2}$. Equation (iv) is now $\dot{\lambda}(t) - r\lambda(t) = cu^*(t) - \lambda(t)[a - be^{rt}/u^*(t)]$, which reduces to

$$\dot{\lambda}(t) = (r - a)\lambda(t) + 2\sqrt{bc} e^{\frac{1}{2}rt} (\lambda(t))^{1/2} \quad (\text{vii})$$

Finally, (v) reduces to

$$\lambda(T) = a_0 + a_1 x^*(T) + a_1 x^*(T) = a_0 + 2a_1 x^*(T) \quad (\text{viii})$$

The standard trick for solving the Bernoulli equation (vii) is to introduce a new variable z defined by $z = \lambda^{1/2}$. (See (5.6.2).) Then $\lambda = z^2$, so $\dot{\lambda} = 2z\dot{z}$, and (vii) yields

$$2z\dot{z} = (r - a)z^2 + 2\sqrt{bc} e^{\frac{1}{2}rt} z \quad (\text{ix})$$

Because we are looking for a solution with $z(t) = \sqrt{\lambda(t)} > 0$ for all t , (ix) implies that $\dot{z} = \frac{1}{2}(r - a)z + \sqrt{bc} e^{\frac{1}{2}rt}$. According to (5.4.4) this has the solution

$$z = Ae^{\frac{1}{2}(r-a)t} + \sqrt{bc} e^{\frac{1}{2}(r-a)t} \int e^{\frac{1}{2}(s-r+a)t} dt = Ae^{\frac{1}{2}(r-a)t} + \frac{2\sqrt{bc}}{s-r+a} e^{\frac{1}{2}st}$$

where A is a constant. Since $u^*(t) = \sqrt{b/c} e^{\frac{1}{2}rt} z$, we get

$$u^*(t) = A\sqrt{b/c} e^{\frac{1}{2}(s+r-a)t} + \frac{2b}{s-r+a} e^{st}$$

Inserting $u^*(t)$ into (ii) yields a separable differential equation for $x^*(t)$, with a unique solution satisfying $x^*(0) = x_0$. The constant A is finally determined by equation (viii). $\square$

PROBLEMS FOR SECTION 9.10

1. Find the solution to the control problem

$$\max \left\{ \int_0^1 (1 - tu - u^2) dt + 2x(1) + 3 \right\}, \quad \dot{x} = u, \quad x(0) = 1, \quad u \in (-\infty, \infty)$$

2. In a study of savings and inheritance, Atkinson (1971) considers the problem

$$\max \left\{ \int_0^T U(rA(t) + w - u(t))e^{-\rho t} dt + e^{-\rho T} \varphi(A(T)) \right\}, \quad \dot{A} = u, \quad A(0) = A_0$$

An economic interpretation is given in Example 8.5.3, except that the objective function now includes an extra term which measures the individual's discounted benefit from bequeathing $A(T)$. Suppose that $\varphi' > 0$, $\varphi'' < 0$. Give a set of sufficient conditions for the solution of this problem.

3. Solve the following control problem from economic growth theory:

$$\max \left\{ \int_0^{10} (1-s)\sqrt{k} dt + 10\sqrt{k(10)} \right\}, \quad \dot{k} = s\sqrt{k}, \quad k(0) = 1, \quad s \in [0, 1]$$

where $k = k(t)$ is the capital stock, and $s = s(t)$ is the savings ratio. (See Problem 9.7.2.)

4. (a) Solve the problem

$$\max \left\{ \int_0^1 (x - u) dt + \frac{1}{2}x(1) \right\}, \quad \dot{x} = u, \quad x(0) = \frac{1}{2}, \quad x(1) \text{ free}, \quad u \in [0, 1]$$

(b) Solve the problem with the objective function $\int_0^1 (x - u) dt - \frac{1}{4}(x(1) - 2)^2$.

5. Consider the problem:

$$\max \left\{ \int_0^T -u^2 dt - x(T)^2 \right\}, \quad \dot{x} = -x + u, \quad x(0) = x_0, \quad u \in \mathbb{R}$$

(a) Solve the problem using Theorem 9.10.3.

(b) Compute the optimal value function, $V = V(x_0, T)$. Show that $\partial V / \partial x_0 = p(0)$ and $\partial V / \partial T = H^*(T)$.

6. Solve the following problem using the current value formulation

$$\max_{u \in \mathbb{R}} \left\{ \int_0^T -e^{-rt}(x - u)^2 dt - e^{-rT}x(T)^2 \right\} \text{ s.t. } \dot{x} = u - x + a, \quad x(0) = 0, \quad x(T) \text{ free}$$

The constants r , a , and T are all positive.

7. Consider the control problem

$$\max \int_{t_0}^{t_1} [f(t, x, u) + S'(x)g(t, x, u)] dt, \quad \dot{x} = g(t, x, u), \quad x(t_0) = x_0, \quad u \in U$$

(See (3).) Let the Hamiltonian be $H_1 = f + S'(x)g + qg = f + (q + S'(x))g$, with q as the adjoint variable. Then an optimal pair (x^*, u^*) for this problem must satisfy conditions (a) and (b) above Theorem 9.10.1. Define $p = q + S'(x^*)$ and let $H = f + pg$. Prove that properties (a) and (b) imply that u^* maximizes $H(t, x^*, u, p)$ for $u \in U$, while $\dot{p} = -\partial H^*/\partial x$, with $p(t_1) = S'(x^*(t_1))$. Thus conditions (A)–(C) in Theorem 9.10.1 are satisfied.

9.11 Infinite Horizon

Most of the optimal growth models appearing in the economics literature have an infinite time horizon. This is despite the following comment by Nobel laureate Ragnar Frisch (1970) about infinite horizon growth models:

Questions of convergence under an infinite time horizon will depend so much on epsilonic refinements in the system of assumptions—and on the infinite constancy of these refinements—that we are humanly speaking absolutely certain of getting infinite time horizon results which have no relevance to concrete reality. And in particular we are absolutely certain of getting irrelevant results if such epsilonic exercises are made under the assumption of a constant technology. “In the long run we are all dead.” These words by Keynes ought to be engraved in marble and put on the desk of all epsilonologists in growth theory under an infinite horizon.

Clearly, choosing an infinite horizon makes sense in economic models only if the distant future has no significant influence on the optimal path for the near future in which we are most interested. Nevertheless, the infinite horizon assumption often does simplify formulas and conclusions, though at the expense of some new mathematical problems that need to be sorted out.

A typical infinite horizon optimal control problem in economics takes the following form:

$$\max \int_{t_0}^{\infty} f(t, x(t), u(t)) e^{-rt} dt, \quad \dot{x}(t) = g(t, x(t), u(t)), \quad x(t_0) = x_0, \quad u(t) \in U \quad (1)$$

Often no condition is placed on $x(t)$ as $t \rightarrow \infty$, but many problems do impose the constraint

$$\lim_{t \rightarrow \infty} x(t) \geq x_1 \quad (x_1 \text{ is a fixed number}) \quad (2)$$

The pair $(x(t), u(t))$ is *admissible* if it satisfies $\dot{x}(t) = g(t, x(t), u(t))$, $x(t_0) = x_0$, $u(t) \in U$, along with (2) when that is imposed. Suppose the integral (1) converges whenever the pair $(x(t), u(t))$ is admissible. For example, the integral will converge for all admissible $(x(t), u(t))$ if r is a positive constant, and if there exists a number M such that $|f(t, x, u)| \leq M$ for all (x, u) .

One can then show (Halkin (1974)) that all the necessary conditions in the maximum principle (Theorem 9.2.1) hold, except the transversality condition $p(t_1) = 0$. With no transversality condition we get too many solution candidates.

NOTE 1 It is tempting to assume that all results for finite horizon problems can be carried over in a simple way to the infinite horizon case. This is wrong. For example, in a finite horizon problem with $x(t_1)$ free, the transversality condition is $p(t_1) = 0$. However, with no terminal condition, the “natural” transversality condition, $p(t) \rightarrow 0$ as $t \rightarrow \infty$, is not correct. A well-known counterexample is due to Halkin (1974). That example also shows that the condition $p(t)x(t) \rightarrow 0$ is *not* a necessary condition for optimality, contrary to a widespread belief in economic literature, including some popular textbooks. (See also Example 1.)

However, in economic models with $x(\infty)$ free, it is in most cases a sensible working hypothesis that $p(t)$ does tend to 0 as t tends to ∞ . But, ultimately, this must be confirmed.

Because the discount factor e^{-rt} appears in the problem above, it is convenient to use the current value formulation with the current value Hamiltonian

$$H^c(t, x, u, \lambda) = \lambda_0 f(t, x, u) + \lambda g(t, x, u)$$

and with λ as the current value shadow price.

THEOREM 9.11.1 (SUFFICIENT CONDITIONS WITH AN INFINITE HORIZON)

Suppose that an admissible pair $(x^*(t), u^*(t))$ for problem (1), with or without terminal condition (2), satisfies the following conditions for some $\lambda(t)$ for all $t \geq t_0$, with $\lambda_0 = 1$:

- (a) $u^*(t)$ maximizes $H^c(t, x^*(t), u, \lambda(t))$ w.r.t. $u \in U$
- (b) $\dot{\lambda}(t) - r\lambda = -\partial H^c(t, x^*(t), u^*(t), \lambda(t)) / \partial x$
- (c) $H^c(t, x, u, \lambda(t))$ is concave w.r.t. (x, u)
- (d) $\lim_{t \rightarrow \infty} \lambda(t) e^{-rt} [x(t) - x^*(t)] \geq 0$ for all admissible $x(t)$

Then $(x^*(t), u^*(t))$ is optimal.

Proof: For any admissible pair $(x(t), u(t))$ and for all $t \geq t_0$, define

$$D_u(t) = \int_{t_0}^t f(\tau, x^*(\tau), u^*(\tau)) e^{-r\tau} d\tau - \int_{t_0}^t f(\tau, x(\tau), u(\tau)) e^{-r\tau} d\tau = \int_{t_0}^t (f^* - f) e^{-r\tau} d\tau$$

in simplified notation. Now, $f^* = (H^c)^* - \lambda g^* = (H^c)^* - \lambda \dot{x}^*$ and $f = H^c - \lambda \dot{x}$, so

$$D_u(t) = \int_{t_0}^t [(H^c)^* - H^c] e^{-r\tau} d\tau + \int_{t_0}^t \lambda e^{-r\tau} (\dot{x} - \dot{x}^*) d\tau$$

By concavity of H^c , one has

$$\begin{aligned} (H^c)^* - H^c &\geq -\frac{\partial (H^c)^*}{\partial x} (x - x^*) + \frac{\partial (H^c)^*}{\partial u} (u^* - u) \\ &= (\dot{\lambda} - r\lambda)(x - x^*) + \frac{\partial (H^c)^*}{\partial u} (u^* - u) \end{aligned}$$

so

$$D_u(t) \geq \int_{t_0}^t e^{-r\tau} [(\dot{\lambda} - r\lambda)(x - x^*) + \lambda(\dot{x} - \dot{x}^*)] d\tau + \int_{t_0}^t \frac{\partial (H^c)^*}{\partial u} (u^* - u) e^{-r\tau} d\tau$$

As in the proof of Theorem 9.7.1, we see that the second integral is ≥ 0 and so

$$D_u(t) \geq \int_{t_0}^t \frac{d}{d\tau} [e^{-r\tau} \lambda(\tau)(x(\tau) - x^*(\tau))] d\tau = \left[e^{-rt} \lambda(t)(x(t) - x^*(t)) \right]_{t_0}^t$$

The contribution from the lower limit of integration is 0 because $x^*(t_0) - x(t_0) = x_0 - x_0 = 0$, so $D_u(t) \geq e^{-rt} \lambda(t)(x(t) - x^*(t))$. Passing to the limit as $t \rightarrow \infty$ in this inequality and using (d), one concludes that $(x^*(t), u^*(t))$ is optimal. ■

NOTE 2 Condition (d) in Theorem 9.11.1 is well known in the economics literature, but is often not properly checked. Note that the inequality (d) must be shown for *all* admissible $x(t)$, which is often problematic. Suppose for example that $\lim_{t \rightarrow \infty} \lambda(t)e^{-rt} \geq 0$, $\lim_{t \rightarrow \infty} \lambda(t)e^{-rt}x^*(t) = 0$, and $x(t) \geq 0$ for all t . Do these conditions ensure that (d) is satisfied? The answer is no, unless one is also sure that $\lambda(t) \geq 0$ for all (large enough) t . For a counterexample consider what happens when $\lambda(t) = -1$, $r = 1$, $x(t) = e^t$, and $x^*(t) = 1$. Then $\lambda(t)e^{-t}[x(t) - x^*(t)] = -e^{-t}(e^t - 1) = e^{-t} - 1 \rightarrow -1$ as $t \rightarrow \infty$.

NOTE 3 Suppose the terminal condition is $\lim_{t \rightarrow \infty} x(t) \geq x_1$. Rewrite the product in Theorem 9.11.1(d) as

$$\lambda(t)e^{-rt}(x(t) - x_1) + \lambda(t)e^{-rt}(x_1 - x^*(t)) \quad (*)$$

We claim that, provided the following three conditions are all satisfied, then condition (d) is satisfied:

- (A) $\lim_{t \rightarrow \infty} \lambda(t)e^{-rt}(x_1 - x^*(t)) \geq 0$;
- (B) there exists a number M such that $|\lambda(t)e^{-rt}| \leq M$; for all $t \geq t_0$;
- (C) there exists a number t' such that $\lambda(t) \geq 0$ for all $t \geq t'$.

Because of (A), in order to prove (d), it suffices to show that the first term in (*) tends to a number ≥ 0 . If $\lim_{t \rightarrow \infty} x(t) = x_1$, then $x(t) - x_1$ tends to 0 as t tends to ∞ , so because of (B), the first term in (*) tends to 0. If $\lim_{t \rightarrow \infty} x(t) > x_1$, then $x(t) - x_1 > 0$ for t sufficiently large. Then, because of (C), $\lambda(t)e^{-rt}(x(t) - x_1)$ tends to a number ≥ 0 . We conclude that, if (A)–(C) are all satisfied for all admissible pairs, then (d) holds. $\square$

NOTE 4 Suppose that we introduce additional conditions on $x(t)$ in Problem (1). Then the inequality in Theorem 9.11.1(d) need only hold for pairs $(x(t), x^*(t))$ satisfying the additional conditions.

In particular, if it is required that $x(t) \geq x_1$ for all t , then it suffices to check conditions (A) and (C) in Note 3. This result is referred to as the **Malinvaud transversality condition**.

EXAMPLE 1 Consider the problem

$$\max \int_0^\infty -u^2 e^{-at} dt, \quad \dot{x} = ue^{-at}, \quad x(0) = 0, \quad \lim_{t \rightarrow \infty} x(t) \geq K, \quad u \in \mathbb{R}$$

The constants r , a , and K are positive, with $a > r/2$. Find the optimal solution.

Solution: The current value Hamiltonian is $H^c = -u^2 + \lambda ue^{-at}$, which is obviously concave in x and u . We find $\partial H^c / \partial x = 0$ and $\partial H^c / \partial u = -2u + \lambda e^{-at}$. It follows that $u^* = \frac{1}{2} \lambda e^{-at}$. The differential equation for λ is $\dot{\lambda} - r\lambda = -\partial H^c / \partial x = 0$, with the solution $\lambda = Ae^{rt}$, where A is a constant. Thus $u^* = \frac{1}{2} Ae^{(r-a)t}$. The differential equation for x then becomes

$$\dot{x}^* = ue^{-at} = \frac{1}{2} Ae^{(r-2a)t}, \quad x^*(0) = 0, \quad \text{with the solution } x^*(t) = \frac{A}{2(2a-r)}(1 - e^{(r-2a)t})$$

Thus, $x^*(t)$ converges to $A/2(2a-r)$ as t approaches ∞ . So admissibility requires that $A/2(2a-r) \geq K$, or $A \geq 2K(2a-r)$. In particular, $A \geq 0$, so (C) in Note 3 is satisfied. To check condition (A) in Note 3 requires considering

$$\lambda(t)e^{-rt}(K - x^*(t)) = Ae^{rt}e^{-rt} \left[K - \frac{A}{2(2a-r)}(1 - e^{(r-2a)t}) \right]$$

which tends to $A(K - A/2(2a-r))$ as t tends to ∞ . We conclude from Note 4 that if we choose $A = 2K(2a-r)$, then all the conditions in Theorem 9.11.1 are satisfied and we have found an optimal solution. Note that $p(t) = \lambda e^{-rt} = 2K(2a-r)$, which does not tend to 0 as t tends to ∞ . Nor does $p(t)x^*(t)$. $\square$

EXAMPLE 2 Consider the following version of Example 8.5.3:

$$\max \int_0^\infty \frac{1}{1-\delta} [rA(t) + w - u(t)]^{1-\delta} e^{-\rho t} dt$$

$$\dot{A}(t) = u(t), \quad A(0) = A_0 > 0, \quad \lim_{t \rightarrow \infty} A(t) \geq -w/r, \quad u \in \mathbb{R}$$

Assume that $0 < \delta < 1$ and $0 < r < \rho$, and then solve the problem.

Solution: The current value Hamiltonian is $H^c = \frac{1}{1-\delta}(rA + w - u)^{1-\delta} + \lambda u$, and the differential equation for $\lambda(t)$ is

$$\dot{\lambda}(t) - \rho\lambda(t) = -\frac{\partial(H^c)^*}{\partial A} = -r[rA^*(t) + w - u^*(t)]^{-\delta} \quad (i)$$

The control function $u^*(t)$ maximizes

$$\varphi(u) = \frac{1}{1-\delta}[rA^*(t) + w - u]^{1-\delta} + \lambda u \quad \text{for } u \in \mathbb{R} \quad (ii)$$

Now the function H^c is concave in (A, u) , as the sum of a linear function (λu) and a concave function ($c^{1-\delta}/(1-\delta)$) of a linear function ($c = rA + w - u$). (Alternatively, look at the Hessian.) In particular, $\varphi(u)$ is concave in u , so $u^*(t)$ maximizes $\varphi(u)$ provided $\varphi'(u^*(t)) = 0$, i.e.

$$-[rA^*(t) + w - u^*(t)]^{-\delta} + \lambda(t) = 0, \quad \text{or } u^*(t) = rA^*(t) + w - \lambda(t)^{-1/\delta} \quad (iii)$$

Combining (i) and (iii), it follows that $\dot{\lambda}(t) - \rho\lambda(t) = -r\lambda(t)$, so $\dot{\lambda} = (\rho - r)\lambda(t)$, with solution

$$\lambda(t) = C_1 e^{(\rho-r)t} \quad (iv)$$

for some constant C_1 . Because $\dot{A}^* = u^*$, it follows that

$$\dot{A}^*(t) - rA^*(t) = w - C_1^{-1/\delta} e^{-at} \quad \text{where } a = (\rho - r)/\delta$$

The general solution of this linear differential equation is

$$A^*(t) = C_2 e^{rt} - w/r + C_1^{-1/\delta} e^{-at} / (a + r)$$

We must now find suitable values of the constants C_1 and C_2 . It seems reasonable to assume that $\lim_{t \rightarrow \infty} A^*(t) = -w/r$. This is only possible if $C_2 = 0$. Then C_1 is determined by the condition $A^*(0) = A_0$, which gives $A_0 = -w/r + C_1^{-1/5}/(a+r)$. Hence, we find that $C_1^{-1/5}/(a+r) = A_0 + w/r$. We therefore have the following candidate for an optimum:

$$A^*(t) = (A_0 + w/r)e^{-at} - w/r, \quad u^*(t) = -a(A_0 + w/r)e^{-at}, \quad \lambda(t) = \bar{\lambda}e^{-(\rho+r)t} \quad (v)$$

where $\bar{\lambda} = ((a+r)(A_0 + w/r))^{-5}$.

It remains to verify (d) in Theorem 9.11.1. According to Note 3 it suffices to show that conditions (A), (B), and (C) are satisfied. In our case (A) holds because

$$\lim_{t \rightarrow \infty} \lambda(t)(w/r + A^*(t)) = \bar{\lambda}(A_0 + w/r) \lim_{t \rightarrow \infty} e^{-(r+a)t} = 0$$

and (B) and (C) are evidently satisfied. Hence we have shown that $(A^*(t), u^*(t))$ solves the problem. $\blacksquare$

Many economists seem to believe that, for problems with an infinite horizon, no necessary transversality conditions are generally valid. This is wrong. But certain growth conditions are needed for such conditions to hold. A special result of this type is given in the next theorem. (See Seierstad and Sydsæter (1987), Section 3.9, Theorem 16 for a more general result. There is a misprint in that theorem: please replace $b > k$ by $b > (n-m)k$.)

THEOREM 9.11.2 (NECESSARY CONDITION FOR AN INFINITE HORIZON)

Assume that $(x^*(t), u^*(t))$ is optimal in problem (1), with no condition on the limiting behaviour of $x(t)$ as $t \rightarrow \infty$. Assume that $\int_0^\infty |f(t, x(t), u(t))|e^{-rt} dt < \infty$ for all admissible $(x(t), u(t))$. Suppose too that there exist positive constants A and k with $k < r$ such that

$$|\partial f(t, x, u^*(t))/\partial x| \leq A \quad \text{for all } x \quad (3)$$

and

$$\partial g(t, x, u^*(t))/\partial x \leq k \quad \text{for all } x \quad (4)$$

Then there exists a continuous function $\lambda(t)$ such that, with $\lambda_0 = 1$,

$$H^c(t, x^*(t), u, \lambda(t)) \leq H^c(t, x^*(t), u^*(t), \lambda(t)) \quad \text{for all } u \text{ in } U \quad (5)$$

and such that $\lambda(t) = \lim_{T \rightarrow \infty} \lambda(t, T)$, where $\lambda(t, T)$ is the solution of

$$\dot{\lambda} - r\lambda = -\partial H^c(t, x^*(t), u^*(t), \lambda)/\partial x, \quad \lambda(T, T) = 0 \quad (6)$$

NOTE 5 Suppose that, for each t , the set $G(t)$ contains all admissible $x(t)$, and that N and M are two positive numbers such that $|x_0| < N$ and $\sup_{|x| \leq N} |g(t, x, u^*(t))| \leq kM$. Then, for any t , (3) need only hold for x in $B(x^*(t); N + Me^{kt}) \cap G(t)$ and (4) need only hold for $|x| \geq N, x \in G(t)$.

PROBLEMS FOR SECTION 9.11

SM 1. Solve the problem

$$\max \int_0^\infty (\ln u)e^{-\rho t} dt, \quad \dot{x} = 0.1x - u, \quad x(0) = 10, \quad \lim_{t \rightarrow \infty} x(t) \geq 0, \quad u > 0$$

using Theorem 9.11.1 and Note 3.

SM 2. Find the only possible solution to the problem

$$\max \int_0^\infty x(2-u)e^{-t} dt, \quad \dot{x} = ux e^{-t}, \quad x(0) = 1, \quad x(\infty) \text{ is free}, \quad u \in [0, 1]$$

3. Compute the optimal value V of the objective function in Example 2. How does V change when ρ increases and when w increases? Show that $\partial V/\partial A_0 = \lambda(0)$.

SM 4. Solve the problem

$$\max \int_{-1}^\infty (x-u)e^{-t} dt, \quad \dot{x} = ux e^{-t}, \quad x(-1) = 0, \quad x(\infty) \text{ is free}, \quad u \in [0, 1]$$

9.12 Phase Diagrams

Consider the following problem

$$\max \int_0^t f(x, u)e^{-rt} dt, \quad \dot{x} = g(x, u), \quad x(t_0) = x_0, \quad u \in U \subseteq \mathbb{R} \quad (1)$$

with the standard end constraints, and with t_1 finite or ∞ . In this case the functions f and g do not depend explicitly on t . Nor, therefore, does the current value Hamiltonian H^c .

Suppose that $u = u(x, \lambda)$ maximizes $H^c = f(x, u) + \lambda g(x, u)$ w.r.t. u for $u \in U$. Replacing u by $u = u(x, \lambda)$ in the differential equations for x and λ gives

$$\begin{aligned} \dot{x} &= F(x, \lambda) \\ \dot{\lambda} &= G(x, \lambda) \end{aligned} \quad (2)$$

This is an *autonomous* system that is simpler to handle than one in which $\dot{x}$ and $\dot{\lambda}$ depend explicitly on t as well as on x and λ . In particular, *phase plane analysis* (see Section 6.7) can be used to shed light on the evolution of an autonomous system even when explicit solutions are not obtainable. Example 9.9.1 showed a simple case.

We study two examples.

EXAMPLE 1

Write down the system of equations (2), and draw a phase diagram for the problem

$$\max \int_0^\infty (x-u^2)e^{-0.1t} dt, \quad \dot{x} = -0.4x + u, \quad x(0) = 1, \quad x(\infty) \text{ is free}, \quad u \in (0, \infty)$$

Try to find the solution of the problem. (See Example 9.10.2.)

Solution: In this case the Hamiltonian $H^c(t, x, u, \lambda) = (x - u^2) + \lambda(-0.4x + u)$ is concave in (x, u) . The maximization of H^c w.r.t. u gives $u = 0.5\lambda$, assuming that λ is > 0 . (We justify this assumption later.) Hence, $\dot{x} = -0.4x + 0.5\lambda$. System (2) is here

$$\begin{aligned}\dot{x} &= -0.4x + 0.5\lambda, & x(0) &= 1 \\ \dot{\lambda} &= 0.5\lambda - 1\end{aligned}\quad (*)$$

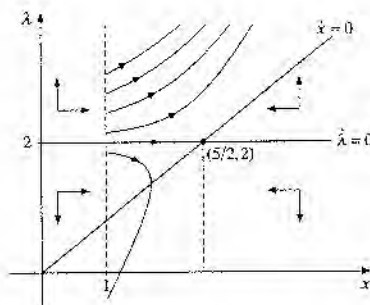


Figure 1 Phase diagram for system (*) in Example 1.

Figure 1 shows a phase diagram for (*). Any path $(x(t), \lambda(t))$ that solves the problem must start at some point on the vertical line $x = 1$, but no restrictions are imposed on $x(t)$ as $t \rightarrow \infty$. If we start above or below the line $\lambda = 2$, it appears that $(x(t), \lambda(t))$ will "wander off to infinity", which makes it difficult to satisfy requirement (d) in Theorem 9.11.1.

In fact, the general solution of (*) is $x^*(t) = \frac{5}{9}Ae^{0.5t} + \frac{5}{2} - (\frac{5}{9}A + \frac{3}{2})e^{-0.4t}$ and $\lambda(t) = Ae^{0.5t} + 2$. The expression we need to consider in Theorem 9.11.1(d) is the difference of the two terms $\lambda(t)e^{-0.1t}x(t)$ and $\lambda(t)e^{-0.1t}x^*(t)$. For large values of t , the latter product is dominated by the term $\frac{5}{9}A^2e^{0.9t}$, which tends to infinity as t tends to infinity when $A \neq 0$; that does not seem promising. Alternatively, it approaches 0 as t approaches infinity if $A = 0$ (then $\lambda = 2$), and then the product is equal to $5e^{-0.1t} - 3e^{-0.5t}$, which does approach 0 as t approaches infinity. It is easy to see that $x(t) > 0$ for all $t \geq 0$, so $\lambda(t)e^{-0.1t}x(t)$ is > 0 for all $t \geq 0$. It follows that condition (d) in Theorem 9.11.1 is satisfied, and $x^*(t) = -\frac{3}{2}e^{-0.4t} + \frac{5}{2}$ is therefore optimal.

Coming back to the phase diagram, if we start at the point $(x, \lambda) = (1, 2)$, then $\lambda(t) \equiv 2$, while $x(t)$ converges to the value $\frac{5}{2}$, which is the x -coordinate of the point of intersection between the curves $\dot{\lambda} = 0$ and $\dot{x} = 0$. The phase diagram therefore suggests the optimal solution to the problem.

The point $(\frac{5}{2}, 2)$ is an equilibrium point for system (*). Let us see what Theorem 6.9.1 says about this equilibrium point. Defining $f(x, \lambda) = -0.4x + 0.5\lambda$ and $g(x, \lambda) = 0.5\lambda - 1$, we find that the determinant of the Jacobian matrix in Theorem 6.9.1 is

$$\begin{vmatrix} -0.4 & 0.5 \\ 0 & 0.5 \end{vmatrix} = -0.2 < 0$$

so $(\frac{5}{2}, 2)$ is a saddle point.

EXAMPLE 2

Consider an economy with capital stock $K = K(t)$ and production per unit of time $Y = Y(t)$, where $Y = aK - bK^2$, with a and b as positive constants. Consumption is $C > 0$, whereas $Y - C = aK - bK^2 - C$ is investment. Over the period $[0, \infty)$, the objective is to maximize total discounted utility. In particular we consider the problem

$$\int_0^\infty \frac{1}{1-v} C^{1-v} e^{-rt} dt, \quad \dot{K} = aK - bK^2 - C, \quad K(0) = K_0 > 0$$

where $a > r > 0$ and $v > 0$, with C as the control variable. We require that

$$K(t) \geq 0 \text{ for all } t$$

The current value Hamiltonian is $H^c = \frac{1}{1-v} C^{1-v} + \lambda(aK - bK^2 - C)$. An interior maximum of H^c requires $\partial H^c / \partial C = 0$, i.e.

$$C^{-v} = \lambda \quad (i)$$

The differential equation for $\lambda = \lambda(t)$ is $\dot{\lambda} = -\lambda(a - 2bK) + r\lambda$, or

$$\dot{\lambda} = \lambda(r - a + 2bK) = 2b\lambda \left(K - \frac{a-r}{2b} \right) \quad (ii)$$

Now (i) implies that $C = \lambda^{-1/v}$, which inserted into the differential equation for K yields

$$\dot{K} = aK - bK^2 - \lambda^{-1/v} \quad (iii)$$

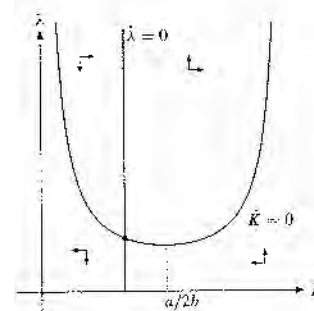


Figure 2

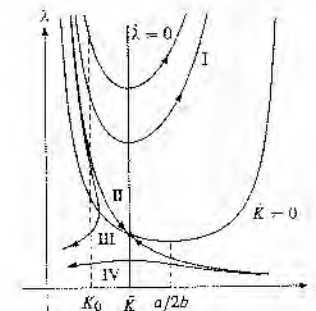


Figure 3

Figure 2 presents a phase diagram for the system given by (ii) and (iii). We see that $\dot{K} = 0$ for $\lambda = (aK - bK^2)^{-v}$, with $v > 0$. Here $z = aK - bK^2$ represents a concave parabola with $z = 0$ for $K = 0$ and for $K = a/b$. For $z = 0$, one has $\lambda = \infty$. The graph of $\dot{K} = 0$ is symmetrical about $K = a/2b$. Note that $\lambda = 0$ when $K = (a-r)/2b$, which gives a straight line parallel to the λ -axis. Because $0 < (a-r)/2b < a/2b$, the graph of $\dot{\lambda} = 0$ will be as suggested in the figure. The equilibrium point $(\bar{K}, \bar{\lambda})$ is given by $\bar{K} = (a-r)/2b$, $\bar{\lambda} = [(a^2 - r^2)/4b]^{-v}$.

In Fig. 2 the $K\lambda$ -plane is divided into four parts. The arrows indicate the directions of the integral curves in each of these four parts. From (ii) we see that $K > (a-r)/2b$ implies $\dot{\lambda} > 0$, whereas $K < (a-r)/2b$ implies $\dot{\lambda} < 0$. Also, the right-hand side of (iii), $aK - bK^2 - \lambda^{-1/v}$, increases as λ increases for each fixed K , so that $\dot{K} > 0$ above the curve $\dot{K} = 0$, and $\dot{K} < 0$ below this curve.

Figure 3 shows some integral curves that $(K(t), \lambda(t))$ could follow as t increases. In this figure we have assumed that $K_0 < \bar{K}$. Of particular interest are paths that start at $K = K_0$, but other curves, which start with other values of K , are also drawn. Note that, although K_0 is known, the quantity $\lambda(0)$ must be regarded as an unknown parameter. In this particular problem $\lambda(0)$ can be determined as follows: If $\lambda(0)$ is large, the point $(K(t), \lambda(t))$ starts high up on the line $K = K_0$ and moves along a curve like that marked I in Fig. 3. If $\lambda(0)$ is small, then $(K(t), \lambda(t))$ starts low down on the line $K = K_0$ and moves along a curve like III in the figure. If $\lambda(0)$ is even smaller, and $(K_0, \lambda(0))$ lies below the curve $\dot{K} = 0$, then $(K(t), \lambda(t))$ moves steadily "south-west", like curve IV. At some point on the line $K = K_0$, continuity suggests that there should be some particular value $\lambda^*(0)$ of $\lambda(0)$ such that the resultant curve is of type II, which converges to the stationary point $(\bar{K}, \bar{\lambda})$.

Here is a more precise argument: Curve I was obtained using a high initial value for $\lambda(0)$. Along curve I the point $(K(t), \lambda(t))$ moves down to the right until it reaches a minimum point where it crosses the line $\dot{\lambda} = 0$. Let $\lambda(0)$ decrease. Then curve I shifts downwards. Its minimum point on the line $\dot{\lambda} = 0$ will then shift downwards towards the equilibrium point $(\bar{K}, \bar{\lambda})$. Actually, $\lambda^*(0)$ is precisely that value of $\lambda(0)$ which makes this minimum occur at the point $(\bar{K}, \bar{\lambda})$. This initial value $\lambda^*(0)$ leads to a special path $(K^*(t), \lambda^*(t))$. Both $\dot{K}^*(t)$ and $\dot{\lambda}^*(t)$ approach zero as $t \rightarrow \infty$. Note that $(K^*(t), \lambda^*(t))$ is never equal to $(\bar{K}, \bar{\lambda})$ for any finite t , but $(K^*(t), \lambda^*(t)) \rightarrow (\bar{K}, \bar{\lambda})$ as $t \rightarrow \infty$.

So far we have argued that the conditions of the maximum principle are satisfied along a curve $(K^*(t), \lambda^*(t))$ of type II in Fig. 3, where $K^*(t) \rightarrow \bar{K}$ and $\lambda^*(t) \rightarrow \bar{\lambda}$ as $t \rightarrow \infty$. Let us prove that this candidate solution is optimal.

The current value Hamiltonian H^c is concave as a function of (K, C) . With $\lambda^*(t)$ given and $C^*(t) = \lambda^*(t)^{-1/v}$, the first-order condition for a maximum of H^c is satisfied, and because H^c is concave in C , it reaches a maximum at $C^*(t)$. Moreover, because $\lambda^*(t)e^{-rt}K(t) \geq 0$, $e^{-rt} \rightarrow 0$, and $(K^*(t), \lambda^*(t)) \rightarrow (\bar{K}, \bar{\lambda})$ as $t \rightarrow \infty$, it follows that

$$\lim_{t \rightarrow \infty} \lambda^*(t)e^{-rt}[K(t) - K^*(t)] \geq 0 \quad \text{for all admissible } K(t)$$

This verifies all the sufficient conditions in Theorem 9.11.1, so $(K^*(t), C^*(t))$ is optimal.

Any solution of the system (ii) and (iii) will depend on K_0 and on $\lambda(0) = \lambda^0$, so it can be denoted by $K(t) = K(t; K_0, \lambda^0)$ and $\lambda(t) = \lambda(t; K_0, \lambda^0)$. In this problem, K_0 is given, whereas λ^0 is determined by the requirement that $\lim_{t \rightarrow \infty} \lambda(t; K_0, \lambda^0) = \bar{\lambda}$. Figure 3 actually shows two curves of type II that converge to $(\bar{K}, \bar{\lambda})$. The alternative solution of the differential equations converges to $(\bar{K}, \bar{\lambda})$ from the "southeast". This path does not solve the optimization problem, however, because it must start from a wrong value of K at time $t = 0$. (It does solve the problem when $K_0 > \bar{K}$, however.)

The equilibrium point $(\bar{K}, \bar{\lambda}) = ((a-r)/2b, [(a^2-r^2)/4b]^{-1/v})$ is an example of a saddle point (see Section 6.9). We show this by applying Theorem 6.9.1. To do so, define the functions $f(K, \lambda) = aK - bK^2 - \lambda^{-1/v}$ and $g(K, \lambda) = 2b\lambda(K - (a-r)/2b)$ corresponding

to the right-hand sides of (iii) and (ii) respectively. Then at the point $(\bar{K}, \bar{\lambda})$ one has $\partial f/\partial K = a - 2b\bar{K} = r$, $\partial f/\partial \lambda = (1/v)\bar{\lambda}^{-1/v-1}$, $\partial g/\partial K = 2b\bar{\lambda}$ and $\partial g/\partial \lambda = 2b(\bar{K} - (a-r)/2b) = 0$. The determinant of the matrix A in Theorem 6.9.1 is therefore

$$\begin{vmatrix} r & (1/v)\bar{\lambda}^{-1/v-1} \\ 2b\bar{\lambda} & 0 \end{vmatrix} = -\frac{2b}{v}\bar{\lambda}^{-1/v} < 0$$

This confirms that $(\bar{K}, \bar{\lambda})$ really is a saddle point.

PROBLEMS FOR SECTION 9.12

SM 1. (a) Consider the problem

$$\max \int_0^{\infty} (ax - \frac{1}{2}u^2)e^{-rt} dt, \quad \dot{x} = -bx + u, \quad x(0) = x_0, \quad x(\infty) \text{ free}, \quad u \in \mathbb{R}$$

where a , r , and b are all positive. Write down the current value Hamiltonian H^c for this problem, and determine the system (2). What is the equilibrium point?

(b) Draw a phase diagram for $(x(t), \lambda(t))$ and show that for the two solutions converging to the equilibrium point, $\lambda(t)$ must be a constant. Use sufficient conditions to solve the problem.

(c) Show that $\partial V/\partial x_0 = \lambda(0)$, where V is the optimal value function.

SM 2. In Example 9.2.2 we studied a problem closely related to

$$\max \int_0^T (-x^2 - \frac{1}{2}u^2)e^{-2t} dt, \quad \dot{x} = x + u, \quad x(0) = 1, \quad x(T) \geq 0, \quad u \in \mathbb{R}$$

Solve this problem in the case $T = \infty$. (Hint: $\lim_{t \rightarrow \infty} p(t) = 0$.)

SM 3. (a) Consider the problem

$$\max \int_0^T e^{-rt} \ln C(t) dt$$

$$\dot{K}(t) = AK(t)^\alpha - C(t), \quad K(0) = K_0, \quad K(T) = K_T$$

where the constants A and r are positive, and $\alpha \in (0, 1)$. Here $K(t)$ denotes the capital stock of an economy and the control variable $C(t)$ denotes consumption at time t . The horizon T is fixed and finite. Prove that if $K = K^*(t) > 0$ and $C = C^*(t) > 0$ solve the problem, then

$$(i) \quad \dot{K} = AK^\alpha - C \quad (ii) \quad \dot{C} = C(\alpha AK^{\alpha-1} - r)$$

(b) Suppose $A = 2$, $\alpha = 1/2$, and $r = 0.05$. Prove that the equilibrium is a saddle point. In Problem 6.7.4 you were asked to draw a phase diagram of the system.

(c) Indicate in the diagram for Problem 6.7.4 a possible integral curve for the case $K_0 = 100$ and $K_T = 600$. What is the solution when $K_0 = 100$ and $T = \infty$, with $K(t) > 0$ for all t ?

4. Consider the problem

$$\max_{u \in \mathbf{R}} \int_0^{\infty} \left[-(x-1)^2 - \frac{1}{2}u^2 \right] e^{-t} dt, \quad \dot{x} = x + u, \quad x(0) = \frac{1}{2}, \quad x(\infty) \text{ free}$$

- (a) Solve the problem qualitatively by a saddle point argument.
 (b) Find an explicit solution.

10

CONTROL THEORY
WITH MANY VARIABLES

To be sure, mathematics can be extended to any branch of knowledge, including economics, provided the concepts are so clearly defined as to permit accurate symbolic representation. That is only another way of saying that in some branches of discourse it is desirable to know what you are talking about.
 —D. MacDouglas (1956)

This chapter begins by extending the optimal control theory developed in the previous chapter to problems with several state and control variables. In Section 10.1 the main emphasis is on appropriate generalizations of results from Chapter 9. The explanations are brief because the essential motivation was given in Chapter 9. However, we give a proof of the Arrow sufficiency theorem in the case of several state and control variables.

Section 10.2 contains examples illustrating the theory.

Section 10.3 extends the infinite horizon theory of Section 9.11. In fact, the majority of the control models that appear in economics literature assume an infinite horizon. A good treatment of infinite horizon control theory with a large number of economic applications can be found in Weitzman (2003).

If neither the Mangasarian nor the Arrow concavity conditions are satisfied, then we need some assurance that the control problem has a solution. Even if the maximum principle generates a unique candidate for an optimal solution, there might still be no optimal solution. This difficulty is discussed in Section 10.4. The Filippov–Cesari theorem gives sufficient conditions for existence of an optimal control in a rather general control problem. The section ends with a formulation of precise sensitivity results in control theory. These are seldom spelled out except in specialized literature.

Section 10.5 offers a heuristic proof of the maximum principle, which, at least in the case of a free end, is close to a proper proof. Proving the maximum principle is much harder when there are terminal constraints on the state variables.

Section 10.6 has a short discussion of control problems with mixed constraints of the type $h(t, x, u) \geq 0$, and Section 10.7 considers pure state constraints that can be written in the vector form $h(t, x) \geq 0$. In fact, many of the control problems that economists have considered involve mixed and/or pure constraints. Finally, Section 10.8 briefly discusses some generalizations

10.1 Several Control and State Variables

Chapter 9 studied control problems with only one state and one control variable. In this section most of the results from Chapter 9 are generalized to control problems with an arbitrary number of state and control variables.

The **standard end-constrained problem** with n state variables and r controls is to find a pair of vector functions $(\mathbf{x}(t), \mathbf{u}(t)) = (x_1(t), \dots, x_n(t), u_1(t), \dots, u_r(t))$ defined on the fixed time interval $[t_0, t_1]$ in order to maximize

$$\int_{t_0}^{t_1} f(t, \mathbf{x}(t), \mathbf{u}(t)) dt \quad (1)$$

subject to the dynamic constraints

$$\begin{aligned} \frac{dx_1(t)}{dt} &= g_1(t, \mathbf{x}(t), \mathbf{u}(t)) \\ &\dots\dots\dots \text{or} \quad \dot{\mathbf{x}} = \mathbf{g}(t, \mathbf{x}(t), \mathbf{u}(t)) \quad (2) \\ \frac{dx_n(t)}{dt} &= g_n(t, \mathbf{x}(t), \mathbf{u}(t)) \end{aligned}$$

the initial conditions

$$x_i(t_0) = x_i^0, \quad i = 1, \dots, n \quad (\mathbf{x}^0 = (x_1^0, \dots, x_n^0) \text{ is a given point in } \mathbb{R}^n) \quad (3)$$

the terminal conditions

$$\begin{aligned} \text{(a)} \quad x_i(t_1) &= x_i^1, & i &= 1, \dots, l \\ \text{(b)} \quad x_i(t_1) &\geq x_i^1, & i &= l+1, \dots, m \\ \text{(c)} \quad x_i(t_1) &\text{ free,} & i &= m+1, \dots, n \end{aligned} \quad (4)$$

and the control variable restrictions

$$\mathbf{u}(t) = (u_1(t), \dots, u_r(t)) \in U \subseteq \mathbb{R}^r \quad (U \text{ is a given set in } \mathbb{R}^r) \quad (5)$$

In (2) the system of differential equations is also written as a vector differential equation, where $\dot{\mathbf{x}} = (dx_1/dt, dx_2/dt, \dots, dx_n/dt)$, and $\mathbf{g} = (g_1, g_2, \dots, g_n)$ is a vector function. These n equations describe the rate of growth of each state variable. In (4), the numbers t_1 and x_i^1 are fixed.

The pair $(\mathbf{x}(t), \mathbf{u}(t))$ is **admissible** if $u_1(t), \dots, u_r(t)$ are all piecewise continuous, $\mathbf{u}(t)$ takes values in U and $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))$ is the corresponding continuous and piecewise differentiable vector function that satisfies (2), (3), and (4). The functions f and $g_1, \dots, g_n$ and their partial derivatives w.r.t. the variables x_i are assumed to be continuous in all the $n+r+1$ variables.

By analogy with the single variable problem in Section 9.4, associate an adjoint function $p_i(t)$ with each state variable $x_i(t)$. The Hamiltonian $H = H(t, \mathbf{x}, \mathbf{u}, \mathbf{p})$, with $\mathbf{p} = (p_1, \dots, p_n)$, is then defined by

$$H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) = p_0 f(t, \mathbf{x}, \mathbf{u}) + \mathbf{p} \cdot \mathbf{g}(t, \mathbf{x}, \mathbf{u}) = p_0 f(t, \mathbf{x}, \mathbf{u}) + \sum_{i=1}^n p_i g_i(t, \mathbf{x}, \mathbf{u}) \quad (6)$$

The Maximum Principle

The maximum principle for the problem gives *necessary* conditions for optimality, but the conditions are far from sufficient. For a proof see Fleming and Rishel (1975).

THEOREM 10.1.1: THE MAXIMUM PRINCIPLE: STANDARD END CONSTRAINTS

Suppose that $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is an optimal pair for the standard end-constrained problem (1)–(5). Then there exist a constant p_0 , with $p_0 = 0$ or $p_0 = 1$, and a continuous and piecewise differentiable adjoint vector function $\mathbf{p}(t) = (p_1(t), \dots, p_n(t))$, such that for all t in $[t_0, t_1]$, one has $(p_0, \mathbf{p}(t)) \neq (0, \mathbf{0})$, and:

(A) The control function $\mathbf{u}^*(t)$ maximizes the Hamiltonian $H(t, \mathbf{x}^*(t), \mathbf{u}, \mathbf{p}(t))$ for $\mathbf{u} \in U$, i.e.

$$H(t, \mathbf{x}^*(t), \mathbf{u}, \mathbf{p}(t)) \leq H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t)) \quad \text{for all } \mathbf{u} \text{ in } U \quad (7)$$

(B) Wherever $\mathbf{u}^*(t)$ is continuous, the adjoint functions satisfy

$$\dot{p}_i(t) = -\frac{\partial H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t))}{\partial x_i}, \quad i = 1, \dots, n \quad (8)$$

(C) Corresponding to the terminal conditions (a), (b), and (c) in (4), one has the respective transversality conditions:

$$\begin{aligned} \text{(a')} \quad p_i(t_1) &\text{ no condition,} & i &= 1, \dots, l \\ \text{(b')} \quad p_i(t_1) &\geq 0 \quad (p_i(t_1) = 0 \text{ if } x_i^*(t_1) > x_i^1), & i &= l+1, \dots, m \\ \text{(c')} \quad p_i(t_1) &= 0, & i &= m+1, \dots, n \end{aligned} \quad (9)$$

NOTE 1 For any inequality in (4)(b) that is reversed, the corresponding inequality in (9)(b') is reversed as well.

NOTE 2 One can show the following additional properties:

(a) The Hamiltonian

$$H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t)) \quad \text{is continuous for all } t \quad (10)$$

(b) If the partial derivatives $\partial f / \partial t$ and $\partial g_i / \partial t$, $i = 1, \dots, n$, exist and are continuous, then

$$\frac{d}{dt} H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t)) = \frac{\partial H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t))}{\partial t} \quad (11)$$

at all points of continuity of $\mathbf{u}^*(t)$. (See Problem 2.)

(c) Moreover,

$$U \text{ convex and } H \text{ strictly concave in } \mathbf{u} \implies \mathbf{u}^*(t) \text{ continuous for all } t \quad (12)$$

NOTE 3 Suppose that the terminal condition is that all state variables $x_i(t_1)$ are free, for $i = 1, \dots, n$. In this case, (9)(c') yields $p(t_1) = 0$, and then $p_0 = 1$.

NOTE 4 The adjoint variables in Theorem 10.1.1 can be given price interpretations corresponding to those in Section 9.6 for the case $n = r = 1$. Indeed, let $\mathbf{x}^1 = (x_1^1, \dots, x_m^1)$ and define the value function V associated with the standard problem as

$$V(\mathbf{x}^0, \mathbf{x}^1, t_0, t_1) = \max \left\{ \int_{t_0}^{t_1} f(t, \mathbf{x}(t), \mathbf{u}(t)) dt : (\mathbf{x}(t), \mathbf{u}(t)) \text{ admissible} \right\} \quad (13)$$

Then, under precise assumptions stated in Section 10.4, for $i = 1, 2, \dots, n$,

$$\frac{\partial V}{\partial x_i^0} = p_i(t_0), \quad \frac{\partial V}{\partial x_i^1} = -p_i(t_1), \quad \frac{\partial V}{\partial t_0} = -H^*(t_0), \quad \frac{\partial V}{\partial t_1} = H^*(t_1) \quad (14)$$

where H^* denotes the Hamiltonian evaluated along the optimal path.

Sufficient Conditions

The simplest general sufficiency theorem is the following:

THEOREM 10.1.2 (MANGASARIAN)

Consider the standard end-constrained problem (1)–(5) with U convex, and suppose that the partial derivatives $\partial f / \partial u_j$ and $\partial g_i / \partial u_j$ all exist and are continuous. If the pair $(\mathbf{x}^*(t), \mathbf{u}^*(t))$, together with a continuous and piecewise differentiable adjoint function $\mathbf{p}(t)$, satisfies all the conditions in Theorem 10.1.1 with $p_0 = 1$, and if

$$H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}(t)) \text{ is concave in } (\mathbf{x}, \mathbf{u}) \text{ for all } t \text{ in } [t_0, t_1] \quad (15)$$

then $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ solves the problem.

If the function $H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}(t))$ is strictly concave in $(\mathbf{x}, \mathbf{u})$, then $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is the unique solution to the problem.

NOTE 5 Because a sum of concave functions is concave, the concavity condition (15) is satisfied if f and $p_1 g_1, \dots, p_n g_n$ are all concave in $(\mathbf{x}, \mathbf{u})$.

At this point the reader might want to study Example 10.2.1 and then do Problem 10.2.1.

The proof of Theorem 10.1.2 is very similar to the proof of Theorem 9.7.1, so we skip it. Instead, we take a closer look at a generalization of Mangasarian's theorem due to Arrow (see Arrow and Kurz (1970)). Define the maximized Hamiltonian as

$$\hat{H}(t, \mathbf{x}, \mathbf{p}) = \max_{\mathbf{u} \in U} H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) \quad (16)$$

assuming that the maximum value is attained. Then the appropriate generalization of Theorem 9.7.2 is this:

THEOREM 10.1.3 (ARROW)

Suppose that $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is an admissible pair in the standard end-constrained problem (1)–(5) that, together with the continuous and piecewise differentiable adjoint (vector) function $\mathbf{p}(t)$, satisfies all the conditions in Theorem 10.1.1 with $p_0 = 1$. Suppose further that

$$\hat{H}(t, \mathbf{x}, \mathbf{p}(t)) \text{ is concave in } \mathbf{x} \text{ for all } t \text{ in } [t_0, t_1] \quad (17)$$

Then $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ solves the problem.

Proof: Let $(\mathbf{x}, \mathbf{u}) = (\mathbf{x}(t), \mathbf{u}(t))$ be an arbitrary admissible pair. We must show that $D_{\mathbf{u}} = \int_{t_0}^{t_1} f(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) dt - \int_{t_0}^{t_1} f(t, \mathbf{x}(t), \mathbf{u}(t)) dt \geq 0$. Let us simplify the notation by letting f^* denote $f(t, \mathbf{x}^*(t), \mathbf{u}^*(t))$, f denote $f(t, \mathbf{x}(t), \mathbf{u}(t))$, H^* denote $H(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t))$, etc. As in the proof of Theorem 9.7.1, it is easy to see that

$$D_{\mathbf{u}} = \int_{t_0}^{t_1} (H^* - H) dt + \int_{t_0}^{t_1} \mathbf{p}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t)) dt \quad (i)$$

Integration by parts yields

$$\begin{aligned} \int_{t_0}^{t_1} \mathbf{p}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t)) dt &= \left| \int_{t_0}^{t_1} \mathbf{p}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t)) dt - \int_{t_0}^{t_1} \dot{\mathbf{p}}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t)) dt \right. \\ &\quad \left. \geq - \int_{t_0}^{t_1} \dot{\mathbf{p}}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t)) dt \right. \end{aligned} \quad (ii)$$

To explain the last inequality, note first that because $\mathbf{x}(t_0) = \mathbf{x}^*(t_0)$ we get

$$\int_{t_0}^{t_1} \mathbf{p}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t)) dt = \mathbf{p}(t_1) \cdot (\mathbf{x}(t_1) - \mathbf{x}^*(t_1)) = \sum_{i=1}^n p_i(t_1)(x_i(t_1) - x_i^*(t_1)) \quad (iii)$$

We claim that this sum is ≥ 0 , which will imply the inequality in (ii). In fact, for $i = 1, 2, \dots, l$, we have $x_i(t_1) = x_i^*(t_1) = x_i^1$, so the corresponding terms are 0. Also, for $i = m+1, \dots, n$, the corresponding terms in the sum in (iii) are 0 because by (9)(c'), $p_i(t_1) = 0$. If $i = l+1, \dots, m$ and $x_i^*(t_1) > x_i^1$, the corresponding terms are 0 because by (9)(b'), $p_i(t_1) = 0$. Finally, if $x_i^*(t_1) = x_i^1$, then $x_i(t_1) - x_i^*(t_1) \geq 0$ and, by (9)(b'), $p_i(t_1) \geq 0$, so the corresponding terms are ≥ 0 . All in all, this proves that the sum in (iii) is ≥ 0 .

To proceed, note that by the definition of $\hat{H}$,

$$H^* = \hat{H}^* \quad \text{and} \quad H \leq \hat{H} \quad (iv)$$

It follows from (i)–(iv) that

$$D_{\mathbf{u}} \geq \int_{t_0}^{t_1} [\hat{H}^* - \hat{H} - \dot{\mathbf{p}}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t))] dt \quad (v)$$

But (8) implies that $-\dot{\mathbf{p}}(t)$ is the (partial) gradient vector $\nabla_{\mathbf{x}} H^*$, which must equal $\nabla_{\mathbf{x}} \hat{H}^*$ by the envelope theorem (Theorem 3.10.2). Because $\hat{H}$ is concave w.r.t. $\mathbf{x}$, it follows from Theorem 2.4.1 that

$$\hat{H} - \hat{H}^* \leq -\dot{\mathbf{p}}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t)), \quad \text{or} \quad \hat{H}^* - \hat{H} \geq \dot{\mathbf{p}}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t))$$

This implies that the integral on the right-hand side of (v) is nonnegative for all t in $[t_0, t_1]$, so $D_{\mathbf{u}} \geq 0$ as required. ■

NOTE 6 The result in Problem 3 shows that condition (15) implies (17). Thus Theorem 10.1.3 generalizes Theorem 10.1.2.

NOTE 7 Suppose that in the standard end-constrained problem (1)–(5) one requires that $\mathbf{x}(t) \in A(t)$ for all t , where $A(t)$ for each t is a given convex set in $\mathbb{R}^n$. Suppose also that $\mathbf{x}^*(t)$ is an interior point of $A(t)$ for each t . The conclusion in Theorem 10.1.3 is then valid, and $\mathbf{x} \mapsto \hat{H}(t, \mathbf{x}, \mathbf{p}(t))$ need only be concave in the set $A(t)$.

Variable Final Time

Consider problem (1)–(5) with variable final time t_1 . The problem is among all control functions $\mathbf{u}(t)$ that during the time interval $[t_0, t_1]$ steer the system from $\mathbf{x}^0$ along a time path satisfying (2) to a point where the boundary conditions in (4) are satisfied, to find one which maximizes the integral in (1). The time t_1 at which the process stops is not fixed, as the different admissible control functions can be defined on different time intervals. Theorem 9.8.1 has then the following immediate generalization:

THEOREM 10.1.4 (THE MAXIMUM PRINCIPLE WITH A VARIABLE FINAL TIME)

Suppose that $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is an admissible pair defined on $[t_0, t_1^*]$ that solves problem (1)–(5) with t_1 free ($t_1 \in (t_0, \infty)$). Then all the conditions in the maximum principle (Theorem 10.1.1) are satisfied on $[t_0, t_1^*]$ and, in addition,

$$H(t_1^*, \mathbf{x}^*(t_1^*), \mathbf{u}^*(t_1^*), \mathbf{p}(t_1^*)) = 0 \quad (18)$$

For a proof, see Hestenes (1966). Neither the Mangasarian nor the Arrow theorem applies to variable final time problems. For sufficiency results, see Seierstad and Sydsæter (1987).

Current Value Formulations with Scrap Values

The theorems in Section 9.9 on current value formulations of optimal control problems with scrap value functions can easily be generalized to the following problem involving several state and control variables:

$$\max_{\mathbf{u} \in U \subseteq \mathbb{R}^r} \left\{ \int_{t_0}^{t_1} f(t, \mathbf{x}, \mathbf{u}) e^{-rt} dt + S(\mathbf{x}(t_1)) e^{-rt_1} \right\}, \quad \dot{\mathbf{x}}(t) = \mathbf{g}(t, \mathbf{x}, \mathbf{u}), \quad \mathbf{x}(t_0) = \mathbf{x}^0 \quad (19)$$

$$\begin{aligned} \text{(a)} \quad & x_i(t_1) = x_i^1, \quad i = 1, \dots, l \\ \text{(b)} \quad & x_i(t_1) \geq x_i^1, \quad i = l+1, \dots, m \\ \text{(c)} \quad & x_i(t_1) \text{ free}, \quad i = m+1, \dots, n \end{aligned} \quad (20)$$

Here r denotes a discount factor. The current value Hamiltonian is by definition

$$H^c(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}) = \lambda_0 f(t, \mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda} \cdot \mathbf{g}(t, \mathbf{x}, \mathbf{u}) \quad (21)$$

THEOREM 10.1.5 (THE MAXIMUM PRINCIPLE)

Suppose that $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is an optimal pair for the problem (19)–(20). Then there exist a continuous and piecewise continuously differentiable vector function $\boldsymbol{\lambda}(t) = (\lambda_1(t), \dots, \lambda_n(t))$ and a constant λ_0 , with $\lambda_0 = 0$ or $\lambda_0 = 1$, such that $(\lambda_0, \boldsymbol{\lambda}(t)) \neq (0, \mathbf{0})$ for all t in $[t_0, t_1]$, and such that:

(A) For all t in $[t_0, t_1]$,

$$\mathbf{u} = \mathbf{u}^*(t) \text{ maximizes } H^c(t, \mathbf{x}^*(t), \mathbf{u}, \boldsymbol{\lambda}(t)) \text{ for } \mathbf{u} \in U \quad (22)$$

(B) Wherever $\mathbf{u}^*(t)$ is continuous,

$$\dot{\lambda}_i(t) - r\lambda_i(t) = -\frac{\partial H^c(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \boldsymbol{\lambda}(t))}{\partial x_i}, \quad i = 1, \dots, n \quad (23)$$

(C) Finally, corresponding to the terminal conditions (20)(a), (b), and (c), one has the respective transversality conditions:

$$\begin{aligned} \text{(a')} \quad & \lambda_i(t_1) \text{ no condition}, \quad i = 1, \dots, l \\ \text{(b')} \quad & \lambda_i(t_1) \geq \lambda_0 \frac{\partial S(\mathbf{x}^*(t_1))}{\partial x_i} \quad (\text{with } = \text{ if } x_i^*(t_1) > x_i^1), \quad i = l+1, \dots, m \\ \text{(c')} \quad & \lambda_i(t_1) = \lambda_0 \frac{\partial S(\mathbf{x}^*(t_1))}{\partial x_i}, \quad i = m+1, \dots, n \end{aligned} \quad (24)$$

THEOREM 10.1.6 (SUFFICIENT CONDITIONS - ARROW)

The conditions in Theorem 10.1.5 are sufficient (with $\lambda_0 = 1$) if

$$\hat{H}^c(t, \mathbf{x}, \boldsymbol{\lambda}(t)) = \max_{\mathbf{u} \in U} H^c(t, \mathbf{x}, \mathbf{u}, \boldsymbol{\lambda}(t)) \text{ is concave in } \mathbf{x} \quad (25)$$

and

$$S(\mathbf{x}) \text{ is concave in } \mathbf{x}. \quad (26)$$

The problems for this section are of a theoretical nature. Non-theoretical exercises are found at the end of the next section.

PROBLEMS FOR SECTION 10.1

1. Consider the variational problem with an integral constraint

$$\max_{\mathbf{x}} \int_{t_0}^{t_1} F(t, \mathbf{x}, \dot{\mathbf{x}}) dt, \quad \mathbf{x}(t_0) = \mathbf{x}^0, \quad \mathbf{x}(t_1) = \mathbf{x}^1, \quad \int_{t_0}^{t_1} G(t, \mathbf{x}, \dot{\mathbf{x}}) dt = K$$

Transform the problem to a control problem with one control variable ($\mathbf{u} = \dot{\mathbf{x}}$) and two state variables $\mathbf{x} = \mathbf{x}(t)$ and $\mathbf{y}(t) = \int_{t_0}^t G(\tau, \mathbf{x}(\tau), \dot{\mathbf{x}}(\tau)) d\tau$.

2. Prove (11) assuming that $u^*(t)$ is differentiable and $u^*(t)$ belongs to the interior of U . (Hint: Differentiate $H(t, x^*(t), u^*(t), p(t))$ totally w.r.t. t .)
3. Let S and U be convex sets in $\mathbb{R}^n$ and $\mathbb{R}^r$, respectively, and let $F(x, u)$ be a real-valued concave function of (x, u) , $x \in S$, $u \in U$. Define

$$f(x) = \max_{u \in U} F(x, u) \quad (*)$$

where we assume that the maximum value exists for each $x \in S$. Prove that f is concave in S . (Hint: Let $x_1, x_2 \in S$, $\lambda \in [0, 1]$ and choose u_1, u_2 in U such that $f(x_1) = F(x_1, u_1)$, $f(x_2) = F(x_2, u_2)$.)

Let B be a convex set in $\mathbb{R}^n \times \mathbb{R}^r$ and define the set $U_x = \{u : (x, u) \in B\}$. Prove that $g(x) = \max_{u \in U_x} F(x, u)$ is concave.

4. Rewrite the following problem as one of the type (1)–(5):

$$\max \int_{t_0}^{t_1} f(t, x, u) dt, \quad \dot{x} = g(t, x, u), \quad x(t_0) = x^0, \quad u \in U, \quad \int_{t_0}^{t_1} h(t, x, u) dt = K$$

Here t_0, t_1, x^0 , and K are given numbers, f, g , and h are given functions, and U is a subset of $\mathbb{R}^r$.

10.2 Some Examples

In this section the theory from the previous section is used to solve some multidimensional control problems. The first is intended to be simple enough for you to be able to make a real effort to solve it before looking at the suggested solution.

EXAMPLE 1 Solve the problem

$$\max_{u(t) \in U} \int_0^T (x(t) + y(t) - \frac{1}{2}u(t)^2) dt, \quad \begin{cases} \dot{x}(t) = y(t), & x(0) = 0, & x(T) \text{ is free} \\ \dot{y}(t) = u(t), & y(0) = 0, & y(T) \text{ is free} \end{cases}$$

Verify that the last equality in (10.1.14) is satisfied.

Solution: Suppose that $(x^*(t), y^*(t), u^*(t))$ solves the problem. With the two adjoint variables p_1 and p_2 , the Hamiltonian is $H = x + y - \frac{1}{2}u^2 + p_1 y + p_2 u$, which is clearly concave in (x, y, u) . (Because $x(T)$ and $y(T)$ are free, Note 10.1.3 implies $p_0 = 1$.) We see that $H'_x = 1$, $H'_y = 1 + p_1$, and $H'_u = -u + p_2$.

The differential equations for p_1 and p_2 are $\dot{p}_1(t) = -1$ with $p_1(T) = 0$, and $\dot{p}_2(t) = -1 - p_1(t)$ with $p_2(T) = 0$. It follows that $p_1(t) = T - t$. Hence, $\dot{p}_2(t) = -1 + t - T$ and therefore $p_2(t) = -t + \frac{1}{2}t^2 - Tt + A$. The requirement $p_2(T) = 0$ implies that $A = \frac{1}{2}T^2 + T$. Thus

$$p_1(t) = T - t, \quad p_2(t) = \frac{1}{2}(T - t)^2 + T - t$$

Now H is concave in u and $u \in \mathbb{R}$, so H has its maximum when $H'_u = 0$. This gives $u^*(t) = p_2(t) = \frac{1}{2}(T - t)^2 + T - t$. Since $\dot{y}^*(t) = u^*(t) = \frac{1}{2}(T - t)^2 + T - t$, we find by integration that $y^*(t) = -\frac{1}{6}(T - t)^3 + Tt - \frac{1}{2}t^2 + B$. The initial condition $y^*(0) = 0$ gives $B = \frac{1}{6}T^3$. From $\dot{x}^*(t) = y^*(t)$ we get $x^*(t) = \frac{1}{24}(T - t)^4 + \frac{1}{2}Tt^2 - \frac{1}{6}t^3 + \frac{1}{6}T^3t + C$. The requirement $x^*(0) = 0$ gives $C = -\frac{1}{24}T^4$. Hence the optimal choices for x^* and y^* are

$$x^*(t) = \frac{1}{24}(T - t)^4 + \frac{1}{2}Tt^2 - \frac{1}{6}t^3 + \frac{1}{6}T^3t - \frac{1}{24}T^4, \quad y^*(t) = -\frac{1}{6}(T - t)^3 + Tt - \frac{1}{2}t^2 + \frac{1}{6}T^3$$

Mangasarian's theorem shows that we have found the optimal solution.

The value function is $V(T) = \int_0^T (x^*(t) + y^*(t) - \frac{1}{2}u^*(t)^2) dt$, and a rather tedious computation (using Leibniz's formula) yields $V'(T) = \frac{1}{2}T^2 + \frac{1}{2}T^3 + \frac{1}{8}T^4$. On the other hand, $H^*(T) = x^*(T) + y^*(T) - \frac{1}{2}(u^*(T))^2 + p_1(T)y^*(T) + p_2(T)u^*(T)$ is easily seen to equal $\frac{1}{2}T^2 + \frac{1}{2}T^3 + \frac{1}{8}T^4$, confirming (10.1.14). $\square$

EXAMPLE 2 (Two-sector model) (This is related to a model of Mahalanobis.) Consider an economy that is divided into two sectors. Sector 1 produces investment goods, while sector 2 produces consumption goods. Define

$x_i(t)$ = output in sector i per unit of time, $i = 1, 2$

$u(t)$ = the fraction of output in sector 1 that is invested in sector 1

Assume that each unit of investment in either sector increases output in that sector by a units. It follows that $\dot{x}_1 = aux_1$ and $\dot{x}_2 = a(1 - u)x_1$, where a is a positive constant. By definition, one has $0 \leq u(t) \leq 1$. Finally, if the planning period starts at time $t = 0$, then $x_1(0)$ and $x_2(0)$ are historically given.

We consider the problem of maximizing total consumption in a given planning period $[0, T]$. The problem is then, with a, T, x_1^0 , and x_2^0 as positive constants:

$$\max_{u(t) \in [0, 1]} \int_0^T x_2(t) dt, \quad \begin{cases} \dot{x}_1(t) = au(t)x_1(t), & x_1(0) = x_1^0, & x_1(T) \text{ is free} \\ \dot{x}_2(t) = a(1 - u(t))x_1(t), & x_2(0) = x_2^0, & x_2(T) \text{ is free} \end{cases}$$

The Hamiltonian is $H = x_2 + p_1aux_1 + p_2a(1 - u)x_1$, where p_1 and p_2 are the adjoint variables associated with the two differential equations. (Because both terminal stocks are free, Note 10.1.3 implies $p_0 = 1$.)

Suppose that $(x_1^*(t), x_2^*(t))$ and $u^*(t)$ solve the problem. According to Theorem 10.1.1 there exists a continuous vector function $(p_1(t), p_2(t))$ such that for all t in $[0, T]$, $u^*(t)$ is the value of u in $[0, 1]$ which maximizes $x_2^*(t) + p_1(t)aux_1^*(t) + p_2(t)a(1 - u)x_1^*(t)$. Collecting the terms in H which depend on u , note that $u^*(t)$ must be chosen as that value of u in $[0, 1]$ which maximizes $a(p_1(t) - p_2(t))x_1^*(t)u$. Now, $x_1^*(0) = x_1^0 > 0$, and because $\dot{x}_1^*(t) = au^*(t)x_1^*(t)$, it follows that $x_1^*(t) > 0$ for all t . The maximum condition therefore implies that $u^*(t)$ should be chosen as

$$u^*(t) = \begin{cases} 1 & \text{if } p_1(t) > p_2(t) \\ 0 & \text{if } p_1(t) < p_2(t) \end{cases} \quad (i)$$

The function $p_2(t)$ satisfies $\dot{p}_2(t) = -\partial H^*/\partial x_2 = -1$ with $p_2(T) = 0$. Hence,

$$p_2(t) = T - t$$

The function $p_1(t)$ satisfies $\dot{p}_1(t) = -\partial H^*/\partial x_1 = -p_1(t)au^*(t) - p_2(t)a(1 - u^*(t))$, with $p_1(T) = 0$. Because $p_1(T) = p_2(T) = 0$, one has $\dot{p}_1(T) = 0$. From $\dot{p}_2(t) = -1$, it follows that $p_1(t) < p_2(t)$ in an interval immediately to the left of T . (See Fig. 1.) If we let $t^* = \inf\{t \in [0, T] : p_1(t) < p_2(t) \text{ for } t \text{ in } (t, T)\}$, then (t^*, T) is the largest such interval. (Possibly, $t^* = 0$.) Using (i) it follows that $u^*(t) = 0$ in (t^*, T) . Hence, $\dot{p}_1(t) = -ap_2(t) = -a(T - t)$ in (t^*, T) . Integration yields $p_1(t) = -aTt + \frac{1}{2}at^2 + C_1$. But $p_1(T) = 0$, so $C_1 = \frac{1}{2}aT^2$ and hence

$$p_1(t) = -aTt + \frac{1}{2}at^2 + \frac{1}{2}aT^2 = \frac{1}{2}a(T - t)^2, \quad t \in [t^*, T]$$

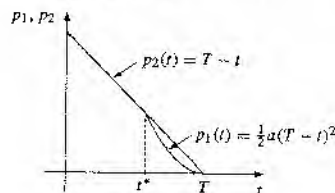


Figure 1 The behaviour of p_1 and p_2 .

Unless $p_1(t) < p_2(t)$ for all t in $[0, T]$, the number t^* is determined by the equation $p_1(t^*) = p_2(t^*)$. Using the expressions found for $p_1(t)$ and $p_2(t)$, it follows that

$$t^* = T - 2/a \text{ if } T > 2/a, \text{ otherwise } t^* = 0$$

Consider the case when $T > 2/a$, so $t^* > 0$. How does $p_1(t)$ behave in the interval $[0, t^*]$? Note first that

$$\dot{p}_1(t) = \begin{cases} -ap_1(t) & \text{if } p_1(t) > p_2(t) \\ -ap_2(t) & \text{if } p_1(t) \leq p_2(t) \end{cases}$$

If $p_1(t) > p_2(t)$, then $-ap_1(t) < -ap_2(t)$. Whatever is the relationship between $p_1(t)$ and $p_2(t)$, we always have

$$\dot{p}_1(t) \leq -ap_2(t) = a(t - T)$$

In particular, if $t < t^*$, then $\dot{p}_1(t) \leq a(t - T) < a(t^* - T) = -2$. Because $\dot{p}_2(t) = -1$ for all t and $p_1(t^*) = p_2(t^*)$, we conclude that $p_1(t) > p_2(t)$ for $t < t^*$. Hence, $u^*(t) = 1$ for t in $[0, t^*]$. The maximum principle therefore yields the following candidate for an optimal control, in the case when $T > 2/a$:

$$u^*(t) = \begin{cases} 1 & \text{if } t \in [0, T - 2/a] \\ 0 & \text{if } t \in (T - 2/a, T) \end{cases} \quad (T > 2/a) \quad (\text{ii})$$

For t in $[0, T - 2/a]$, we have $u^*(t) = 1$ and so $\dot{p}_1(t) = -ap_1(t)$, i.e. $p_1(t) = Ce^{-at}$. Because $p_1(t^*) = p_2(t^*) = T - t^* = 2/a$, this yields

$$p_1(t) = (2/a)e^{-at+T-2}, \quad t \in [0, T - 2/a]$$

It is easy to find explicit expressions for $x_1^*(t)$ and $x_2^*(t)$. (See Problem 2.)

In the other case, when $T \leq 2/a$, one has $t^* = 0$, so the candidate for an optimal control is

$$u^*(t) = 0 \quad \text{for all } t \text{ in } [0, T] \quad (T \leq 2/a)$$

In this example the maximum principle yields only one candidate for an optimal control (in each of the cases $T > 2/a$ and $T \leq 2/a$).

The Hamiltonian is not concave in (x_1, x_2, u) (because of the product ux_1). Thus the Mangasarian theorem does not apply. For $x_1 \geq 0$ and $x_2 \geq 0$, however, the maximized Hamiltonian $\hat{H}$ defined in (10.1.16) is

$$\hat{H}(t, x_1, x_2, p_1, p_2) = \begin{cases} x_2 + ap_1x_1 & \text{if } p_1 > p_2 \\ x_2 + ap_2x_1 & \text{if } p_1 \leq p_2 \end{cases}$$

For each t in $[0, T]$, the function $\hat{H}$ is linear in (x_1, x_2) . It is therefore concave in the set $A = \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0\}$. According to Theorem 10.1.3 and Note 10.1.7, the solution to the two-sector problem has been found. ■

PROBLEMS FOR SECTION 10.2

1. Solve the problem

$$\max_{u \in [-1, 1]} \int_0^4 (10 - x_1 + u) dt, \quad \begin{cases} \dot{x}_1(t) = x_2(t), & x_1(0) = 2, & x_1(4) \text{ is free} \\ \dot{x}_2(t) = u(t), & x_2(0) = 4, & x_2(4) \text{ is free} \end{cases}$$

SM 2. In Example 2, for the case when $T > 2/a$, find the functions $x_1^*(t)$ and $x_2^*(t)$ corresponding to the control function given in (ii).

SM 3. (a) Solve the problem

$$\max \int_0^T (\frac{1}{2}x_1 + \frac{1}{3}x_2 - u_1 - u_2) dt, \quad \begin{cases} \dot{x}_1(t) = u_1(t), & x_1(0) = 0, & x_1(T) \text{ is free} \\ \dot{x}_2(t) = u_2(t), & x_2(0) = 0, & x_2(T) \text{ is free} \end{cases}$$

with $0 \leq u_1(t) \leq 1$, $0 \leq u_2(t) \leq 1$, and with T as a fixed number greater than 5.

(b) Replace the objective functional by $\int_0^T (\frac{1}{2}x_1 + \frac{1}{3}x_2 - u_1 - u_2) dt + 3x_1(T) + 2x_2(T)$ and find the solution in this case.

SM 4. Solve the problem

$$\begin{aligned} &\max \int_0^T (x_2 + c(1 - u_1 - u_2)) dt \\ &\dot{x}_1(t) = au_1(t), \quad x_1(0) = x_1^0, \quad x_1(T) \text{ free} \\ &\dot{x}_2(t) = au_2(t) + bx_1(t), \quad x_2(0) = x_2^0, \quad x_2(T) \text{ free} \\ &0 \leq u_1, \quad 0 \leq u_2, \quad u_1 + u_2 \leq 1 \end{aligned}$$

where T , a , b , and c are positive constants and $T - c/a > T - 2/b > 0$. (Compared with Example 2, an extra flow of income amounting to one unit (say 1 billion per year) can be divided between extra capital investment in either the investment or consumption goods sectors, or consumed directly.)

5. Solve the problem

$$\max_{u \in [0, u^0]} \int_0^T (x_1 - cx_2 + u^0 - u) dt, \quad \begin{cases} \dot{x}_1 = u, & x_1(0) = x_1^0, & x_1(T) \text{ is free} \\ \dot{x}_2 = bx_1, & x_2(0) = x_2^0, & x_2(T) \text{ is free} \end{cases}$$

where T , b , c , and u^0 are positive constants with $bcT > 2$ and $2bc < 1$. (Economic interpretation: Oil is produced at the rate of u^0 per unit of time. The proceeds can be used to increase the capacity x_1 in the sector producing consumption goods. By adjusting the physical units, assume $\dot{x}_1 = u$. The production of consumption goods is proportional to x_1 and, by adjusting the time unit, the constant of proportionality is chosen as 1. The production of consumption goods increases the stock of pollution, x_2 , at a constant rate per unit. This subtracts cx_2 from utility per unit of time.)

6. Consider the problem

$$\max \int_0^T U(c(t))e^{-rt} dt, \quad \begin{cases} \dot{K}(t) = f(K(t), u(t)) - c(t), & K(0) = K_0, & K(T) = K_T \\ \dot{x}(t) = -u(t), & x(0) = x_0, & x(T) = 0 \end{cases}$$

where $u(t) \geq 0$, $c(t) \geq 0$. Here $K(t)$ denotes capital stock, $x(t)$ is the stock of a natural resource, $c(t)$ is consumption, and $u(t)$ is the rate of extraction. Moreover, U is a utility function and f is the production function. The constants T , K_0 , K_T , and x_0 are positive. Assume that $U' > 0$, $U'' \leq 0$, $f'_K > 0$, $f'_u > 0$, and that $f(K, u)$ is concave in (K, u) . This problem has two state variables (K and x) and two control variables (u and c).

(a) Write down the conditions in Theorem 10.1.1, assuming that $u(t) > 0$ and $c(t) > 0$ at the optimum.

(b) Derive from these conditions that

$$\frac{\dot{c}}{c} = \frac{r - f'_K(K, u)}{\tilde{\omega}}, \quad \frac{d}{dt}(f'_u(K, u)) = f'_K(K, u)f'_u(K, u)$$

where $\tilde{\omega}$ is the elasticity of the marginal utility. See Section 8.4.

7. Solve the problem

$$\max_{u \in [0, 1]} \int_0^2 (x - \frac{1}{2}u) dt, \quad \begin{cases} \dot{x} = u, & x(0) = 1, & x(2) \text{ is free} \\ \dot{y} = u, & y(0) = 0, & y(2) \leq 1 \end{cases}$$

10.3 Infinite Horizon

Infinite horizon control problems were introduced in Section 9.11. This section extends the analysis in several directions. Consider as a point of departure the problem

$$\max_{u(t) \in U} \int_0^\infty f(t, x(t), u(t)) dt, \quad \dot{x}(t) = g(t, x(t), u(t)), \quad x(0) = x^0, \quad \lim_{t \rightarrow \infty} x(t) = x^1 \quad (1)$$

where x^1 is a fixed vector in $\mathbb{R}^n$. Suppose the integral converges whenever $(x(t), u(t))$ satisfies the differential equation and $x(t)$ tends to the limit x^1 as t tends to ∞ . For this

problem the maximum principle holds. If we replace the condition $\lim_{t \rightarrow \infty} x(t) = x^1$ with $\lim_{t \rightarrow \infty} x(t) \geq x^1$ or $\lim_{t \rightarrow \infty} x(t)$ free, then the maximum principle again holds, except for the transversality conditions.

When the integral in (1) does not converge for all admissible pairs, what is a reasonable optimality criterion? Suppose $(x(t), u(t))$ is an arbitrary admissible pair, and $(x^*(t), u^*(t))$ is a pair we wish to test for optimality. Define

$$D_u(t) = \int_0^t f(\tau, x^*(\tau), u^*(\tau)) d\tau - \int_0^t f_0(\tau, x(\tau), u(\tau)) d\tau \quad (2)$$

There are several optimality criteria in economics literature which differ in how $D_u(t)$ behaves for large values of t . The simplest of these criteria is:

OVERTAKING OPTIMAL

The pair $(x^*(t), u^*(t))$ is **OT optimal** if for each admissible pair $(x(t), u(t))$ there exists a number T_u such that $D_u(t) \geq 0$ for all $t \geq T_u$. (3)

More important than overtaking optimality is the next criterion:

CATCHING-UP OPTIMAL

The pair $(x^*(t), u^*(t))$ is **CU optimal** if for each admissible pair $(x(t), u(t))$ and every $\varepsilon > 0$ there exists a number $T_{u,\varepsilon}$ such that $D_u(t) \geq -\varepsilon$ whenever $t \geq T_{u,\varepsilon}$. (4)

NOTE 1 In general, let $f(t)$ be a function defined for all $t \geq t_0$. Following the discussion of upper and lower limits in Sections A.3 and A.4, define $F(t) = \inf \{ f(\tau) : \tau \geq t \}$. Then $F(t)$ is a nondecreasing function of t , and we define the lower limit of $f(t)$ as t tends to infinity as

$$\lim_{t \rightarrow \infty} f(t) = \lim_{t \rightarrow \infty} F(t) = \lim_{t \rightarrow \infty} (\inf \{ f(\tau) : \tau \geq t \}) \quad (5)$$

Here we allow $\lim_{t \rightarrow \infty} F(t) = \infty$. The following characterization is useful and quite straightforward to prove.

$$\lim_{t \rightarrow \infty} f(t) \geq a \iff \begin{cases} \text{For each } \varepsilon > 0 \text{ there exists a } t' \\ \text{such that } f(t) \geq a - \varepsilon \text{ for all } t \geq t' \end{cases} \quad (6)$$

With this definition the requirement in (4) can be formulated as:

$$(x^*(t), u^*(t)) \text{ is CU optimal} \iff \lim_{t \rightarrow \infty} D_u(t) \geq 0 \text{ for all admissible pairs } (x(t), u(t))$$

We turn next to the behaviour of $\mathbf{x}(t)$ as t approaches infinity. The requirement that $\mathbf{x}(t)$ tends to a limit as t approaches infinity is often too restrictive. So is the alternative requirement that $\lim_{t \rightarrow \infty} \mathbf{x}(t) \geq \mathbf{x}^1$ because it excludes paths where $\mathbf{x}(t)$ oscillates indefinitely. Among many possible terminal conditions consider the following:

$$\lim_{t \rightarrow \infty} x_i(t) \text{ exists and is equal to } x_i^1, \quad i = 1, \dots, l \quad (7a)$$

$$\lim_{t \rightarrow \infty} x_i(t) \geq x_i^1, \quad i = l+1, \dots, m \quad (7b)$$

$$\text{no conditions imposed on } x_i(t) \text{ as } t \rightarrow \infty, \quad i = m+1, \dots, n \quad (7c)$$

One can show the following theorem (Halkin (1974)):

THEOREM 10.3.1 (THE MAXIMUM PRINCIPLE INFINITE HORIZON)

Suppose the pair $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ satisfies the differential equation in (1), the initial condition $\mathbf{x}(t_0) = \mathbf{x}^0$, and the terminal conditions (7). If this pair is OT or CU optimal, then it must satisfy all the conditions in Theorem 10.1.1 except the transversality conditions.

The problem with this theorem is that when $l < n$ it gives too many solution candidates, because it includes no transversality condition.

Here is a result that gives sufficient conditions for CU optimality:

THEOREM 10.3.2 (SUFFICIENT CONDITIONS INFINITE HORIZON)

Consider problem (1) with the terminal conditions (7). Suppose that the pair $(\mathbf{x}^*(t), \mathbf{u}^*(t))$, together with the continuous and piecewise differentiable adjoint function $\mathbf{p}(t)$, satisfy the conditions (A) and (B) of Theorem 10.1.1, with $p_0 = 1$, for all $t \geq t_0$. Suppose too that U is convex, that

$$H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}(t)) \text{ is concave in } (\mathbf{x}, \mathbf{u}) \quad (8)$$

and

$$\lim_{t \rightarrow \infty} [\mathbf{p}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t))] \geq 0 \text{ for all admissible } \mathbf{x}(t) \quad (9)$$

Then the pair $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is CU optimal.

Proof: Applying the arguments in the proof of Theorem 9.7.1 and putting $t_1 = t$, we obtain $D_{\mathbf{u}}(t) \geq \mathbf{p}(t) \cdot (\mathbf{x}(t) - \mathbf{x}^*(t))$. Taking $\lim$ on both sides, it follows that $\lim_{t \rightarrow \infty} D_{\mathbf{u}}(t) \geq 0$. ■

The following conditions are *sufficient* for (9) to hold (see Seierstad and Sydsæter (1987), Section 3.7, Note 16). For all admissible $\mathbf{x}(t)$:

$$\lim_{t \rightarrow \infty} [p_i(t)(x_i^1 - x_i^*(t))] \geq 0 \quad i = 1, \dots, m \quad (10a)$$

$$\text{There exists a constant } M \text{ such that} \quad |p_i(t)| \leq M \text{ for all } t \geq t_0 \quad i = 1, \dots, m \quad (10b)$$

$$\begin{aligned} &\text{Either there exists a number } t' \geq t_0 \text{ such that} \\ &p_i(t) \geq 0 \text{ for all } t \geq t', \text{ or there exists a} \\ &\text{number } P \text{ such that } |x_i(t)| \leq P \text{ for all } t \geq t_0 \quad i = l+1, \dots, m \quad (10c) \\ &\text{and } \lim_{t \rightarrow \infty} p_i(t) \geq 0 \end{aligned}$$

$$\begin{aligned} &\text{There exists a number } Q \text{ such that } |x_i(t)| < Q \\ &\text{for all } t \geq t_0, \text{ and } \lim_{t \rightarrow \infty} p_i(t) = 0 \quad i = m+1, \dots, n \quad (10d) \end{aligned}$$

NOTE 2 (Malinvaud's transversality conditions) If the terminal conditions 7(a)–(c) are replaced by the conditions $x_i(t) \geq x_i^1$ for all $t \geq t_0$ and all $i = 1, \dots, n$, then the inequalities $\mathbf{p}(t) \geq \mathbf{0}$ for all $t \geq t_0$ and 10(a) are sufficient for (9) to hold.

PROBLEMS FOR SECTION 10.3

SM 1. Given $r \in (0, 1)$, solve the problem

$$\max_{u \in [0, 1]} \int_0^\infty (x - u)e^{-rt} dt, \quad \dot{x} = ue^{-t}, \quad x(0) = x_0 \geq 0, \quad u \in [0, 1]$$

SM 2. (a) Solve the following problem when $r > a > 0$:

$$\max_{u \in [0, 1]} \int_0^\infty x_2 e^{-rt} dt, \quad \begin{cases} \dot{x}_1 = aux_1, & x_1(0) = x_1^0 > 0 \\ \dot{x}_2 = a(1 - u)x_1, & x_2(0) = x_2^0 = 0 \end{cases}$$

(b) Show that the problem has no solution when $r < a$.

10.4 Existence Theorems and Sensitivity

At the end of Section 9.3 we mentioned the role played by existence theorems in optimal control theory. Not every control problem has an optimal solution. For example, in most control problems in economics it is easy to impose requirements on the final state that are entirely unattainable. These are trivial examples of problems without optimal solutions. Moreover, when the control region U is open or unbounded, it is frequently the case that no optimal solution exists. Even if U is compact and there exist admissible pairs, there is no guarantee that an optimal pair exists.

As a practical control problem without an optimal solution, think of trying to keep a pan of boiling water at a constant temperature of 100°C for one hour when it is being heated on an electric burner whose only control is an on/off switch. If we disregard the cost of switching, there is no limit to the number of times we should turn the burner on and off.

In applications one often sees the argument that practical physical or economic considerations strongly suggest the existence of an optimum. Such considerations may be useful as heuristic arguments, but they can never replace a proper mathematical existence proof. In general, a necessary condition for a mathematical optimization problem to give a realistic representation of physical or economic reality is that the problem has a solution. If a practical problem appears to have no solution, the fault may lie with the mathematical description used to model it.

Consider the standard end-constrained problem (10.1.1)–(10.1.5). For every $(t, \mathbf{x})$ in $\mathbb{R}^{n+1}$, define the set

$$N(t, \mathbf{x}) = \{ (f(t, \mathbf{x}, \mathbf{u}) + \gamma, g_1(t, \mathbf{x}, \mathbf{u}), \dots, g_n(t, \mathbf{x}, \mathbf{u})) : \gamma \leq 0, \mathbf{u} \in U \} \quad (1)$$

This is a set in $\mathbb{R}^{n+1}$ generated by letting γ take all values ≤ 0 , while $\mathbf{u}$ varies in U .

The next theorem requires the set $N(t, \mathbf{x})$ to be convex. This implies that when the system starts in position $\mathbf{x}$ at time t , if either of the two velocity vectors $\dot{\mathbf{x}}_1$ and $\dot{\mathbf{x}}_2$ are feasible, then so is any convex combination of them. Moreover, the associated value of f is no smaller than the convex combination of the values associated with $\dot{\mathbf{x}}_1$ and $\dot{\mathbf{x}}_2$. (For a proof see Cesari (1983).)

THEOREM 10.4.1 (FILIPPOV-CESARI EXISTENCE THEOREM)

Consider the standard end-constrained problem (10.1.1)–(10.1.5). Suppose that there exists an admissible pair, and suppose further that:

- (a) $N(t, \mathbf{x})$ in (1) is convex for every $(t, \mathbf{x})$.
- (b) U is compact.
- (c) There exists a number $b > 0$ such that $\|\dot{\mathbf{x}}(t)\| \leq b$ for all t in $[t_0, t_1]$ and all admissible pairs $(\mathbf{x}(t), \mathbf{u}(t))$.

Then there exists an optimal pair $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ (where the control function $\mathbf{u}^*(t)$ is measurable).

NOTE 1 Condition (a) in Theorem 10.4.1 can be dropped if all the functions g_i are of the form $g_i(t, \mathbf{x}, \mathbf{u}) = h_i(t, \mathbf{x}) + k_i(t, \mathbf{u})$, where the functions h_i are linear in $\mathbf{x}$.

NOTE 2 Condition (c) in the theorem is implied by the following sufficient condition:

$$\text{There exist continuous functions } a(t) \text{ and } b(t) \text{ such that} \\ \|g(t, \mathbf{x}, \mathbf{u})\| \leq a(t)\|\mathbf{x}\| + b(t) \text{ for all } (t, \mathbf{x}, \mathbf{u}), \mathbf{u} \in U \quad (2)$$

NOTE 3 For an existence theorem for infinite horizon problems, see Seierstad and Sysdsæter (1987), Section 3.7, Theorem 15.

NOTE 4 Consider problem (10.1.1)–(10.1.5) where t_1 is free to take values in an interval $[T_1, T_2]$ with $T_1 \geq t_0$. Then Theorem 10.4.1 is still valid if the requirements are satisfied for all t in $[t_0, T_2]$.

A technical problem with the Filippov-Cesari existence theorem is that, in order to ensure the existence of an optimal control, the class of admissible control functions must be enlarged to include “measurable” functions. These can be “much more discontinuous” than piecewise continuous functions. But as long as they are bounded, they will still yield integrable functions of t along the optimal path. (For a brief survey, see Lee and Markus (1967), pp. 55–56.) In almost all control problems encountered in applications one can assume that if there is a measurable control that solves the problem, then there exists a piecewise continuous control that solves the problem.

EXAMPLE 1 Consider the problem $\max \int_0^1 x^2 dt$, $\dot{x} = 1 - u^2$, $x(0) = x(1) = 4$, $u \in [-1, 2] = U$. The Hamiltonian $H = x^2 + p(1 - u^2)$ is not concave in (x, u) and Arrow’s sufficiency condition also fails. In Problem 3 you are asked to find a unique solution candidate by using the maximum principle. Use Theorem 10.4.1 to prove that this candidate is optimal.

Solution: Note first that $(x(t), u(t)) \equiv (4, 1)$ is an admissible pair. Also,

$$N(t, x) = \{(x^2 + \gamma, 1 - u^2) : \gamma \leq 0, u \in [-1, 2]\}$$

which does not depend on t . As u varies in $[-1, 2]$, the second coordinate takes all values between 1 and -3 . For fixed x , the first coordinate takes all values less than or equal to x^2 . The set $N(t, x)$ is therefore as illustrated in Fig. 1.

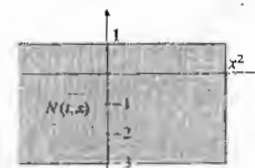


Figure 1 The set $N(t, x)$ in Example 1 is convex.

Obviously, $N(t, x)$ is convex as an “infinite rectangle”, so (a) is satisfied. The set $U = [-1, 2]$ is compact. Since $|\dot{x}(t)| = |1 - u^2(t)| \leq 3$ for all admissible $u(t)$, any admissible $x(t)$ satisfies $1 \leq x(t) \leq 7$ for all t in $[0, 1]$, which takes care of (c). We conclude that the unique pair satisfying the conditions in the maximum principle is optimal. ■

EXAMPLE 2 Show the existence of an optimal control for Example 10.2.2. (Hint: Use Note 2.)

Solution: Clearly, $u(t) \equiv 0$ gives an admissible solution, and the set $U = [0, 1]$ is compact. The set $N = N(t, \mathbf{x})$ is here

$$N(t, x_1, x_2) = \{(x_2 + \gamma, aux_1, a(1 - u)x_1) : \gamma \leq 0, u \in [0, 1]\}$$

This is the set of points (ξ_1, ξ_2, ξ_3) in $\mathbb{R}^3$ with $\xi_1 \leq x_2$ and (ξ_2, ξ_3) lying on the line segment that joins $(0, ax_1)$ to $(ax_1, 0)$ in $\mathbb{R}^2$. Hence N is convex.

The inequality in (2) is also satisfied because

$$\|g(t, x_1, x_2, u)\| = \|(aux_1, a(1 - u)x_1)\| = \sqrt{(aux_1)^2 + (a(1 - u)x_1)^2} \\ = a|x_1|\sqrt{2u^2 - 2u + 1} \leq a|x_1| = a\sqrt{x_1^2} \leq a\sqrt{x_1^2 + x_2^2} = a\|(x_1, x_2)\|$$

using the fact that $2u^2 - 2u + 1 = 2u(u - 1) + 1 \leq 1$ for all u in $[0, 1]$. The existence of an optimal control follows from Theorem 10.4.1. $\square$

Precise Sensitivity Results

Here we shall briefly present precise conditions for the sensitivity results in (10.1.14) to hold. Consider the standard end-constrained problem (10.1.1)–(10.1.5) and assume that admissible pairs exist. Let $\mathbf{x}^1 = (x_1^1, \dots, x_n^1)$ and define

$$V(\mathbf{x}^0, \mathbf{x}^1, t_0, t_1) = \sup \left\{ \int_{t_0}^{t_1} f(t, \mathbf{x}(t), \mathbf{u}(t)) dt : (\mathbf{x}(t), \mathbf{u}(t)) \text{ admissible} \right\} \quad (3)$$

(If $m = 0$, the right end point is free and V will not have $\mathbf{x}^1$ as an argument.) The function V is called the (optimal) value function of the problem. It is defined only for those $(\mathbf{x}^0, \mathbf{x}^1, t_0, t_1)$ for which admissible pairs exist. If for a given $(\mathbf{x}^0, \mathbf{x}^1, t_0, t_1)$ an optimal pair exists, then V is finite and equal to the integral in (10.1.1) evaluated along the optimal pair. (This was the case studied in Section 9.6.) If the set in (3) is not bounded above, then $V = \infty$.

Suppose that $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ solves problem (10.1.1)–(10.1.5) with $\mathbf{x}^0 = \bar{\mathbf{x}}^0, \mathbf{x}^1 = \bar{\mathbf{x}}^1, t_0 = \bar{t}_0, t_1 = \bar{t}_1$ for $p_0 = 1$, with corresponding adjoint function $\mathbf{p}(t)$. The next theorem gives sufficient conditions for V to be defined in a neighbourhood of $(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1, \bar{t}_0, \bar{t}_1)$, and for V to be differentiable at $(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1, \bar{t}_0, \bar{t}_1)$ with the following partial derivatives:

$$\frac{\partial V(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1, \bar{t}_0, \bar{t}_1)}{\partial x_i^0} = p_i(\bar{t}_0), \quad i = 1, \dots, n \quad (4)$$

$$\frac{\partial V(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1, \bar{t}_0, \bar{t}_1)}{\partial x_i^1} = -p_i(\bar{t}_1), \quad i = 1, \dots, n \quad (5)$$

$$\frac{\partial V(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1, \bar{t}_0, \bar{t}_1)}{\partial t_0} = -H(\bar{t}_0, \mathbf{x}^*(\bar{t}_0), \mathbf{u}^*(\bar{t}_0), \mathbf{p}(\bar{t}_0)) \quad (6)$$

$$\frac{\partial V(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1, \bar{t}_0, \bar{t}_1)}{\partial t_1} = H(\bar{t}_1, \mathbf{x}^*(\bar{t}_1), \mathbf{u}^*(\bar{t}_1), \mathbf{p}(\bar{t}_1)) \quad (7)$$

THEOREM 10.4.2

Consider the standard end-constrained problem (10.1.1)–(10.1.5) with a compact control region U . Suppose that

- $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is a unique optimal solution.
- $\mathbf{p}(t)$ is uniquely determined by the necessary conditions given $\mathbf{x}^*(t), \mathbf{u}^*(t)$, and $p_0 = 1$.
- There exist continuous functions $a(t)$ and $b(t)$ such that

$$\|f(t, \mathbf{x}, \mathbf{u})\| \leq a(t)\|\mathbf{x}\| + b(t) \quad \text{for all } (t, \mathbf{x}, \mathbf{u}) \text{ with } \mathbf{u} \in U \quad (8)$$

- The set $N(t, \mathbf{x})$ in (1) is convex for each $(t, \mathbf{x})$.

Then (4)–(7) are all valid.

For a proof of this theorem see Clarke (1983).

NOTE 5 Assume in this note that the uniqueness condition in (b) is replaced by the condition that the function $\mathbf{x} \mapsto \bar{H}(t, \mathbf{x}, \mathbf{p}(t))$ is concave. Then the function V is defined for $t_0 = \bar{t}_0, t_1 = \bar{t}_1$, and $(\mathbf{x}^0, \mathbf{x}^1)$ in a neighbourhood of $(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1)$, and the partial derivatives are given by (4) and (5) at $(\bar{\mathbf{x}}^0, \bar{\mathbf{x}}^1)$. If $I = n$ (and so the end point is fixed), or if $\mathbf{x} \mapsto \bar{H}(t, \mathbf{x}, \mathbf{p}(t))$ is strictly concave, then all the partial derivatives (including those in (6) and (7)) exist. (For further details see Seierstad and Sydsæter (1987), Section 3.5.)

PROBLEMS FOR SECTION 10.4

- Show the existence of an optimal control and draw a picture of the set $N(t, x)$ for the problem

$$\max \int_0^1 x(t) dt, \quad \dot{x}(t) = x(t) + u(t), \quad x(0) = 0, \quad x(1) \geq 1, \quad u \in [-1, 1]$$

- Solve the problem: $\max_{u \in [0, 1]} \int_0^1 (1-u)x^2 dt, \quad \dot{x} = ux, \quad x(0) = x_0 > 0, \quad x(1) \text{ free}.$

- Find the unique solution candidate in Example 1 using the maximum principle. (Hint: Argue why $u^*(t)$ can only take the values 0 and 2, and why any admissible $x(t)$ is > 0 in $[0, 1]$.)

10.5 A Heuristic Proof of the Maximum Principle

A full proof of the general maximum principle is quite demanding and draws on several advanced results in the theory of differential equations which are not in the toolkit of most economists. The heuristic arguments for the main results given below, although not precise, give a good indication of why the maximum principle is correct. We restrict our attention to problems with one state and one control variable.

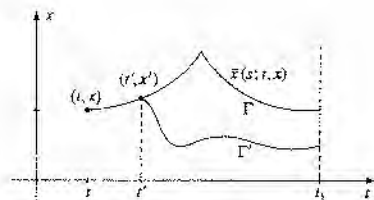
Consider the following control problem with two alternative terminal conditions:

$$\max_{u \in U} \int_{t_0}^{t_1} f(t, x, u) dt, \quad \dot{x} = g(t, x, u), \quad x(t_0) = x_0, \quad \begin{cases} x(t_1) \text{ free} \\ x(t_1) = x_1 \end{cases} \quad (i)$$

Think of $x = x(t)$ as a firm's capital stock and $\int_{t_0}^{t_1} f(t, x, u) dt$ as the total profit over the planning period $[t_0, t_1]$, in line with our general economic interpretation in Section 9.6. Define the value function by

$$V(t, x) = \max_{u \in U} \left\{ \int_t^{t_1} f(s, x(s), u(s)) ds : \dot{x}(s) = g(s, x(s), u(s)), \quad x(t) = x, \quad \begin{cases} x(t_1) \text{ free} \\ x(t_1) = x_1 \end{cases} \right\} \quad (ii)$$

Thus $V(t, x)$ is the maximum profit obtainable if we start at time t with the capital stock x . Suppose the problem in (ii) has a unique solution, which we denote by $\bar{u}(s; t, x)$, $\bar{x}(s; t, x)$, for $t_0 \leq t \leq s \leq t_1$. Then, by definition, $\bar{x}(t; t, x) = x$.

Figure 1 The case of $x(t_1)$ free.

Consider any starting point (t', x') that lies on the optimal path Γ defined by the original solution $\tilde{x}(s; t, x)$. If there were a better path Γ' starting at (t', x') , it would have been optimal for the solution starting at (t, x) to follow Γ' over the time interval $[t', t_1]$.¹ (See Fig. 1, which deals with the case when $x(t_1)$ is free.) For this reason, an optimal solution starting at (t, x) is automatically an optimal solution from (t', x') as well: the "tail" of an optimal solution must be optimal. Using the uniqueness of $(\tilde{x}(s; t, x), \tilde{u}(s; t, x))$ for all (t, x) , this implies the relations

$$\tilde{u}(s; t', x') = \tilde{u}(s; t, x), \quad \tilde{x}(s; t', x') = \tilde{x}(s; t, x)$$

whenever $t' \in [t, t_1]$ and $x' = \tilde{x}(t'; t, x)$. Hence,

$$V(t', \tilde{x}(t'; t, x)) = \int_{t'}^{t_1} f(s, \tilde{x}(s; t, x), \tilde{u}(s; t, x)) ds$$

Differentiate this equation w.r.t. t' at $t' = t$. Because $d\tilde{x}(t'; t, x)/dt' = g(t', \tilde{x}(t'; t, x), \tilde{u}(t'; t, x))$, we have

$$V'_t(t, x) + V'_x(t, x)g(t, x, \tilde{u}(t; t, x)) = -f(t, x, \tilde{u}(t; t, x)) \quad (\text{iii})$$

Hence, if we define

$$\tilde{p}(t, x) = V'_x(t, x)$$

and introduce the Hamiltonian function $H(t, x, u, p) = f(t, x, u) + p g(t, x, u)$, then equation (iii) can be written in the form

$$V'_t(t, x) + H(t, x, \tilde{u}(t; t, x), \tilde{p}(t, x)) = 0 \quad (\text{iv})$$

Starting at the point (t, x) , consider an alternative control which is a constant v on an interval $[t, t + \Delta t]$ and optimal thereafter. Let the corresponding state variable be $x^v(s)$ for s in $[t, t + \Delta t]$. Then

$$V(t, x) \geq \int_t^{t+\Delta t} f(s, x^v(s), v) ds + V(t + \Delta t, x^v(t + \Delta t))$$

and so

$$V(t + \Delta t, x^v(t + \Delta t)) - V(t, x) + \int_t^{t+\Delta t} f(s, x^v(s), v) ds \leq 0 \quad (\text{v})$$

Dividing this inequality by Δt and letting $\Delta t \rightarrow 0^+$, we get $\frac{d}{dt}V(t, x^v(t)) + f(t, x, v) \leq 0$. Now, $\frac{d}{dt}V(t, x) = V'_t(t, x) + V'_x(t, x)\dot{x}$. Since $V'_x(t, x) = \tilde{p}(t, x)$ and $\dot{x}^v(t) = g(t, x, v)$, we must have

$$V'_t(t, x) + \tilde{p}(t, x)g(t, x, v) + f(t, x, v) \leq 0$$

¹ "Better path" Γ' is intuitive language. It means that there exists an admissible pair $(x(s), u(s))$ (with corresponding path Γ') that gives a higher value to the integral of f over $[t', t_1]$ when $(x(s), u(s))$ is inserted, as compared with the value resulting from $(\tilde{x}(s; t, x), \tilde{u}(s; t, x))$.

Thus for all v in U ,

$$V'_t(t, x) + H(t, x, v, \tilde{p}(t, x)) \leq 0$$

Because of (iv), this implies that the optimal control $\tilde{u}(t; t, x)$ must maximize $H(t, x, u, \tilde{p}(t, x))$ w.r.t. $u \in U$. In addition,

$$V'_t(t, x) + \max_{u \in U} H(t, x, u, V'_x(t, x)) = 0 \quad (\text{vi})$$

This is called the **Hamilton-Jacobi-Bellman equation**.

Next, define $x^*(t) = \tilde{x}(t; t_0, x_0)$ and $u^*(t) = \tilde{u}(t; t_0, x_0)$. These functions give the optimal solution to the original problem. Also, let $p(t) = \tilde{p}(t, x^*(t))$. Then $\tilde{u}(t; t, x^*(t)) = u^*(t)$, and therefore

$$u = u^*(t) \text{ maximizes } H(t, x^*(t), u, p(t)) \text{ w.r.t. } u \in U \quad (\text{vii})$$

Finally, differentiating (iv) w.r.t. x and using the envelope theorem (see Section 3.10), we get

$$V'_{tx} + H'_x + H'_p \tilde{p}'_x = 0$$

Because $V'_t = \tilde{p}$ and $H'_p = g$, this yields $\tilde{p}'_t + \tilde{p}'_x g(t, x, \tilde{u}(t; t, x)) = -H'_x$. If we let $x = x^*(t)$ and use $u^*(t) = \tilde{u}(t; t, x)$, then $\dot{p} = \tilde{p}'_t + \tilde{p}'_x \dot{x} = \tilde{p}'_t + \tilde{p}'_x g(t, x, u^*(t))$, so

$$\dot{p}(t) = -H'_x(t, x^*(t), u^*(t), p(t)) \quad (\text{viii})$$

By definition of V , if $x(t_1)$ is free, then $V(t_1, x) = 0$ for all x . Thus $\tilde{p}(t_1, x) = 0$, and so we have the transversality condition

$$p(t_1) = 0 \quad (\text{ix})$$

Conditions (vii) to (ix) are the necessary conditions in the maximum principle (with t_1 fixed and $x(t_1)$ free). If $x(t_1)$ is fixed, condition (ix) is not valid (and not needed).

We have shown that

$$V'_t = -H^*(t_0), \quad V'_{x_0} = p(t_0) \quad (\text{x})$$

In fact, the first equality follows from (iv) and the second one from the definitions of the functions $\tilde{p}$ and p . These are two of the formulas in (10.1.14). Reversing time gives the other two formulas:

$$V'_t = H^*(t_1), \quad V'_{x_1} = -p(t_1) \quad (\text{xi})$$

Variable Final Time Problems

Consider problem (i) with t_1 free. Suppose $(x^*(t), u^*(t))$ is an optimal solution defined on $[t_0, t_1^*]$. Then conditions (vi)–(viii) must be valid on the interval $[t_0, t_1^*]$, because $(x^*(t), u^*(t))$ must be an optimal pair for the corresponding fixed time problem with $t_1 = t_1^*$. Moreover, at the terminal time t_1^* the value function's derivative w.r.t. t_1 must be 0. (As a function of t_1 it has a maximum at t_1^* .) Because of (xi), this implies that

$$H^*(t_1^*) = H(t_1^*, x^*(t_1^*), u^*(t_1^*), p(t_1^*)) = V'_t(t_1^*, x^*(t_1^*)) = 0 \quad (\text{xii})$$

This equation gives an extra condition for determining t_1^* , and is precisely condition (9.8.2).

NOTE 1 In the above heuristic "proof" of the maximum principle, differentiability of the function V was assumed without proof.

10.6 Mixed Constraints

This section describes control problems where the admissible pairs $(\mathbf{x}, \mathbf{u})$ are required to satisfy additional constraints involving the state variable of the form $\mathbf{h}(t, \mathbf{x}, \mathbf{u}) \geq \mathbf{0}$. Such restrictions often occur in economic models. If the control variable $\mathbf{u}$ as well as the state vector $\mathbf{x}$ appear in the function $\mathbf{h}$, the restriction is often referred to as a "mixed constraint", while restrictions of the type $\mathbf{h}(t, \mathbf{x}) \geq \mathbf{0}$ are called "pure state constraints".

Whether or not mixed constraints are present in a given control problem is partly a question of the form in which the problem is stated. Consider the following problem.

EXAMPLE 1 Consider the growth problem

$$\max_u \int_0^T U((1-u)f(K)) dt, \quad \dot{K} = u, \quad K(0) = K_0, \quad K(T) = K_T, \quad u \geq 0, \quad f(K) - u \geq 0$$

Here there are two constraints for each t —namely, the simple constraint $h_1(t, K, u) = u \geq 0$ and the mixed constraint $h_2(t, K, u) = f(K) - u \geq 0$. However, if we specify a new control variable v so that $\dot{K} = vf(K)$, then the simple constraint $0 \leq v \leq 1$ replaces the mixed constraints. (If we require $f(K) - u \geq k > 0$, this trick does not work.)

We consider the general **mixed constraints problem**

$$\max_u \int_{t_0}^{t_1} f(t, \mathbf{x}, \mathbf{u}) dt, \quad \dot{\mathbf{x}} = \mathbf{g}(t, \mathbf{x}, \mathbf{u}), \quad \mathbf{x}(t_0) = \mathbf{x}^0 \quad (1)$$

$$\mathbf{h}(t, \mathbf{x}, \mathbf{u}) \geq \mathbf{0} \quad \text{for all } t \quad (2)$$

with the terminal conditions

$$\begin{aligned} \text{(a)} \quad & x_i(t_1) = x_i^1, \quad i = 1, \dots, l \\ \text{(b)} \quad & x_i(t_1) \geq x_i^1, \quad i = l+1, \dots, m \\ \text{(c)} \quad & x_i(t_1) \text{ free}, \quad i = m+1, \dots, n \end{aligned} \quad (3)$$

As usual, $\mathbf{x}$ is n -dimensional and $\mathbf{u}$ is r -dimensional, while $\mathbf{h}$ is an s -dimensional vector function, so that the inequality $\mathbf{h}(t, \mathbf{x}, \mathbf{u}) \geq \mathbf{0}$ represents the s inequalities

$$h_k(t, \mathbf{x}(t), \mathbf{u}(t)) \geq 0, \quad k = 1, \dots, s \quad (4)$$

All the restrictions on $\mathbf{u}(t)$ are assumed to have been incorporated into (2). Thus, no additional requirement of the form $\mathbf{u} \in U$ is imposed. In addition to the usual requirements on f and $\mathbf{g}$, it is assumed that $\mathbf{h}$ is a C^1 function in $(t, \mathbf{x}, \mathbf{u})$. The pair $(\mathbf{x}(t), \mathbf{u}(t))$ is admissible if $u_1(t), \dots, u_r(t)$ are all piecewise continuous, and $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))$ is the corresponding continuous and piecewise differentiable vector function that satisfies $\dot{\mathbf{x}} = \mathbf{g}(t, \mathbf{x}, \mathbf{u})$, $\mathbf{x}(t_0) = \mathbf{x}^0$, (2), and (3). The theorem below gives sufficient conditions for the solution of the mixed constraints problem (1)–(3). To economists, it will come as no

surprise that we associate multipliers $q_1(t), \dots, q_s(t)$ with the constraints (2) and define the **Lagrangian function**, with $\mathbf{q} = (q_1, \dots, q_s)$, as

$$\mathcal{L}(t, \mathbf{x}, \mathbf{u}, \mathbf{p}, \mathbf{q}) = H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) + \sum_{k=1}^s q_k h_k(t, \mathbf{x}, \mathbf{u}) \quad (5)$$

with the Hamiltonian $H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) = f(t, \mathbf{x}, \mathbf{u}) + \sum_{i=1}^n p_i g_i(t, \mathbf{x}, \mathbf{u})$ (with $p_0 = 1$).

In the following theorem $\partial \mathcal{L}^* / \partial u_j$ and $\partial \mathcal{L}^* / \partial x_i$ in (6) and (8) are partial derivatives of $\mathcal{L}$ evaluated at $(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t), \mathbf{q}(t))$.

THEOREM 10.6-1 (SUFFICIENT CONDITIONS)

Suppose $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is an admissible pair in the mixed constraints problem (1)–(3). Suppose further that there exist functions $\mathbf{p}(t) = (p_1(t), \dots, p_n(t))$ and $\mathbf{q}(t) = (q_1(t), \dots, q_s(t))$, where $\mathbf{p}(t)$ is continuous, while $\dot{\mathbf{p}}(t)$ and $\mathbf{q}(t)$ are piecewise continuous, such that the following requirements are all satisfied:

$$\frac{\partial \mathcal{L}^*}{\partial u_j} = 0, \quad j = 1, \dots, r \quad (6)$$

$$q_k(t) \geq 0, \quad \text{and } q_k(t) = 0 \text{ if } h_k(t, \mathbf{x}^*(t), \mathbf{u}^*(t)) > 0, \quad k = 1, \dots, s \quad (7)$$

$$\dot{\mathbf{p}}(t) = -\frac{\partial \mathcal{L}^*}{\partial \mathbf{x}_i} \text{ at all continuity points of } \mathbf{u}^*(t), \quad i = 1, \dots, n \quad (8)$$

$$\text{No conditions on } p_i(t_1), \quad i = 1, \dots, l \quad (9a)$$

$$p_i(t_1) \geq 0, \quad \text{and } p_i(t_1) = 0 \text{ if } x_i^*(t_1) > x_i^1, \quad i = l+1, \dots, m \quad (9b)$$

$$p_i(t_1) = 0, \quad i = m+1, \dots, n \quad (9c)$$

$$H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}(t)) \text{ is concave in } (\mathbf{x}, \mathbf{u}) \quad (10)$$

$$h_k(t, \mathbf{x}, \mathbf{u}) \text{ is quasiconcave in } (\mathbf{x}, \mathbf{u}), \quad k = 1, \dots, s \quad (11)$$

Then $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ solves the problem.

A proof of this theorem is given in Seierstad and Sydsæter (1987), Section 4.3, which also discusses necessary conditions, generalizations, and examples, and has further references to other literature. A simpler treatment can be found in Léonard and Long (1992), Chapter 6. Note that as in nonlinear programming a constraint qualification is often necessary for the existence of a pair $(x(t), u(t))$ of the type occurring in the theorem. The constraint qualification, more or less, requires that the control u appears in each constraint.

EXAMPLE 2 Solve the mixed constraints problem

$$\max \int_0^T u dt, \quad \dot{x} = ax - u, \quad x(0) = x^0, \quad x(T) \text{ free}, \quad \begin{cases} h_1(t, x, u) = u - c \geq 0 \\ h_2(t, x, u) = ax - u \geq 0 \end{cases}$$

Here x is the capital stock, u is consumption, and c is a subsistence level. The constants T , a , c , and x^0 are positive, with $T > 1/a$ and $ax^0 > c$.

Solution: The Hamiltonian and the Lagrangian are

$$H = u + p(ax - u), \quad \mathcal{L} = H + q_1(u - c) + q_2(ax - u)$$

Here H as well as h_1 and h_2 are linear and hence concave in (x, u) . The following conditions from Theorem 10.6.1 are therefore sufficient for optimality:

$$\frac{\partial \mathcal{L}^*}{\partial u} = 1 - p(t) + q_1(t) - q_2(t) = 0 \quad (\text{i})$$

$$q_1(t) \geq 0, \quad \text{and } q_1(t) = 0 \text{ if } u^*(t) > c \quad (\text{ii})$$

$$q_2(t) \geq 0, \quad \text{and } q_2(t) = 0 \text{ if } ax^*(t) > u^*(t) \quad (\text{iii})$$

$$\dot{p}(t) = -\frac{\partial \mathcal{L}^*}{\partial x} = -ap(t) - aq_2(t), \quad p(T) = 0 \quad (\text{iv})$$

$$u^*(t) \geq c, \quad ax^*(t) - u^*(t) \geq 0 \quad (\text{v})$$

Because $x^*(0) = x^0 > 0$ and $\dot{x}^*(t) = ax^*(t) - u^*(t) \geq 0$ for all t , one has $x^*(t) \geq x^0$ for all t . If $u^*(t) = c$, then $ax^*(t) - u^*(t) = ax^*(t) - c \geq ax^0 - c > 0$, and then from (iii), $q_2(t) = 0$. On the other hand, if $u^*(t) > c$, then (ii) implies $q_1(t) = 0$. Hence, for all t in $[0, T]$, at least one of $q_1(t)$ and $q_2(t)$ is 0.

Because $p(T) = 0$, it follows that $p(t) < 1$ for t sufficiently close to (and less than) T . Equation (i) shows that for such t we have $q_2(t) = 1 - p(t) + q_1(t) \geq 1 - p(t) > 0$. Then $q_1(t)$ must be 0, so $q_2(t) = 1 - p(t)$ and (iv) gives $\dot{p}(t) = -a$ whenever $p(t) < 1$.

We see from Problem 5.4.8 that the solution of (iv) is $p(t) = \int_t^T a q_2(\tau) e^{-a(T-\tau)} d\tau$, which is clearly > 0 for $t < T$. Hence, $\dot{p}(t) = -ap(t) - aq_2(t) \leq -ap(t) < 0$ for $t < T$, so $p(t)$ is strictly decreasing in the interval $[0, T]$.

Suppose $p(0) \leq 1$. Then for all t in $(0, T]$ we get $p(t) < 1$ and $\dot{p}(t) = -a$. This in turn implies $p(t) = p(0) - at$ and, in particular $p(T) = p(0) - aT < 1 - aT$. But this is impossible because $T > 1/a$ and condition (iv) says that $p(T) = 0$. It follows that we must have $p(0) > 1$. Since p is strictly decreasing, there is a unique point t^* in $(0, T)$ such that $p(t^*) = 1$. For t in $(t^*, T]$ we have $\dot{p}(t) = -a$, so $p(t) = a(T - t)$. By continuity, $a(T - t^*) = p(t^*) = 1$, so $t^* = T - 1/a$.

For t in $[0, t^*)$ we have $p(t) > 1$, so (i) implies that $q_1(t) > q_2(t)$ and, because at least one of $q_1(t)$ and $q_2(t)$ is 0, in fact $q_2(t) = 0$. Then from (ii), $u^*(t) = c$ so that $\dot{x}^*(t) = ax^*(t) - c$, with $x^*(0) = x^0$. Solving this linear differential equation yields $x^*(t) = (x^0 - c/a)e^{at} + c/a$. The differential equation for $p(t)$ is $\dot{p} = -ap$ because $q_2 = 0$. Hence, $p(t) = Ae^{-at}$ with $p(t^*) = 1$, so $p(t) = e^{-a(t-t^*)}$.

Since $x^*(t)$ is continuous also at t^* , and $\dot{x}^* = 0$ in (t^*, T) , $x^*(t)$ has the constant value $(x^0 - c/a)e^{at^*} + c/a$ for t in (t^*, T) .

We have found the following candidate for an optimal solution, with $t^* = T - 1/a$:

	$u^*(t)$	$x^*(t)$	$p(t)$	$q_1(t)$	$q_2(t)$
$t \in [0, t^*]$	c	$(x^0 - c/a)e^{at} + c/a$	$e^{-a(t-t^*)}$	$e^{-a(t-t^*)} - 1$	0
$t \in (t^*, T]$	$ax^*(t)$	$(x^0 - c/a)e^{at^*} + c/a$	$a(T - t)$	0	$1 - a(T - t)$

Theorem 10.6.1 implies that this candidate is optimal. Note that in this example the multipliers $q_1(t)$ and $q_2(t)$ are continuous.

PROBLEMS FOR SECTION 10.6

5M 1. (a) Write down the conditions in Theorem 10.6.1 for the problem

$$\max \int_0^2 (x - \frac{1}{2}u^2) dt, \quad \dot{x} = -u, \quad x(0) = 1, \quad x(2) \text{ free}, \quad x \geq u$$

(b) Solve the problem. (Hint: Guess that $u^*(t) = x^*(t)$ on some interval $[0, t^*]$, and $u^*(t) < x^*(t)$ on $(t^*, 2]$. Then $q(t^{*+}) = 0$, and $u^*(t^{*-}) = x^*(t^*) \geq u^*(t^{*+})$. We can use the following argument² to show that $q(t^{*-}) = 0$: From $\partial \mathcal{L}^*/\partial u = 0$ we get $q(t) = -p(t) - u^*(t)$. In particular, $q(t^{*-}) = -p(t^*) - u^*(t^{*-}) \leq -p(t^*) - u^*(t^{*+}) = q(t^{*+}) = 0$.)

5M 2. Solve the problem

$$\max \int_0^2 (x - \frac{1}{2}u^2) dt, \quad \dot{x} = u, \quad x(0) = 1, \quad x(2) \text{ free}, \quad x \geq u$$

(Hint: Guess that $u^*(t) = x^*(t)$ on some interval $[0, t^*]$, and $u^*(t) < x^*(t)$ on $(t^*, 2]$. As in Problem 1, $q(t^{*-}) = 0$.)

5M 3. Solve the following variant of Example 2:

$$\max \int_0^T u dt, \quad \dot{x} = ax - u, \quad x(0) = x^0 > 0, \quad x(T) \geq x_T, \quad c \leq u \leq ax$$

where $a > 0$, $c > 0$, $T > 1/a$, $ax^0 > c$, and $x^0 \leq x_T < (x^0 - c/a)e^{aT} + c/a$. (This model can be interpreted as a simple growth model with a subsistence level c .)

5M 4. Solve the problem

$$\max \int_0^1 x dt, \quad \dot{x} = x + u, \quad x(0) = 0, \quad x(1) \text{ free}, \quad \begin{cases} h_1(t, x, u) = 1 - u \geq 0 \\ h_2(t, x, u) = 1 + u \geq 0 \\ h_3(t, x, u) = 2 - x - u \geq 0 \end{cases}$$

(Hint: See the solution to Example 9.4.1. Try with $u^*(t) = 1$, $x^*(t) = e^t - 1$ in the beginning.

10.7 Pure State Constraints

This section briefly discusses a result giving sufficient conditions for a pure state-constrained problem. It gives an indication of the type of results that can be proved, but we refer to the control theory literature for proofs, examples, and generalizations.

Consider the following **pure state-constrained problem**.

$$\max \int_{t_0}^{t_1} f(t, x, u) dt, \quad \dot{x} = g(t, x, u), \quad x(t_0) = x^0, \quad u(t) \in U \subseteq \mathbb{R}^r \quad (1)$$

$$h(t, x) \geq 0 \quad \text{for all } t \quad (2)$$

² The same argument is useful in other problems also, for example in Problem 2.

with the terminal conditions

$$\begin{aligned} \text{(a)} \quad & x_i(t_1) = x_i^1, \quad i = 1, \dots, l \\ \text{(b)} \quad & x_i(t_1) \geq x_i^1, \quad i = l+1, \dots, m \\ \text{(c)} \quad & x_i(t_1) \text{ free}, \quad i = m+1, \dots, n \end{aligned} \quad (3)$$

Note that, in contrast to the mixed constraints case, we now allow a restriction of the form $\mathbf{u} \in U$. The vector function $\mathbf{h}$ is s -dimensional, and the pure state constraint (2) can be written

$$h_k(t, \mathbf{x}(t)) \geq 0, \quad k = 1, \dots, s \quad (4)$$

The sufficient conditions given in the next theorem are somewhat more complicated than those in Theorem 10.6.1. In particular, the adjoint functions may have jumps at the terminal time.

The Lagrangian associated with this problem is

$$\mathcal{L}(t, \mathbf{x}, \mathbf{u}, \mathbf{p}, \mathbf{q}) = H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) + \sum_{k=1}^s q_k h_k(t, \mathbf{x}) \quad (5)$$

with $H(t, \mathbf{x}, \mathbf{u}, \mathbf{p})$ as the usual Hamiltonian (with $p_0 = 1$).

THEOREM 10.7.1 (SUFFICIENT CONDITIONS)

Suppose $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ is admissible in problem (1)–(3), and that there exist vector functions $\mathbf{p}(t)$ and $\mathbf{q}(t)$, where $\mathbf{p}(t)$ is continuous and $\dot{\mathbf{p}}(t)$ and $\mathbf{q}(t)$ are piecewise continuous in $[t_0, t_1]$, and numbers β_k , $k = 1, \dots, s$, such that the following conditions are satisfied with $p_0 = 1$:

$$\mathbf{u} = \mathbf{u}^*(t) \text{ maximizes } H(t, \mathbf{x}^*(t), \mathbf{u}, \mathbf{p}(t)) \text{ for } \mathbf{u} \text{ in } U \quad (6)$$

$$q_k(t) \geq 0, \quad \text{and } q_k(t) = 0 \text{ if } h_k(t, \mathbf{x}^*(t)) > 0, \quad k = 1, \dots, s \quad (7)$$

$$\dot{p}_i(t) = -\frac{\partial \mathcal{L}^*}{\partial x_i} \text{ at all continuity points of } \mathbf{u}^*(t), \quad i = 1, \dots, n \quad (8)$$

At t_1 , $p_i(t)$ can have a jump discontinuity, in which case

$$p_i(t_1^-) - p_i(t_1) = \sum_{k=1}^s \beta_k \frac{\partial h_k(t_1, \mathbf{x}^*(t_1))}{\partial x_i}, \quad i = 1, \dots, n \quad (9)$$

$$\beta_k \geq 0, \quad \text{with } \beta_k = 0 \text{ if } h_k(t_1, \mathbf{x}^*(t_1)) > 0, \quad k = 1, \dots, s \quad (10)$$

$$\text{No conditions on } p_i(t_1), \quad i = 1, \dots, l \quad (11a)$$

$$p_i(t_1) \geq 0, \quad \text{and } p_i(t_1) = 0 \text{ if } x_i^1(t_1) > x_i^1, \quad i = l+1, \dots, m \quad (11b)$$

$$p_i(t_1) = 0, \quad i = m+1, \dots, n \quad (11c)$$

$$\hat{H}(t, \mathbf{x}, \mathbf{p}(t)) = \max_{\mathbf{u} \in U} H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}(t)) \text{ is concave in } \mathbf{x} \quad (12)$$

$$h_k(t, \mathbf{x}) \text{ is quasiconcave in } \mathbf{x}, \quad k = 1, \dots, s \quad (13)$$

Then $(\mathbf{x}^*(t), \mathbf{u}^*(t))$ solves the problem.

Here $\mathbf{p}(t) = (p_1(t), \dots, p_n(t))$ and $\mathbf{q}(t) = (q_1(t), \dots, q_s(t))$, while $\partial \mathcal{L}^* / \partial x_i$ denotes the value of $\partial \mathcal{L} / \partial x_i$ at $(t, \mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t), \mathbf{q}(t))$.

NOTE 1 The conditions in this theorem are somewhat restrictive. In particular, sometimes the solution requires $\mathbf{p}(t)$ to have discontinuities at interior points of $[t_0, t_1]$. For details and a proof, see Seierstad and Sydsæter (1987), Chapter 5.

EXAMPLE 1 Solve the problem

$$\max \int_0^4 (x - (u - 2)^2) dt, \quad \dot{x} = u \in \mathbb{R}, \quad x(0) = 0, \quad x(4) \text{ free}, \quad x(t) \leq 1$$

Solution: The Lagrangian is $\mathcal{L} = H + q(1 - x) = x - (u - 2)^2 + pu + q(1 - x)$. Here H is concave in (x, u) and $h(t, x) = 1 - x$ is quasiconcave. The conditions (i)–(iv) below are therefore sufficient for optimality. Equation (i) below results from the observation that H is concave in u and $u \in \mathbb{R}$, so condition (6) is equivalent to the condition $\partial H^* / \partial u = 0$.

$$u^*(t) = \frac{1}{2} p(t) + 2 \quad (i)$$

$$q(t) \geq 0, \quad \text{and } q(t) = 0 \text{ if } x^*(t) < 1 \quad (ii)$$

$$\dot{p}(t) = -\frac{\partial \mathcal{L}^*}{\partial x} = -1 + q(t), \quad p(4) = 0 \quad (iii)$$

$$p(4^-) - p(4) = -\beta \leq 0, \quad \text{and } \beta = 0 \text{ if } x^*(4) < 1 \quad (iv)$$

We can guess the form of the solution as long as we eventually verify that all the conditions in the theorem are satisfied. Accordingly, we guess that $x^*(t) < 1$ in an interval $[0, t^*)$ and that $x^*(t) = 1$ in $(t^*, 4]$. Then in $(t^*, 4)$, $u^*(t) = \dot{x}^*(t) = 0$, and from (i), $p(t) = -4$. But then from (iii) and (iv), $\beta = p(4) - p(4^-) = 4$. On $[0, t^*)$, from (ii) and (iii), $\dot{p}(t) = -1$. Since $p(t)$ is continuous at t^* , $p(t^*) = -4$. Hence, $p(t) = -4 + (t^* - t)$, and from (i), $u^*(t) = \frac{1}{2}(t^* - t)$. Integrating $\dot{x}^*(t) = \frac{1}{2}(t^* - t)$ yields $x^*(t) = -\frac{1}{4}(t^* - t)^2 + C$ on $[0, t^*)$. Since $x^*(t^*) = 1$, we get $x^*(t^*) = C = 1$. But $x^*(0) = 0$, so $t^* = 2$. Our suggestion is therefore:

$$\text{In } [0, 2]: u^*(t) = 1 - \frac{1}{2}t, \quad x^*(t) = 1 - \frac{1}{4}(2 - t)^2, \quad p(t) = -t - 2, \quad \text{and } q(t) = 0.$$

$$\text{In } (2, 4]: u^*(t) = 0, \quad x^*(t) = 1, \quad p(t) = -4 \text{ (except } p(4) = 0), \quad \text{and } q(t) = 1 \text{ with } \beta = 4.$$

You should now verify that all the conditions (i)–(iv) are satisfied. Note that $p(t)$ has a jump at $t = 4$, from -4 to 0 .

PROBLEMS FOR SECTION 10.7

501. Solve the problem

$$\min \int_0^5 (u + x) dt, \quad \dot{x} = u - t, \quad x(0) = 1, \quad x(5) \text{ free}, \quad x \geq 0, \quad u \geq 0$$

(Hint: It seems a good idea to keep $x(t)$ as low as possible all the time.)

2. Solve the problem

$$\max \int_0^2 (1-x) dt, \quad \dot{x} = u, \quad x(0) = 1, \quad x(2) \text{ free}, \quad x \geq 0, \quad u \in [-1, 1]$$

(Hint: Start by reducing $x(t)$ as quickly as possible until $x(t) = 0$.)

3. Solve the problem

$$\max \int_0^{10} (-u^2 - x) dt, \quad \dot{x} = u, \quad x(0) = 1, \quad x(10) \text{ free}, \quad x \geq 0, \quad u \in \mathbb{R}$$

4. Consider the problem

$$\max \int_0^3 (4-t)u dt, \quad \dot{x} = u \in [0, 2], \quad x(0) = 1, \quad x(3) = 3, \quad t+1-x \geq 0 \quad (*)$$

(a) Solve the problem when the constraint $t+1-x \geq 0$ is not imposed.

(b) Solve problem (*).

10.8 Generalizations

In Chapter 9 and the previous sections of this chapter we have discussed how to solve the most commonly encountered problems in optimal control theory. Nevertheless, many important economic problems require methods that go beyond those presented so far.

More General Terminal Conditions

In some dynamical optimization problems the standard terminal conditions are replaced by the requirement that $x(t)$ at time t_1 hits a **target** defined as a certain curve or surface in $\mathbb{R}^n$.

The optimal path in such a problem must end at some point $\mathbf{x}^1$ and therefore, in particular, will solve the corresponding control problem where *all* the admissible paths end at $\mathbf{x}^1$. The conditions in Theorem 10.1.1 are therefore still valid, except the transversality conditions, which must be modified. See e.g. Seierstad and Sydsæter (1987), Chapter 3.

Markov Controls

The optimal solutions we have been looking for have been functions of time, $\mathbf{u}^*(t)$ and $\mathbf{x}^*(t)$. Such control functions are called **open-loop controls**. Faced with the problem of steering an economic system optimally, such open-loop controls are often inadequate. The problem is that "disturbances" of many types will almost always occur, which will divert the

system from the optimal path initially computed. If one still uses the "old" control $\mathbf{u}^*(t)$, the resulting development of the economy may be far from optimal, and it may end up at an undesirable final state.

This problem is partly resolved if we are able to "synthesize" the optimal control, in the sense of expressing the optimal control as a function of the present time s and the present state $\mathbf{y}$. In this case, for each time s and each point $\mathbf{y}$ in the state space, we specify the optimal control $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(s, \mathbf{y})$ to use. Such controls are called **closed-loop** or **Markov controls**. We can find such Markov controls by solving the control problem with an arbitrary start point $(s, \mathbf{y})$, $s \in [t_0, t_1]$. The controls $\mathbf{u}^*(t)$ obtained will depend on the starting point $(s, \mathbf{y})$, $\mathbf{u}^*(t) = \mathbf{u}_{s, \mathbf{y}}^*(t)$. Of course, at time s , the control $\tilde{\mathbf{u}}(s, \mathbf{y}) = \mathbf{u}_{s, \mathbf{y}}^*(s)$ is used. Then $\tilde{\mathbf{u}}(s, \mathbf{y})$ is the required Markov control.

But these Markov controls are only conditionally optimal. They tell us which control to use after a disturbance has occurred, but they are optimal only in the absence of further disturbances.

If we stipulate the probability of future disturbances and then want to optimize the expected value of the objective functional, this gives a stochastic control problem, in which optimal Markov controls are determined by a different set of necessary conditions. Discrete time versions of stochastic control problems are discussed extensively in Chapter 12 of this book.

Jumps in State Variables

So far we have assumed that the control functions are piecewise continuous and that the state variables are continuous. In certain applications (e.g. in the theory of investment), the optimum may require sudden jumps in the state variables. For example, somebody who buys an apartment worth, say, \$200,000 will experience a corresponding downward jump in their bank balance. See e.g. Seierstad and Sydsæter (1987), Chapter 3.

11

DIFFERENCE
EQUATIONS

*He (an economist) must study the present in the light of the past
for the purpose of the future.*

—J. N. Keynes¹

Many of the quantities economists study (such as income, consumption, and savings) are recorded at fixed time intervals (for example, each day, week, quarter, or year). Equations that relate such quantities at different discrete moments of time are called **difference equations**. For example, such an equation might relate the amount of national income in one period to the national income in one or more previous periods. In fact difference equations can be viewed as the discrete time counterparts of the differential equations in continuous time that were studied in Chapters 5–7.

Section 11.1 introduces first-order difference equations. The almost trivial fact is pointed out that, for given initial conditions, such equations always have a unique solution, provided the relevant function is defined everywhere. This is followed by a systematic study of linear equations, which can be solved explicitly.

Section 11.2 starts with a discussion of a cobweb model of the hog cycle that has received a great deal of attention in economics. Economic applications of difference equations to mortgage repayments and compound interest follow.

Second-order equations are introduced in Section 11.3. Their solution depends on a double initial condition. Then attention turns to linear equations, and we see in this section and the next that the theory closely resembles the corresponding theory for linear differential equations.

In Section 11.5 the theory in the preceding section is generalized in a straightforward way to higher-order equations, except that the stability criteria in Theorem 11.5.4 probably would strike you as less than obvious.

Section 11.6 is concerned with systems of difference equations. It explains the obvious matrix formulation of linear systems, and discusses their (global) stability properties.

The final Section 11.7 gives some stability results for nonlinear first-order difference equations. Some results on periodic solutions complete the chapter.

¹ John Neville Keynes (1852–1949) was the father of John Maynard Keynes.

11.1 First-Order Difference Equations

Let $t = 0, 1, 2, \dots$ denote different discrete time periods or moments of time. We usually call $t = 0$ the *initial period*. If $x(t)$ is a function defined for $t = 0, 1, 2, \dots$, we often use $x_0, x_1, x_2, \dots$ to denote $x(0), x(1), x(2), \dots$, and in general, we write x_t for $x(t)$.

Let $f(t, x)$ be a function defined for all positive integers t and all real numbers x . A first-order difference equation (or recurrence relation) in x_t can usually be written in the form

$$x_{t+1} = f(t, x_t), \quad t = 0, 1, 2, \dots \quad (1)$$

This is a first-order equation because it relates the value of a function in period $t + 1$ to the value of the same function in the previous period t only.²

Suppose x_0 is given. Then repeated application of equation (1) yields

$$\begin{aligned} x_1 &= f(0, x_0) \\ x_2 &= f(1, x_1) = f(1, f(0, x_0)) \\ x_3 &= f(2, x_2) = f(2, f(1, f(0, x_0))) \end{aligned}$$

and so on. For a given value of x_0 , we can compute x_t for any value of t . We call this the “insertion method” of solving (1).

Sometimes we can find a simple formula for x_t , but often this is not possible. A **general solution** of (1) is a function of the form $x_t = g(t; A)$ that satisfies (1) for every value of A , where A is an arbitrary constant. For each choice of x_0 there is usually one value of A such that $g(0, A) = x_0$.

We have seen that when x_0 is given, the successive values of x_t can be computed for any natural number t . Does this not tell us the whole story? In fact, economists often need to know more. In many economic applications, we are interested in establishing qualitative results such as the behaviour of the solution when t becomes very large, or how the solution depends on some parameters that might influence the difference equation. Such questions are difficult or impossible to handle if we rely only on the above insertion method.

Actually, the insertion method suffers from another defect even as a numerical procedure. For example, suppose that we have a difference equation like (1), and we want to compute x_{100} . A time-consuming process of successive insertions will finally yield an expression for x_{100} . However, computational errors can easily occur, and if we work with approximate numbers (as we are usually forced to do in serious applications), the approximation error might well explode and in the end give an entirely misleading answer. So there really is a need for a more systematic theory of difference equations. Ideally, the solutions should be expressed in terms of elementary functions. Unfortunately, this is possible only for rather restricted classes of equations.

² It would be more appropriate to call (1) a “recurrence relation”, and to reserve the term “difference equation” for an equation of the form $\Delta x_t = \tilde{f}(t, x_t)$, where Δx_t denotes the difference $x_{t+1} - x_t$. However, it is obvious how to transform a difference equation into an equivalent recurrence relation, and vice versa, so we make no distinction between the two kinds of equation.

A Simple Case

Consider the difference equation

$$x_{t+1} = ax_t, \quad t = 0, 1, \dots \quad (2)$$

where a is a constant. It is called homogeneous because if x_t is any solution, so is ax_t for any constant.

Suppose x_0 is given. Repeatedly applying (2) gives first $x_1 = ax_0$, next $x_2 = ax_1 = a \cdot ax_0 = a^2x_0$, then $x_3 = ax_2 = a \cdot a^2x_0 = a^3x_0$, and so on. In general,

$$x_t = a^t x_0, \quad t = 0, 1, \dots \quad (3)$$

The function $x_t = a^t x_0$ satisfies (2) for all t , as can be verified directly. For each given value of x_0 , there is clearly no other function that satisfies the equation.

EXAMPLE 1 Find the solution of the following difference equation which has $x_0 = 5$:

$$x_{t+1} = -3x_t, \quad t = 0, 1, \dots$$

Solution: From (3) we immediately get $x_t = 5(-3)^t$, $t = 0, 1, \dots$

EXAMPLE 2 (A multiplier–accelerator model of growth) Let Y_t denote national income, I_t total investment, and S_t total saving—all in period t . Suppose that savings are proportional to national income, and that investment is proportional to the change in income from period t to $t + 1$. Then, for $t = 0, 1, 2, \dots$,

$$(i) S_t = \alpha Y_t \quad (ii) I_{t+1} = \beta(Y_{t+1} - Y_t) \quad (iii) S_t = I_t$$

The last equation is the familiar equilibrium condition that saving equals investment in each period. Here α and β are positive constants, and we assume that $0 < \alpha < \beta$. Deduce a difference equation determining the path of Y_t , given Y_0 , and solve it.

Solution: From (i) and (iii), $I_t = \alpha Y_t$, and so $I_{t+1} = \alpha Y_{t+1}$. Inserting this into (ii) yields $\alpha Y_{t+1} = \beta(Y_{t+1} - Y_t)$, or $(\alpha - \beta)Y_{t+1} = -\beta Y_t$. Thus,

$$Y_{t+1} = \frac{\beta}{\beta - \alpha} Y_t = \left(1 + \frac{\alpha}{\beta - \alpha}\right) Y_t, \quad t = 0, 1, 2, \dots \quad (*)$$

Using (3) gives the solution

$$Y_t = \left(1 + \frac{\alpha}{\beta - \alpha}\right)^t Y_0, \quad t = 0, 1, 2, \dots$$

The difference equation (*) constitutes an instance of the equation

$$Y_{t+1} = (1 + g)Y_t, \quad t = 0, 1, 2, \dots$$

which describes growth at the constant proportional rate g each period. The solution of the equation is $Y_t = (1 + g)^t Y_0$. Note that $g = (Y_{t+1} - Y_t)/Y_t$.

Linear First-Order Equations with Constant Coefficients

Consider next the inhomogeneous difference equation

$$x_{t+1} = ax_t + b, \quad t = 0, 1, \dots \quad (4)$$

where a and b are constants. The homogeneous equation (2) is the special case with $b = 0$.

Starting with a given x_0 , we can calculate x_t algebraically for small t . Indeed

$$\begin{aligned} x_1 &= ax_0 + b \\ x_2 &= ax_1 + b = a(ax_0 + b) + b = a^2x_0 + (a+1)b \\ x_3 &= ax_2 + b = a(a^2x_0 + (a+1)b) + b = a^3x_0 + (a^2 + a + 1)b \end{aligned}$$

and so on. This makes the pattern clear. In general we have

$$x_t = a^t x_0 + (a^{t-1} + a^{t-2} + \dots + a + 1)b$$

According to the summation formula for a geometric series, $1 + a + a^2 + \dots + a^{t-1} = (1 - a^t)/(1 - a)$, for $a \neq 1$. Thus, for $t = 0, 1, 2, \dots$,

$$x_{t+1} = ax_t + b \iff x_t = a^t \left(x_0 - \frac{b}{1-a} \right) + \frac{b}{1-a} \quad (a \neq 1) \quad (5)$$

For $a = 1$, we have $1 + a + \dots + a^{t-1} = t$ and $x_t = x_0 + tb$ for $t = 1, 2, \dots$

EXAMPLE 3 Solve the following difference equations:

$$(a) \quad x_{t+1} = \frac{1}{2}x_t + 3, \quad (b) \quad x_{t+1} = -3x_t + 4$$

Solution: (a) Using (5) we obtain the solution $x_t = \left(\frac{1}{2}\right)^t (x_0 - 6) + 6$.

(b) In this case, (5) gives $x_t = (-3)^t (x_0 - 1) + 1$.

Equilibrium States and Stability

Consider the solution of $x_{t+1} = ax_t + b$ given in (5). If $x_0 = b/(1-a)$, then $x_t = b/(1-a)$ for all t . In fact, if $x_s = b/(1-a)$ for any $s \geq 0$, then $x_{s+1} = a(b/(1-a)) + b = b/(1-a)$, and again $x_{s+2} = b/(1-a)$, and so on. We conclude that if x_t ever becomes equal to $b/(1-a)$ at some time s , then x_t will remain at this constant level for each $t \geq s$. The constant $x^* = b/(1-a)$ is called an **equilibrium** (or **stationary**) state for $x_{t+1} = ax_t + b$.

An alternative way of finding an equilibrium state x^* is to seek a solution of $x_{t+1} = ax_t + b$ with $x_t = x^*$ for all t . Such a solution must satisfy $x_{t+1} = x_t = x^*$ and so $x^* = ax^* + b$. Therefore, for $a \neq 1$, we get $x^* = b/(1-a)$ as before.

Suppose the constant a in (5) is less than 1 in absolute value—that is, $-1 < a < 1$. Then $a^t \rightarrow 0$ as $t \rightarrow \infty$, so (5) implies that

$$x_t \rightarrow x^* = b/(1-a) \quad \text{as} \quad t \rightarrow \infty \quad (6)$$

Hence, if $|a| < 1$, the solution converges to the equilibrium state as $t \rightarrow \infty$. The equation is then called **globally asymptotically stable**.

EXAMPLE 4 Equation (a) in Example 3 is globally asymptotically stable because $a = 1/2$. The equilibrium state is $b/(1-a) = 3/(1-1/2) = 6$. We see from the solution given in that example that $x_t \rightarrow 6$ as $t \rightarrow \infty$.

Equation (b) in Example 3 is not stable because $|a| = |-3| = 3 > 1$. The solution does *not* converge to the equilibrium state $x^* = 1$ as $t \rightarrow \infty$, except if $x_0 = 1$ —in fact, there are explosive oscillations.

Two kinds of stability are shown in Figs. 1(a) and (b). In the first case, x_t decreases monotonically and converges to the equilibrium state x^* . In the second case, x_t exhibits decreasing fluctuations or **damped oscillations** around x^* .

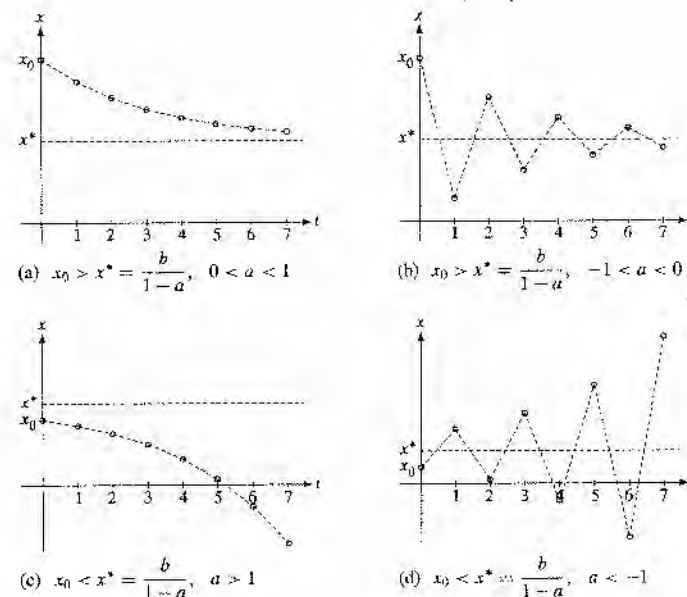


Figure 1

If $|a| > 1$, then the absolute value of a^t tends to ∞ as $t \rightarrow \infty$. From (5), it follows that x_t moves farther and farther away from the equilibrium state, except when $x_0 = b/(1-a)$.

Two versions of this phenomenon are illustrated in Figs. 1 (c) and (d). In the first case, x_t tends to $-\infty$, and in the second case, x_t exhibits increasing fluctuations or **explosive oscillations** around the equilibrium state.

Variable Right-Hand Side

Consider briefly the case when the constant b in equation (4) is replaced by an arbitrary given function of t :

$$x_{t+1} = ax_t + b_t, \quad t = 0, 1, \dots \quad (7)$$

where a is still a constant. Starting with a given x_0 , we can again calculate x_t algebraically for small t . Indeed

$$\begin{aligned} x_1 &= ax_0 + b_0 \\ x_2 &= ax_1 + b_1 = a(ax_0 + b_0) + b_1 = a^2x_0 + ab_0 + b_1 \\ x_3 &= ax_2 + b_2 = a(a^2x_0 + ab_0 + b_1) + b_2 = a^3x_0 + a^2b_0 + ab_1 + b_2 \end{aligned}$$

and so on. This makes the pattern clear. In each case, the formula for x_t begins with the term $a^t x_0$, and then adds the terms $a^{t-1}b_0, a^{t-2}b_1, \dots, ab_{t-2}, b_{t-1}$ in turn. We thus arrive at the following general result (which can be proved by induction):

$$x_{t+1} = ax_t + b_t \iff x_t = a^t x_0 + \sum_{k=1}^t a^{t-k} b_{k-1}, \quad t = 1, 2, \dots \quad (8)$$

Linear Equations with a Variable Coefficient

Sometimes economists need to consider a more general form of the linear difference equation (7), where the coefficient a can vary over time. This will be the case in Example 11.2.3 concerning the present value of an income stream when the interest rate varies.

The general first-order linear difference equation takes the form

$$x_{t+1} = a_t x_t + b_t, \quad t = 0, 1, 2, \dots \quad (9)$$

where, unlike in (7), the coefficient a_t depends on t . Proceeding as before, we calculate x_t explicitly for the first few values of t , starting with a given x_0 when $t = 0$. We have

$$\begin{aligned} x_1 &= a_0 x_0 + b_0 \\ x_2 &= a_1 x_1 + b_1 = a_1(a_0 x_0 + b_0) + b_1 = a_1 a_0 x_0 + a_1 b_0 + b_1 \end{aligned}$$

Then, omitting the details of the next two calculations, we have

$$\begin{aligned} x_3 &= a_2 a_1 a_0 x_0 + a_2 a_1 b_0 + a_2 b_1 + b_2 \\ x_4 &= a_3 a_2 a_1 a_0 x_0 + a_3 a_2 a_1 b_0 + a_3 a_2 b_1 + a_3 b_2 + b_3 \end{aligned}$$

This is considerably more complicated than when a_t was independent of t , yet you should still be able to discern a pattern. Indeed, the successive coefficients of x_0 are

$$a_0, \quad a_0 a_1, \quad a_0 a_1 a_2, \quad a_0 a_1 a_2 a_3$$

Using an obvious notation for the product we have

$$\prod_{s=0}^{t-1} a_s \quad \text{for } t = 0, 1, 2, 3 \quad (10)$$

In fact, for $t = 1, 2, \dots$, the general formula for x_t becomes

$$x_t = \left(\prod_{s=0}^{t-1} a_s \right) x_0 + \left(\prod_{s=1}^{t-1} a_s \right) b_0 + \left(\prod_{s=2}^{t-1} a_s \right) b_1 + \dots + \left(\prod_{s=t-1}^{t-1} a_s \right) b_{t-2} + b_{t-1}$$

This can be written as

$$x_t = \left(\prod_{s=0}^{t-1} a_s \right) x_0 + \sum_{k=0}^{t-1} \left(\prod_{s=k+1}^{t-1} a_s \right) b_k \quad (11)$$

provided we agree that the product $\prod_{s=t}^{t-1} a_s$ of zero terms is 1. A formal proof of (11.11) can be given by mathematical induction.

NOTE 1 (Differential versus difference equations) Consider the analogous differential equation $\dot{x}(t) = ax(t) + b$ studied in Section 5.4. If we use the (rough) approximation $\dot{x}(t) \approx x_{t+1} - x_t$, the equation becomes $x_{t+1} - x_t = ax_t + b$, or $x_{t+1} = (1+a)x_t + b$, whose solution according to (5) is $x_t = (1+a)^{t-1}(x_0 - b/a) + b/a$. The equilibrium state is b/a for both equations. The differential equation $\dot{x}(t) = ax(t) + b$ converges to b/a provided $a < 0$. The difference equation $x_{t+1} = (1+a)x_t + b$ converges to b/a provided $|1+a| < 1$, i.e. $-2 < a < 0$. So convergence of the solution to the differential equation implies the same for the difference equation, but not conversely — when $a \leq -2$ the differential equation has a convergent solution even though the difference equation does not.

PROBLEMS FOR SECTION 11.1

1. Find the solutions of the following difference equations with the given values of x_0 :

- (a) $x_{t+1} = 2x_t + 4, \quad x_0 = 1$ (b) $3x_{t+1} = x_t + 2, \quad x_0 = 2$
 (c) $2x_{t+1} + 3x_t + 2 = 0, \quad x_0 = -1$ (d) $x_{t+1} - x_t + 3 = 0, \quad x_0 = 3$

2. Consider the difference equation $x_{t+1} = ax_t + b$ in (4) and explain how its solution behaves in each of the following cases, with $x^* = b/(1-a)$ (for $a \neq 1$):

- (a) $0 < a < 1, x_0 < x^*$ (b) $-1 < a < 0, x_0 < x^*$ (c) $a > 1, x_0 > x^*$
 (d) $a < -1, x_0 > x^*$ (e) $a \neq 1, x_0 = x^*$ (f) $a = -1, x_0 \neq x^*$
 (g) $a = 1, b > 0$ (h) $a = 1, b < 0$ (i) $a = 1, b = 0$

3. By substituting $y_t = x_t - b/(1-a)$ transform equation (4) into a homogeneous difference equation in y_t . Solve it and find a new confirmation of (5).

4. (a) Consider the difference equation

$$y_{t+1}(a + by_t) = cy_t, \quad t = 0, 1, \dots$$

where a , b , and c are positive constants, and $y_0 > 0$. Show that $y_t > 0$ for all t .

- (b) Define a new function x_t by $x_t = 1/y_t$. Show that by using this substitution, the new difference equation is of the type in (4). Next solve the difference equation $y_{t+1}(2 + 3y_t) = 4y_t$, assuming that $y_0 = 1/2$. What is the limit of y_t as $t \rightarrow \infty$?
5. Consider the difference equation $x_t = \sqrt{x_{t-1}} - 1$ with $x_0 = 5$. Compute x_1 , x_2 , and x_3 . What about x_4 ? (This problem illustrates that a solution may not exist if the domain of the function f in (1) is restricted in any way.)

11.2 Economic Applications

In this section we consider several interesting applications of the theory studied in the previous section.

EXAMPLE 1

(The hog cycle: a cobweb model) Assume that the total cost of raising q pigs is $C(q) = \alpha q + \beta q^2$. Suppose there are N identical pig farms. Let the demand function for pigs be given by $D(p) = \gamma - \delta p$, as a function of the price p , where the constants α , β , γ , and δ are all positive. Suppose, too, that each farmer behaves competitively, taking the price p as given and maximizing profits $\pi(q) = pq - C(q) = pq - \alpha q - \beta q^2$.

The quantity $q > 0$ maximizes profits only if

$$\pi'(q) = p - \alpha - 2\beta q = 0 \quad \text{and so} \quad q = (p - \alpha)/2\beta$$

We see that $\pi'(q) > 0$ for $q < (p - \alpha)/2\beta$, and $\pi'(q) < 0$ for $q > (p - \alpha)/2\beta$. Thus, $q = (p - \alpha)/2\beta$ maximizes profits provided $p > \alpha$. In aggregate, the total supply of pigs from all N farms is the function

$$S = N(p - \alpha)/2\beta \quad (p > \alpha)$$

of the price p . Now, suppose it takes one period to raise each pig, and that when choosing the number of pigs to raise for sale at time $t + 1$, each farmer remembers the price p_t at time t and expects p_{t+1} to be the same as p_t . Then the aggregate supply at time $t + 1$ will be $S(p_t) = N(p_t - \alpha)/2\beta$.

Equilibrium of supply and demand in all periods requires that $S(p_t) = D(p_{t+1})$, which implies that $N(p_t - \alpha)/2\beta = \gamma - \delta p_{t+1}$, $t = 0, 1, \dots$. Solving for p_{t+1} in terms of p_t and the parameters gives the difference equation

$$p_{t+1} = -\frac{N}{2\beta\delta}p_t + \frac{\alpha N + 2\beta\gamma}{2\beta\delta}, \quad t = 1, 2, \dots \quad (*)$$

This is a special case of equation (11.1.4), with p_t replacing x_t , with $a = -N/2\beta\delta$, and with $b = (\alpha N + 2\beta\gamma)/2\beta\delta$. The solution of (*) can be expressed as

$$p_t = p^* + (-a)^t(p_0 - p^*) \quad (a = N/2\beta\delta)$$

where p^* is the equilibrium price $p^* = b/(1 - a) = (\alpha N + 2\beta\gamma)/(2\beta\delta + N)$. Equation (*) is stable if $|-a| < 1$, which happens when $N < 2\beta\delta$. In this case, $p_t \rightarrow p^*$ as $t \rightarrow \infty$. The solution in this case is illustrated in Fig. 1.

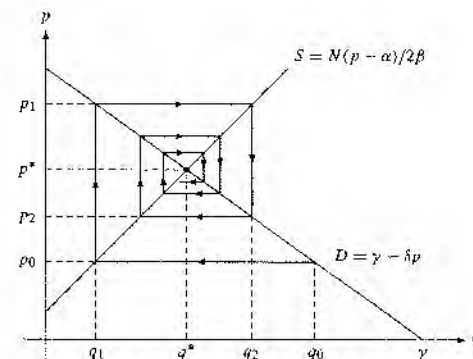


Figure 1 The cobweb model in Example 1—the stable case.

Here, q_0 is the supply of pigs at time 0. The price at which all these can be sold is p_0 . This determines the supply q_1 one period later. The resulting price at which they sell is p_1 , and so on.

In the stable case when $N < 2\beta\delta$ the resulting price cycles are damped, and both price and quantity converge to a steady-state equilibrium at (q^*, p^*) . This is also an equilibrium of supply and demand. If $N > 2\beta\delta$, however, then the oscillations explode, and eventually p_t becomes less than α . Then some pig farms must go out of business, and the solution has to be described in a different way. There is no convergence to a steady state in this case. A third, intermediate, case occurs when $N = 2\beta\delta$ and $a = 1$. Then the pair (q_t, p_t) oscillates perpetually between the two values $(\gamma - \delta p_0, p_0)$ and $(\delta(p_0 - \alpha), \alpha + \gamma/\delta - p_0)$ in even- and odd-numbered periods, respectively.

EXAMPLE 2

(Mortgage repayments) A particular case of the difference equation (11.1.4) occurs when a family borrows an amount K at time 0 as a home mortgage. Suppose there is a fixed interest rate r per period (usually a month rather than a year). Suppose too that there are equal repayments of amount a each period, until the mortgage is paid off after n periods (for example, 360 months = 30 years). The outstanding balance or *principal* b_t on the loan in period t satisfies the difference equation $b_{t+1} = (1 + r)b_t - a$, with $b_0 = K$ and $b_n = 0$. This difference equation can be solved by using (11.1.5), which gives

$$b_t = (1 + r)^t(K - a/r) + a/r$$

But $b_t = 0$ when $t = n$, so $0 = (1+r)^n(K - a/r) + a/r$. Solving for K yields

$$K = \frac{a}{r} [1 - (1+r)^{-n}] = a \sum_{t=1}^n (1+r)^{-t} \quad (*)$$

The original loan, therefore, is equal to the present discounted value of n equal repayments of amount a each period, starting in period 1. Solving for the payment a each period instead yields

$$a = \frac{rK}{1 - (1+r)^{-n}} = \frac{rK(1+r)^n}{(1+r)^n - 1} \quad (**)$$

Formulas (*) and (**) are the same as those derived by a more direct argument in EMEA, Section 10.6. $\square$

EXAMPLE 3 (Compound interest and PDVs with constant interest rate) The result in (11.1.8) can be applied to describe the changes over time in a savings account whose balance is subject to compound interest.

Let w_t denote the value of the assets held in the account at the end of period t . Further, let c_t be the amount withdrawn for consumption and y_t the amount deposited as income during period t . If the interest rate per period is a constant r , the relevant difference equation is

$$w_{t+1} = (1+r)w_t + y_{t+1} - c_{t+1}, \quad t = 0, 1, 2, \dots \quad (1)$$

Using (11.1.8), the solution of (1) is

$$w_t = (1+r)^t w_0 + \sum_{k=1}^t (1+r)^{t-k} (y_k - c_k), \quad t = 1, 2, \dots \quad (2)$$

Let us multiply each term in (2) by $(1+r)^{-t}$, which is a factor of sufficient economic importance to have earned a standard name, namely the **discount factor**. The result is

$$(1+r)^{-t} w_t = w_0 + \sum_{k=1}^t (1+r)^{-k} (y_k - c_k) \quad (3)$$

If time 0 is now, then the left-hand side is the **present discounted value** (PDV) of the assets in the account at time t . Equation (3) says that this is equal to

- (a) initial assets w_0 ;
- (b) plus the total PDV of all future deposits, $\sum_{k=1}^t (1+r)^{-k} y_k$;
- (c) minus the total PDV of all future withdrawals, $\sum_{k=1}^t (1+r)^{-k} c_k$.

If time t is now, the formula for w_t in (2) can be interpreted as follows:

Current assets w_t reflect the interest earned on initial assets w_0 , with adjustments for the interest earned on all later deposits, or foregone because of later withdrawals. $\square$

EXAMPLE 4 (Compound interest and PDVs with variable interest rates) We modify the compound interest and present discounted value calculations in the previous example to allow interest rates that vary from period to period. The relevant difference equation becomes

$$w_{t+1} = (1+r_{t+1})w_t + y_{t+1} - c_{t+1}, \quad t = 0, 1, 2, \dots \quad (4)$$

Formula (11.1.11) yields

$$w_t = \left[\prod_{s=0}^{t-1} (1+r_{s+1}) \right] w_0 + \sum_{k=0}^{t-1} \left[\prod_{s=k+1}^{t-1} (1+r_{s+1}) \right] (y_{k+1} - c_{k+1})$$

or, equivalently,

$$w_t = \left[\prod_{s=1}^t (1+r_s) \right] w_0 + \sum_{k=1}^t \left[\prod_{s=k+1}^t (1+r_s) \right] (y_k - c_k) \quad (5)$$

Define the **discount factor** D_t by

$$D_t = \frac{1}{\prod_{s=1}^t (1+r_s)} = \prod_{s=1}^t (1+r_s)^{-1} \quad (6)$$

Note that when $r_s = r$ for all s , then $D_t = (1+r)^{-t}$, the discount factor used in Example 3 (see (3)). In the special case of no deposits or withdrawals, equation (5) reduces to

$$w_t = w_0 \prod_{s=1}^t (1+r_s)$$

just as one would expect. After all, w_0 invested initially becomes $w_0(1+r_1)$ after one period, then $w_0(1+r_1)(1+r_2)$ after two periods, and so on. So the discount factors have an entirely appropriate form.

Multiplying each term of (5) by the discount factor D_t yields

$$D_t w_t = w_0 + \sum_{k=1}^t \left[\frac{\prod_{s=k+1}^t (1+r_s)}{\prod_{s=1}^t (1+r_s)} \right] (y_k - c_k)$$

But

$$\begin{aligned} \frac{\prod_{s=k+1}^t (1+r_s)}{\prod_{s=1}^t (1+r_s)} &= \frac{(1+r_{k+1}) \cdots (1+r_t)}{(1+r_1) \cdots (1+r_k)(1+r_{k+1}) \cdots (1+r_t)} \\ &= \frac{1}{(1+r_1) \cdots (1+r_k)} = D_k \end{aligned}$$

Hence,

$$D_t w_t = w_0 + \sum_{k=1}^t D_k (y_k - c_k), \quad t = 1, 2, \dots \quad (7)$$

The interpretation in terms of present discounted values (PDVs) is exactly the same as before (see formula (3)).

Introduce the **interest factor** R_k , defined by

$$R_k = D_k/D_1 = \prod_{s=k+1}^T (1+r_s) \quad (8)$$

Then formula (5) can be written as

$$w_T = R_0 w_0 + \sum_{k=1}^T R_k (y_k - c_k) \quad (9)$$

which is the appropriate generalization of formula (2).

PROBLEMS FOR SECTION 11.2

- Find the solution of (1) for $r = 0.2$, $w_0 = 1000$, $y_t = 100$, and $c_t = 50$.
- Suppose that at time $t = 0$, you borrow \$100,000 at a fixed interest rate r of 7% per year. You are supposed to repay the loan in 30 equal annual repayments so that after $n = 30$ years, the mortgage is paid off. How much is each repayment?
- (SM) 3. (a) A loan of amount $\$L$ is taken out on January 1 of year 0. Instalment payments for the principal and interest are paid annually, commencing on January 1 of year 1. Let the interest rate be $r < 2$, so that the interest amounts to rL for the first payment. The contract states that the principal share of the repayment will be half the size of the interest share. Show that the debt after January 1 of year n is $(1 - r/2)^n L$.
 (b) Find r when it is known that exactly half the original loan is paid after 10 years.
 (c) What will the remaining payments be each year if the contract is not changed?
- (SM) 4. Work through the mortgage example in Example 2 when interest rates are variable. Note that, in practice, variable interest mortgages have repayments that increase when the interest rate increases, and decrease when it decreases. Why is this? What would happen if there were a large enough unforeseen increase in interest rates without any increase in repayments?

11.3 Second-Order Difference Equations

So far this chapter has considered first-order difference equations, in which each value x_{t+1} of a function is expressed in terms of its value x_t in the previous period only. Next we present a typical example from economics where it is necessary to consider second-order difference equations.

EXAMPLE 1 (A multiplier-accelerator growth model) Let Y_t denote national income, C_t total consumption, and I_t total investment in a country at time t . Assume that for $t = 0, 1, \dots$,

$$(i) \ Y_t = C_t + I_t \quad (ii) \ C_{t+1} = aY_t + b \quad (iii) \ I_{t+1} = c(C_{t+1} - C_t)$$

where a , b , and c are positive constants.

Equation (i) simply states that national income is divided between consumption and investment. Equation (ii) expresses the assumption that consumption in period $t + 1$ is a linear function of national income in the previous period. This is the “multiplier” part of the model. Finally, equation (iii) states that investment in period $t + 1$ is proportional to the change in consumption from the previous period. The idea is that the existing capital stock provides enough capacity for production to meet current consumption. So investment is only needed when consumption increases. This is the “accelerator” part of the model. The combined “multiplier-accelerator” model has been studied by several economists, notably P. A. Samuelson.

Assume that consumption C_0 and investment I_0 are known in the initial period $t = 0$. Then by (i), $Y_0 = C_0 + I_0$, and by (ii), $C_1 = aY_0 + b$. From (iii), we obtain $I_1 = c(C_1 - C_0)$, and then (i) in turn gives $Y_1 = C_1 + I_1$. Hence, Y_1 , C_1 , and I_1 are all known. Turning to (ii) again, we find C_2 , then (iii) gives us the value of I_2 , and (i) in turn produces the value of Y_2 . Obviously, in this way, we can obtain expressions for C_t , Y_t , and I_t for all t in terms of C_0 , I_0 , and the constants a , b , and c . However, the expressions derived get increasingly complicated.

Another method of studying the system is usually more enlightening. It consists of eliminating two of the unknown functions so as to end up with one difference equation in one unknown. Here we use this method to end up with a difference equation in Y_t . To do so, note that equations (i) to (iii) are valid for all $t = 0, 1, \dots$. Replace t with $t + 1$ in (ii) and (iii), and t with $t + 2$ in (i) to obtain

$$(iv) \ C_{t+2} = aY_{t+1} + b \quad (v) \ I_{t+2} = c(C_{t+2} - C_{t+1}) \quad (vi) \ Y_{t+2} = C_{t+2} + I_{t+2}$$

Inserting (iv) and (ii) into (v) yields $I_{t+2} = ac(Y_{t+1} - Y_t)$. Inserting this result and (iv) into (vi) gives $Y_{t+2} = aY_{t+1} + b + ac(Y_{t+1} - Y_t)$. Rearranging we have

$$Y_{t+2} - a(1+c)Y_{t+1} + acY_t = b, \quad t = 0, 1, \dots \quad (vii)$$

This is a second-order linear difference equation with Y_t as the unknown function, and with constant coefficients of Y_{t+1} and Y_t . The next section sets out a general method for solving such equations. (See Problem 11.4.3.)

The typical second-order difference equation can be written in the form

$$x_{t+2} = f(t, x_t, x_{t+1}), \quad t = 0, 1, \dots \quad (1)$$

Suppose that f is defined for all possible values of the variables (t, x_t, x_{t+1}) . Suppose x_0 and x_1 have fixed values. Letting $t = 0$ in (1), we see that $x_2 = f(0, x_0, x_1)$. Letting $t = 1$ yields $x_3 = f(1, x_1, f(0, x_0, x_1))$. By successively inserting $t = 2, t = 3, \dots$ into (1), the values of x_t for all t are uniquely determined in terms of x_0 and x_1 . Note in particular that there are infinitely many solutions, and that the solution of the equation is uniquely determined by its values in the first two periods. By definition, a **general solution** of (1) is a function of the form

$$x_t = g(t; A, B) \quad (2)$$

that satisfies (1) and has the property that every solution of (1) can be obtained from (2) by choosing appropriate values of A and B .

Linear Equations

The general second-order linear difference equation is

$$x_{t+2} + a_t x_{t+1} + b_t x_t = c_t \quad (3)$$

where a_t , b_t , and c_t are given functions of t , with $b_t \neq 0$. The associated **homogeneous equation**

$$x_{t+2} + a_t x_{t+1} + b_t x_t = 0 \quad (4)$$

is obtained from (3) by replacing c_t with 0. Compare these equations with the linear differential equations (6.2.1) and (6.2.2). By arguments which are much the same as for differential equations (but simpler), the following results are easy to establish:

THEOREM 11.3.1

- (a) The **general solution** of the homogeneous difference equation

$$x_{t+2} + a_t x_{t+1} + b_t x_t = 0 \quad \text{is} \quad x_t = Au_t^{(1)} + Bu_t^{(2)}$$

where $u_t^{(1)}$ and $u_t^{(2)}$ are any two solutions that are not proportional, and A and B are arbitrary constants.

- (b) The **general solution** of the nonhomogeneous difference equation

$$x_{t+2} + a_t x_{t+1} + b_t x_t = c_t \quad \text{is} \quad x_t = Au_t^{(1)} + Bu_t^{(2)} + u_t^*$$

where $Au_t^{(1)} + Bu_t^{(2)}$ is the general solution of the associated homogeneous equation (with c_t replaced by zero), and u_t^* is any particular solution of the nonhomogeneous equation.

NOTE 1 In order to use the theorem, we need to know when two solutions of (4) are linearly independent, i.e. not proportional. The following necessary and sufficient condition is easy to apply (and generalizes easily to the case of n functions):

$$u_t^{(1)} \text{ and } u_t^{(2)} \text{ are linearly independent} \iff \begin{vmatrix} u_0^{(1)} & u_0^{(2)} \\ u_1^{(1)} & u_1^{(2)} \end{vmatrix} \neq 0 \quad (5)$$

See Problem 5 for a proof.

A General Solution

There is no universally applicable method of discovering the two linearly independent solutions of (4) that we need in order to find the general solution of the equation. But if we know two linearly independent solutions $u_t^{(1)}$ and $u_t^{(2)}$ of the homogeneous equation (4) and thereby its general solution, then it is always possible to find the general solution of the nonhomogeneous equation (3).

Indeed, define

$$D_t = u_t^{(1)} u_{t+1}^{(2)} - u_{t+1}^{(1)} u_t^{(2)}$$

Then, provided $D_t \neq 0$ for all $t = 1, 2, \dots$, the general solution of (3) is given by

$$x_t = Au_t^{(1)} + Bu_t^{(2)} - u_t^{(1)} \sum_{k=1}^t \frac{c_{k-1} u_k^{(2)}}{D_k} + u_t^{(2)} \sum_{k=1}^t \frac{c_{k-1} u_k^{(1)}}{D_k} \quad (6)$$

where A and B are arbitrary constants. (See Hildebrand (1968).)

When the coefficients a_t and b_t in (4) are constants independent of t , then it is always possible to find a simple formula for the general solution of (4). The next section shows how to do this.

PROBLEMS FOR SECTION 11.3

- SM** 1. Prove by direct substitution that the following functions of t are solutions of the associated difference equation (A and B are constants):
 - (a) $x_t = A + B 2^t$, $x_{t+2} - 3x_{t+1} + 2x_t = 0$
 - (b) $x_t = A 3^t + B 4^t$, $x_{t+2} - 7x_{t+1} + 12x_t = 0$
2. Prove that $x_t = A + B t$ is the general solution of $x_{t+2} - 2x_{t+1} + x_t = 0$.
3. Prove that $x_t = A 3^t + B 4^t$ is the general solution of $x_{t+2} - 7x_{t+1} + 12x_t = 0$.
4. Prove that $x_t = A 2^t + B t 2^t + 1$ is the general solution of $x_{t+2} - 4x_{t+1} + 4x_t = 1$.
- SM** 5. Prove the equivalence in (5). (*Hint:* If the determinant is zero, then the two columns are linearly dependent, and, since both $u_t^{(1)}$ and $u_t^{(2)}$ are solutions of equation (4), this dependence will propagate to $u_t^{(1)}$ and $u_t^{(2)}$ for all t .)
- SM** 6. (a) Find an expression for the solution of $x_{t+2} - 2x_{t+1} + x_t = c_t$, using the answer to Problem 2 along with equation (6).
 (b) Find the solution when $c_t = t$. (*Hint:* Prove that $\sum_{k=1}^t (k-1)k = \frac{1}{3}(t-1)t(t+1)$.)

11.4 Linear Equations with Constant Coefficients

Consider the homogeneous linear equation

$$x_{t+2} + ax_{t+1} + bx_t = 0 \quad (1)$$

where a and b are arbitrary constants, $b \neq 0$, and x_t is the unknown function. According to Theorem 11.3.1(a), finding the general solution of (1) requires us to discover two solutions $u_t^{(1)}$ and $u_t^{(2)}$ that are linearly independent. On the basis of experience gained in some of the previous problems, it should not be surprising that we try to find solutions to (1) of the form $x_t = m^t$. Then $x_{t+1} = m^{t+1} = m \cdot m^t$ and $x_{t+2} = m^{t+2} = m^2 \cdot m^t$. Inserting these expressions into (1) yields $m^t(m^2 + am + b) = 0$. If $m \neq 0$, then m^t satisfies (1) provided that

$$m^2 + am + b = 0 \quad (2)$$

This is the **characteristic equation** of the difference equation. Its solutions are

$$m_1 = -\frac{1}{2}a + \frac{1}{2}\sqrt{a^2 - 4b}, \quad m_2 = -\frac{1}{2}a - \frac{1}{2}\sqrt{a^2 - 4b} \quad (3)$$

There are three different cases, which are summed up in the following theorem:

THEOREM 11.4.1

The general solution of

$$x_{t+2} + ax_{t+1} + bx_t = 0 \quad (b \neq 0)$$

is as follows:

(I) If $a^2 - 4b > 0$ (the characteristic equation has two distinct real roots),

$$x_t = Am_1^t + Bm_2^t, \quad m_{1,2} = -\frac{1}{2}a \pm \frac{1}{2}\sqrt{a^2 - 4b}$$

(II) If $a^2 - 4b = 0$ (the characteristic equation has one real double root),

$$x_t = (A + Bt)m^t, \quad m = -\frac{1}{2}a$$

(III) If $a^2 - 4b < 0$ (the characteristic equation has no real roots),

$$x_t = r^t(A \cos \theta t + B \sin \theta t), \quad r = \sqrt{b}, \quad \cos \theta = -\frac{a}{2\sqrt{b}}, \quad \theta \in [0, \pi]$$

NOTE 1 If x_0 and x_1 are given numbers, then in all three cases the constants A and B are uniquely determined. For instance, in case (I), A and B are uniquely determined by the simultaneous equations $x_0 = A + B$ and $x_1 = Am_1 + Bm_2$.

NOTE 2 The solution in case (III) can be expressed as

$$x_t = Cr^t \cos(\theta t + \omega) \quad (4)$$

where ω and C are arbitrary constants. (See the corresponding case for differential equations in Section 6.3.)

Proof of Theorem 11.4.1: (I) The case $a^2 - 4b > 0$ is the simplest. Then the roots m_1 and m_2 of (2) are real and different, and $u_t^{(1)} = m_1^t$ and $u_t^{(2)} = m_2^t$ are both solutions of (1). The determinant in (11.3.5) has the value $m_2 - m_1 \neq 0$, so the two solutions are linearly independent, and the general solution is consequently as given in (I).

(II) If $a^2 - 4b = 0$, then $m = -\frac{1}{2}a$ is a double root of (2). This means that $m^2 + am + b = (m + \frac{1}{2}a)^2$. In addition to $u_t^{(1)} = m^t$, the function $u_t^{(2)} = tm^t$ also satisfies (1) (see Problem 6). Moreover, these two functions are linearly independent because the determinant in (11.3.5) is equal to $m = -\frac{1}{2}a$. (Note that $a \neq 0$ because $b \neq 0$.) The general solution is, therefore, as indicated in (II).

(III) If $a^2 - 4b < 0$, the roots of (2) are complex. The two functions $u_t^{(1)} = r^t \cos \theta t$ and $u_t^{(2)} = r^t \sin \theta t$ are linearly independent. Indeed, the determinant in (11.3.5) is

$$\begin{vmatrix} 1 & 0 \\ r \cos \theta & r \sin \theta \end{vmatrix} = r \sin \theta = \sqrt{b} \sqrt{1 - \cos^2 \theta} = \sqrt{b} \sqrt{1 - a^2/4b} = \frac{1}{2} \sqrt{4b - a^2} > 0$$

Moreover, direct substitution shows that both these functions satisfy (1).

Indeed, let us show that $u_t^{(1)} = r^t \cos \theta t$ satisfies (1). We find that $u_{t+1}^{(1)} = r^{t+1} \cos \theta(t+1)$ and $u_{t+2}^{(1)} = r^{t+2} \cos \theta(t+2)$. Hence, using the formula (B.1.8) for the cosine of a sum, we get

$$\begin{aligned} u_{t+2}^{(1)} + au_{t+1}^{(1)} + bu_t^{(1)} &= r^{t+2} \cos \theta(t+2) + ar^{t+1} \cos \theta(t+1) + br^t \cos \theta t \\ &= r^t [r^2 (\cos \theta t \cos 2\theta - \sin \theta t \sin 2\theta) + ar (\cos \theta t \cos \theta - \sin \theta t \sin \theta) + b \cos \theta t] \\ &= r^t [(r^2 \cos 2\theta + ar \cos \theta + b) \cos \theta t - (r^2 \sin 2\theta + ar \sin \theta) \sin \theta t] \end{aligned}$$

Here the coefficients of $\cos \theta t$ and $\sin \theta t$ are both equal to 0 because $r^2 \cos 2\theta + ar \cos \theta + b = r^2(2\cos^2 \theta - 1) + ar \cos \theta + b = b(2a^2/4b - 1) + a\sqrt{b}(-a/2\sqrt{b}) + b = 0$, and likewise $r^2 \sin 2\theta + ar \sin \theta = 2r^2 \sin \theta \cos \theta + ar \sin \theta = 2r^2(-a/2r) \sin \theta + ar \sin \theta = 0$. This shows that $u_t^{(1)} = r^t \cos \theta t$ satisfies equation (1), and a similar argument shows that so does $u_t^{(2)} = r^t \sin \theta t$. ■

NOTE 3 An alternative argument for the solution in (III) relies on properties of the complex exponential function. In trigonometric form the roots in (3) are $m_1 = \alpha + i\beta = r(\cos \theta + i \sin \theta)$ and $m_2 = \alpha - i\beta = r(\cos \theta - i \sin \theta)$, with $\theta \in [0, \pi]$, $r = \sqrt{a^2 + \beta^2} = \sqrt{b}$, $\cos \theta = \alpha/r = -a/2\sqrt{b}$, and $\sin \theta = \beta/r = (\sqrt{b - a^2/4})/\sqrt{b}$.

By de Moivre's formula (B.3.8), $m_1^t = r^t(\cos \theta t + i \sin \theta t)$ and $m_2^t = r^t(\cos \theta t - i \sin \theta t)$. The complex functions m_1^t and m_2^t both satisfy (1), and so does every complex linear combination of them. In particular, $\frac{1}{2}(m_1^t + m_2^t) = r^t \cos \theta t$ and $\frac{1}{2i}(m_1^t - m_2^t) = r^t \sin \theta t$ both satisfy (1). The general solution of (1) is therefore as given in case (III).

We see that when the characteristic equation has complex roots, the solution of (1) involves oscillations. The number r is the **growth factor**. Note that when $|r| < 1$, then $|Ar^t| \rightarrow 0$ as $t \rightarrow \infty$ and the oscillations are **damped**. If $|r| > 1$, the oscillations are **explosive**, and in the case $|r| = 1$, we have undamped oscillations.

Let us now consider some examples of difference equations of the form (1).

EXAMPLE 1 Find the general solutions of

$$(a) \ x_{t+2} - 5x_{t+1} + 6x_t = 0 \quad (b) \ x_{t+2} - 6x_{t+1} + 9x_t = 0 \quad (c) \ x_{t+2} - x_{t+1} + x_t = 0$$

Solution: (a) The characteristic equation is $m^2 - 5m + 6 = 0$, whose roots are $m_1 = 2$ and $m_2 = 3$, so the general solution is

$$x_t = A2^t + B3^t$$

(b) The characteristic equation is $m^2 - 6m + 9 = (m - 3)^2 = 0$, so $m = 3$ is a double root. The general solution is

$$x_t = (A + Bt)3^t$$

(c) The characteristic equation is $m^2 - m + 1 = 0$, with complex roots $m_1 = \frac{1}{2}(1 + i\sqrt{3})$ and $m_2 = \frac{1}{2}(1 - i\sqrt{3})$. Here $r = \sqrt{b} = 1$ and $\cos \theta = 1/2$, so $\theta = \frac{1}{3}\pi$. The general solution is

$$x_t = A \cos \frac{\pi}{3}t + B \sin \frac{\pi}{3}t$$

The frequency is $(\pi/3)/(2\pi) = 1/6$ and the growth factor is $\sqrt{b} = 1$, so the oscillations are undamped.

The Nonhomogeneous Case

Now consider the nonhomogeneous equation

$$x_{t+2} + ax_{t+1} + bx_t = c_t \quad (b \neq 0) \quad (5)$$

According to Theorem 11.3.1(b), its general solution is

$$x_t = Au_t^{(1)} + Bu_t^{(2)} + u_t^* \quad (6)$$

where $Au_t^{(1)} + Bu_t^{(2)}$ is the general solution of the associated homogeneous equation (1), and u_t^* is a particular solution of (5). Theorem 11.4.1 tells us how to find $Au_t^{(1)} + Bu_t^{(2)}$. How do we find u_t^* ? The general formula in (11.3.6) gives one answer, but it involves a lot of work, even when c_t is a simple function.

In some cases it is much easier. For example, suppose $c_t = c$, where c is a constant. Then (5) takes the form

$$x_{t+2} + ax_{t+1} + bx_t = c \quad (c \text{ is a constant}) \quad (7)$$

We look for a solution of the form $x_t = C$, where C is a constant. Then $x_{t+1} = x_{t+2} = C$, so inserting $x_t = C$ into (7) gives $C + aC + bC = c$. Provided $1 + a + b \neq 0$, we get $C = c/(1 + a + b)$. Hence,

$$u_t^* = \frac{c}{1 + a + b} \quad \text{is a particular solution of (7) when } 1 + a + b \neq 0 \quad (8)$$

(If $1 + a + b = 0$, no constant function satisfies (7). To handle this case, see Problem 4.)

Consider more generally the case in which c_t in (5) is a linear combination of terms of the form

$$a^t, \quad t^m, \quad \cos qt, \quad \text{or} \quad \sin qt$$

or products of such terms. Then the method of undetermined coefficients can be used to obtain a particular solution of (5). (If the function c_t in (5) happens to satisfy the homogeneous equation, the procedures described below must be modified.)³

EXAMPLE 2

Solve the equation $x_{t+2} - 5x_{t+1} + 6x_t = 4^t + t^2 + 3$.

Solution: According to Example 1(a), the associated homogeneous equation has the general solution $A2^t + B3^t$. To find a particular solution we look for constants C , D , E , and F such that

$$u_t^* = C4^t + Dt^2 + Et + F$$

is a solution. (You cannot put $E = 0$, even though there is no t term on the right-hand side of the original equation.) Inserting u_t^* into the given equation yields

$$\begin{aligned} C4^{t+2} + D(t+2)^2 + E(t+2) + F - 5[C4^{t+1} + D(t+1)^2 + E(t+1) + F] \\ + 6(C4^t + Dt^2 + Et + F) = 4^t + t^2 + 3 \end{aligned}$$

Expanding and rearranging, $2C4^t + 2Dt^2 + (-6D + 2E)t + (-D - 3E + 2F) = 4^t + t^2 + 3$. For this to hold for all $t = 0, 1, \dots$ one must have $2C = 1$, $2D = 1$, $-6D + 2E = 0$, and $-D - 3E + 2F = 3$. It follows that $C = 1/2$, $D = 1/2$, $E = 3/2$, and $F = 4$. The general solution of the equation is, therefore,

$$x_t = A2^t + B3^t + \frac{1}{2}4^t + \frac{1}{2}t^2 + \frac{3}{2}t + 4$$

Stability

Suppose an economy evolves according to some difference equation (or system of difference equations). If the right number of initial conditions are imposed, the system has a unique solution. Also, if one or more initial conditions are changed, the solution changes. An important question is this: Will a small change in the initial conditions have any effect on the long-run behaviour of the solution, or will its effect die out as $t \rightarrow \infty$? In the latter case, the system is called **stable**. On the other hand, if a small change in the initial conditions might lead to significant differences in the long-run behaviour of the solution, then the system is **unstable**. Because an initial state cannot be pinpointed exactly, but only approximately, stability in the sense indicated above is sometimes a minimum requirement for a model to be economically useful.

For the remainder of this section, $u_t^{(1)}$ and $u_t^{(2)}$ will denote the two solutions of (1) that emerge in the proof of Theorem 11.4.1.

Consider the second-order nonhomogeneous difference equation (5) whose general solution is of the form $x_t = Au_t^{(1)} + Bu_t^{(2)} + u_t^*$. Equation (5) is called **globally asymptotically stable** if the general solution $Au_t^{(1)} + Bu_t^{(2)}$ of the associated homogeneous equation tends to 0 as $t \rightarrow \infty$, for all values of A and B . So the effect of the initial conditions which determine A and B dies out as $t \rightarrow \infty$.

³ For more details, we refer to Goldberg (1958) or Gandolfo (1980).

If $Au_t^{(1)} + Bu_t^{(2)}$ tends to 0 as $t \rightarrow \infty$ for all values of A and B , then in particular $u_t^{(1)} \rightarrow 0$ as $t \rightarrow \infty$ (choose $A = 1, B = 0$), and $u_t^{(2)} \rightarrow 0$ as $t \rightarrow \infty$ (choose $A = 0, B = 1$). On the other hand, these two conditions are obviously sufficient for $Au_t^{(1)} + Bu_t^{(2)}$ to approach 0 as $t \rightarrow \infty$.

We claim that $u_t^{(1)} \rightarrow 0$ and $u_t^{(2)} \rightarrow 0$ as $t \rightarrow \infty$ if and only if the moduli of the roots of $m^2 + am + b = 0$ are both less than 1.⁴

First, in the case when the characteristic polynomial has two distinct real roots $m_1 \neq m_2$, the two solutions are $u_t^{(1)} = m_1^t$ and $u_t^{(2)} = m_2^t$. In this case, we see that $u_t^{(1)} \rightarrow 0$ and $u_t^{(2)} \rightarrow 0$ as $t \rightarrow \infty$ if and only if $|m_1| < 1$ and $|m_2| < 1$.

Second, when the characteristic polynomial has a double root $m = -a/2$, then the two linearly independent solutions are m^t and tm^t . Again, $|m| < 1$ is a necessary and sufficient condition for these two solutions to approach 0 as $t \rightarrow \infty$.

Third, suppose the characteristic polynomial has complex roots $m = \alpha \pm i\beta$. Then $\alpha = -\frac{1}{2}a$ and $\beta = \frac{1}{2}\sqrt{4b - a^2}$. So the modulus of each root is equal to $|m| = \sqrt{\alpha^2 + \beta^2} = \sqrt{b}$. We argued before that the two solutions $r^t \cos \theta t$ and $r^t \sin \theta t$ tend to 0 as t tends to infinity if and only if $r = \sqrt{b} < 1$ —that is, if and only if $b < 1$.

To summarize, we have the following result:

THEOREM 11.4.2

The equation

$$x_{t+2} + ax_{t+1} + bx_t = c_t$$

is globally asymptotically stable if and only if the following two equivalent conditions are satisfied:

- (A) The roots of the characteristic equation $m^2 + am + b = 0$ have moduli strictly less than 1.
- (B) $|a| < 1 + b$ and $b < 1$.

It remains to prove that (B) is equivalent to (A). Assume first that $b > a^2/4$. Then the characteristic equation has complex roots $m_{1,2} = \alpha \pm i\beta$. Hence, $|m_1| = |m_2| = \sqrt{b}$, so (B) obviously implies (A).

On the other hand, since $f(m) = m^2 + am + b$ is never zero when $b > a^2/4$, and since $f(0) = b$ is positive, the parabola $y = f(m)$ is always above the m -axis, so $f(m)$ must be positive for all m . In particular $f(1) = 1 + a + b > 0$ and $f(-1) = 1 - a + b > 0$. But these conditions together are equivalent to $|a| < 1 + b$, so (A) implies that the conditions in (B) are also necessary. Problem 11 asks you to analyse the case of real roots.

EXAMPLE 3 Investigate the stability of the equation $x_{t+2} - \frac{1}{6}x_{t+1} - \frac{1}{6}x_t = c_t$.

⁴ See Section B.3. Note that, if m is a real number, the modulus of m equals the absolute value of m .

Solution: In this case $a = -1/6$ and $b = -1/6$, so $|a| = 1/6$ and $1 + b = 5/6$. Thus, according to Theorem 11.4.2, the equation is stable. This conclusion can be confirmed by looking at the general solution of the associated homogeneous equation, which is $x_t = A(1/2)^t + B(-1/3)^t$. Clearly, $x_t \rightarrow 0$ irrespective of the values of A and B , so the given equation is globally asymptotically stable. $\square$

EXAMPLE 4 Investigate the stability of equation (vii) in Example 11.3.1, where a and c are positive,

Solution: From Theorem 11.4.2 the equation is stable if and only if $a(1 + c) < 1 + ac$ and $ac < 1$ —that is, if and only if $a < 1$ and $ac < 1$. (See also Problem 3.) $\square$

PROBLEMS FOR SECTION 11.4

Find the general solutions of the difference equations in Problems 1 and 2.

1. (a) $x_{t+2} - 6x_{t+1} + 8x_t = 0$ (b) $x_{t+2} - 8x_{t+1} + 16x_t = 0$
(c) $x_{t+2} + 2x_{t+1} + 3x_t = 0$ (d) $3x_{t+2} + 2x_t = 4$
2. (a) $x_{t+2} + 2x_{t+1} + x_t = 9 \cdot 2^t$ (b) $x_{t+2} - 3x_{t+1} + 2x_t = 3 \cdot 5^t + \sin(\frac{1}{2}\pi t)$

3. (a) Consider the difference equation (vii) in Example 11.3.1, with $a > 0$, $c > 0$, and $a \neq 1$. Find a special solution of the equation.
(b) Find the characteristic equation of the associated homogeneous equation and determine when it has two different real roots, or a double real root, or two complex roots.

4. Consider equation (7) and assume that $1 + a + b = 0$. If $a \neq -2$, find a constant D such that Dt satisfies (7). If $a = -2$, find a constant D such that Dt^2 satisfies (7).

5. A model of location uses the difference equation

$$D_{n+2} - 4(ab + 1)D_{n+1} + 4a^2b^2D_n = 0, \quad n = 0, 1, \dots$$

where a and b are constants, and D_n is the unknown function. Find the solution of this equation assuming that $1 + 2ab > 0$.

6. Consider equation (1) assuming that $\frac{1}{3}a^2 - b = 0$, so that the characteristic equation has a real double root $m = -a/2$. Let $x_t = u_t(-a/2)^t$ and prove that x_t satisfies (1) provided that u_t satisfies the equation $u_{t+2} - 2u_{t+1} + u_t = 0$. Use the result in Problem 11.3.2 to find x_t .

7. Investigate the global asymptotic stability of the following equations:

$$(a) \quad x_{t+2} - \frac{1}{3}x_t = \sin t \quad (b) \quad x_{t+2} - x_{t+1} - x_t = 0 \quad (c) \quad x_{t+2} - \frac{1}{8}x_{t+1} + \frac{1}{6}x_t = t^2 e^t$$

8. (a) A model due to R. J. Ball and E. Smolensky is based on the following system:

$$C_t = cY_{t-1}, \quad K_t = \sigma Y_{t-1}, \quad Y_t = C_t + K_t - K_{t-1}$$

Here C_t denotes consumption, K_t capital stock, Y_t net national product, whereas c and σ are positive constants. Give an economic interpretation of the equations.

- (b) Derive a difference equation of the second order for Y_t . Find necessary and sufficient conditions for the solution of this equation to have explosive oscillations.

59. (a) A model by J. R. Hicks uses the following difference equation:

$$Y_{t+2} - (b+k)Y_{t+1} + kY_t = a(1+g)^t, \quad t = 0, 1, \dots$$

where a , b , g , and k are constants. Find a particular solution Y_t^* of the equation.

- (b) Give conditions for the characteristic equation to have two complex roots.
 (c) Find the growth factor r of the oscillations when the conditions obtained in part (b) are satisfied, and determine when the oscillations are damped.
10. (a) In their study of the "wage-price spiral" of inflation, Frisch, Haavelmo, Nøregård-Rasmussen, and Zeuthen considered the following system for $t = 0, 1, \dots$:

$$(i) \quad \frac{W_{t+2} - W_{t+1}}{W_{t+1}} = \frac{P_{t+1} - P_t}{P_t} \quad (ii) \quad P_t = \gamma + \beta W_t$$

Here W_t denotes the wage level and P_t the price index at time t , whereas γ and β are constants. Give economic interpretations for the two equations.

- (b) Deduce from (i) and (ii) the following equation for W_t :

$$\frac{W_{t+2}}{\gamma + \beta W_{t+1}} = \frac{W_{t+1}}{\gamma + \beta W_t}, \quad t = 0, 1, \dots \quad (iii)$$

- (c) Use (iii) to prove that $W_{t+1} = c(\gamma + \beta W_t)$, $t = 0, 1, \dots$, where $c = W_1/P_0$, and find a general expression for W_t when $c\beta \neq 1$. Under what conditions will the equation be globally asymptotically stable, and what is then the limit of W_t as $t \rightarrow \infty$?

HARDER PROBLEMS

60. 11. Prove that the conditions in (B) in Theorem 11.4.2 are equivalent to the condition in (A) for the case when the characteristic polynomial has real roots, by studying the parabola $y = f(m) = m^2 + am + b$. (Hint: Consider the values of $f(-1)$, $f(1)$, $f'(-1)$, and $f'(1)$.)

11.5 Higher-Order Equations

In this section we briefly record some results for general n th-order difference equations,

$$x_{t+n} = f(t, x_t, x_{t+1}, \dots, x_{t+n-1}), \quad t = 0, 1, \dots \quad (1)$$

Suppose f is defined for all values of the variables. If we require that $x_0, x_1, \dots, x_{n-1}$ have given fixed values, then by letting $t = 0$ in (1), we find that $x_n = f(0, x_0, x_1, \dots, x_{n-1})$ is uniquely determined. Then letting $t = 1$ in (1) yields $x_{n+1} = f(1, x_1, x_2, \dots, x_n) = f(1, x_1, x_2, \dots, f(0, x_0, x_1, \dots, x_{n-1}))$. And so on. Thus the solution of equation (1) for all $t \geq n$ (if it exists) is uniquely determined by the values x_t takes in the first n periods, $0, 1, \dots, n-1$.

The **general solution** of (1) is a function $x_t = g(t; C_1, \dots, C_n)$ that depends on n arbitrary constants $C_1, \dots, C_n$, satisfies (1), and has the property that every solution of (1) can be obtained by giving $C_1, \dots, C_n$ appropriate values.

Linear Equations

The general theory for second-order linear difference equations is easily generalized to n th-order linear equations.

THEOREM 11.5.1

The general solution of the homogeneous linear difference equation

$$x_{t+n} + a_1(t)x_{t+n-1} + \dots + a_{n-1}(t)x_{t+1} + a_n(t)x_t = 0$$

with $a_n(t) \neq 0$ is given by

$$x_t = C_1 u_t^{(1)} + \dots + C_n u_t^{(n)}$$

where $u_t^{(1)}, \dots, u_t^{(n)}$ are n linearly independent solutions of the equation, and $C_1, \dots, C_n$ are arbitrary constants.

THEOREM 11.5.2

The general solution of the nonhomogeneous linear difference equation

$$x_{t+n} + a_1(t)x_{t+n-1} + \dots + a_{n-1}(t)x_{t+1} + a_n(t)x_t = b_t$$

with $a_n(t) \neq 0$ is given by

$$x_t = C_1 u_t^{(1)} + \dots + C_n u_t^{(n)} + u_t^*$$

where $C_1 u_t^{(1)} + \dots + C_n u_t^{(n)}$ is the general solution of the corresponding homogeneous equation, and u_t^* is a particular solution of the nonhomogeneous equation.

NOTE 1 To use these theorems, it helps to know the following generalization of (11.3.5): If $u_t^{(1)}, \dots, u_t^{(n)}$ are solutions of the homogeneous difference equation in Theorem 11.5.1, then

$$u_t^{(1)}, \dots, u_t^{(n)} \text{ are linearly independent} \iff \begin{vmatrix} u_0^{(1)} & \dots & u_0^{(n)} \\ u_1^{(1)} & \dots & u_1^{(n)} \\ \vdots & \vdots & \vdots \\ u_{n-1}^{(1)} & \dots & u_{n-1}^{(n)} \end{vmatrix} \neq 0 \quad (2)$$

Constant Coefficients

The general linear difference equation of n th order with constant coefficients takes the form

$$x_{t+n} + a_1 x_{t+n-1} + \cdots + a_{n-1} x_{t+1} + a_n x_t = b_t, \quad t = 0, 1, \dots \quad (3)$$

The corresponding homogeneous equation is

$$x_{t+n} + a_1 x_{t+n-1} + \cdots + a_{n-1} x_{t+1} + a_n x_t = 0, \quad t = 0, 1, \dots \quad (4)$$

We try to find solutions to (4) of the form $x_t = m^t$. Inserting this solution and cancelling the common factor m^t yields the **characteristic equation**

$$m^n + a_1 m^{n-1} + \cdots + a_{n-1} m + a_n = 0 \quad (5)$$

According to the fundamental theorem of algebra, this equation has exactly n roots, when each is counted according to its multiplicity.

Suppose first that equation (5) has n different real roots $m_1, m_2, \dots, m_n$. Then $m_1^t, m_2^t, \dots, m_n^t$ all satisfy (4). These functions are moreover linearly independent, so the general solution of (4) in this case is

$$x_t = C_1 m_1^t + C_2 m_2^t + \cdots + C_n m_n^t$$

This is *not* the general solution of (4) if equation (5) has multiple roots and/or complex roots. The general method for finding n linearly independent solutions of (4) is as follows:

Find the roots of equation (5) together with their multiplicity.

- (A) A real root m_i with multiplicity 1 gives the one solution m_i^t .
- (B) A real root m_j with multiplicity $p > 1$ gives the p solutions $m_j^t, t m_j^t, \dots, t^{p-1} m_j^t$.
- (C) A pair of complex roots $\alpha \pm i\beta$, each with multiplicity 1, gives the two solutions $r^t \cos \theta t$, $r^t \sin \theta t$, where $r = \sqrt{\alpha^2 + \beta^2}$, and $\theta \in [0, \pi]$ satisfies $\cos \theta = \alpha/r$, $\sin \theta = \beta/r$.
- (D) A pair of complex roots $\alpha \pm i\beta$, each with multiplicity $q > 1$, gives the $2q$ solutions $u, v, tu, tv, \dots, t^{q-1}u, t^{q-1}v$, with $u = r^t \cos \theta t$ and $v = r^t \sin \theta t$, where $r = \sqrt{\alpha^2 + \beta^2}$, and $\theta \in [0, \pi]$ satisfies $\cos \theta = \alpha/r$ and $\sin \theta = \beta/r$.

In order to find the general solution of the nonhomogeneous equation (3), it remains to find a particular solution u_t^* of (3). If b_t is a linear combination of products of terms of the form $\alpha^t, t^m, \cos qt$ and $\sin qt$, as in Section 11.4, the method of undetermined coefficients again can be used.

Stability

Equation (3) is **globally asymptotically stable** if the general solution $C_1 u_t^{(1)} + \cdots + C_n u_t^{(n)}$ of the associated homogeneous equation (4) tends to 0 as $t \rightarrow \infty$, for all values of the constants $C_1, \dots, C_n$. Then the effect of the initial conditions “dies out” as $t \rightarrow \infty$.

As in the case $n = 2$, equation (3) is globally asymptotically stable if and only if $u_t^{(i)} \rightarrow 0$ as $t \rightarrow \infty$ for all $i = 1, \dots, n$. Each u_i corresponds to a root m_i of the characteristic polynomial. Again, $u_t^{(i)} \rightarrow 0$ as $t \rightarrow \infty$ if and only if modulus of the corresponding solution of the characteristic equation is < 1 .

THEOREM 11.5.3

A necessary and sufficient condition for (3) to be globally asymptotically stable is that all roots of the characteristic polynomial of the equation have moduli strictly less than 1.

The following result gives a stability condition based directly on the coefficients of the characteristic equation. (The dashed lines have been included to make it easier to see the partitioned structure of the determinants.) See Chipman (1950) for discussion.

THEOREM 11.5.4 (SCHUR)

Let

$$m^n + a_1 m^{n-1} + \cdots + a_{n-1} m + a_n$$

be a polynomial of degree n with real coefficients. A necessary and sufficient condition for all roots of the polynomial to have moduli less than 1 is that

$$\begin{vmatrix} 1 & a_n \\ a_n & 1 \end{vmatrix} > 0, \quad \begin{vmatrix} 1 & 0 & a_n & a_{n-1} \\ a_1 & 1 & 0 & a_n \\ a_n & 0 & 1 & a_1 \\ a_{n-1} & a_n & 0 & 1 \end{vmatrix} > 0, \quad \dots,$$

$$\begin{vmatrix} 1 & 0 & \cdots & 0 & a_n & a_{n-1} & \cdots & a_1 \\ a_1 & 1 & \cdots & 0 & 0 & a_n & \cdots & a_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n-1} & a_{n-2} & \cdots & 1 & 0 & 0 & \cdots & a_n \\ a_n & 0 & \cdots & 0 & 1 & a_1 & \cdots & a_{n-1} \\ a_{n-1} & a_n & \cdots & 0 & 0 & 1 & \cdots & a_{n-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_1 & a_2 & \cdots & a_n & 0 & 0 & \cdots & 1 \end{vmatrix} > 0$$

In the case when $n = 1$, Theorem 11.5.4 says that $m + a_1 = 0$ has a root with modulus < 1 if and only if $\begin{vmatrix} 1 & a_1 \\ a_1 & 1 \end{vmatrix} > 0$, i.e. if and only if $a_1^2 < 1$. (Of course, this is clear without

using the theorem.) Now, $a_1^2 < 1 \iff |a_1| < 1$, so

$$x_{t+1} + a_1 x_t = c_t \text{ is globally asymptotically stable} \iff |a_1| < 1 \quad (6)$$

When $n = 2$, Theorem 11.5.4 says that both roots of $m^2 + a_1m + a_2 = 0$ have moduli < 1 if and only if

$$D_1 = \begin{vmatrix} 1 & a_2 \\ a_2 & 1 \end{vmatrix} > 0 \quad \text{and} \quad D_2 = \begin{vmatrix} 1 & 0 & a_2 & a_1 \\ a_1 & 1 & 0 & a_2 \\ a_2 & 0 & 1 & a_1 \\ a_1 & a_2 & 0 & 1 \end{vmatrix} > 0 \quad (*)$$

Evaluating the determinants yields

$$D_1 = 1 - a_2^2 \quad \text{and} \quad D_2 = (1 - a_2)^2(1 + a_1 + a_2)(1 - a_1 + a_2)$$

Here $D_1 > 0 \iff |a_2| < 1$. If $D_1 > 0$, then $1 + a_2 > 0$ and $1 - a_2 > 0$, so that

$$D_2 > 0 \iff (1 + a_1 + a_2)(1 - a_1 + a_2) > 0 \iff |a_1| < 1 + a_2$$

$$\iff -(1 + a_2) < a_1 < 1 + a_2 \iff 1 + a_1 + a_2 > 0 \text{ and } 1 - a_1 + a_2 > 0$$

Hence, if $D_1 > 0$ and $D_2 > 0$, then

$$1 + a_1 + a_2 > 0 \quad \text{and} \quad 1 - a_1 + a_2 > 0 \quad \text{and} \quad 1 - a_2 > 0 \quad (**)$$

On the other hand, if these inequalities are satisfied, then adding the first two shows that $2 + 2a_2 > 0$, i.e. $1 + a_2 > 0$. But then we see that (**) implies that D_1 and D_2 defined by (*) are both positive. Thus the conditions in (*) are equivalent to the conditions in (**). Since $1 + a_1 + a_2 > 0$ and $1 - a_1 + a_2 > 0$ are equivalent to $|a_1| < 1 + a_2$, we see that Theorem 11.4.2 is the particular case of Theorem 11.5.4 that holds when $n = 2$.

PROBLEMS FOR SECTION 11.5

- 1.** Solve the following difference equations:

(a) $x_{i+3} - 3x_{i+1} + 2x_i = 0$ (b) $x_{i+4} + 2x_{i+2} + x_i = 8$

2. Examine the stability of the following difference equations:

(a) $x_{t+2} - \frac{1}{3}x_t = \sin t$ (b) $x_{t+2} - x_{t+1} - x_t = 0$
(c) $x_{t+2} - \frac{1}{6}x_{t+1} + \frac{1}{6}x_t = t^2 e^t$ (d) $x_{t+2} + 3x_{t+1} - 4x_t = t - 1$

3. In the a_1a_2 -plane, describe the domain defined by the inequalities (**).

4. Examine when the equation in Problem 11.4.9 is globally asymptotically stable, assuming $k > 0$ and $b > 0$.

5. A paper by Akerlof and Stiglitz studies the equation

$$K_{i+2} + (\sigma\beta/\alpha - 2)K_{i+1} + (1 - \sigma\beta)K_i = d$$

where the constants α , β , and σ are positive. Find a condition for both roots of the characteristic polynomial to be complex, and find a necessary and sufficient condition for stability.

11.6 Systems of Difference Equations

A system of first-order difference equations in the n unknown functions $x_1(t), \dots, x_n(t)$ can usually be expressed in the **normal form**:⁵

$$\begin{aligned} x_1(t+1) &= f_1(t, x_1(t), \dots, x_n(t)) \\ &\dots\dots\dots, \\ x_n(t+1) &= f_n(t, x_1(t), \dots, x_n(t)) \end{aligned} \quad t = 0, 1, \dots \quad (1)$$

If $x_1(0), \dots, x_n(0)$ are specified, then $x_1(1), \dots, x_n(1)$ are found by substituting $t = 0$ in (1), next $x_1(2), \dots, x_n(2)$ are found by substituting $t = 1$, etc. Thus the values of $x_1(t), \dots, x_n(t)$ are uniquely determined for all t (assuming that $f_1, \dots, f_n$ are defined for all values of the variables). Thus the solution of (1) is uniquely determined by the values of $x_1(0), \dots, x_n(0)$.

The general solution of (1) is given by n functions

$$x_1 = g_1(t; C_1, \dots, C_n), \dots, x_n = g_n(t; C_1, \dots, C_n) \quad (*)$$

with the property that an arbitrary solution $(x_1(t), \dots, x_n(t))$ is obtained from (*) by giving $C_1, \dots, C_n$ appropriate values.

Of course, there are no general methods that lead to explicit solutions of (1) in "closed" form. Only in some special cases can we find closed form solutions.

EXAMPLE 3 Find the general solution of the system

$$(i) \quad x_{t+1} = \frac{1}{4}x_t + \frac{1}{3}y_t, \quad (ii) \quad y_{t+1} = \frac{1}{2}x_t + \frac{2}{3}y_t, \quad t = 0, 1, \dots$$

Solution: Guided by the method we used in Section 6.5 to solve systems of two differential equations, we try to derive a second-order difference equation with x_t as the only unknown. From (i) we obtain (iii) $y_t = 3x_{t+1} - \frac{3}{2}x_t$, which inserted into (ii) yields (iv) $y_{t-1} = 2x_t + 1 - \frac{1}{2}x_t$. Replacing t by $t+1$ in (i), we get (v) $x_{t+2} = \frac{1}{2}x_{t+1} + \frac{1}{3}y_{t+1}$. Inserting (iv) into (v), then rearranging, we get

$$x_{t+2} - \frac{2}{6}x_{t+1} + \frac{1}{6}x_t = 0$$

⁵ In this section, the argument τ is usually included in parentheses, when subscripts are needed to indicate different variables in the system.

The characteristic equation is $m^2 - \frac{7}{6}m + \frac{1}{6} = 0$, with the roots $m_1 = 1$, $m_2 = \frac{1}{6}$. The general solution for $x(t)$ is then easily found. In turn, (iii) is used to find y . The result is

$$x_i = A + B\left(\frac{1}{6}\right)^i, \quad y_i = \frac{3}{2}A - B\left(\frac{1}{6}\right)^i$$

Matrix Formulation of Linear Systems

If the functions $f_1, \dots, f_n$ in (1) are linear, we obtain the system

$$\begin{aligned} x_1(t+1) &= a_{11}(t)x_1(t) + \cdots + a_{1n}(t)x_n(t) + b_1(t) \\ &\vdots \\ x_n(t+1) &= a_{n1}(t)x_1(t) + \cdots + a_{nn}(t)x_n(t) + b_n(t) \end{aligned} \quad , \quad t = 0, 1, \dots \quad (2)$$

Suppose we define

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \quad \mathbf{A}(t) = \begin{pmatrix} a_{11}(t) & \dots & a_{1n}(t) \\ \vdots & \ddots & \vdots \\ a_{n1}(t) & \dots & a_{nn}(t) \end{pmatrix}, \quad \mathbf{b}(t) = \begin{pmatrix} b_1(t) \\ \vdots \\ b_n(t) \end{pmatrix}$$

Then (2) is equivalent to the matrix equation

$$\mathbf{x}(t+1) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{b}(t), \quad t = 0, 1, \dots, \quad (3)$$

The method suggested in Example 1 allows one, in general, to derive a linear n th order difference equation in one of the unknowns, say x_1 . When x_1 has been found, we can also find $x_2, \dots, x_n$.

If all the coefficients $a_{ij}(t)$ are constants, $a_{ij}(t) = a_{ij}$, then $A(t)$ is a constant matrix A . In this case, with constant coefficients, (3) reduces to

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}(t), \quad t = 0, 1, \dots \quad (4)$$

Inserting $t = 0, 1, \dots$, we get successively $\mathbf{x}(1) = \mathbf{A}\mathbf{x}(0) + \mathbf{b}(0)$, $\mathbf{x}(2) = \mathbf{A}\mathbf{x}(1) + \mathbf{b}(1) = \mathbf{A}^2\mathbf{x}(0) + \mathbf{A}\mathbf{b}(0) + \mathbf{b}(1)$, $\mathbf{x}(3) = \mathbf{A}\mathbf{x}(2) + \mathbf{b}(2) = \mathbf{A}^3\mathbf{x}(0) + \mathbf{A}^2\mathbf{b}(0) + \mathbf{A}\mathbf{b}(1) + \mathbf{b}(2)$, and, in general, $\mathbf{x}(t) = \mathbf{A}^t\mathbf{x}(0) + \mathbf{A}^{t-1}\mathbf{b}(0) + \mathbf{A}^{t-2}\mathbf{b}(1) + \dots + \mathbf{b}(t-1)$, or, equivalently,

$$\mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0) + \sum_{k=1}^t \mathbf{A}^{t-k} \mathbf{b}(k-1) \quad (5)$$

which is obviously an n -dimensional version of (11.1.8) (with $\mathbf{A}^0 = \mathbf{I}$ as the identity matrix). If $\mathbf{b}(t) = \mathbf{0}$ for all t , then

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) \iff \mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0), \quad t = 0, 1, \dots \quad (6)$$

Stability of Linear Systems

The linear system (4) is said to be **globally asymptotically stable** if, no matter what the initial conditions, the general solution of the corresponding homogeneous system $\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t)$

tends to 0 as t tends to infinity. According to (6), the homogeneous system has the solution $\mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0)$. Hence (4) is globally asymptotically stable if and only if $\mathbf{A}^t$ tends to 0 as $t \rightarrow \infty$, for each choice of initial vector $\mathbf{x}(0) = \mathbf{x}_0$.

The $n \times n$ matrix A^t is said to **converge** to $0_{n \times n}$ as $t \rightarrow \infty$ if and only if each component of A^t converges to 0. Obviously, $A^t \rightarrow 0$ implies that $A^t x_0 \rightarrow 0$ for every x_0 in $\mathbb{R}^n$. Conversely, if $A^t x_0 \rightarrow 0$ for every x_0 in $\mathbb{R}^n$, then in particular $A^t e_j \rightarrow 0$ for each unit vector $e_j = (0, \dots, 1, \dots, 0)$ in $\mathbb{R}^n$. But $A^t e_j$ is just the j th column of A^t , so we infer that $A^t \rightarrow 0_{n \times n}$. Thus we have proved that

$$A'x_0 \xrightarrow{t \rightarrow \infty} 0 \quad \text{for all } x_0 \text{ in } \mathbb{R}^n \iff A' \xrightarrow{t \rightarrow \infty} 0 \quad (7)$$

A necessary and sufficient condition for this is:

$$A^t \xrightarrow{t \rightarrow \infty} 0 \iff \text{all the eigenvalues of } A \text{ have moduli less than } 1 \quad (8)$$

Proof of (8) if $\mathbf{A}$ is diagonalizable: According to Example 1.6.2, if $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $\mathbf{A}$, then $\mathbf{A}^t = \mathbf{P} \operatorname{diag}(\lambda_1^t, \dots, \lambda_n^t) \mathbf{P}^{-1}$. The conclusion follows. $\blacksquare$

The following result follows immediately:

THEOREM 13.6.1

A necessary and sufficient condition for system $\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}(t)$ to be globally asymptotically stable is that all the eigenvalues of the matrix $\mathbf{A}$ have moduli (strictly) less than 1.

Suppose in particular that the vector $\mathbf{b}(t)$ is independent of t , $\mathbf{b}(t) = \mathbf{b}$. According to (5) the solution of the system is

$$\mathbf{x}(t) = \mathbf{A}^t \mathbf{x}(0) + (\mathbf{A}^{t-1} + \mathbf{A}^{t-2} + \dots + \mathbf{A} + \mathbf{I})\mathbf{b} \quad (9)$$

Suppose that the system is globally asymptotically stable so that all the eigenvalues of $\mathbf{A}$ have moduli less than 1. Now, expanding the left-hand side,

$$(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \dots + \mathbf{A}^{t-1})(\mathbf{I} - \mathbf{A}) = \mathbf{I} - \mathbf{A}^t \quad (10)$$

Since $\lambda = 1$ cannot be an eigenvalue of A (it has modulus *equal to* 1), we have $|I - A| \neq 0$, so $(I - A)^{-1}$ exists. Multiplying (10) on the right by $(I - A)^{-1}$ yields $I + A + A^2 + \dots + A^{t-1} = (I - A^t)(I - A)^{-1}$. As $t \rightarrow \infty$, because (8) implies that $A^t \rightarrow 0$, we get

$$I + A + A^2 + \dots + A^{t-1} \rightarrow (I - A)^{-1} \quad \text{as } t \rightarrow \infty \quad (11)$$

We conclude that:

THEOREM 11.6.2

If all the eigenvalues of $A = (a_{ij})_{n \times n}$ have moduli (strictly) less than 1, the difference equation

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + \mathbf{b}, \quad t = 0, 1, \dots$$

is globally asymptotically stable, and every solution $\mathbf{x}(t)$ of the equation converges to the constant vector $(I - A)^{-1}\mathbf{b}$.

The following theorem can often be used to show that a matrix has only eigenvalues with moduli less than 1 (see e.g. Corollary 6.1.5 in Horn and Johnson (1985)):

THEOREM 11.6.3

Let $A = (a_{ij})$ be an arbitrary $n \times n$ matrix and suppose that

$$\sum_{j=1}^n |a_{ij}| < 1 \quad \text{for all } i = 1, \dots, n$$

Then all the eigenvalues of A have moduli less than 1.

PROBLEMS FOR SECTION 11.6

1. Find the solutions of the following systems of difference equations with the given initial conditions (in each case $t = 0, 1, \dots$):

$$\begin{array}{ll} \text{(a)} & \begin{array}{l} x_{t+1} = 2y_t \\ y_{t+1} = \frac{1}{2}x_t \end{array} \quad \begin{array}{l} x_0 = y_0 = 1 \\ x_0 = y_0 = 0 \end{array} \\ \text{(b)} & \begin{array}{l} x_{t+1} = -y_t - z_t + 1 \\ y_{t+1} = -x_t - z_t + t \\ z_{t+1} = -x_t - y_t + 2t \end{array} \quad \begin{array}{l} x_0 = y_0 = 0 \\ z_0 = 1 \end{array} \end{array}$$

2. Find the general solutions of the systems when $a > 0$ and $b > 0$.

$$\begin{array}{ll} \text{(a)} & \begin{array}{l} x_{t+1} = ax_t \\ y_{t+1} = bx_t \end{array} \\ \text{(b)} & \begin{array}{l} x_{t+1} = ay_t + ck^t \\ y_{t+1} = bx_t + dk^t \end{array} \quad (k^2 \neq ab) \end{array}$$

3. A study of the US economy by R. J. Ball and E. Smolensky uses the system

$$y_t = 0.49y_{t-1} + 0.68i_{t-1}, \quad i_t = 0.032y_{t-1} + 0.43i_{t-1}$$

where y_t denotes production and i_t denotes investment at time t .

- (a) Derive a difference equation of order 2 for y_t , and find its characteristic equation.
(b) Find approximate solutions of the characteristic equation, and indicate the general solution of the system.

11.7 Stability of Nonlinear Difference Equations

Stability of an equilibrium state for a first-order linear difference equation with constant coefficients was considered in Section 11.1. In the present section we take a brief look at the nonlinear case, and also the possibility of cycles of order 2.

Consider an autonomous first-order difference equation in one variable

$$x_{t+1} = f(x_t) \quad (1)$$

where $f: I \rightarrow I$ is defined on an interval I in $\mathbb{R}$. An **equilibrium** or **stationary state** for (1) is a number x^* such that $x^* = f(x^*)$, i.e. the constant function $x_t = x^*$ is a solution of (1). In the language of Chapter 14, x^* is a fixed point of f . As in the case of differential equations, equilibrium states for (1) may be stable or unstable.

An equilibrium state x^* for (1) is called **locally asymptotically stable** if every solution that starts close enough to x^* converges to x^* , i.e. there exists an $\varepsilon > 0$ such that if $|x_0 - x^*| < \varepsilon$, then $\lim_{t \rightarrow \infty} x_t = x^*$. The equilibrium state x^* is **unstable** if a solution that starts close to x^* tends to move away from x^* , at least to begin with. More precisely, x^* is unstable if there exists an $\varepsilon > 0$ such that for every x with $0 < |x - x^*| < \varepsilon$ one has $|f(x) - x^*| > |x - x^*|$.

The following result is analogous to (5.7.2):

THEOREM 11.7.1

Let x^* be an equilibrium state for the difference equation (1), and suppose that f is C^1 in an open interval around x^* .

- (a) If $|f'(x^*)| < 1$, then x^* is locally asymptotically stable.
(b) If $|f'(x^*)| > 1$, then x^* is unstable.

Proof: The mean value theorem says that, for some c between x_t and x^* , one has

$$|x_{t+1} - x^*| = |f(x_t) - f(x^*)| = |f'(c)(x_t - x^*)| \quad (*)$$

(a) Since f' is continuous and $|f'(x^*)| < 1$, there exist an $\varepsilon > 0$ and a positive number $k < 1$ such that $|f'(x)| \leq k$ for all x in $(x^* - \varepsilon, x^* + \varepsilon)$. Provided that $|x_t - x^*| < \varepsilon$, we infer from (*) that $|x_{t+1} - x^*| \leq k|x_t - x^*|$. By induction on t , it follows that $|x_t - x^*| \leq k^t|x_0 - x^*|$ for all $t \geq 0$, and so $x_t \rightarrow x^*$ as $t \rightarrow \infty$.

(b) Now suppose that $|f'(x^*)| > 1$. By continuity there exist an $\varepsilon > 0$ and a $K > 1$ such that $|f'(x)| \geq K$ for all x in $(x^* - \varepsilon, x^* + \varepsilon)$. By (*), if $x_t \in (x^* - \varepsilon, x^* + \varepsilon)$, then

$$|x_{t+1} - x^*| = |f(x_t) - f(x^*)| \geq K|x_t - x^*|$$

Thus if x_t is close to but not equal to x^* , the distance between the solution x and the equilibrium x^* is magnified by a factor K or more at each step as long as x_t remains in $(x^* - \varepsilon, x^* + \varepsilon)$. ■

NOTE 1 If $|f'(x^*)| < 1$ for all x in I , then x^* is actually globally asymptotically stable in the obvious sense.

An equilibrium state x^* of equation (1) corresponds to a point (x^*, x^*) where the graph $y = f(x)$ of f intersects the straight line $y = x$. Figures 1 and 2 show two possible configurations around a stable equilibrium. In Fig. 1, $f'(x^*)$ is positive and the sequence $x_0, x_1, \dots$ converges monotonically to x^* , whereas in Fig. 2, $f'(x^*)$ is negative and we get a cobweb-like behaviour with x_i alternating between values above and below the equilibrium state $x^* = \lim_{i \rightarrow \infty} x_i$. In both cases the sequence of points $P_i = (x_i, x_{i+1}) = (x_i, f(x_i))$, $i = 0, 1, 2, \dots$, on the graph of f converges towards the point (x^*, x^*) .

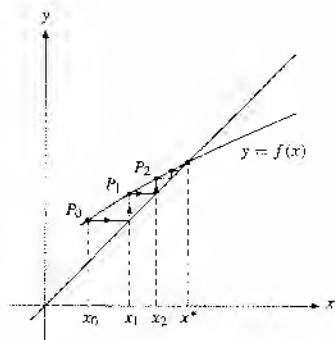


Figure 1 x^* stable, $f'(x^*) \in (0, 1)$.

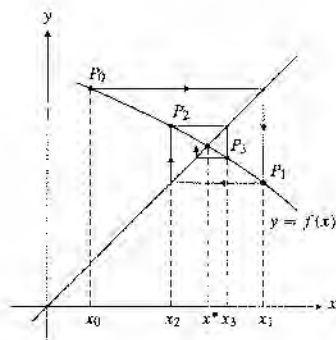


Figure 2 x^* stable, $f'(x^*) \in (-1, 0)$.

In Fig. 3, the graph of f near the equilibrium is too steep for convergence. Figure 4 shows that an equation of the form (1) may have solutions that exhibit cyclic behaviour, in this case a cycle of period 2. This is the topic of the next subsection.

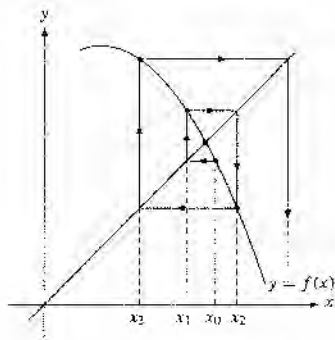


Figure 3 x^* unstable, $|f'(x^*)| > 1$.

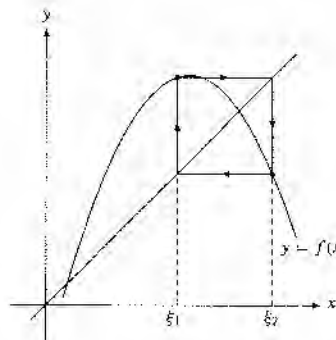


Figure 4 A cycle of period 2.

Cycles of Period 2

A **cycle** or **periodic solution** of (1) with period 2 is a solution x_i for which $x_{i+2} = x_i$ for all i , but $x_{i+1} \neq x_i$. In other words, $x_1 \neq x_0$, but $x_0 = x_2 = x_4 = \dots$ and $x_1 = x_3 = x_5 = \dots$.

Thus equation (1) admits a cycle of period 2 if and only if there exist distinct cycle points ξ_1 and ξ_2 such that $f(\xi_1) = \xi_2$ and $f(\xi_2) = \xi_1$. If we let $F = f \circ f$, it is clear that ξ_1 and ξ_2 must be fixed points of F , i.e. they are equilibria of the difference equation

$$y_{i+1} = F(y_i) = f(f(y_i)) \quad (2)$$

Such a cycle is said to be **locally asymptotically stable** if every solution of (1) that comes close to ξ_1 or ξ_2 converges to the cycle. Thus the cycle is locally asymptotically stable if and only if ξ_1 is a locally asymptotically stable equilibrium of equation (2), or equivalently, if and only if ξ_2 is such an equilibrium. The cycle is **unstable** if ξ_1 and ξ_2 are unstable equilibria of $f \circ f$. By the chain rule, $F'(x) = f'(f(x))f'(x)$, and so $F'(\xi_1) = f'(\xi_2)f'(\xi_1) = F'(\xi_2)$. Theorem 11.7.1 therefore implies the following:

If equation (1) admits a cycle of period 2, alternating between the values ξ_1 and ξ_2 , then:

- If $|f'(\xi_1)f'(\xi_2)| < 1$, the cycle is locally asymptotically stable.
- If $|f'(\xi_1)f'(\xi_2)| > 1$, the cycle is unstable.

(3)

The Quadratic Case

A linear difference equation $x_{i+1} = ax_i + b$ with constant coefficients has no interesting cycles. The simplest nonlinear case is the case of a quadratic polynomial. So let $f(x) = ax^2 + bx + c$ (with $a \neq 0$) and consider the difference equation

$$x_{i+1} = f(x_i) = ax_i^2 + bx_i + c \quad (4)$$

The equilibrium states of (4), if any, are the solutions

$$x_1 = \frac{1-b+\sqrt{(b-1)^2-4ac}}{2a}, \quad x_2 = \frac{1-b-\sqrt{(b-1)^2-4ac}}{2a}$$

of the quadratic equation $x = f(x)$, i.e. $ax^2 + (b-1)x + c = 0$. These solutions exist if and only if $(b-1)^2 \geq 4ac$, and they are distinct if and only if $(b-1)^2 > 4ac$. The values of f' at these points are $f'(x_{1,2}) = 2ax_{1,2} + b = 1 \pm \sqrt{(b-1)^2 - 4ac}$. It follows that if the equilibrium points exist and are distinct, then x_1 is always unstable, while x_2 is locally asymptotically stable if $(b-1)^2 - 4ac < 4$, and unstable if $(b-1)^2 - 4ac > 4$. (If $(b-1)^2 - 4ac = 4$, then x_2 is "locally asymptotically stable on one side" and unstable on the other side.)

Equation (4) admits a cycle of period 2 if there exist distinct numbers ξ_1 and ξ_2 such that $f(\xi_1) = \xi_2$ and $f(\xi_2) = \xi_1$. These numbers must be solutions of the equation $x = f(f(x))$. Since $f(f(x))$ is a polynomial of degree 4, it seems at first sight that we have to solve a rather difficult equation in order to find ξ_1 and ξ_2 . Fortunately the equation simplifies because any solution of $x = f(x)$ is also a solution of $x = f(f(x))$, so $x - f(x)$ is a factor of the polynomial $x - f(f(x))$. A simple but tedious computation shows that $x - f(f(x)) = (x - f(x))g(x)$, where

$$g(x) = a^2x^2 + a(b+1)x + ac + b + 1$$

The cycle points are the roots of the equation $g(x) = 0$, which are

$$\xi_{1,2} = \frac{-(b+1) \pm \sqrt{(b+1)^2 - 4ac - 4}}{2a}$$

These roots exist and are distinct if and only if $(b+1)^2 > 4ac + 4$. Hence, if there is a cycle of period 2, the equilibrium points x_1 and x_2 also exist, and are both unstable. (See also Problem 1.)

Because $f'(\xi) = 2a\xi + b$, while $\xi_1 + \xi_2 = -(b+1)/a$ and $\xi_1\xi_2 = (ac + b + 1)/a^2$, a simple calculation shows that $f'(\xi_1)f'(\xi_2) = 4ac - (b+1)^2 + 5$. Then

$$|f'(\xi_1)f'(\xi_2)| < 1 \iff 4 < (b+1)^2 - 4ac < 6 \quad (5)$$

It follows that if both inequalities on the right are satisfied, then equation (4) admits a stable cycle of period 2. (The first inequality on the right is precisely the necessary and sufficient condition for a period 2 cycle to exist.)

PROBLEMS FOR SECTION 11.7

1. Show that if $f: I \rightarrow I$ is continuous and the difference equation $x_{t+1} = f(x_t)$ admits a cycle ξ_1, ξ_2 of period 2, it also has at least one equilibrium solution between ξ_1 and ξ_2 . (Hint: Consider the function $f(x) - x$ over the interval with endpoints ξ_1 and ξ_2 .)
2. A solution x^* of the equation $x = f(x)$ can be viewed as an equilibrium solution of the difference equation

$$x_{t+1} = f(x_t) \quad (*)$$
 If this equilibrium is stable and x_0 is a sufficiently good approximation to x^* , then the solution $x_0, x_1, x_2, \dots$ of (*) starting from x_0 will converge to x^* .
 - (a) Use this technique to determine the negative solution of $x = e^x - 3$ to at least three decimal places.
 - (b) The equation $x = e^x - 3$ also has a positive solution, but this is an unstable equilibrium of $x_{t+1} = e^{x_t} - 3$. Explain how nevertheless we can find the positive solution by rewriting the equation and using the same technique as above.
3. The function f in Fig. 4 is given by $f(x) = -x^2 + 4x + 4/5$. Find the values of the cycle points ξ_1 and ξ_2 , and use (5) to determine whether the cycle is stable. It is clear from the figure that the difference equation $x_{t+1} = f(x_t)$ has two equilibrium states. Find these equilibria, show that they are both unstable, and verify the result in Problem 1.

12

DISCRETE TIME OPTIMIZATION

In science, what is susceptible to proof must not be believed without proof.¹
—R. Dedekind (1887)

This chapter gives a brief introduction to *discrete time dynamic optimization problems*. The term *dynamic* is used because the problems involve systems evolving over time. Time is here measured by the number of whole periods (say weeks, quarters, or years) that have passed since time 0. So we speak of *discrete time*. In this case it is natural to study dynamic systems whose development is governed by difference equations.

If the horizon is finite, such dynamic problems can be solved, in principle, using classical calculus methods. There are, however, solution techniques that take advantage of the special structure of discrete dynamic optimization problems. Section 12.1 on dynamic programming studies a standard problem with one state and one control variable.

In the economics literature a dynamic programming version of the Euler equation in continuous time control theory is much used. Section 12.2 gives a brief description.

When discussing optimization problems in discrete time, economists often prefer models with an infinite time horizon, just as they do in continuous time. Section 12.3 treats such models. The fundamental result is the Bellman equation.

When a discrete time optimization problem has restrictions on the terminal values of the state variable, there is a discrete time version of the maximum principle which may work better than dynamic programming. Sections 12.4 and 12.5 set out the relevant discrete time maximum principle, first for a single state variable, then for many. In contrast to the continuous time maximum principle, the Hamiltonian is not necessarily maximized at the optimal control. Section 12.5 also presents a very brief discussion of infinite horizon problems in this setting.

Section 12.6 offers an introduction to stochastic dynamic programming, including the stochastic Euler equation that plays such a prominent role in current macroeconomic theory. The concluding Section 12.7 is devoted to the important case of stationary problems with an infinite horizon. (Sections 12.6 and 12.7 are the only parts of the book that rely on some knowledge of probability theory, though only at a basic level.)

¹ There is no ideal English translation of the German original: "Was beweisbar ist, soll in der Wissenschaft nicht ohne Beweis geglaubt werden."

12.1 Dynamic Programming

Consider a system that changes at discrete times $t = 0, 1, \dots, T$. Suppose the state of the system at time t is characterized by a real number x_t . For example, x_t might be the quantity of grain that is stockpiled at time t . Assume that the initial state x_0 is historically given, and that from then on the system evolves through time under the influence of a sequence of controls u_t , which can be chosen freely from a given set U , called the **control region**. For example, u_t might be the quantity of grain removed from the stock x_t at time t . The controls influence the evolution of the system through a difference equation

$$x_{t+1} = g(t, x_t, u_t), \quad x_0 \text{ given}, \quad u_t \in U \quad (1)$$

where g is a given function. Thus, we assume that the state of the system at time $t + 1$ depends explicitly on the time t , on the state x_t in the preceding period t , and on u_t , the value chosen for the control at time t .

Suppose that we choose values for $u_0, u_1, \dots, u_{T-1}$. Then (1) gives $x_1 = g(0, x_0, u_0)$. Since x_1 is now determined, so too is $x_2 = g(1, x_1, u_1)$, then next $x_3 = g(2, x_2, u_2)$, etc. In this way, (1) can be used to compute recursively the successive states $x_1, x_2, \dots, x_T$ in terms of the initial state, x_0 , and the controls, $u_0, \dots, u_{T-1}$. Each choice of $(u_0, u_1, \dots, u_{T-1})$ gives rise to a sequence $(x_1, x_2, \dots, x_T)$, for instance path 1 in Fig. 1. A different choice of $(u_0, u_1, \dots, u_{T-1})$ would give another path, such as path 2 in the figure. Such controls u_t that depend only on time are often called **open-loop controls**.

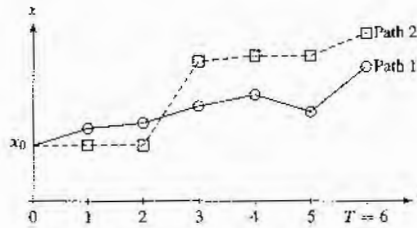


Figure 1 Different evolutions of system (1)

Different paths will usually have different utility or value. Assume that there is a function $f(t, x, u)$ of three variables such that the utility associated with a given path is represented by the sum

$$\sum_{t=0}^{T-1} f(t, x_t, u_t) \quad (*)$$

The sum is called the **objective function**, and it represents the sum of utilities (values) obtained at each point of time.

NOTE 1 The objective function is sometimes specified as $\sum_{t=0}^{T-1} f(t, x_t, u_t) + S(x_T)$, where S measures the net value associated with the terminal period. This is a special case of $(*)$ in which $f(T, x_T, u_T) = S(x_T)$. (S is often called a **scrap value function**.)

Suppose that we choose values for $u_0, u_1, \dots, u_{T-1}, u_T$, all from the set U , as specified in (1). The initial state x_0 is given, and as explained above, (1) gives us $x_1, \dots, x_T$. Let us denote corresponding pairs $(x_0, \dots, x_T), (u_0, \dots, u_T)$ by $(\{x_t\}, \{u_t\})$, and call them **admissible sequence pairs**. For each admissible sequence pair the objective function has a definite value. We shall study the following problem:

Among all admissible sequence pairs $(\{x_t\}, \{u_t\})$ find one, $(\{x_t^\}, \{u_t^*\})$, that makes the value of the objective function as large as possible.*

Such an admissible sequence pair is called an **optimal pair**, and the corresponding control sequence $\{u_t^*\}_{t=0}^{T-1}$ is called an **optimal control**. The discrete time optimization problem can be briefly formulated as

$$\max \sum_{t=0}^{T-1} f(t, x_t, u_t) \quad \text{subject to} \quad x_{t+1} = g(t, x_t, u_t), \quad x_0 \text{ given}, \quad u_t \in U \quad (2)$$

EXAMPLE 1

Let x_t be an individual's wealth at time t . At each point of time t , the individual has to decide the proportion u_t of x_t to consume, leaving the remaining proportion $1 - u_t$ for savings. Assume that wealth earns interest at the rate $\rho - 1 > 0$. After $u_t x_t$ has been withdrawn for consumption, the remaining stock of wealth is $(1 - u_t)x_t$. Because of interest, this grows to the amount $x_{t+1} = \rho(1 - u_t)x_t$ at the beginning of period $t + 1$. This equation holds for $t = 0, \dots, T-1$, with x_0 a positive constant. Suppose that the utility of consuming $c_t = u_t x_t$ is $U(t, c_t)$. Then the total utility over periods $t = 0, \dots, T$ is $\sum_{t=0}^{T-1} U(t, u_t x_t)$. The problem facing the individual is therefore the following:

$$\max \sum_{t=0}^{T-1} U(t, u_t x_t) \quad \text{subject to} \quad x_{t+1} = \rho(1 - u_t)x_t, \quad t = 0, \dots, T-1 \quad (*)$$

with x_0 given and with u_t in $[0, 1]$ for $t = 0, \dots, T$. This is a standard dynamic optimization problem of the type described above. (See Problems 2, 3, and 8.)

The Value Function and its Properties

Returning to the general problem described by (2), suppose that at time $t = s$ the state of the system is x (any given real number). The best we can do in the remaining periods is to choose $u_s, u_{s+1}, \dots, u_T$ (and thereby also $x_{s+1}, \dots, x_T$) to maximize $\sum_{t=s}^{T-1} f(t, x_t, u_t)$ with $x_s = x$. We define the (optimal) **value function** for the problem at time s by²

$$J_s(x) = \max_{u_s, \dots, u_T \in U} \sum_{t=s}^{T-1} f(t, x_t, u_t) \quad (3)$$

where

$$x_t = x \quad \text{and} \quad x_{t+1} = g(t, x_t, u_t) \quad \text{for } t > s, \quad u_t \in U \quad (4)$$

At the terminal time $t = T$, definition (3) implies that $J_T(x) = \max_{u_T \in U} f(T, x, u_T)$.

² We assume that the maximum in (3) is attained. This is true if, for example, the functions f and g are continuous and U is compact.

We now prove an important property of the value function. Suppose that at time $t = s$ ($s < T$) we are in state $x_s = x$. What is the optimal choice for u_s ? If we choose $u_s = u$, then at time $t = s$ we obtain the immediate reward $f(s, x, u)$, and, according to (4), the state at time $s + 1$ will be $x_{s+1} = g(s, x, u)$. Using definition (3) again, the highest obtainable value of the total reward $\sum_{t=s+1}^T f(t, x_t, u_t)$ from time $s + 1$ to time T , starting from the state x_{s+1} , is $J_{s+1}(x_{s+1}) = J_{s+1}(g(s, x, u))$. Hence, the best choice of $u = u_s$ at time s must be a value of u that maximizes the sum

$$f(s, x, u) + J_{s+1}(g(s, x, u))$$

This leads to the following general result:

THEOREM 12.1.1 (FUNDAMENTAL EQUATIONS OF DYNAMIC PROGRAMMING)

For each $s = 0, 1, \dots, T-1, T$, let $J_s(x)$ be the value function (3) for the problem

$$\max \sum_{t=0}^T f(t, x_t, u_t) \quad \text{subject to} \quad x_{t+1} = g(t, x_t, u_t), \quad u_t \in U \quad (5)$$

with x_0 given. Then the sequence of value functions satisfies the equations

$$J_s(x) = \max_{u \in U} [f(s, x, u) + J_{s+1}(g(s, x, u))], \quad s = 0, 1, \dots, T-1 \quad (6)$$

$$J_T(x) = \max_{u \in U} f(T, x, u) \quad (7)$$

NOTE 2 If we minimize rather than maximize the sum in (5), then Theorem 12.1.1 holds with “max” replaced by “min” in (3), (6) and (7). This is because minimizing f is equivalent to maximizing $-f$.

NOTE 3 Let $\mathcal{X}_t(x_0)$ denote the range of all possible values of the state x_t that can be generated by the difference equation (1) if we start in state x_0 and then go through all possible values of $u_0, \dots, u_{t-1}$. Of course only the values of $J_t(x)$ for $x \in \mathcal{X}_t(x_0)$ are relevant.

Theorem 12.1.1 is the basic tool for solving dynamic optimization problems. It is used as follows: First find the function $J_T(x)$ by using (7). The maximizing value of u depends (usually) on x , and is denoted by $u_T^*(x)$. The next step is to use (6) to determine $J_{T-1}(x)$ and the corresponding maximizing control $u_{T-1}^*(x)$. Then work backwards in this fashion to determine recursively all the value functions $J_T(x), \dots, J_0(x)$ and the maximizers $u_T^*(x), \dots, u_0^*(x)$. This allows us to construct the solution of the original optimization problem: Since the state at $t = 0$ is x_0 , the best choice of u_0 is $u_0^*(x_0)$. After $u_0^*(x_0)$ is found, the difference equation in (1) determines the state at time 1 as $x_1^* = g(0, x_0, u_0^*(x_0))$. Then $u_1^*(x_1^*)$ is the best choice of u_1 , and this choice determines x_2^* by (1). Then again, $u_2^*(x_2^*)$ is the best choice of u_2 , and so on.

EXAMPLE 2 Use Theorem 12.1.1 to solve the problem

$$\max \sum_{t=0}^3 (1 + x_t - u_t^2), \quad x_{t+1} = x_t + u_t, \quad t = 0, 1, 2, \quad x_0 = 0, \quad u_t \in \mathbb{R}$$

Solution: Here $T = 3$, $f(t, x, u) = 1 + x - u^2$, and $g(t, x, u) = x + u$. Consider first (7) and note that $J_3(x)$ is the maximum value of $1 + x - u^2$ for $u \in (-\infty, \infty)$. This maximum value is obviously attained for $u = 0$. Hence, in the notation introduced above,

$$J_3(x) = 1 + x, \quad \text{with } u_3^*(x) = 0 \quad (*)$$

For $s = 2$, the function to be maximized in (6) is $h_2(u) = 1 + x - u^2 + J_3(x + u)$, where $(*)$ implies that $J_3(x + u) = 1 + (x + u)$. Thus, $h_2(u) = 1 + x - u^2 + 1 + (x + u) = 2 + 2x + u - u^2$. The function h_2 is concave in u , and $h_2'(u) = 1 - 2u = 0$ for $u = 1/2$, so this is the optimal choice of u . The maximum value of $h_2(u)$ is $h_2(1/2) = 2 + 2x + 1/2 - 1/4 = 9/4 + 2x$. Hence,

$$J_2(x) = \frac{9}{4} + 2x, \quad \text{with } u_2^*(x) = \frac{1}{2}$$

For $s = 1$, the function to be maximized in (6) is given by $h_1(u) = 1 + x - u^2 + J_2(x + u) = 1 + x - u^2 + 9/4 + 2(x + u) = 13/4 + 3x + 2u - u^2$. Because h_1 is concave and $h_1'(u) = 2 - 2u = 0$ for $u = 1$, the maximum value of $h_1(u)$ is $13/4 + 3x + 2 - 1 = 17/4 + 3x$, so

$$J_1(x) = \frac{17}{4} + 3x, \quad \text{with } u_1^*(x) = 1$$

Finally, for $s = 0$, the function to be maximized is $h_0(u) = 1 + x - u^2 + J_1(x + u) = 1 + x - u^2 + 17/4 + 3(x + u) = 21/4 + 4x + 3u - u^2$. The function h_0 is concave and $h_0'(u) = 3 - 2u = 0$ for $u = 3/2$, so the maximum value of $h_0(u)$ is $h_0(3/2) = 21/4 + 4x + 9/2 - 9/4 = 15/2 + 4x$. Thus,

$$J_0(x) = \frac{15}{2} + 4x, \quad \text{with } u_0^*(x) = \frac{3}{2}$$

In this particular case the optimal choices of the controls are constants, independent of the states. The corresponding optimal values of the state variables are $x_1 = x_0 + u_0 = 3/2$, $x_2 = x_1 + u_1 = 3/2 + 1 = 5/2$, $x_3 = x_2 + u_2 = 5/2 + 1/2 = 3$. The maximum value of the objective function is $15/2$.

Alternative solution: In simple cases like this, a dynamic optimization problem can be solved quite easily by ordinary calculus methods. By letting $t = 0, 1$, and 2 in the difference equation $x_{t+1} = x_t + u_t$, we get $x_1 = x_0 + u_0 = u_0$, $x_2 = x_1 + u_1 = u_0 + u_1$, and $x_3 = x_2 + u_2 = u_0 + u_1 + u_2$. Using these results, the objective function becomes the following function of u_0, u_1, u_2 , and u_3 :

$$\begin{aligned} I &= (1 - u_0^2) + (1 + u_0 - u_1^2) + (1 + u_0 + u_1 - u_2^2) + (1 + u_0 + u_1 + u_2 - u_3^2) \\ &= 4 + 3u_0 - u_0^2 + 2u_1 - u_1^2 + u_2 - u_2^2 - u_3^2 \end{aligned}$$

The problem has been reduced to that of maximizing I with respect to the control variables u_0, u_1, u_2 , and u_3 . We see that I is a sum of concave functions and so is concave. Hence, a stationary point will maximize I . The first-order derivatives of I are

$$\frac{\partial I}{\partial u_0} = 3 - 2u_0, \quad \frac{\partial I}{\partial u_1} = 2 - 2u_1, \quad \frac{\partial I}{\partial u_2} = 1 - 2u_2, \quad \frac{\partial I}{\partial u_3} = -2u_3$$

Equating these partial derivatives to zero yields the unique stationary point $(u_0, u_1, u_2, u_3) = (\frac{3}{2}, 1, \frac{1}{2}, 0)$. This gives the same solution as the one obtained by using Theorem 12.1.1. ■

In principle, all deterministic finite horizon dynamic problems can be solved in this alternative way using ordinary calculus. But the method becomes very unwieldy if the horizon T is large, or if there is a stochastic optimization problem of the kind considered in Sections 12.6–12.7 below.

In the next example the terminal time is an arbitrarily given natural number and the optimal control turns out to depend on the state of the system.

EXAMPLE 3 Solve the following problem:

$$\max \left(\sum_{t=0}^{T-1} -\frac{2}{3}u_t x_t + \ln x_T \right), \quad x_{t+1} = x_t(1 + u_t x_t), \quad x_0 \text{ positive constant}, \quad u_t \geq 0 \quad (*)$$

Solution: Because $x_0 > 0$ and $u_t \geq 0$, we have $x_t > 0$ for all t . Now $f(T, x, u) = \ln x$ is independent of u , so $J_T(x) = \ln x$, and any u_T is optimal.

Next, putting $s = T - 1$ in (6) yields

$$J_{T-1}(x) = \max_{u \geq 0} \left[-\frac{2}{3}ux + J_T(x(1 + ux)) \right] = \max_{u \geq 0} \left[-\frac{2}{3}ux + \ln x + \ln(1 + ux) \right]$$

The maximum of the concave function $h(u) = -\frac{2}{3}ux + \ln x + \ln(1 + ux)$ is at the point where its derivative is 0. This gives $h'(u) = -\frac{2}{3}x + x/(1 + ux) = 0$, or (since $x > 0$), $u = 1/2x$. Then $h(1/2x) = \ln x - 1/3 + \ln(3/2)$. Hence,

$$J_{T-1}(x) = h(1/2x) = \ln x + C, \quad \text{with } C = -1/3 + \ln(3/2), \quad \text{and } u_{T-1}^*(x) = 1/2x$$

The next step is to use (6) for $s = T - 2$:

$$J_{T-2}(x) = \max_{u \geq 0} \left[-\frac{2}{3}ux + J_{T-1}(x(1 + ux)) \right] = \max_{u \geq 0} \left[-\frac{2}{3}ux + \ln x + \ln(1 + ux) + C \right]$$

Again $u = u_{T-2}^*(x) = 1/2x$ gives the maximum because the first-order condition is the same, and we get

$$J_{T-2}(x) = \ln x + 2C, \quad \text{with } C = -1/3 + \ln(3/2), \quad \text{and } u_{T-2}^*(x) = 1/2x$$

This pattern continues and so, for $k = 0, 1, \dots, T$ we get

$$J_{T-k}(x) = \ln x + kC, \quad \text{with } C = -1/3 + \ln(3/2), \quad \text{and } u_{T-k}^*(x) = 1/2x$$

So far we have been working backwards from time T to time 0. Putting $t = T - k$ for each k , we find that $J_t(x) = \ln x + (T - t)C$ and $u_t^* = 1/2x$ for $t = 0, 1, \dots, T$.

Finally, inserting $u_t^* = 1/2x_t^*$ in the difference equation gives $x_{t+1}^* = (\frac{3}{2})x_t^*$. So $x_t^* = (\frac{3}{2})^t x_0$, with $u_t^* = (\frac{3}{2})^t / 2x_0$ as optimal control values. ■

NOTE 4 Theorem 12.1.1 also holds if the control region is not a fixed set U , but instead a set $U(t, x)$ that depends on (t, x) . Then the maximization in (2), (3), and (5) is carried out for u_t in $U(t, x_t)$. In (6) and (7), the maximization is carried out for $u \in U(s, x)$ and $u \in U(T, x)$, respectively. Frequently, the set $U(t, x)$ is determined by one or more inequalities of the form $h(t, x, u) \leq 0$, for some function h that is continuous in (x, u) . If $U(t, x)$ is empty, then by convention, the maximum over $U(t, x)$ is set equal to $-\infty$.

NOTE 5 In the above formulation, the state x and the control u may well be vectors, in say $\mathbb{R}^n$ and $\mathbb{R}^r$, respectively. Then g must be a vector function as well, and the difference equation is a system of difference equations, one for each component of x . No changes are then needed in Theorem 12.1.1 (except that we would use boldface letters for x , u , and g).

EXAMPLE 4 Let x_t denote the value of an investor's assets at time t , and let u_t be consumption. Suppose that assets at time $t + 1$ are proportional to savings $x_t - u_t$ at t , with a factor of proportionality depending on t , i.e.

$$x_{t+1} = a_t(x_t - u_t), \quad a_t \text{ given positive numbers}$$

Assume that the initial assets, x_0 , are positive. The utility associated with a level of consumption u during one period is supposed to be $u^{1-\gamma}$, while the utility of the assets at time T is $Ax_T^{1-\gamma}$. Here A is a positive constant and $\gamma \in (0, 1)$. The investor wants to maximize the discounted value of the sum of utility from consumption and terminal assets. Define $\beta = 1/(1 + r)$, where r is the rate of discount. Assume that both savings and consumption must be positive each period, so $0 < u_t < x_t$. The investor's problem is thus:

$$\max \left[\sum_{t=0}^{T-1} \beta^t u_t^{1-\gamma} + \beta^T A x_T^{1-\gamma} \right], \quad x_{t+1} = a_t(x_t - u_t), \quad u_t \in (0, x_t) \quad (i)$$

Solution: We apply Theorem 12.1.1, as amended by Note 4, with the control region $U(t, x)$ given by the open interval $(0, x)$. So $f(t, x, u) = \beta^t u^{1-\gamma}$ for $t = 0, 1, \dots, T - 1$, whereas $f(T, x, u) = \beta^T A x^{1-\gamma}$. Since this function does not depend on u , (7) yields

$$J_T(x) = \max_{u \in (0, x)} \beta^T A x^{1-\gamma} = \beta^T A x^{1-\gamma} \quad (ii)$$

and any u_T in $(0, x)$ is optimal. Moreover, equation (6) yields

$$J_s(x) = \max_{u \in (0, x)} \left[\beta^s u^{1-\gamma} + J_{s+1}(a_s(x-u)) \right] \quad (\text{iii})$$

In particular, (ii) gives $J_T(a_{T-1}(x-u)) = \beta^T A a_{T-1}^{1-\gamma} (x-u)^{1-\gamma}$, so

$$J_{T-1}(x) = \beta^{T-1} \max_{u \in (0, x)} \left[u^{1-\gamma} + \beta A a_{T-1}^{1-\gamma} (x-u)^{1-\gamma} \right] \quad (\text{iv})$$

Let $h(u) = u^{1-\gamma} + c^\gamma (x-u)^{1-\gamma}$ denote the maximand, as a function of u in $(0, x)$, where $c^\gamma = \beta A a_{T-1}^{1-\gamma}$. Then $h'(u) = (1-\gamma)u^{-\gamma} - (1-\gamma)c^\gamma (x-u)^{-\gamma} = 0$ when $u^{-\gamma} = c^\gamma (x-u)^{-\gamma}$ and so $u = (x-u)/c$, implying that

$$u_{T-1} = u = x/w, \quad \text{where } w = 1 + c = 1 + (\beta A a_{T-1}^{1-\gamma})^{1/\gamma} = C_{T-1}^{1/\gamma} \quad (\text{v})$$

for a suitably defined constant C_{T-1} . Because $\gamma \in (0, 1)$ and $c^\gamma > 0$, the function h is easily seen to be concave over $(0, x)$. So the value of u given in (v) does maximize $h(u)$. Then, because $\beta A a_{T-1}^{1-\gamma} = c^\gamma = (w-1)^\gamma$, choosing the value x/w of u_{T-1} gives

$$\begin{aligned} h(x/w) &= x^{1-\gamma} w^{\gamma-1} + (w-1)^\gamma [x(1-w^{-1})]^{1-\gamma} \\ &= x^{1-\gamma} [w^{\gamma-1} + (w-1)^\gamma (w-1)^{1-\gamma} / w^{1-\gamma}] = x^{1-\gamma} w^\gamma = x^{1-\gamma} C_{T-1} \end{aligned}$$

Hence, by (iv),

$$J_{T-1}(x) = \beta^{T-1} C_{T-1} x^{1-\gamma} \quad (\text{vi})$$

Notice that $J_{T-1}(x)$ has the same form as $J_T(x)$. Next, substitute $s = T-2$ in (iii) to get

$$J_{T-2}(x) = \beta^{T-2} \max_{u \in (0, x)} \left[u^{1-\gamma} + \beta C_{T-1} a_{T-2}^{1-\gamma} (x-u)^{1-\gamma} \right]$$

Comparing this with (iv), from (v) we see that the maximum value is attained for

$$u_{T-2} = u = x/C_{T-2}^{1/\gamma}, \quad \text{where } C_{T-2}^{1/\gamma} = 1 + (\beta C_{T-1} a_{T-2}^{1-\gamma})^{1/\gamma}$$

and that $J_{T-2}(x) = \beta^{T-2} C_{T-2} x^{1-\gamma}$.

We can obviously go backwards repeatedly in this way and, for every t , obtain

$$J_t(x) = \beta^t C_t x^{1-\gamma} \quad (\text{vii})$$

From (ii), $C_T = A$, while C_t for $t < T$ is determined by backward recursion using the first-order difference equation

$$C_t^{1/\gamma} = 1 + (\beta C_{t+1} a_t^{1-\gamma})^{1/\gamma} = 1 + (\beta a_t^{1-\gamma})^{1/\gamma} C_{t+1}^{1/\gamma} \quad (\text{viii})$$

that is linear in $C_{t+1}^{1/\gamma}$. The optimal control is

$$u_t^*(x) = x/C_t^{1/\gamma}, \quad t < T \quad (\text{ix})$$

We find the optimal path by successively inserting $u_0^*, u_1^*, \dots$ into the difference equation (i) for x_t .

We can obtain an explicit solution in the special case when $a_t = a$ for all t . Then (viii) reduces to

$$C_{t+1}^{1/\gamma} - \frac{1}{\omega} C_t^{1/\gamma} = -\frac{1}{\omega}, \quad \text{where } \omega = (\beta a^{1-\gamma})^{1/\gamma} \quad (\text{x})$$

This is a first-order linear difference equation with constant coefficients. Using $C_T = A$, and solving the equation for $C_t^{1/\gamma}$, we obtain

$$C_t^{1/\gamma} = A^{1/\gamma} \omega^{T-t} + \frac{1 - \omega^{T-t}}{1 - \omega}, \quad t = T, T-1, \dots, 0$$

NOTE 6 Controls $u_t(x)$ that depend on the state x of the system are called **closed-loop controls**, whereas controls u_t that only depend on time are called **open-loop controls**.

Except in rare special cases, the controls $u_0^*, \dots, u_T^*$ that yield the maximum value $J_s(x)$ in (3) evidently do depend on x . In particular the first control u_0^* does so, i.e. $u_0^* = u_0^*(x)$. So, determining the functions $J_s(x)$ defined in (3) requires finding optimal closed-loop controls $u_t^*(x)$, $s = 0, 1, \dots, T$.

Given the initial state x_0 and any sequence of closed-loop controls $u_t^*(x)$, the evolution of the state x_t is uniquely determined by the difference equation

$$x_{t+1} = g(t, x_t, u_t(x_t)), \quad x_0 \text{ given} \quad (*)$$

Let us denote by $\bar{u}_t = u_t(x_t)$ the control values (numbers) generated by this particular sequence of states $\{x_t\}$. Next insert these numbers $\bar{u}_t$ into the difference equation:

$$x_{t+1} = g(t, x_t, \bar{u}_t), \quad x_0 \text{ given} \quad (**)$$

This obviously has exactly the same solution as equation (*).

Hence, we get the same result whether we insert the closed-loop controls $u_t^*(x)$ or the equivalent open-loop controls $\bar{u}_t$. In fact, once we have used the closed-loop controls to calculate the equivalent open-loop controls, it would seem that we can forget about the former. It may nevertheless be useful not to forget entirely the form of each closed-loop control. For suppose that at some time τ , there is an unexpected disturbance to the state x_τ^* obtained from the difference equation, which has the effect of changing the state to $\hat{x}_\tau$. Then $u_\tau^*(\hat{x}_\tau)$ still gives the optimal control to be used at that time, provided we assume that no further disturbances will occur.

PROBLEMS FOR SECTION 12.1

1. (a) Use Theorem 12.1.1 to solve the problem

$$\max_{u_t} \sum_{t=0}^2 [1 - (x_t^2 + 2u_t^2)], \quad x_{t+1} = x_t - u_t, \quad t = 0, 1 \quad (*)$$

where $x_0 = 5$ and $u_t \in \mathbb{R}$. (Compute $J_s(x)$ and $u_s^*(x)$ for $s = 2, 1, 0$.)

(b) Use the difference equation in (*) to compute x_1 and x_2 in terms of u_0 and u_1 (with $x_0 = 5$), and find the sum in (*) as a function S of u_0, u_1 , and u_2 . Next, maximize this function as in Example 2.

2. Consider the problem

$$\max_{u_t \in [0,1]} \sum_{t=0}^T \left(\frac{1}{1+r} \right)^t \sqrt{u_t x_t}, \quad x_{t+1} = \rho(1-u_t)x_t, \quad t=0, \dots, T-1, \quad x_0 > 0$$

where r is the rate of discount. Compute $J_t(x)$ and $u_t^*(x)$ for $s = T, T-1, T-2$.

3. (a) Replace the utility function in Problem 2 by $\sum_{t=0}^T (1+r)^{-t} u_t x_t$. Compute $J_T(x)$, $u_T^*(x)$, $J_{T-1}(x)$, and $u_{T-1}^*(x)$ for $x \geq 0$.
- (b) Prove that there exist constants P_s (depending on ρ and r) such that $J_s(x) = P_s x$ for $s = 0, 1, \dots, T$. Find $J_0(x)$ and optimal values of $u_0, u_1, \dots, u_T$.

4. Consider the problem

$$\max_{u_t \in [0,1]} \sum_{t=0}^T (3-u_t)x_t^2, \quad x_{t+1} = u_t x_t, \quad t=0, \dots, T-1, \quad x_0 \text{ is given}$$

- (a) Compute the value functions $J_T(x)$, $J_{T-1}(x)$, $J_{T-2}(x)$, and the corresponding control functions, $u_T^*(x)$, $u_{T-1}^*(x)$, and $u_{T-2}^*(x)$.
- (b) Find an expression for $J_{T-n}(x)$ for $n = 0, 1, \dots, T$, and the corresponding optimal controls.

5. Solve the problem

$$\max_{u_t \in [0,1]} \left[\sum_{t=0}^{T-1} \left(-\frac{2}{3} u_t \right) + \ln x_T \right], \quad x_{t+1} = x_t(1+u_t), \quad t=0, \dots, T-1, \quad x_0 > 0 \text{ given}$$

6. (a) Write down the fundamental equations for the problem

$$\max_{u_t \in \mathbb{R}} \sum_{t=0}^T (x_t - u_t^2), \quad x_{t+1} = 2(x_t + u_t), \quad t=0, 1, \dots, T-1, \quad x_0 = 0$$

- (b) Prove that the value function for the problem is given by

$$J_{T-n}(x) = (2^{n+1} - 1)x + \sum_{j=0}^n (2^j - 1)^2, \quad n=0, 1, \dots, T$$

Determine the optimal controls $u_t = u_t^*$ and the maximum value $V = J_0(0)$.

7. (a) Consider the problem

$$\max_{u_t \in \mathbb{R}} \left[\sum_{t=0}^{T-1} (-e^{-\gamma u_t}) - \alpha e^{-\gamma x_T} \right], \quad x_{t+1} = 2x_t - u_t, \quad t=0, 1, \dots, T-1, \quad x_0 \text{ given}$$

where α and γ are positive constants. Compute $J_T(x)$, $J_{T-1}(x)$, and $J_{T-2}(x)$.

- (b) Prove that $J_t(x)$ can be written in the form $J_t(x) = -\alpha_t e^{-\gamma x}$, and find a difference equation for α_t .

8. Consider the following special case of Problem 2, where
- $r = 0$
- :

$$\max_{u_t \in [0,1]} \sum_{t=0}^T \sqrt{u_t x_t}, \quad x_{t+1} = \rho(1-u_t)x_t, \quad t=0, \dots, T-1, \quad x_0 > 0$$

- (a) Compute $J_T(x)$, $J_{T-1}(x)$, $J_{T-2}(x)$. (Hint: Prove that $\max_{u \in [0,1]} [\sqrt{u} + A\sqrt{1-u}] = \sqrt{1+A^2}$ with $u = 1/(1+A^2)$.)
- (b) Show that the optimal control function is $u_s(x) = 1/(1 + \rho + \rho^2 + \dots + \rho^{T-s})$, and find the corresponding $J_s(x)$, $s = 1, 2, \dots, T$.

12.2 The Euler Equation

The economics literature sometimes considers the following formulation of the basic dynamic programming problem without an explicit control variable (e.g. Stokey et al. (1989))

$$\max \sum_{t=0}^T F(t, x_t, x_{t+1}), \quad x_0 \text{ given and } x_1, x_2, \dots, x_T, x_{T+1} \text{ vary freely in } \mathbb{R} \quad (1)$$

Here the instantaneous reward $F(t, x_t, x_{t+1})$ at time t depends on t and on the values of the state variable at adjacent times t and $t+1$.

If we define $u_t = x_{t+1}$, then (1) becomes a standard dynamic programming problem with $U = \mathbb{R}$. On the other hand, the dynamic optimization problem (12.1.2) can usually be formulated as a problem of the type (1). Suppose, in particular, that for every choice of x_t and x_{t+1} the equation $x_{t+1} = g(t, x_t, u_t)$ has a unique solution u_t in U , which we denote by $u_t = \varphi(t, x_t, x_{t+1})$. Now define the function F by $F(t, x_t, x_{t+1}) = f(t, x_t, \varphi(t, x_t, x_{t+1}))$ for $t < T$, and $F(T, x_T, x_{T+1}) = \max_{u \in U} f(T, x_T, u)$. Then problem (12.1.2) becomes precisely the same as problem (1).

If there is more than one value of u in U such that $g(t, x_t, u) = x_{t+1}$, let u_t be a value of u that maximizes $f(t, x_t, u)$, i.e. choose the best u that leads from x_t to x_{t+1} . Then, for each $t = 0, 1, \dots, T$, we have

$$F(t, x_t, x_{t+1}) = \max_u \{ f(t, x_t, u) : x_{t+1} = g(t, x_t, u), u \in U \} \quad (2)$$

Let $(x_0^*, \dots, x_{T+1}^*)$ be an optimal solution of problem (1). Then $(x_1^*, \dots, x_{T+1}^*)$ is a maximum point for the objective function $S(x_1, \dots, x_{T+1}) = \sum_{t=0}^T F(t, x_t, x_{t+1})$, and by the usual first-order condition we must have $S'_t(x_1^*, \dots, x_{T+1}^*) = 0$ for $t = 1, \dots, T+1$. (Remember that $x_0^* = x_0$ is given.) Hence, $(x_1^*, \dots, x_{T+1}^*)$ must satisfy the Euler equation

$$F'_2(t, x_t, x_{t+1}) + F'_3(t-1, x_{t-1}, x_t) = 0, \quad t=1, \dots, T \quad (3)$$

$$F'_3(T-1, x_{T-1}, x_T) = 0, \quad t=T+1$$

(If x_{T+1} does not appear explicitly in $F(T, x_T, x_{T+1})$, the last equation becomes trivial.)

This is a second-order difference equation analogous to the Euler equation in the classical calculus of variations. (See Section 8.2.) Note carefully that the partial derivatives in (3) are evaluated at different triples.

EXAMPLE 1 Consider the problem

maximize {sum_{t=0}^{T-1} ln c_t + ln x_T} subject to x_{t+1} = alpha(x_t - c_t) for t = 0, ..., T-1

Here x_t is wealth at time t, with x_0 fixed. An amount c_t is subtracted for consumption, and the remaining amount x_t - c_t is deposited in an account and increases to x_{t+1} = alpha(x_t - c_t) at time t + 1, where alpha > 1. Formulate the problem without explicit control variables, and use the Euler equation to solve it.

Solution: Define beta = 1/alpha. Because c_t = x_t - beta x_{t+1}, the formulation without control variables is

max {sum_{t=0}^{T-1} ln(x_t - beta x_{t+1}) + ln x_T} (*)

For t = T, the Euler equation is F'_2(T, x_T, x_{T+1}) + F'_3(T-1, x_{T-1}, x_T) = 0, with F(T, x_T, x_{T+1}) = ln x_T and F(T-1, x_{T-1}, x_T) = ln(x_{T-1} - beta x_T). Therefore, the Euler equation reduces to 1/x_T - beta/(x_{T-1} - beta x_T) = 0, so x_{T-1} = 2beta x_T.

For t = 1, 2, ..., T-1, the Euler equation gives

1/(x_t - beta x_{t+1}) - beta/(x_{t-1} - beta x_t) = 0

Solving this for x_{t-1} gives the (reverse) second-order difference equation x_{t-1} = 2beta x_t - beta^2 x_{t+1}. In particular, for t = T-2 this gives x_{T-2} = 2beta x_{T-1} - beta^2 x_T = 4beta^2 x_T - beta^2 x_T = 3beta^2 x_T. More generally, given x_T and x_{T-1} = 2beta x_T, we can show by backward induction that x_t = (T+1-t)beta^{T-t} x_T. This implies that x_0 = (T+1)beta^T x_T, so x_T = x_0 beta^{-T}/(T+1). We conclude that the optimal solution of the problem is

x_t^* = (T+1-t)/(T+1) beta^{-t} x_0, c_t^* = x_t^* - beta x_{t+1}^* = beta^{-t} x_0 / (T+1)

We see that optimal consumption is steadily decreasing as t increases.

NOTE 1 In Example 1 some might prefer to equate the partial derivatives of the maximand in (*) to 0 directly, rather than introducing the function F. In particular, equating the partial derivative w.r.t. x_T to 0 yields -beta/(x_{T-1} - beta x_T) + 1/x_T = 0 again; equating each partial derivative w.r.t. x_t to 0 yields -beta/(x_{t-1} - beta x_t) + 1/(x_t - beta x_{t+1}) = 0 for t = 1, 2, ..., T-1; etc.

NOTE 2 Here is a general solution procedure for problem (1), similar to that used in Section 12.1. First, for t = T and for each fixed x_T, find x_{T+1}^*(x_T) to maximize F(T, x_T, x_{T+1}); the associated first-order condition is F'_3(T, x_T, x_{T+1}) = 0, the appropriate version of (3). Next, x_{T+1}^*(x_T) is inserted into F(t, x_t, x_{t+1}) + F(t+1, x_{t+1}, x_{t+2}) for t = T-1, and this expression is maximized w.r.t. x_T, yielding x_T^*(x_{T-1}), using the first-order condition (3) for t = T-1. Then x_T = x_T^*(x_{T-1}) is inserted into the expression

F(t, x_t, x_{t+1}) + F(t+1, x_{t+1}, x_{t+2}) for t = T-2, and this expression is maximized w.r.t. x_{T-1}, yielding x_{T-1}^*(x_{T-2}), using (3) at time t = T-2. We continue to work backwards in this manner for t = T, T-1, ..., 2, 1, until at the last step we construct the function x_1^*(x_0). Since x_0 is given, we can work forward again to determine first x_1 = x_1^*(x_0), then x_2 = x_2^*(x_1), and so on. (In the example above we used another approach.)

PROBLEMS FOR SECTION 12.2

- 1. (a) Transform Problem 12.1.1 to the form (1). (b) Derive the corresponding Euler equation, and find its solution. Compare with the answer to Problem 12.1.1.
- 2. (a) Transform the problem in Example 12.1.3 to the form (1). (b) Derive the corresponding Euler equation, and find its solution. Compare with the answer in Example 12.1.3.

12.3 Infinite Horizon

Economists often study dynamic optimization problems over an infinite horizon. This avoids specifying what happens after a finite horizon is reached. It also avoids having the horizon as an extra exogenous variable that features in the solution. This section considers how dynamic programming methods can be used to study the following infinite horizon version of the problem set out in (12.1.2):

max sum_{t=0}^{inf} beta^t f(x_t, u_t), x_{t+1} = g(x_t, u_t), t = 0, 1, 2, ..., x_0 given, u_t in U subset of R (1)

Here f and g are given functions of two variables, there is a constant discount factor beta in (0, 1), and x_0 is a given number in R. Having beta in (0, 1) is essential for the subsequent analysis of the problem in this section. Note that, apart from replacing the horizon T by inf as the upper limit of the sum, the two functions f(t, x_t, u_t) and g(t, x_t, u_t) in (12.1.2) have been replaced by beta^t f(x_t, u_t) and g(x_t, u_t) respectively. Because neither the new function f nor g depends explicitly on t, problem (1) is called autonomous or stationary.

The sequence pair ({x_t}, {u_t}) is called admissible provided that each control satisfies u_t in U, the initial state x_0 has the given value, and the difference equation in (1) is satisfied for all t = 0, 1, 2, ...

For simplicity, we begin by assuming that f satisfies the boundedness condition

M_1 <= f(x, u) <= M_2 for all (x, u) with u in U (2)

where M_1 and M_2 are given numbers. Because 0 < beta < 1, the sum in (1) will then always converge.

For any given starting time s with $s = 0, 1, 2, \dots$ and any given state x at that time, take any control sequence $\pi_s = (u_s, u_{s+1}, \dots)$, where $u_t \in U$ for $t = s, s+1, \dots$. The successive states generated by this control sequence are found by solving $x_{t+1} = g(x_t, u_t)$, with $x_s = x$. With this notation, the discounted sum of the infinite utility (or benefit) sequence that is obtained from applying the control sequence π_s starting from state x at time s is

$$V_s(x, \pi_s) = \sum_{t=s}^{\infty} \beta^t f(x_t, u_t) = \beta^s V^s(x, \pi_s)$$

where

$$V^s(x, \pi_s) = \sum_{t=s}^{\infty} \beta^{t-s} f(x_t, u_t) \quad (3)$$

The difference between V_s and V^s is that in V_s all benefits from time s on are discounted back to the fixed initial time 0, whereas in V^s they are discounted back to the variable starting time s . Now let

$$J_s(x) = \max_{\pi_s} V_s(x, \pi_s) = \beta^s J^s(x), \quad \text{where} \quad J^s(x) = \max_{\pi_s} V^s(x, \pi_s) \quad (4)$$

and where the maxima are taken over all sequences $\pi_s = (u_s, u_{s+1}, \dots)$ with $u_{s+k} \in U$.³ Thus, $J_s(x)$ is the maximum total discounted utility (or benefit) that can be obtained over all the periods from $t = s$ to ∞ , given that the system starts in state x at time $t = s$. We call $J_s(x)$ the **(optimal) value function** for problem (1).

We next claim that the function $J^s(x)$ satisfies the important property

$$J^0(x) = J^s(x) \quad (5)$$

Intuitively, this equality is obvious. Because the problem is autonomous and we start in the same state x , the future looks exactly the same at either time 0 or time s . So, finding either $J^s(x) = \max_{\pi_s} V^s(x, \pi_s)$ or $J^0(x) = \max_{\pi_0} V^0(x, \pi_0)$ requires solving essentially the same optimization problem, which therefore gives the same maximum value in each case. A more precise argument for (5) is given in Note 3 below.

Equations (4) and (5) together imply that

$$J_s(x) = \beta^s J^0(x), \quad s = 0, 1, \dots \quad (6)$$

Define

$$J(x) = J_0(x) = J^0(x) \quad (7)$$

From (6) it follows that if we know $J_0(x) = J(x)$, then we know $J_s(x)$ for all s . The main result in this section is the following:

THEOREM 12.3.1 (FUNDAMENTAL EQUATION FOR INFINITE HORIZON)

The value function $J(x) = J_0(x)$ in (7) for problem (1) satisfies the equation

$$J(x) = \max_{u \in U} [f(x, u) + \beta J(g(x, u))] \quad (\text{the Bellman equation}) \quad (8)$$

³ The existence of this maximum is discussed later in Note 4.

A rough argument for (8) resembles the argument for Theorem 12.1.1: Suppose we start in state x at time $t = 0$. If we choose the control u , the immediate reward is $\beta^0 f(x, u) = f(x, u)$, and at time $t = 1$ we move to state $x_1 = g(x, u)$. Choosing an optimal control sequence from $t = 1$ on gives a total reward over all subsequent periods that equals $J_1(g(x, u)) = \beta J(g(x, u))$. Hence, the best choice of u at $t = 0$ is one that maximizes the sum $f(x, u) + \beta J(g(x, u))$. The maximum of this sum is therefore $J(x)$.

We call (8) a "functional equation" because the unknown is the function $J(x)$ that appears on both sides. Under the boundedness condition (2), together with the assumptions that the maximum in (8) is attained and that $0 < \beta < 1$, equation (8) always has one and only one bounded solution $J(x)$, which must therefore be the optimal value function for the problem. The value $u(x)$ of the control $u \in U$ that maximizes the right-hand side of (8) is the optimal control, which is therefore independent of t .

In general it is difficult to use equation (8) to find $J(x)$. The reason is that maximizing the right-hand side of (8) requires the function $J(x)$ to be known beforehand.

EXAMPLE 1 Consider the infinite horizon analogue of problem (i) in Example 12.1.4 in the case when $a_t = a$ for all t , independent of t . We also introduce a new control v defined by $u = vx$. Thus, v is the proportion of wealth x that is spent in the current period.

The former constraint $u \in (0, x)$ is then replaced by $v \in (0, 1)$. So the problem becomes

$$\max \sum_{t=0}^{\infty} \beta^t (v_t x_t)^{1-\gamma}, \quad x_{t+1} = a(1-v_t)x_t, \quad t = 0, 1, \dots, \quad v_t \in (0, 1) \quad (i)$$

where a and x_0 are positive constants, $\beta \in (0, 1)$, $\gamma \in (0, 1)$, and $\beta a^{1-\gamma} < 1$. Because the horizon is infinite, we may think of x_t as the assets of some institution like a university or a government that suffers from "immortality illusion" and so regards itself as timeless.

In the notation of problem (1), we have $f(x, v) = (vx)^{1-\gamma}$ and $g(x, v) = a(1-v)x$. Equation (8) therefore yields

$$J(x) = \max_{v \in (0, 1)} [(vx)^{1-\gamma} + \beta J(a(1-v)x)] \quad (ii)$$

In the closely related problem in Example 12.1.4, the value function was proportional to $x^{1-\gamma}$. A reasonable guess in the present case is that $J(x) = kx^{1-\gamma}$ for some positive constant k . We try this as a solution. Then, after cancelling the factor $x^{1-\gamma}$, (ii) reduces to

$$k = \max_{v \in (0, 1)} [v^{1-\gamma} + \beta k a^{1-\gamma} (1-v)^{1-\gamma}] \quad (iii)$$

Put $\varphi(v) = v^{1-\gamma} + \beta k a^{1-\gamma} (1-v)^{1-\gamma}$, defined on the interval $[0, 1]$. Note that $\varphi(v)$ is the sum of two functions that are concave in v . A helpful trick is to define the new constant $\rho > 0$ so that $\beta a^{1-\gamma} = \rho^\gamma$, and therefore $\varphi(v) = v^{1-\gamma} + k\rho^\gamma (1-v)^{1-\gamma}$. The first-order condition for maximizing φ is then

$$\varphi'(v) = (1-\gamma)v^{-\gamma} - (1-\gamma)k\rho^\gamma(1-v)^{-\gamma} = 0$$

implying that $v^{-\gamma} = k\rho^\gamma(1-v)^{-\gamma}$. Raising each side to the power $-1/\gamma$ and then solving for v , we see that the maximum of φ is attained at

$$v = \frac{1}{1 + \rho k^{1/\gamma}}, \quad \text{where } \rho = (\beta a^{1-\gamma})^{1/\gamma} \quad (\text{iv})$$

Then equation (iii) implies that k satisfies the equation

$$k = \frac{1}{(1 + \rho k^{1/\gamma})^{1-\gamma}} + k \rho^\gamma \frac{\rho^{1-\gamma} k^{(1-\gamma)/\gamma}}{(1 + \rho k^{1/\gamma})^{1-\gamma}} = (1 + \rho k^{1/\gamma})^\gamma$$

Raise each side to the power $1/\gamma$, and solve for $k^{1/\gamma}$ to obtain $k^{1/\gamma} = 1/(1 - \rho)$, or $k = (1 - \rho)^{-\gamma}$. Inserting this into (iv) gives $v = 1 - \rho$, so ρ is the constant fraction of current assets that are saved in each period. Because $J(x) = kx^{1-\gamma}$, we have

$$J(x) = (1 - \rho)^{-\gamma} x^{1-\gamma}, \quad \text{with } v = 1 - \rho, \quad \text{where } \rho = (\beta a^{1-\gamma})^{1/\gamma} \quad (\text{v})$$

Note that ρ increases with the discount factor β and with the return a to saving, as an economist would expect.

In this example the boundedness assumption (2) is not valid without a simple transformation. Note that $a^t x_0$ is the maximum wealth the consumer could have accumulated by time t by spending nothing, i.e. if $v_s = 0$ for $s = 0, 1, \dots, t-1$. Now define the modified state variable $y_t = x_t/(a_0 a^t)$, which is the proportion of this maximum wealth that remains. Obviously $y_0 = 1$, and y_t satisfies the difference equation $y_{t+1} = (1 - v_t)y_t$, so $1 \geq y_1 \geq y_2 \geq \dots \geq y_t \geq y_{t+1} \geq \dots \geq 0$. The new objective function is $\sum_{t=0}^{\infty} \beta^t (x_0 v_t y_t)^{1-\gamma}$ where $\tilde{\beta} = \beta a^{1-\gamma}$ and so $0 < \tilde{\beta} < 1$. The transformed Bellman equation is

$$\hat{J}(y) = \max_{v \in (0,1)} [(x_0 v y)^{1-\gamma} + \tilde{\beta} \hat{J}((1-v)y)]$$

This is easily seen to have $\hat{J}(y) = J(ax_0 y) = (1 - \rho)^{-\gamma} (ax_0 y)^{1-\gamma}$ as a solution, with the same optimal control $v = 1 - \rho$.

The transformed problem satisfies the restricted boundedness condition in Note 2 below because the modified state y_t remains within the interval $[0, 1]$ for all t , and so $0 \leq (x_0 y_t v_t)^{1-\gamma} \leq x_0^{1-\gamma}$ for all t and for all v_t in $[0, 1]$. Therefore the control v defined in (iv) really is optimal and the transformed problem is solved. So is the original problem, of course.

NOTE 1 As pointed out in Note 12.1.5, the same theory applies without change when x_t , u_t , and g are vector functions. Moreover, U may depend on the state, $U = U(x)$ (but not explicitly on time).

NOTE 2 It suffices to assume that condition (2) holds for all x in $\mathcal{X}(x_0) = \bigcup_{t=0}^{\infty} \mathcal{X}_t(x_0)$, where $\mathcal{X}_t(x_0)$ is defined in Note 12.1.3. The function $J(x)$ need only be defined on $\mathcal{X}(x_0)$.

NOTE 3 To show (5) more formally, let x be any fixed state.

First, consider any policy sequence $\pi_s = (u_s, u_{s+1}, \dots)$ that starts at time s . Now define the corresponding sequence $\pi_s^0 = (u_s^0, u_{s+1}^0, \dots)$ shifted earlier so that it starts at time 0 instead of at time s . Thus, $u_s^0 = u_s$, $u_{s+1}^0 = u_{s+1}$, and generally $u_t^0 = u_{t+s}$ for $t = 0, 1, \dots$

Then, given the same starting state x , if x_t and x_t^0 denote the states reached at time t by following π_s and π_s^0 starting at times s and 0 respectively, a moment's reflection leads to the conclusion that $x_t^0 = x_{s+t}$ and so $f(x_t^0, u_t^0) = f(x_{s+t}, u_{s+t})$ for $t = 0, 1, \dots$. It follows from (3) that $V^0(x, \pi_s^0) = V^s(x, \pi_s)$. But every shifted admissible policy π_s^0 is also admissible at time 0, so we can use (4) to obtain

$$J^0(x) = \max_{\pi_0} V^s(x, \pi_0) \geq \max_{\pi_s^0} V^0(x, \pi_s^0) = \max_{\pi_s} V^s(x, \pi_s) = J^s(x) \quad (9)$$

On the other hand, consider any policy sequence $\pi_0 = (u_0, u_1, \dots)$ that starts at time 0, and let $\pi_s^0 = (u_s^0, u_{s+1}^0, \dots)$ be the corresponding sequence shifted to a later time s so that $u_s^0 = u_0$, $u_{s+1}^0 = u_1$ and generally $u_t^0 = u_{t-s}$ for $t = s, s+1, \dots$. Again, given the same starting point x , the states x_t and x_t^0 reached at time t by following π_0 and π_s^0 starting at times 0 and s respectively will satisfy $x_t = x_{t-s}^0$ and so $f(x_t, u_t) = f(x_{t-s}^0, u_{t-s}^0)$ for $t = s, s+1, \dots$. Then (3) and (4) imply that $V^0(x, \pi_0) = V^s(x, \pi_s^0)$, so

$$J^0(x) = \max_{\pi_0} V^0(x, \pi_0) = \max_{\pi_0^s} V^s(x, \pi_0^s) \leq \max_{\pi_s} V^s(x, \pi_s) = J^s(x) \quad (10)$$

From (9) and (10) we conclude that $J^0(x) = J^s(x)$, which is (5).

NOTE 4 Whenever we wrote "max" above, it was implicitly assumed that the maximum exists. Of course, without further conditions on the system, this may not be true. Under the boundedness condition (2), the same assumptions as in the finite horizon case (f and g are continuous and U is compact) ensure that the maxima in (4) and (8) do exist.

Many economic applications, however, do not satisfy the boundedness condition (2). So let us investigate what happens when we replace max with sup in (4), as well as when the set $U(x)$ depends on x , as in Note 1. In fact, suppose the sum $\sum_{t=0}^{\infty} \beta^t f(x_t, u_t)$ always exists (possibly with an infinite value). Then $J_0(x_0) = \sup_{\pi_0} V_0(x_0, \pi_0)$ must exist. By the result (A.4.7) on iterated suprema, we have

$$\begin{aligned} J_0(x_0) &= \sup_{u_0, u_1, \dots} \sum_{t=0}^{\infty} \beta^t f(x_t, u_t) = \sup_{u_0 \in U(x)} [f(x_0, u_0) + \sup_{u_1, u_2, \dots} \sum_{t=1}^{\infty} \beta^t f(x_t, u_t)] \\ &= \sup_{u_0 \in U(x)} [f(x_0, u_0) + J_1(g(x_0, u_0))] = \sup_{u_0 \in U(x)} [f(x_0, u_0) + \beta J_0(g(x_0, u_0))] \end{aligned} \quad (*)$$

So the modification

$$J(x) = \sup_{u \in U(x)} [f(x, u) + \beta J(g(x, u))] \quad (11)$$

of the Bellman equation (8) still holds even if no maximum exists.

Next, let us use the contraction mapping theorem 14.3.1 to prove that version (11) of the Bellman equation has a unique solution.

Indeed, define the operator T on the domain $\mathcal{B}$ of all bounded functions $I(x)$ so that

$$T(I)(x) = \sup_{u \in U(x)} [f(x, u) + \beta I(g(x, u))] \quad (**)$$

for all t and all x . As in Section 14.3, the distance between any two bounded functions $\bar{J}$ and $\tilde{J}$ is defined as $d(\bar{J}, \tilde{J}) = \sup_z |\bar{J}(z) - \tilde{J}(z)|$. Then

$$\begin{aligned} T(\bar{J})(x) &= \sup_{u \in U(x)} [f(x, u) + \beta \bar{J}(g(x, u)) + \beta (\bar{J}(g(x, u)) - \tilde{J}(g(x, u)))] \\ &\leq \sup_{u \in U(x)} [f(x, u) + \beta \bar{J}(g(x, u)) + \beta d(\bar{J}, \tilde{J})] = T(\tilde{J})(x) + \beta d(\bar{J}, \tilde{J}) \end{aligned}$$

Symmetrically, $T(\tilde{J})(x) \leq T(\bar{J})(x) + \beta d(\tilde{J}, \bar{J})$. So $|T(\bar{J})(x) - T(\tilde{J})(x)| \leq \beta d(\bar{J}, \tilde{J})$, implying that

$$d(T(\bar{J}), T(\tilde{J})) = \sup_x |T(\bar{J})(x) - T(\tilde{J})(x)| \leq \beta d(\bar{J}, \tilde{J}) \quad (***)$$

Because $0 < \beta < 1$, this confirms that T is a contraction mapping, so the proof is complete.

Finally, we check that any control $u = \hat{u}$ that yields a maximum in the Bellman equation (11) is optimal. To see this, let $T^{\hat{u}}$ be the operator on $\mathcal{B}$ which is defined by (**) when $U(x)$ takes the form $\{\hat{u}(x)\}$ (leaving no choice except $u = \hat{u}(x)$!). By definition of $\hat{u}$, the unique solution J of the Bellman equation also satisfies $T^{\hat{u}}(J) = J$. Also, because $J^{\hat{u}}$ satisfies (*) for $U(x) = \{\hat{u}(x)\}$, we have $T^{\hat{u}}(J^{\hat{u}}) = J^{\hat{u}}$. But $T^{\hat{u}}$, like T itself, is a contraction mapping, so $T^{\hat{u}}(\tilde{J}) = \tilde{J}$ has a unique solution. It follows that $J = J^{\hat{u}}$, and in particular, the supremum $J(x)$ for any x is equal to $J^{\hat{u}}(x)$, the value attained by following the control policy $\hat{u}$.

PROBLEMS FOR SECTION 12.3

- SM 1. Consider the problem

$$\max_{u_t \in (-\infty, \alpha)} \sum_{t=0}^{\infty} \beta^t (-e^{-u_t} - \frac{1}{2} e^{-2u_t}), \quad x_{t+1} = 2x_t - u_t, \quad t = 0, 1, \dots, \quad x_0 \text{ given}$$

where $\beta \in (0, 1)$. Find a constant $\alpha > 0$ such that $J(x) = -\alpha e^{-x}$ solves the Bellman equation, and show that α is unique.

- SM 2. (a) Consider the following problem with $\beta \in (0, 1)$:

$$\max_{u_t \in (-\infty, \infty)} \sum_{t=0}^{\infty} \beta^t (-\frac{2}{3} x_t^2 - u_t^2), \quad x_{t+1} = x_t + u_t, \quad t = 0, 1, \dots, \quad x_0 \text{ given}$$

Suppose that $J(x) = -\alpha x^2$ solves the Bellman equation. Find a quadratic equation for α . Then find the associated value of u^* .

- (b) By looking at the objective function, show that, given any starting value x_0 , it is reasonable to ignore any policy that fails to satisfy both $|x_t| \leq |x_{t-1}|$ and $|u_t| \leq |x_{t-1}|$ for $t = 1, 2, \dots$. Does Note 2 then apply?

12.4 The Maximum Principle

Dynamic programming is the most frequently used method for solving discrete time dynamic optimization problems. An alternative solution technique is based on the so called maximum principle. The actual calculations needed are often rather similar. However, when there are terminal restrictions on the state variables, the maximum principle is often preferable. The corresponding principle for optimization problems in continuous time is studied in more detail in Chapters 9 and 10, because for such problems it is the most important method.

EXAMPLE 1 Consider first the discrete time dynamic optimization problem with one state, one control variable and a free end state:

$$\max_{u_t \in U \subseteq \mathbb{R}} \sum_{t=0}^T f(t, x_t, u_t), \quad x_{t+1} = g(t, x_t, u_t), \quad t = 0, \dots, T-1, \quad x_0 \text{ is given}, \quad x_T \text{ free} \quad (1)$$

Here we assume that the control region U is convex, i.e. an interval. The state variable x_t evolves from the initial state x_0 according to the law of motion in (1), with u_t as a control that is chosen at each $t = 0, \dots, T$. Define the **Hamiltonian** by

$$H(t, x, u, p) = \begin{cases} f(t, x, u) + pg(t, x, u) & \text{for } t < T \\ f(t, x, u) & \text{for } t = T \end{cases} \quad (2)$$

where p is called an **adjoint variable** (or **co-state variable**).

THEOREM 12.4.1 (THE MAXIMUM PRINCIPLE NECESSARY CONDITIONS)

Suppose $(\{x_t^*\}, \{u_t^*\})$ is an optimal sequence pair for problem (1), and let H be defined by (2). Then there exist numbers p_t , with $p_T = 0$, such that for all $t = 0, \dots, T$,

$$H_u^t(t, x_t^*, u_t^*, p_t)(u - u_t^*) \leq 0 \quad \text{for all } u \in U \quad (3)$$

(Note that if u_t^* is an interior point of U , (3) implies that $H_u^t(t, x_t^*, u_t^*, p_t) = 0$.) Furthermore, p_t is a solution of the difference equation

$$p_{t-1} = H_x^t(t, x_t^*, u_t^*, p_t), \quad t = 1, \dots, T \quad (4)$$

NOTE 1 In Theorem 12.4.1 there are no terminal conditions. When there are terminal conditions, Theorem 12.5.1 gives necessary conditions for the case of several variables. For a proof of these two theorems see Arkin and Evstigneev (1987). A closer analogy with the continuous time maximum principle comes from writing the equation of motion as $x_{t+1} - x_t = g(t, x_t, u_t)$. If we redefine the Hamiltonian accordingly, then (4) is replaced by $p_t - p_{t-1} = -H_x^t(t, x_t^*, u_t^*, p_t)$, which corresponds to equation (9.2.5).

Sufficient conditions are given in the following theorem. The proof is similar to the proof of the corresponding theorem in continuous time.

THEOREM 12.4.2 (SUFFICIENT CONDITIONS)

Suppose that the sequence triple $(\{x_t^*\}, \{u_t^*\}, \{p_t\})$ satisfies all the conditions in Theorem 12.4.1, and suppose further that $H(t, x, u, p_t)$ is concave with respect to (x, u) for every t . Then the sequence triple $(\{x_t^*\}, \{u_t^*\}, \{p_t\})$ is optimal.

NOTE 2 Suppose that admissible pairs are also required to satisfy the constraints $(x_t, u_t) \in A_t$, $t = 0, 1, \dots, T$, where A_t is a convex set for all t . Then Theorem 12.4.2 is still valid, and H need only be concave in A_t .

NOTE 3 If U is compact and the functions f and g are continuous, there will always exist an optimal solution. (This result can be proved by using the extreme value theorem.)

NOTE 4 A weaker sufficient condition for optimality than in Theorem 12.4.2 is that for each t the pair (x_t^*, u_t^*) maximizes $H(t, x, u, p_t) - p_{t-1}x$ as a function of $u \in U$ and $x \in \mathbb{R}$.

EXAMPLE 2 Apply Theorem 12.4.2 to the problem in Example 12.1.2.

$$\max_{u_t \geq 0} \sum_{t=0}^3 (1 + x_t - u_t^2), \quad x_{t+1} = x_t + u_t, \quad x_0 = 0, \quad t = 0, 1, 2, \quad u_t \in \mathbb{R}$$

Solution: For $t < 3$, the Hamiltonian is $H = 1 + x - u^2 + p(x + u)$, so $H'_u = -2u + p$ and $H'_x = 1 + p$. For $t = 3$, $H = 1 + x - u^2$, so $H'_u = -2u$ and $H'_x = 1$. Note that the Hamiltonian is concave in (x, u) . The control region is open, so (3) implies that $H'_u = 0$, i.e., $-2u_t^* + p_t = 0$ for $t = 0, 1, 2$, and $-2u_3^* = 0$ for $t = 3$. Thus $u_0^* = \frac{1}{2}p_0$, $u_1^* = \frac{1}{2}p_1$, and $u_2^* = \frac{1}{2}p_2$.

The difference equation (4) for p_t is $p_{t-1} = 1 + p_t$ for $t = 1, 2$, and so $p_0 = 1 + p_1$, $p_1 = 1 + p_2$. For $t = 3$, (4) yields $p_2 = 1$, and we know from Theorem 12.4.1 that $p_3 = 0$. It follows that $p_2 = 1$, $p_1 = 1 + p_2 = 2$, and $p_0 = 1 + p_1 = 3$. This implies the optimal choices $u_0^* = 3/2$, $u_1^* = 1$, $u_2^* = 1/2$, and $u_3^* = 0$ for the controls, which is the same result as in Example 12.1.2.

EXAMPLE 3 Consider an oil field in which $x_0 > 0$ units of extractable oil remain at time $t = 0$. Let $u_t \geq 0$ be the rate of extraction and let x_t be the remaining stock at time t . Then $u_t = x_t - x_{t+1}$. Let $C(t, x_t, u_t)$ denote the cost of extracting u_t units in period t when the stock is x_t . Let w be the price per unit of oil and let r be the discount rate, with $\beta = 1/(1+r) \in (0, 1)$ the corresponding discount factor. If T is the fixed end of the planning period, the problem of maximizing total discounted profit can be written as

$$\max_{u_t \geq 0} \sum_{t=0}^T \beta^t [wu_t - C(t, x_t, u_t)], \quad x_{t+1} = x_t - u_t, \quad t = 0, 1, \dots, T-1, \quad x_0 > 0 \quad (i)$$

assuming also that

$$u_t \leq x_t, \quad t = 0, 1, \dots, T \quad (ii)$$

because the amount extracted cannot exceed the stock.

Because of restriction (ii), this is not a dynamic optimization problem of the type described by (1). However, if we define a new control v_t by $u_t = v_t x_t$, then restriction (ii) combined with $u_t \geq 0$ reduces to the control restriction $v_t \in [0, 1]$, and we have a standard dynamic optimization problem. Assuming that $C(t, x, u) = u^2/x$ and $0 < w < 1$, apply the maximum principle to find the only possible solution of the problem

$$\max_{v_t \in [0, 1]} \sum_{t=0}^T \beta^t (wv_t x_t - v_t^2 x_t), \quad x_{t+1} = x_t(1 - v_t), \quad x_0 > 0, \quad v_t \in [0, 1] \quad (iii)$$

with x_T free.

Solution: We denote the adjoint function by p_t . We know that $p_T = 0$. The Hamiltonian is $H = \beta^t (wv_t x_t - v_t^2 x_t) + p_t x_t(1 - v_t)$. (This is valid also for $t = T$, because then $p_T = 0$.) Then $H'_v = \beta^t (wx_t - 2v_t x_t) - p_t x_t$ and $H'_x = \beta^t (wv_t - v_t^2) + p_t(1 - v_t)$. So (3) implies that, for $(\{x_t^*\}, \{v_t^*\})$ to solve the problem, there must exist numbers p_t , with $p_T = 0$, such that, for all $t = 0, \dots, T$,

$$[\beta^t x_t^* (w - 2v_t^*) - p_t x_t^*](v - v_t^*) \leq 0 \quad \text{for all } v \text{ in } [0, 1] \quad (iv)$$

For $t = T$, with $p_T = 0$, this condition reduces to

$$\beta^T x_T^* (w - 2v_T^*)(v - v_T^*) \leq 0 \quad \text{for all } v \text{ in } [0, 1] \quad (v)$$

Having $v_T^* = 0$ would imply that $wv \leq 0$ for all v in $[0, 1]$, which is impossible because $w > 0$. Suppose instead that $v_T^* = 1$. Then (v) reduces to $\beta^T x_T^* (w - 2)(v - 1) \leq 0$ for all v in $[0, 1]$, which is impossible because $w - 2 < 0$ (put $v = 0$). Hence, $v_T^* \in (0, 1)$. For $t = T$, condition (v) then reduces to $\beta^T x_T^* (w - 2v_T^*) = 0$, and so

$$v_T^* = \frac{1}{2}w \quad (vi)$$

According to (4), for $t = 1, \dots, T$,

$$p_{t-1} = \beta^t v_t^* (w - v_t^*) + p_t(1 - v_t^*) \quad (vii)$$

For $t = T$, because $p_T = 0$ and $v_T^* = \frac{1}{2}w$, this equation reduces to

$$p_{T-1} = \beta^T v_T^* (w - v_T^*) = \frac{1}{4}w^2 \beta^T \quad (viii)$$

For $t = T - 1$, the expression within square brackets in (iv) is

$$\beta^{T-1} x_{T-1}^* (w - 2v_{T-1}^*) - p_{T-1} x_{T-1}^* = \beta^{T-1} x_{T-1}^* [w(1 - \frac{1}{4}\beta w) - 2v_{T-1}^*] \quad (ix)$$

Because $0 < w < 1$ and $\beta \in (0, 1)$, one has $1 > \frac{1}{4}\beta w$. It follows that both $v_{T-1}^* = 0$ and $v_{T-1}^* = 1$ are impossible as optimal choices in (iv). So the maximizer $v_{T-1}^* \in (0, 1)$ in (iv) must be interior, implying that the square bracket in the last line of (ix) is 0. Hence,

$$v_{T-1}^* = \frac{1}{2}w(1 - \frac{1}{4}\beta w)$$

Let us now go k periods backwards in time. Define $q_{T-k} = p_{T-k}/\beta^{T-k}$. We prove by backward induction that at each time $T-k$ we have an interior maximum point u_{T-k}^* in (iv). Then $u_{T-k}^* = \frac{1}{2}(w - q_{T-k})$, which belongs to $(0, 1)$ if $q_{T-k} \in (w-2, w)$. We prove by backward induction that $q_{T-k} \in [0, w]$. Suppose this is valid for k . Let us show it for $k+1$. Using (vii) and the definition of q_{T-k} , we find that $q_{T-(k+1)} = F(q_{T-k})$ where $F(q) = \beta[\frac{1}{4}(w-q)^2 + q] \geq 0$. Note that $q \mapsto F(q)$ is a strictly convex function and, by the assumptions on the parameters, we have $0 < F(q) \leq \max\{F(0), F(w)\} = \max\{\beta w^2/4, \beta w\} < w$ for all q in $[0, w]$. Hence, $q_{T-(k+1)} \in [0, w]$. Because $q_T = 0$, the backward induction can be started. Thus the solution of the problem is given by $u_{T-k}^* = (1/2)(w - q_{T-k})$, where q_{T-k} is determined by $q_{T-(k+1)} = \beta^{-(T-k)} p_{T-(k+1)} = F(q_{T-k})$, with $q_T = 0$.

PROBLEMS FOR SECTION 12.4

1. Consider Problem 12.1.1. Write down the Hamiltonian, condition (3), and the difference equation for p_t . Use the maximum principle to find a unique solution candidate. Verify that the conditions in Theorem 12.4.2 are satisfied, and that you have found the optimal solution.

SM 2. (Bolyanski) Consider the problem

$$\max_{u_t \in [-1, 1]} \sum_{t=0}^T (u_t^2 - 2x_t^2) \quad \text{s.t.} \quad x_{t+1} = u_t, \quad t = 0, 1, \dots, T-1, \quad x_0 = 0$$

- (a) Prove that $u_t^* = 0$ for $t = 0, 1, \dots, T-1$, and $u_T^* = 1$ (or -1) are optimal controls. (Express the objective function as a function of $u_0, u_1, \dots, u_T$ only.)
- (b) Verify that, although the conditions in Theorem 12.4.1 are satisfied, u_t^* does not maximize $H(t, x_t^*, u, p_t)$ subject to $u \in [-1, 1]$.

12.5 More Variables

Consider the following end constrained problem with n state and r control variables:

$$\max \sum_{t=0}^T f(t, \mathbf{x}_t, \mathbf{u}_t), \quad \mathbf{x}_{t+1} = \mathbf{g}(t, \mathbf{x}_t, \mathbf{u}_t), \quad \mathbf{x}_0 \text{ is given}, \quad \mathbf{u}_t \in U \subseteq \mathbb{R}^r \quad (1)$$

Here $\mathbf{x}_t$ is a state vector in $\mathbb{R}^n$ that evolves from the initial state $\mathbf{x}_0$ according to the law of motion in (1), with $\mathbf{u}_t$ as a control vector in U that is chosen at each $t = 0, \dots, T$. We put $\mathbf{x}_t = (x_t^1, \dots, x_t^n)$, $\mathbf{u}_t = (u_t^1, \dots, u_t^r)$, and $\mathbf{g} = (g^1, \dots, g^n)$. We assume that the control region U is convex.

The terminal conditions are assumed to be

$$\begin{aligned} \text{(a)} \quad & x_T^i = \bar{x}^i \quad \text{for } i = 1, \dots, l \\ \text{(b)} \quad & x_T^i \geq \bar{x}^i \quad \text{for } i = l+1, \dots, m \\ \text{(c)} \quad & x_T^i \text{ free} \quad \text{for } i = m+1, \dots, n \end{aligned} \quad (2)$$

where $0 \leq l \leq m \leq n$. Define the **Hamiltonian** by

$$H(t, \mathbf{x}, \mathbf{u}, \mathbf{p}) = \begin{cases} q_0 f(t, \mathbf{x}, \mathbf{u}) + \sum_{i=1}^n p_i^i g_i(t, \mathbf{x}, \mathbf{u}) & \text{for } t < T \\ f(t, \mathbf{x}, \mathbf{u}) & \text{for } t = T \end{cases}$$

where $\mathbf{p} = (p^1, \dots, p^n)$ is called an **adjoint variable** (or **co-state variable**). (For a proof of the following theorem, see Arkin and Evstigneev (1987).)

THEOREM 12.5.1 (THE MAXIMUM PRINCIPLE AND SUFFICIENCY)

Suppose that $(\{\mathbf{x}_t^*\}, \{\mathbf{u}_t^*\})$ is an optimal sequence pair for problem (1)–(2). Then there exist vectors $\mathbf{p}_t$ in $\mathbb{R}^n$ and a number q_0 , with $(q_0, \mathbf{p}_T) \neq (0, 0)$ and with $q_0 = 0$ or 1 , such that for $t = 0, \dots, T$,

$$\sum_{j=1}^r \frac{\partial H}{\partial u^j}(t, \mathbf{x}_t^*, \mathbf{u}_t^*, \mathbf{p}_t) (u_t^j - (u_t^j)^*) \leq 0 \quad \text{for all } \mathbf{u} \in U \quad (3)$$

Also, the vector $\mathbf{p}_t = (p_t^1, \dots, p_t^n)$ is a solution of

$$p_{t-1}^i = \frac{\partial H}{\partial x^i}(t, \mathbf{x}_t^*, \mathbf{u}_t^*, \mathbf{p}_t), \quad t = 1, \dots, T-1 \quad (4)$$

Moreover,

$$p_{T-1}^i = q_0 \frac{\partial f}{\partial x^i}(T, \mathbf{x}_T^*, \mathbf{u}_T^*) + p_T^i \quad (5)$$

where the vector $\mathbf{p}_T = (p_T^1, \dots, p_T^n)$ satisfies

$$\begin{aligned} \text{(a')} \quad & p_T^i \text{ no conditions} & i = 1, \dots, l \\ \text{(b')} \quad & p_T^i \geq 0, \quad \text{with } p_T^i = 0 \text{ if } x_T^{*i} > \bar{x}^i & i = l+1, \dots, m \\ \text{(c')} \quad & p_T^i = 0 & i = m+1, \dots, n \end{aligned} \quad (6)$$

If the conditions above are satisfied with $q_0 = 1$ and if $H(t, \mathbf{x}, \mathbf{u}, \mathbf{p})$ is concave in $(\mathbf{x}, \mathbf{u})$, then $(\{\mathbf{x}_t^*\}, \{\mathbf{u}_t^*\})$ is optimal.

NOTE 1 If $m = 0$ (so that there are no restrictions on the terminal state $\mathbf{x}_T$), then $\mathbf{p}_T = 0$, and it follows from Theorem 12.5.1 that $q_0 = 1$.

NOTE 2 If $\mathbf{u}_t^*$ is an interior point of U , then (3) implies that $H_{u^i}^i(t, \mathbf{x}_t^*, \mathbf{u}_t^*, \mathbf{p}_t) = 0$ for all $i = 1, \dots, r$.

EXAMPLE 1 (A life cycle model) Solve the problem

$$\max \sum_{t=0}^{T-1} \beta^t U(u_t) \quad \text{subject to} \quad x_{t+1} = \alpha(x_t + y_t - u_t), \quad t = 0, 1, \dots, T-1$$

where x_0 is given and we require $x_T \geq 0$. The control region is $(0, \infty)$.

The economic interpretation is that a consumer wants to maximize a sum of discounted utilities $\beta^t U(u_t)$ up to a fixed horizon T . We assume that $U'(u) > 0$ and $U''(u) < 0$ for $u > 0$. The coefficient β is the subjective discount rate. Wealth x_t develops according to the given difference equation, where y_t is income at time t , and $\alpha = (1+r)$ with r as the interest rate. (A consumer who deposits $x_t + y_t - u_t$ in an interest-bearing account at time t receives x_{t+1} at time $t+1$.)

Solution: The Hamiltonian is $H = H(t, x, u, p) = \beta^t U(u) + p\alpha(x + y - u)$ for $t = 0, 1, \dots, T-1$, and $H = H(T, x, u, p) = 0$ for $t = T$. Clearly H is concave in (x, u) , so we use sufficient conditions with $q_0 = 1$. According to (4) and (5) we get $p_{t-1} = \alpha p_t$ for $t < T$ and $p_{T-1} = p_T$. It follows that for $t < T$ we obtain $p_t = \alpha^{T-t-1} p_T$. Because the control region is open, condition (3) reduces to $H'_u(t, x_t^*, u_t^*, p_t) = 0$ for $t < T$ (see Note 2). This means that $\beta^t U'(u_t^*) - \alpha p_T \alpha^{T-t-1} = 0$, so $U'(u_t^*) = p_T \alpha^T (\alpha\beta)^{-t}$. In particular, optimality requires

$$\frac{U'(u_t^*)}{U'(u_{t+1}^*)} = \alpha\beta$$

Thus, the ratio of the marginal utilities from one period to the next is the constant $\alpha\beta$, equal to the discounted one-period rate of return. Note that consumption is constant in case $\alpha\beta = 1$, rising in case $\alpha\beta > 1$, and falling in case $\alpha\beta < 1$.

Because U' is strictly decreasing and so has an inverse, $u_t^* = (U')^{-1}(p_T \alpha^T (\alpha\beta)^{-t})$. In particular we see that $p_T \neq 0$, so $p_T > 0$. Then (6)(b') implies that $x_T^* = 0$. But $x_T^* = \alpha^T x_0 + \sum_{k=1}^T \alpha^{T-k+1} (y_{k-1} - u_{k-1}^*) = \alpha^T x_0 + \sum_{k=1}^T \alpha^{T-k+1} (y_{k-1} - (U')^{-1}(p_T \alpha^T (\alpha\beta)^{1-t}))$, using formula (11.1.8). The equality $x_T^* = 0$ can then be used to determine p_T uniquely. $\blacksquare$

Infinite Horizon

We consider briefly the following infinite horizon version of problem (1)–(2):

$$\max \sum_{t=0}^{\infty} f(t, x_t, u_t), \quad x_t \in \mathbb{R}^n, \quad u_t \in U \subseteq \mathbb{R}^r, \quad U \text{ convex} \quad (7)$$

where we maximize over all sequence pairs $\{(x_t), \{u_t\}\}$, satisfying

$$x_{t+1} = g(t, x_t, u_t), \quad t = 1, 2, \dots, \quad x_0 \text{ given} \quad (8)$$

and the terminal conditions⁴

$$\begin{aligned} \text{(a)} \quad & \lim_{T \rightarrow \infty} x_i(T) = \hat{x}_i, \quad i = 1, \dots, l \\ \text{(b)} \quad & \lim_{T \rightarrow \infty} x_i(T) \geq \hat{x}_i, \quad i = l+1, \dots, m \\ \text{(c)} \quad & \text{no condition,} \quad i = m+1, \dots, n \end{aligned} \quad (9)$$

Note that f and $g = (g_1, \dots, g_n)$ can now depend explicitly on t . Assume that the sum in (7) exists for all admissible sequences. The functions f and g are assumed to be C^1 with respect to all x_t and u_t .

⁴ See Section 10.3 or A.3 for the definition of $\lim$ or $\liminf$.

We merely state a sufficient condition for this problem:

THEOREM 12.5.2 (SUFFICIENT CONDITIONS)

Suppose that the sequence triple $\{(x_t^*), \{u_t^*\}, \{p_t\}\}$ satisfies the conditions (3)–(4) with $q_0 = 1$. Suppose further that U is convex and the Hamiltonian $H(t, x, u, p_t)$ is concave in (x, u) for every t . Then $\{(x_t^*), \{u_t^*\}\}$ is optimal provided that the following transversality condition is satisfied: for all admissible sequence pairs $\{(x_t), \{u_t\}\}$,

$$\lim_{t \rightarrow \infty} p_t \cdot (x_t - x_t^*) \geq 0 \quad (10)$$

NOTE 3 Suppose that any admissible sequence $\{x_t, u_t\}$ is required to satisfy additional constraints. Then (10) needs only to be tested for such sequences.

PROBLEMS FOR SECTION 12.5

5M 1. Consider the problem

$$\max_{u,v \in \mathbb{R}} \sum_{t=0}^2 [1 + x_t - y_t - 2u_t^2 - v_t^2] \quad \text{s.t.} \quad \begin{cases} x_{t+1} = x_t - u_t, & x_0 = 5 \\ y_{t+1} = y_t + v_t, & y_0 = 2 \end{cases} \quad t = 0, 1.$$

- Solve the problem by using the difference equations to express the objective function I as a function of only u_0, u_1, u_2, v_0, v_1 , and v_2 , and then optimize.
- Solve the problem by using dynamic programming. (Hint: Find $J_2(t, y)$, $J_1(x, y)$, and $J_0(x, y)$ and the corresponding optimal controls.)
- Solve the problem by using Theorem 12.5.1.

2. Solve the problem

$$\max \sum_{t=0}^T (-x_t^2 - u_t^2) \quad \text{subject to} \quad x_{t+1} = y_t, \quad y_{t+1} = y_t + u_t, \quad t = 0, 1, \dots, T-1$$

where $x_0 = x^0$ and $y_0 = y^0$ are given numbers and $u_t \in \mathbb{R}$.

3. Solve the problem

$$\max \sum_{t=0}^{\infty} \beta^t \ln(x_t - u_t) \quad \text{subject to} \quad x_{t+1} = u_t, \quad x_0 > 0, \quad u_t > 0$$

where $\beta \in (0, 1)$. Verify that $x_t^* > u_t^*$ for all t .

12.6 Stochastic Optimization

What is the best way of controlling a dynamic system subject to random disturbances? Stochastic dynamic programming is a central tool for tackling this problem.

In deterministic dynamic programming the state develops according to a difference equation $\mathbf{x}_{t+1} = \mathbf{g}(t, \mathbf{x}_t, \mathbf{u}_t)$ that is controlled by appropriate choices of the control variables $\mathbf{u}_t$. In this section, the state $\mathbf{x}_t$ is also influenced by stochastic shocks, so that it becomes a random variable. Following common practice, we typically use capital instead of lower case letters to denote random variables, e.g. $\mathbf{X}_t$ instead of $\mathbf{x}_t$. We assume that $\mathbf{x}_t$ belongs to $\mathbb{R}^n$, and that $\mathbf{u}_t$ is required to belong to a given subset U of $\mathbb{R}^m$.

Suppose now that for $t = 0, 1, \dots, T$ the state equation takes the new form

$$\mathbf{X}_{t+1} = \mathbf{g}(t, \mathbf{X}_t, \mathbf{u}_t, \mathbf{V}_{t+1}), \quad \mathbf{X}_0 = \mathbf{x}_0, \quad \mathbf{V}_0 = \mathbf{v}_0, \quad \text{with } \mathbf{x}_0 \text{ and } \mathbf{v}_0 \text{ given, } \mathbf{u}_t \in U \quad (1)$$

including a stochastic shock $\mathbf{V}_{t+1}$ that makes each $\mathbf{X}_{t+1}$ stochastic as well.

We consider two cases. In the first, $\mathbf{V}_{t+1}$ is a random variable that takes values in a finite set $\mathcal{V}$. It is assumed that the probability that $\mathbf{V}_{t+1} = \mathbf{v} \in \mathcal{V}$ may depend on the outcome $\mathbf{v}_t$ at time t , as well as explicitly on time t . Then we consider the **conditional probability** that $\mathbf{V}_{t+1} = \mathbf{v}$ given the outcome $\mathbf{v}_t$ at time t , which is denoted by $P_t(\mathbf{v} | \mathbf{v}_t)$. The same notation could also be used when $\mathcal{V}$ is a countably infinite set. In the second case, $\mathbf{V}_{t+1}$ may take values anywhere in a Euclidean space. Then the distribution of $\mathbf{V}_{t+1}$ is assumed to be described by a conditional density $p_t(\mathbf{v} | \mathbf{v}_t)$ that is a continuous function of $\mathbf{v}$ and $\mathbf{v}_t$ together.

EXAMPLE 1 Consider first a two-stage decision problem with one state variable x and one control variable u . Assume that one wants to maximize the objective function

$$E[f(0, X_0, u_0) + f(1, X_1, u_1)] = f(0, X_0, u_0) + E[f(1, X_1, u_1)] \quad (*)$$

where E denotes expectation and f is some given reward function. Here the initial state $X_0 = x_0$ and an initial outcome u_0 are given, after which X_1 is determined by the difference equation (1), i.e. $X_1 = g(0, x_0, u_0, V_1)$. We can find the maximum by first maximizing with respect to u_1 , and then with respect to u_0 . When choosing u_1 , we simply maximize $f(1, X_1, u_1)$, assuming that X_1 is known before the maximization is carried out. The maximum point u_1^* is a function $u_1^*(X_1)$ of X_1 . Insert this function instead of u_1 into the objective function (*), and then replace the two occurrences of X_1 by $g(0, x_0, u_0, V_1)$. This yields

$$f(0, X_0, u_0) + E[f(1, g(0, x_0, u_0, V_1), u_1^*(g(0, x_0, u_0, V_1)))] \quad (**)$$

Note that u_0 occurs in both terms of (**). A maximizing value of u_0 is then chosen, taking both these occurrences into account.

When V_1 is uncertain, the following special case shows why it matters whether we can observe X_1 before choosing u_1 . Suppose that $f(0, X_0, u_0) = 0$, $f(1, X_1, u_1) = X_1 u_1$, and $X_1 = V_1$, where V_1 takes the values 1 and -1 with probabilities 1/2, and where the control u must equal one of the two values 1 and -1. If we have to choose u_1 before observing X_1 , implying that u_1 must be a constant, then $E[X_1 u_1] = 0$. But if we can first observe X_1 , then u_1 can depend on X_1 . By choosing $u_1 = u_1(X_1) = X_1$, we can make $E[X_1 u_1] = 1$, which yields a higher value of the objective (*).

In all that follows we shall assume that both X_t and V_t can be observed before choosing u_t . Often this assumption will be satisfied because we define the state X_t and random shock V_t as what can be observed at time t .

Let us turn to the general problem. The process determined by (1) and the values of the random variables $\mathbf{V}_1, \mathbf{V}_2, \dots$ is to be controlled in the best possible manner by appropriate choices of the successive variables $\mathbf{u}_t$. The objective function is now the expectation

$$E \left[\sum_{t=0}^T f(t, \mathbf{X}_t, \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t)) \right] \quad (2)$$

Here several things have to be explained. Each control $\mathbf{u}_t$, $t = 0, 1, 2, \dots, T$ should be a function $\mathbf{u}_t(\mathbf{x}_t, \mathbf{v}_t)$ of the current state $\mathbf{x}_t$ and the outcome $\mathbf{v}_t$. Such functions are called "policies", or more specifically **Markov policies** or **Markov controls**. For many stochastic optimization problems, including those studied here, this is the natural class of policies to consider in order to achieve an optimum. The policies that occur in (2) are of this type.

The expectation in (2) is the sum $\sum_{t=0}^T E[f(t, \mathbf{X}_t, \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t))]$ of the expectations of each successive term. These expectations, of course, depend on the probability of each pair $(\mathbf{X}_t, \mathbf{V}_t)$. To calculate these, first recall that in the discrete random variable case the probability that the events $\mathbf{V}_1 = \mathbf{v}_1$ and $\mathbf{V}_2 = \mathbf{v}_2$ occur jointly, given $\mathbf{V}_0 = \mathbf{v}_0$, equals the conditional probability that $\mathbf{V}_2 = \mathbf{v}_2$ occurs given $\mathbf{V}_1 = \mathbf{v}_1$, times the probability that $\mathbf{V}_1 = \mathbf{v}_1$ occurs given $\mathbf{V}_0 = \mathbf{v}_0$. That is, the joint probability equals $P_1(\mathbf{v}_2 | \mathbf{v}_1)$ times $P_0(\mathbf{v}_1 | \mathbf{v}_0)$. Similarly, the probability of the joint event $\mathbf{V}_1 = \mathbf{v}_1, \mathbf{V}_2 = \mathbf{v}_2, \dots, \mathbf{V}_t = \mathbf{v}_t$, is given by

$$P^t(\mathbf{v}_1, \dots, \mathbf{v}_t) = P_0(\mathbf{v}_1 | \mathbf{v}_0) \cdot P_1(\mathbf{v}_2 | \mathbf{v}_1) \cdots P_{t-1}(\mathbf{v}_t | \mathbf{v}_{t-1}) \quad (3)$$

Note that this probability is unaffected by the choice of the policies $\mathbf{u}_t(\mathbf{x}_t, \mathbf{v}_t)$. In the continuous random variable case, the joint density $p^t(\mathbf{v}_1, \dots, \mathbf{v}_t)$ is determined by the corresponding formula with each P_t in (3) replaced by p_t ; again, this is independent of the policy choices.

On the other hand, the probability distribution over the state $\mathbf{X}_t$ at each time t does depend on the choice of the policies $\mathbf{u}_t(\mathbf{x}_t, \mathbf{v}_t)$, in general, as does the joint distribution of each pair $(\mathbf{X}_t, \mathbf{V}_t)$. Specifically, the sequence $\mathbf{X}_t$, $t = 1, \dots, T$ in (2) is the solution of (1) when $\mathbf{V}_1, \dots, \mathbf{V}_T$ and $\mathbf{u}_t = \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t)$, $t = 0, \dots, T-1$ are inserted successively. Hence, $\mathbf{X}_t$ depends on $\mathbf{V}_1, \dots, \mathbf{V}_t$ and, for each t , the expectation $E[f(t, \mathbf{X}_t, \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t))]$ is calculated by means of the probabilities (or densities) specified in (3).

Though not always necessary, we shall assume that f and g are continuous in $\mathbf{x}, \mathbf{u}$ (or in $\mathbf{x}, \mathbf{u}$, and $\mathbf{v}$ in the case when $\mathbf{v}$ is a random variable with a continuous density function).

The optimization problem is to find a sequence of policies $\mathbf{u}_0^*(\mathbf{x}_0, \mathbf{v}_0), \dots, \mathbf{u}_T^*(\mathbf{x}_T, \mathbf{v}_T)$, that makes the objective (2) as large as possible. We now define

$$J_t(\mathbf{x}_t, \mathbf{v}_t) = \max E \left[\sum_{s=t}^T f(s, \mathbf{X}_s, \mathbf{u}_s(\mathbf{X}_s, \mathbf{V}_s)) \mid \mathbf{x}_t, \mathbf{v}_t \right] \quad (4)$$

Here the expected discounted total reward is maximized over all policy sequences $\mathbf{u}_s = \mathbf{u}_s(\mathbf{x}_s, \mathbf{v}_s)$ ($s = t, \dots, T$). The expectation is taken over all possible sequences of realizations $\mathbf{V}_t$ of the random variables, given that:

- (i) we start equation (1) in state $\mathbf{x}_t$ at time t , as indicated by " $\mathbf{x}_t, \mathbf{v}_t$ " in (4);
- (ii) we apply each control $\mathbf{u}_s(\mathbf{X}_s, \mathbf{V}_s)$ in the sequence when computing the sequence of successive states $\mathbf{X}_s$ ($s = t, \dots, T$).

Note that $\mathbf{X}_s$ will depend on $\mathbf{v}_t, \mathbf{V}_{t+1}, \dots, \mathbf{V}_s$, for each $s = t + 1, \dots, T$.

The central tool in solving optimization problems of the type (1)–(2) is the following **dynamic programming equation** or **optimality equation**:

$$J_t(\mathbf{x}_t, \mathbf{v}_t) = \max_{\mathbf{u}_t} \{f(t, \mathbf{x}_t, \mathbf{u}_t) + E[J_{t+1}(\mathbf{X}_{t+1}, \mathbf{V}_{t+1}) \mid \mathbf{x}_t, \mathbf{v}_t]\} \quad (5)$$

where $\mathbf{X}_{t+1} = \mathbf{g}(t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{V}_{t+1})$. (The notation $\mathbf{x}_t, \mathbf{v}_t$ is just a reminder that inserting this value of $\mathbf{X}_{t+1}$ makes the conditional expectation depend on $\mathbf{x}_t$ as well as on $\mathbf{v}_t$). After this insertion, since $\mathbf{x}_t$ affects the expectation of J_{t+1} only through its effect on $\mathbf{X}_{t+1}$, we can write equation (5) as

$$J_t(\mathbf{x}_t, \mathbf{v}_t) = \max_{\mathbf{u}_t} \{f(t, \mathbf{x}_t, \mathbf{u}_t) + E[J_{t+1}(\mathbf{g}(t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{V}_{t+1}), \mathbf{V}_{t+1}) \mid \mathbf{v}_t]\} \quad (6)$$

Moreover, when $t = T$ we have

$$J_T(\mathbf{x}_T, \mathbf{v}_T) = J_T(\mathbf{x}_T) = \max_{\mathbf{u}_T} f(T, \mathbf{x}_T, \mathbf{u}_T) \quad (7)$$

These equations are similar to (6) and (7) in Theorem 12.1.1 for the deterministic case. The only significant differences are that $\mathbf{v}_t$ appears as an extra state variable, in effect, and that (6) allows for uncertainty by including the conditional expectation of J_{t+1} .

As in the corresponding deterministic problem considered in Section 12.1, first (7) is used to find $\mathbf{u}_T^*(\mathbf{x}_T, \mathbf{v}_T)$. Thereafter (6) is used repeatedly in a backward recursion to find first $\mathbf{u}_{T-1}^*(\mathbf{x}_{T-1}, \mathbf{v}_{T-1})$, then $\mathbf{u}_{T-2}^*(\mathbf{x}_{T-2}, \mathbf{v}_{T-2})$, and so on all the way back to $\mathbf{u}_0^*(\mathbf{x}_0, \mathbf{v}_0)$.

As in the deterministic case, equations (6) and (7) are, essentially, both necessary and sufficient. They are sufficient in the sense that if $\mathbf{u}_t^*(\mathbf{x}_t, \mathbf{v}_t)$ maximizes the right-hand side of (6) for $t = 1, \dots, T$, and also $\mathbf{u}_T^*(\mathbf{x}_T, \mathbf{v}_T)$ maximizes the right-hand side of (7), then $\mathbf{u}_t^*(\mathbf{x}_t, \mathbf{v}_t)$, $t = 0, 1, \dots, T$, is indeed an optimal policy sequence. On the other hand, the same equations are necessary in the sense that, for every pair $(\mathbf{x}_t, \mathbf{v}_t)$ that occurs with positive probability (or has a positive probability density when there is a continuous density function), an optimal control $\mathbf{u}_t^*(\mathbf{x}_t, \mathbf{v}_t)$ must yield a maximum on the right-hand side of (6) for $t = 0, 1, \dots, T-1$, and of (7) for $t = T$.

An important special case occurs when the successive random variables $\mathbf{V}_t$ are independently distributed. Then $P_t(\mathbf{v}_t \mid \mathbf{v}_t)$ (or $p_t(\mathbf{v}_t \mid \mathbf{v}_t)$) does not depend on $\mathbf{v}_t$, which can therefore be dropped from the functions $J_t(\mathbf{x}_t, \mathbf{v}_t)$, $\mathbf{u}_t(\mathbf{x}_t, \mathbf{v}_t)$ and from (6) (or (7) for $t = T$). Intuitively, this is because the conditioning variable $\mathbf{v}_t$ drops out of (5), so both $J_t(\mathbf{x}_t, \mathbf{v}_t)$ and the optimal control $\mathbf{u}_t^*(\mathbf{x}_t, \mathbf{v}_t)$ will not depend on $\mathbf{v}_t$ either. Formally, this independence can be proved by backward induction. Some examples below are of this special form.

The argument above also holds if the control region is a closed set that depends on t and $\mathbf{x}$ —for example, if $U = U(t, \mathbf{x}) = \{\mathbf{u} : h_t(t, \mathbf{x}, \mathbf{u}) \leq 0\}$ where the functions h_t are continuous in $(\mathbf{x}, \mathbf{u})$. Thus the constraint $\mathbf{u}_t \in U(t, \mathbf{x}_t)$ is imposed. If $U(t, \mathbf{x})$ is empty, then the maximum over $U(t, \mathbf{x})$ is set equal to $-\infty$.

EXAMPLE 2

Consider a stochastic version of Example 12.1.4, where at each time $t = 0, 1, \dots, T-1$, the state variable X_t is an investor's wealth, the control variable u_t is consumption, and the certain return a_t to investment in that example is replaced by a random return Z_{t+1} . Moreover, suppose that $\{Z_{t+1}\}_{t=0}^{T-1}$ is a sequence of independently distributed random variables with positive values. In particular, for each $t = 1, 2, \dots, T$ there is either a discrete distribution $P_t(Z_t)$, or a continuous density function $p_t(z_t)$.

More specifically, the state X_t is assumed to evolve according to the stochastic difference equation

$$X_{t+1} = Z_{t+1}(X_t - u_t), \quad u_t \in [0, x_t], \quad x_0 \text{ given} \quad (i)$$

The objective function is the obvious counterpart of that in Example 12.1.4, namely the expected sum of discounted utility, given by

$$E \left[\sum_{t=0}^{T-1} \beta^t u_t^{1-\gamma} + \beta^T A X_T^{1-\gamma} \right] \quad (ii)$$

where $\beta \in (0, 1)$ is a discount factor, while $\gamma \in (0, 1)$ is a taste parameter, and A is a positive constant. Thus, the problem is to maximize (ii) subject to (i). Assume that $E[Z_t^{1-\gamma}] < \infty$ for all t .

Solution: Here $J_T(x_T) = \beta^T A x_T^{1-\gamma}$ exactly as in (ii) in Example 12.1.4. Because the random variables Z_t are independently distributed, the value functions take the form $J_t(x)$. To find $J_{T-1}(x)$, we use the optimality equation

$$J_{T-1}(x) = \max_u \{ \beta^{T-1} u^{1-\gamma} + E[\beta^{T-1} A (Z_T(x - u))^{1-\gamma}] \} \quad (*)$$

The expectation must be calculated by using the probability distribution for Z_T . In fact, the right-hand side of (*) is of the same form as (iv) in Example 12.1.4. To make it exactly the same, define the new constant a_{T-1} so that $a_{T-1}^{1-\gamma} = E[Z_T^{1-\gamma}]$. With this new notation, the optimal control u_{T-1} is given by (v) in Example 12.1.4. Furthermore, equations (vii), (viii), and (ix) of that example still hold provided we define each a_t to satisfy $a_t^{1-\gamma} = E[Z_{t+1}^{1-\gamma}]$ for $t = 0, 1, \dots, T-1$. One may call each a_t the "certainty equivalent" return because the solution is exactly the same as if it replaced the uncertain return described by the random variable Z_{t+1} . Of course, each a_t depends on the taste parameter γ as well as on the distribution Z_t .

EXAMPLE 3

Suppose a skillful gambler repeatedly chooses to bet a certain fraction u of her wealth at even odds, expecting to win back this fraction with probability $p \geq 1/2$. Thus, if wealth at time $t-1$ is x_{t-1} , then x_t is equal to $x_{t-1} + u x_{t-1}$ with probability p , and to $x_{t-1} - u x_{t-1}$ with probability $1-p$. (In the notation of this section, $X_t = X_{t-1} + u_{t-1} V_t X_{t-1}$, where $V_t \in \{-1, 1\}$, $\Pr\{V_t = 1\} = p$, and $\Pr\{V_t = -1\} = 1-p$.) Suppose the gambler plays T times, with the objective of maximizing the expected value of the utility of terminal wealth x_T , assumed to be $f(T, x_T) = \ln x_T$. Note that $f(T, x_T)$ is independent of u_T , so $J_T(x_T) = \ln x_T$. We also have $f(t, x_t) \equiv 0$ for $t < T$.

If the gambler's wealth at time $T-1$ is x_{T-1} and then the amount bet is ux_{T-1} , the resulting expected utility of terminal wealth will be

$$p \ln(x_{T-1} + ux_{T-1}) + q \ln(x_{T-1} - ux_{T-1}) = \ln x_{T-1} + A(u)$$

where $A(u) = p \ln(1+u) + q \ln(1-u)$ (because $p+q=1$). At time $T-1$, therefore, the optimality equation is

$$J_{T-1}(x_{T-1}) = \ln x_{T-1} + \max_{0 \leq u \leq 1} A(u)$$

The function $A(u)$ is concave, so the maximum is attained where

$$A'(u) = p \frac{1}{1+u} - q \frac{1}{1-u} = 0$$

This implies $p(1-u) = q(1+u)$, or $p-q = u(p+q) = u$, so $u_{T-1}^* = p-q$. Inserting this expression for u into the right-hand side gives the maximum value. This is

$$J_{T-1}(x) = \ln x + B$$

where $B = p \ln[1+(p-q)] + q \ln[1-(p-q)] = p \ln(2p) + q \ln(2q) = \ln 2 + p \ln p + q \ln q$.

Starting from x_{T-2} at time $T-2$, next period the gambler ends up with probability p at $x_{T-1} = x_{T-2} + ux_{T-2}$, where $J_{T-1}(x_{T-1}) = \ln(x_{T-2} + ux_{T-2}) + B$; and with probability $1-p$ at $x_{T-1} = x_{T-2} - ux_{T-2}$, where $J_{T-1}(x_{T-1}) = \ln(x_{T-2} - ux_{T-2}) + B$. Therefore,

$$\begin{aligned} J_{T-2}(x_{T-2}) &= \max_{0 \leq u \leq 1} [p \ln(x_{T-2} + ux_{T-2}) + B + q \ln(x_{T-2} - ux_{T-2}) + B] \\ &= \ln x_{T-2} + B + \max_{0 \leq u \leq 1} A(u) \end{aligned}$$

Once again, the maximum value in the latter maximization problem is B , with $u = p-q$. Hence,

$$J_{T-2}(x) = \ln x + 2B, \quad \text{with } u_{T-2}^* = p-q = 2p-1$$

Continuing in this manner for $k=3, 4, \dots$ gives

$$J_{T-k}(x) = \ln x + kB, \quad \text{with } u_{T-k}^* = p-q = 2p-1$$

To conclude, in every round it is optimal for the gambler to bet the same fraction $u = 2p-1$ of his wealth. (If the objective function were $f(T, x_T) = x_T$ instead and $p > 1/2$, it is easy to see that she would bet all her wealth at every stage—see Problem 7.)

The following formal result confirms that the solutions we found in the previous examples, for instance, really are optimal:

THEOREM 12.6.1 (SUFFICIENCY OF THE OPTIMALITY EQUATIONS)

The sequence of policies $\pi = \{u_t(x_t, v_t)\}_{t=0}^T$ solves the problem of maximizing (2) subject to (1) if, together with a sequence of functions $\{J_t(x_t, v_t)\}_{t=0}^T$, it satisfies the optimality equations (6) and (7).

Proof: Let $\pi = \{u_t(x_t, v_t)\}_{t=0}^T$ be an arbitrary control sequence. Define

$$J_t^\pi(x_t, v_t) = E \left[\sum_{s=t}^T f(s, X_s, u_s(X_s, V_s)) \mid x_t, v_t \right]$$

Given the starting time t and state (x_t, v_t) , this is the conditionally expected value of following the process (1) using the control sequence π from time t on. Trivially, $J_t^\pi(x_t, v_t) \leq J_t(x_t, v_t)$, with equality if $u_t(x_t, v_t)$ satisfies (7). By backward induction, let us prove that $J_t^\pi(x_t, v_t) \leq J_t(x_t, v_t)$, with equality if π is such that $u_s(x_s, v_s)$ satisfies (6) for $s = t, t+1, \dots, T-1$, and $u_T(x_T, v_T)$ satisfies (7). As the induction hypothesis, assume that this is true for t . Replacing t by $t-1$ in the above definition gives

$$J_{t-1}^\pi(x_{t-1}, v_{t-1}) = f(t-1, x_{t-1}, u_{t-1}(x_{t-1}, v_{t-1})) + E \left[\sum_{s=t}^T f(s, X_s, u_s(X_s, V_s)) \mid x_{t-1}, v_{t-1} \right]$$

But the law of iterated expectations and the induction hypothesis together imply that

$$\begin{aligned} &E \left[\sum_{s=t}^T f(s, X_s, u_s(X_s, V_s)) \mid x_{t-1}, v_{t-1} \right] \\ &= E \left[E \left[\sum_{s=t}^T f(s, X_s, u_s(X_s, V_s)) \mid X_t, V_t, x_{t-1}, v_{t-1} \right] \mid x_{t-1}, v_{t-1} \right] \\ &= E \left[E \left[\sum_{s=t}^T f(s, X_s, u_s(X_s, V_s)) \mid X_t, V_t \right] \mid x_{t-1}, v_{t-1} \right] \\ &= E[J_t^\pi(X_t, V_t) \mid x_{t-1}, v_{t-1}] \leq E[J_t(x_t, v_t) \mid x_{t-1}, v_{t-1}] \end{aligned}$$

where $X_t = g(t, x_{t-1}, u_{t-1}(x_{t-1}, v_{t-1}), V_t)$, with equality if $u_t(x_t, v_t)$ satisfies (6) for $s = t, t+1, \dots, T-1$, and $u_T(x_T, v_T)$ satisfies (7). Hence,

$$\begin{aligned} J_{t-1}^\pi(x_{t-1}, v_{t-1}) &\leq f(t-1, x_{t-1}, u_{t-1}(x_{t-1}, v_{t-1})) + E[J_t(x_t, v_t) \mid x_{t-1}, v_{t-1}] \\ &\leq \max_u \{ f(t-1, x_{t-1}, u) + E[J_t(g(t, x_{t-1}, u, V_t), V_t) \mid v_{t-1}] \} \\ &= J_{t-1}(x_{t-1}, v_{t-1}) \end{aligned}$$

with equalities if $u_t(x_t, v_t)$ satisfies (6) for $s = t-1, t, t+1, \dots, T-1$, and $u_T(x_T, v_T)$ satisfies (7). This verifies the induction hypothesis for $t-1$, and so completes the proof. ■

In the discrete random variable case, the above proof is easily adapted to show that a policy π^* is optimal only if the optimality equations hold at every time $t = 0, 1, 2, \dots, T$ in any state (x_t, v_t) that is reached with positive probability given the policy π^* and the initial state (x_0, v_0) . In the continuous variable case these necessary conditions become a little bit more complicated: essentially, the optimality equations must hold at every time $t = 0, 1, 2, \dots, T$ in almost every state (x_t, v_t) that has a positive conditional probability density given π^* and the initial state (x_0, v_0) .

The Stochastic Euler Equation

We previously derived the Euler equation for the kind of deterministic problem considered in Section 12.2. There we were able to define the instantaneous reward function $F(t, x_t, x_{t+1})$ each period by using equation (12.2.2) in combination with the function $g(t, x_t, u_t)$ that determines x_{t+1} . In this section, however, the deterministic equation $x_{t+1} = g(t, x_t, u_t)$ is replaced by the stochastic equation $X_{t+1} = g(t, x_t, u_t, V_t)$, where V_t is random and known at time t when the control u_t is to be chosen.⁵ Then the earlier construction works, and leads to a problem of the form

$$\max E \left[\sum_{t=0}^T F(t, X_t, X_{t+1}, V_t) \right], \quad x_0 \text{ given, } x_t, x_2, \dots, x_T \text{ free} \quad (8)$$

This is like problem (12.2.1) that led to the discrete time Euler equation, except that the function F in the criterion contains an extra stochastic variable V_t , where as before $v = V_{t+1}$ is determined given v_t by either a conditional probability distribution $P_t(v | v_t)$ or a conditional density $p_t(v | v_t)$.

In problem (8) we decide the value of each x_{t+1} at time t , after x_t has been determined by previous choices, and after v_t has been observed. Hence, x_{t+1} is allowed to be a function of the pair (x_t, v_t) . Then the Euler equations take the form

$$F'_3(T, x_T, x_{T+1}, v_T) = 0 \quad (9)$$

at time $T+1$, and

$$E[F'_2(t, x_t, x_{t+1}(x_t, V_t), V_t) | v_{t-1}] + F'_3(t-1, x_{t-1}, x_t, v_{t-1}) = 0 \quad (10)$$

at times $t = 1, \dots, T$. This is called the "stochastic Euler equation" because it differs from (12.2.2) by including the random term V_t as an extra argument, and by having the conditional expectation of $F'_2(t, \cdot)$ given v_{t-1} .

In order to solve the problem (8), we first find x_{T+1} from (9), yielding the function $x_{T+1} = x_{T+1}(x_T, v_T)$. Next, this function is inserted into (10) for $t = T$, and the resulting equation is then solved for x_T , yielding $x_T = x_T(x_{T-1}, v_{T-1})$. After this, at the next step we solve for $x_{T-1}(x_{T-2}, v_{T-2})$. Working backwards in this manner, we continue until the function $x_1(x_0, v_0)$ has been constructed. Since x_0 and v_0 are given, the value of x_1 is determined. Then the value of $x_2 = x_2(x_1, v_1)$ is determined for each possible observed value of v_1 , and so on.

EXAMPLE 4 Solve the problem

$$\max E \left[\sum_{t=0}^{T-1} -(X_{t+1} - \frac{1}{2}X_t + V_t)^2 - (X_{T+1} - X_T)^2 - \frac{1}{2}X_T^2 \right]$$

where $V_t, t = 1, \dots, T-1$, are i.i.d. random variables with $E[V_t] = 1$.

⁵ Any problem involving a stochastic difference equation like this can also be solved as a dynamic programming problem by writing $x_{t+1} = g(t, x_t, u_t, y_t)$, where y_t is an extra state variable whose evolution is determined by $y_{t+1} = V_{t+1}, y_0 = v_0$.

Solution: In this example,

$$F(T, x_T, x_{T+1}, v_T) = -(x_{T+1} - x_T)^2 - \frac{1}{2}x_T^2$$

and for $t = 0, 1, \dots, T-1$ we have

$$F(t, x_t, x_{t+1}, v_t) = -(x_{t+1} - \frac{1}{2}x_t + v_t)^2$$

Here equation (9) becomes $-2(x_{T+1} - x_T) = 0$, so $x_{T+1}(x_T, v_T) = x_T$. Next, for $t = T$, equation (10) becomes $E[-x_T] - 2(x_T - \frac{1}{2}x_{T-1} + v_{T-1}) = 0$, so $x_T = \frac{1}{3}x_{T-1} - \frac{2}{3}v_{T-1}$.

Furthermore, for $t = T-1$, equation (10) yields

$$E \left[\frac{1}{3}x_{T-1} - \frac{2}{3}v_{T-1} - \frac{1}{2}x_{T-1} + v_{T-1} \right] - 2(x_{T-1} - \frac{1}{2}x_{T-2} + v_{T-2}) = 0$$

so $x_{T-1} = \frac{6}{13}(x_{T-2} - 2v_{T-2} + \frac{1}{3})$.

We will now prove by backward induction that the optimal policy takes the form

$$x_{t+1}(x_t, v_t) = a_t + b_t x_t + c_t v_t \quad \text{for } t = 0, 1, \dots, T \quad (*)$$

where a_t, b_t , and c_t are suitable constants. This is evidently true for $t = T$ with $(a_T, b_T, c_T) = (0, 1, 0)$, for $t = T-1$ with $(a_{T-1}, b_{T-1}, c_{T-1}) = (0, \frac{1}{3}, -\frac{2}{3})$, and for $t = T-2$ with $(a_{T-2}, b_{T-2}, c_{T-2}) = (\frac{2}{13}, \frac{6}{13}, -\frac{10}{13})$.

For $t = 1, 2, \dots, T-3$, assuming that $(*)$ is true for $t+1$, we have

$$E[F'_2(t, x_t, x_{t+1}(x_t, V_t), V_t) | v_{t-1}] = E[x_{t+1}(x_t, V_t) - \frac{1}{2}x_t + V_t]$$

So equation (10) becomes

$$E[(a_t + b_t x_t + c_t V_t) - \frac{1}{2}x_t + V_t] - 2(x_t - \frac{1}{2}x_{t-1} + v_{t-1}) = 0$$

or $a_t + b_t x_t + c_t - \frac{1}{2}x_t + 1 - 2x_t + x_{t-1} - 2v_{t-1} = 0$, which gives

$$x_t = (a_t + c_t + 1 + x_{t-1} - 2v_{t-1}) / (\frac{5}{2} - b_t)$$

It follows that $(*)$ holds with $a_{t-1} = b_{t-1}(a_t + c_t + 1)$, where $b_{t-1} = 2/(5 - 2b_t)$, and with $c_{t-1} = -2b_{t-1}$. It can also be shown by backward induction that, as t decreases toward 1, so b_t increases but remains less than $\frac{1}{2}$. Moreover, the triple $(a_{T-k}, b_{T-k}, c_{T-k})$ converges rapidly to $(0, \frac{1}{2}, -1)$ as $T \rightarrow \infty$ for each fixed k . $\blacksquare$

EXAMPLE 5 Hall (1978) presents a stochastic version of the life cycle model considered in Example 12.5.1. In it, a consumer faces a random income stream V_t over the life cycle, which extends from period $t = 0$ to T . Specifically, the consumer's objective is to choose consumption c_t as a function $c_t(x_t, v_t)$ of wealth x_t and current income v_t , in order to maximize the expected discounted sum of utility, given by

$$E \left[\sum_{t=0}^{T-1} \beta^t u(C_t) + \beta^T u(X_T) \right] \quad \text{s.t.} \quad X_{t+1} = a(X_t + V_t - C_t), \quad t = 0, 1, \dots, T-1$$

with x_0 given, where $a > 0$ is the return to saving. Finally, it is assumed that the instantaneous utility function $u(C)$ is C^2 and satisfies $u'(C) > 0$, $u''(C) < 0$ for all $C > 0$.

To apply the stochastic Euler equation, we define $F(T, x_T, x_{T+1}, v_T) = \beta^T u(x_T)$ and $F(t, x_t, x_{t+1}, v_t) = \beta^t u(x_t) = \beta^t u(x_t + v_t - a^{-1}x_{t+1})$ for $t = 0, 1, \dots, T-1$. At time T , equation (9) is trivially satisfied because $F(T, \cdot)$ is independent of X_{T+1} . For $t = 1, 2, \dots, T-1$, the stochastic Euler equation (10) takes the form

$$E[\beta^t u'(x_t + v_t - a^{-1}x_{t+1}(x_t, v_t)) | v_{t-1}] - \beta^{t-1} a^{-1} u'(x_{t-1} + v_{t-1} - a^{-1}x_t) = 0 \quad (**)$$

As in the general problem (8), suppose (**) is solved backwards to find successive policy functions $x_{t+1}(x_t, v_t)$. These can be used to determine the consumption expenditure $c_t = x_t + v_t - a^{-1}x_{t+1}(x_t, v_t)$ at each time t and in each state (x_t, v_t) . But then (**) implies that $E[\beta u'(C_t) | v_{t-1}] = u'(c_{t-1})/a$ or, equivalently $E[\beta u'(C_t)/u'(c_{t-1}) | v_{t-1}] = 1/a$. Thus, at time $t-1$ the consumption level c_{t-1} must be set so that the consumer's expected marginal rate of substitution $E[\beta u'(C_t)/u'(c_{t-1}) | v_{t-1}]$ between random consumption C_t in period t and known consumption c_{t-1} in period $t-1$ is equal to $1/a$, the inverse rate of return.

There are several special cases of some note that can be solved explicitly for c_{t-1} as the conditionally expected value of some function of C_t :

Case 1, Quadratic Utility: Here $u(c) = -\frac{1}{2}(\bar{c} - c)^2$ where $\bar{c}$ is a "bliss" level of consumption at which satiation occurs. The Euler equation is $E[-\beta(\bar{c} - C_t) | v_t] = -(\bar{c} - c_{t-1})/a$ and so $c_{t-1} = a\beta E[C_t | v_{t-1}] - (a\beta - 1)\bar{c}$, an affine function of the conditional mean.

Case 2, Logarithmic Utility: Here $u(c) = \ln c$, and the Euler equation is $E[\beta/C_t | v_{t-1}] = 1/(ac_{t-1})$, implying that $c_{t-1} = 1/(a\beta E[1/C_t | v_{t-1}])$, proportional to the conditional harmonic mean.

Case 3, Negative Exponential Utility: Here $u(c) = -e^{-\alpha c}$ where $\alpha > 0$. The Euler equation is $E[\beta e^{-\alpha C_t} | v_{t-1}] = e^{-\alpha c_{t-1}}/a$, whose solution is $c_{t-1} = -\ln(a\beta E[e^{-\alpha C_t} | v_{t-1}])/\alpha$.

Case 4, Isoelastic Utility: Here $u(c) = c^{1-\rho}/(1-\rho)$ where $\rho > 0$ and $\rho \neq 1$. (The elasticity of marginal utility, which is $cu''(c)/u'(c)$, is equal to the constant $-\rho$.) Here the Euler equation is $E[\beta C_t^{-\rho} | v_{t-1}] = c_{t-1}^{-\rho}/a$, implying that $c_{t-1} = (a\beta E[C_t^{-\rho} | v_{t-1}])^{-1/\rho}$.

PROBLEMS FOR SECTION 12.6

1. Consider the stochastic dynamic programming problem

$$\max E \left[-\delta \exp(-\gamma X_T) + \sum_{t=0}^{T-1} -\exp(-\gamma u_t) \right], \quad X_{t+1} = 2X_t - u_t + V_{t+1}, \quad x_0 \text{ given}$$

where u_t are controls taking values anywhere in $\mathbb{R}$, and δ, γ are positive constants. Here the random variables V_{t+1} ($t = 0, 1, 2, \dots, T-1$) are identically and independently distributed. Moreover, suppose that $K = E[\exp(-\gamma V_{t+1})] < \infty$. Show that the optimal value function $J_t(x)$ can be written as $J_t(x) = -\alpha_t \exp(-\gamma x)$, and then find a backward difference equation for α_t . What is α_T ?

2. (Blanchard and Fischer (1989)). Solve the problem

$$\max E \left[\sum_{t=0}^{T-1} (1+\theta)^{-t} \ln C_t + k(1+\theta)^{-T} \ln A_T \right]$$

where w_t and C_t are controls, k and θ are positive constants, subject to

$$A_{t+1} = (A_t - C_t)[(1+r_t)w_t + (1+V_{t+1})(1-w_t)], \quad A_0 \text{ fixed}$$

where r_t is a given sequence of interest rates on a safe asset, and the random returns V_t to a risky asset are independently and identically distributed.

3. Solve the problem: $\max E \left[\sum_{t=0}^{T-1} 2u_t^{1/2} + aX_T \right]$ subject to $u_t \geq 0$, with x_0 fixed and positive, where $a > 0$ and $X_{t+1} = X_t - u_t$ with probability $1/2$, $X_{t+1} = 0$ with probability $1/2$.

4. Solve the problem

$$\max E \left[\sum_{t=0}^{T-1} -u_t^2 - X_T^2 \right] \quad \text{subject to} \quad X_{t+1} = X_t V_{t+1} + u_t, \quad x_0 \text{ fixed}$$

where $V_{t+1} \in \{0, 1\}$ with $\Pr[V_{t+1} = 1 | V_t = 1] = 3/4$, $\Pr[V_{t+1} = 1 | V_t = 0] = 1/4$. (Hint: Try $J_t(x_t, 1) = -a_t x_t^2$ and $J_t(x_t, 0) = -b_t x_t^2$.)

5. Consider the problem

$$\max E \left[\sum_{t=0}^{T-1} ((1-u_t)X_t^2 - u_t) + 2X_T^2 \right] \quad \text{s. t.} \quad X_{t+1} = u_t X_t V_{t+1}, \quad u_t \in U = [0, 1]$$

where $V_{t+1} = 2$ with probability $1/4$ and $V_{t+1} = 0$ with probability $3/4$. Find $J_T(x)$, $J_{T-1}(x)$, and $J_{T-2}(x)$. (Note that the maximand will be convex in the control u , so any maximum will be attained at an endpoint of U .) Then find $J_t(x)$ for general t .

6. Solve the problem

$$\max E \left[\sum_{t=0}^{T-1} u_t^{1/2} + aX_T^{1/2} \right] \quad \text{subject to} \quad X_{t+1} = (X_t - u_t)V_{t+1}$$

with a a given positive number, where $V_{t+1} = 0$ with probability $1/2$, and $V_{t+1} = 1$ with probability $1/2$. (Hint: Try $J_t(x) = 2a_t x^{1/2}$, where $a_t > 0$.)

7. Solve the problem in Example 3 when $f(T, x_T) = (x_T)^{1-\alpha}/(1-\alpha)$, where $\alpha > 0$ with $\alpha \neq 1$. What happens as $\alpha \rightarrow 0$?

8. Use the stochastic Euler equation to solve the problem

$$\max E \left[\sum_{t=0}^2 [1 - (v_{t+1} + X_{t+1} - X_t)^2] + (1 + v_3 + X_3) \right], \quad X_0 = 0, \quad X_1, X_2, X_3 \in \mathbb{R}$$

where the random variables V_t ($t = 0, 1, 2$) are identically and independently distributed, with $E[V_t] = 1/2$.

12.7 Infinite Horizon Stationary Problems

We consider an infinite horizon version of the problem given by (12.6.1) in the previous section. As in the corresponding deterministic infinite horizon problem (12.3.1), however, we assume that the problem is **stationary** or **autonomous**. Specifically, both the conditional probabilities $P(\mathbf{v}_{t+1} | \mathbf{v}_t)$ (or densities $p(\mathbf{v}_{t+1} | \mathbf{v}_t)$) and the transition function $\mathbf{g}(\mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_{t+1})$ are independent of t , whereas the instantaneous reward is $\beta^t f(\mathbf{x}_t, \mathbf{u}_t)$ with $\beta \in (0, 1)$. In the discrete random variable case, the problem takes the form

$$\max_{\pi} E \left[\sum_{t=0}^{\infty} \beta^t f(\mathbf{X}_t, \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t)) \right], \quad \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t) \in U, \quad \Pr[\mathbf{V}_{t+1} = \mathbf{v} | \mathbf{v}_t] = P(\mathbf{v} | \mathbf{v}_t) \quad (1)$$

where $\mathbf{X}_t$ is governed by the stochastic difference equation

$$\mathbf{X}_{t+1} = \mathbf{g}(\mathbf{X}_t, \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t), \mathbf{V}_{t+1}) \quad (2)$$

with $\mathbf{X}_0$ and $\mathbf{V}_0$ given. It is assumed that the functions f and $\mathbf{g}$ are continuous, and that the control functions $\mathbf{u}_t$ take values in a fixed control region U . Among all sequences $\pi = (\mathbf{u}_0(\mathbf{x}_0, \mathbf{v}_0), \mathbf{u}_1(\mathbf{x}_1, \mathbf{v}_1), \dots)$ of Markov controls, we seek one that maximizes the objective function in (1).

We introduce the following *boundedness condition*, as in (12.3.2):

$$M_1 \leq f(\mathbf{x}, \mathbf{u}) \leq M_2 \quad \text{for all } (\mathbf{x}, \mathbf{u}) \in \mathbb{R}^n \times U \quad (3)$$

where M_1 and M_2 are given real numbers. For each admissible control sequence $\pi = (\mathbf{u}_0(\mathbf{x}_0, \mathbf{v}_0), \mathbf{u}_1(\mathbf{x}_1, \mathbf{v}_1), \dots)$, starting time s , and state pair $(\mathbf{x}_s, \mathbf{v}_s)$, let us write

$$W^s(\mathbf{x}_s, \mathbf{v}_s, \pi) = E \left[\sum_{t=s}^{\infty} \beta^t f(\mathbf{X}_t, \mathbf{u}_t(\mathbf{X}_t, \mathbf{V}_t)) \mid \mathbf{x}_s, \mathbf{v}_s \right] \quad (4)$$

where $\mathbf{X}_t$ is the process corresponding to π , starting at $\mathbf{x}_s$ as specified by (2). Define

$$J^s(\mathbf{x}_s, \mathbf{v}_s) = \sup_{\pi} W^s(\mathbf{x}_s, \mathbf{v}_s, \pi)$$

We now claim that

$$J^0(\mathbf{x}, \mathbf{v}) = J^s(\mathbf{x}, \mathbf{v}) \quad (5)$$

The intuitive argument for this is just the same as it was for (12.3.5) in the earlier deterministic case: because time does not enter explicitly in $P(\mathbf{v} | \mathbf{v}_t)$, $\mathbf{g}$, or f , the future looks exactly the same starting in state $(\mathbf{x}, \mathbf{v})$ at time $t = s$ as it does starting in the same state at time $t = 0$. The obvious implication of (5) is that we can define the **optimal value function** $J(\mathbf{x}, \mathbf{v})$ as the common value of $J^s(\mathbf{x}, \mathbf{v})$ for all $s = 0, 1, 2, \dots$. Replacing both J_t and J_{t+1} in (12.6.5) by J , we derive the following **optimality equation** or **Bellman equation**

$$J(\mathbf{x}, \mathbf{v}) = \max_{\mathbf{u}} \{ f(\mathbf{x}, \mathbf{u}) + \beta E[J(\mathbf{X}_1, \mathbf{V}_1) \mid \mathbf{x}, \mathbf{v}] \} \quad \text{where } \mathbf{X}_1 = \mathbf{g}(\mathbf{x}, \mathbf{u}, \mathbf{V}_1) \quad (6)$$

Like (12.3.8), this is a “functional equation” that (we hope) determines the unknown function J that occurs on both sides of the equality sign. Once J is known, for each possible state pair $(\mathbf{x}, \mathbf{v})$ an optimal Markov control $\mathbf{u}(\mathbf{x}, \mathbf{v})$ is obtained from the maximization in (6).

Note especially that $\mathbf{u}(\mathbf{x}, \mathbf{v})$ does not depend on t . This is to be expected. Whether we observe $(\mathbf{x}, \mathbf{v})$ at time 0 or at time t does not matter; the optimal choice of $\mathbf{u}$ should be the same in each case because the future looks exactly the same. A policy like this that is independent of t is said to be a **stationary optimum**.

When the boundedness condition (3) is satisfied, the same argument as in Note 12.3.4 shows that the optimal value function is defined and satisfies the optimality equation. Moreover, the optimality equation has a unique bounded solution $J(\mathbf{x}, \mathbf{v})$. (At least this is so when “max” is replaced by “sup” in the Bellman equation, as it was in (12.3.11) for the corresponding deterministic problem.) Furthermore, $J(\mathbf{x}, \mathbf{v})$ is automatically the optimal value function in the problem, and given this function, any control function $\mathbf{u}(\mathbf{x}, \mathbf{v})$ that maximizes the right-hand side of (6) is a stationary optimum.

NOTE 1 The boundedness condition (3), or the alternatives in Note 3 below, need only hold for all $\mathbf{x}$ in $\mathcal{X}(\mathbf{x}_0) = \bigcup_t \mathcal{X}_t(\mathbf{x}_0)$, where $\mathcal{X}_t(\mathbf{x}_0)$ for all t denotes the set of states that can be reached at time t when starting at $\mathbf{x}_0$ at time 0, considering all controls and all outcomes that can occur with positive probability (in the discrete random variable case we are considering). Furthermore, the feasible set of controls can be a set $U(\mathbf{x})$ that depends on the state $\mathbf{x}$.

NOTE 2 The conclusions drawn above for the case when the boundedness condition (3) is satisfied are also valid if the following weaker condition holds: there exist positive constants M , M^* , α , and δ with $\beta\delta^\alpha < 1$ such that for all $\mathbf{x} \in \mathcal{X}(\mathbf{x}_0)$ and $\mathbf{u} \in U$, one has $|f(\mathbf{x}, \mathbf{u})| \leq M^*(1 + \|\mathbf{x}\|^\alpha)$ and $\|\mathbf{g}(\mathbf{x}, \mathbf{u}, \mathbf{v})\| \leq M + \delta\|\mathbf{x}\|$.

NOTE 3 (Alternative boundedness conditions) Complications arise when the boundedness condition (3) fails. First, the Bellman equation might then have more than one solution, or perhaps none. Even if it has one or more solutions, it is possible that none of them is the optimal value function. There are nevertheless two cases where some results can be obtained. In both cases we must allow infinite values for the optimal value function. (Of course, both $\hat{J}(\mathbf{x}, \mathbf{v}) \equiv \infty$ and $\hat{J}(\mathbf{x}, \mathbf{v}) \equiv -\infty$ satisfy the Bellman equation in a sense, though they may well be “false” solutions that fail to correspond to an optimal policy.) Both cases include a subcase where $\beta = 1$.⁶

Throughout the rest of this note, let $(\hat{J}(\mathbf{x}, \mathbf{v}), \hat{\mathbf{u}}(\mathbf{x}, \mathbf{v}))$ be a pair satisfying the Bellman equation (6) (so $\hat{\mathbf{u}}(\mathbf{x}, \mathbf{v})$ yields the maximum in the equation, given $\hat{J}$). Also let $J(\mathbf{x}, \mathbf{v})$ denote the optimal value function. Finally, let $J^a(\mathbf{x}, \mathbf{v})$ denote the value of the objective derived from using the stationary policy $\mathbf{u}(\mathbf{x}, \mathbf{v})$ all the time. In finding $J^a(\mathbf{x}, \mathbf{v})$, it is sometimes useful to know that it is the limit as $T \rightarrow \infty$ of the value $J^a(0, \mathbf{x}, \mathbf{v}, T) = E \left[\sum_{t=0}^T \beta^t f(\mathbf{X}_t, \mathbf{u}(\mathbf{X}_t, \mathbf{V}_t)) \right]$ derived from using the stationary policy $\mathbf{u}(\mathbf{x}, \mathbf{v})$ all the time from $t = 0$ until $t = T$.

⁶ For the results in this note see Bertsekas (1976) and Hernández-Lerma and Lasserre (1996).

Case A: There exists a lower bound γ such that $f(x, u) \geq \gamma$ for all $(x, u) \in \mathbb{R}^n \times U$; if $\beta = 1$, then $\gamma = 0$. In this case it is possible that $J(x, v) = +\infty$ for some, or all, (x, v) . Provided that $\hat{J} = J^{\hat{u}}$, the policy function $\hat{u}(x, v)$ is optimal.

Case B: There exists an upper bound γ such that $f(x, u) \leq \gamma$ for all $(x, u) \in \mathbb{R}^n \times U$; if $\beta = 1$, then $\gamma = 0$. In this case it is possible that $J(x, v) = -\infty$ for some, or all, (x, v) . Unlike Case A, even if $\hat{J} = J^{\hat{u}}$, the policy function $\hat{u}(x, v)$ may not be optimal. So a more complicated test is needed.

Suppose we are able to prove that the Bellman equation has a unique solution $\hat{J}$ satisfying $\hat{J}(x, v) \leq \gamma(1 - \beta)$ for all state pairs (x, v) . Then this is the optimal value function $J(x, v)$. (Recall that $J(x, v)$ is known to satisfy the Bellman equation, in both cases A and B.)

Another sufficient condition for optimality is the following: Suppose we have solved the modified problem where the upper limit of the sum in (1) is the finite horizon T instead of ∞ . Assume that U is compact, that the functions $f(x, u)$, $g(x, u, v)$ are continuous in (x, u) for each v , and that V is finite. Denote the optimal value function in this problem by $J(0, x, v, T)$. Then the limit $\lim_{T \rightarrow \infty} J(0, x, v, T)$ exists and equals the optimal value function. This is true not only in case B, but also in the cases A and (3).

To sum up, what should we do after finding a pair $(\hat{J}(x, v), \hat{u}(x, v))$ that satisfies the Bellman equation (6)? In case A, we try to check that $J^{\hat{u}} = \hat{J}$. If it is, then the optimum has indeed been found. In case B, after first checking that $\hat{J} \leq \gamma(1 - \beta)$, we try to show that either $\hat{J}$ is the unique solution of the Bellman equation, or $J(0, x, v, T) \rightarrow \hat{J}(x, v)$ as $T \rightarrow \infty$. If either of these tests is passed, then the optimum has indeed been found.

NOTE 4 (Transversality conditions) An alternative test, based on a transversality condition, is sometimes useful. Consider a pair $(\hat{J}(x, v), \hat{u}(x, v))$ satisfying the Bellman equation. Note that, in case A in Note 3, $\hat{J} \equiv J^{\hat{u}}$ automatically holds (so $\hat{u}(x, v)$ is an optimal policy) provided that, for any solution sequence X_t , $t = 0, 1, \dots$, of the stochastic difference equation (2) that starts from any given pair $(\hat{x}_0, \hat{v}_0) \in \mathcal{X}(x_0) \times V$ and follows the particular policy $u_t = \hat{u}(x_t, v_t)$ for $t = 0, 1, \dots$, one has $\beta^t E[\hat{J}(X_t, V_t) | \hat{x}_0, \hat{v}_0] \rightarrow 0$ as $t \rightarrow \infty$. In case B in Note 3, $\hat{J}$ is the optimal value function (so $\hat{u}(x, v)$ is optimal) provided that the same transversality condition holds even for arbitrary choices of the controls $u_t(x, v)$ at each time t .

EXAMPLE 1 Consider the following stochastic version of Example 12.3.1:

$$\max_{w \in (0, 1)} E \left[\sum_{t=0}^{\infty} \beta^t (w_t X_t)^{1-\gamma} \right] \quad (i)$$

$$X_{t+1} = V_{t+1}(1 - w_t)X_t, \quad x_0 \text{ is a positive constant, } 0 < \gamma < 1 \quad (ii)$$

Here $w \in (0, 1)$ is the control, whereas $V_1, V_2, \dots$ are identically and independently distributed nonnegative stochastic variables. Define $D = E[V^{1-\gamma}]$, where V denotes any of the V_t . It is assumed that

$$\beta \in (0, 1), \quad \gamma \in (0, 1), \quad D < \infty, \quad \rho = (\beta D)^{1/\gamma} < 1 \quad (iii)$$

In the notation of problem (1)–(3), $f(x, w) = (wx)^{1-\gamma}$ and $g(x, w, V) = V(1 - w)x$. The optimality equation (6) yields

$$J(x) = \max_{w \in (0, 1)} [(wx)^{1-\gamma} + \beta E[J(V(1 - w)x)]] \quad (iv)$$

We guess that $J(x)$ has the form $J(x) = kx^{1-\gamma}$ for some constant k . (After all, the optimal value function had a similar form in the finite horizon version of this problem discussed in the previous section, as well as in the deterministic infinite horizon version of Example 12.3.1.) Then, cancelling the factor $x^{1-\gamma}$, (iv) reduces to

$$k = \max_{w \in (0, 1)} [w^{1-\gamma} + \beta k D (1 - w)^{1-\gamma}] \quad (v)$$

where $D = E[V^{1-\gamma}]$. Note that equation (v) is the same as equation (iii) in Example 12.3.1, except that $\alpha^{1-\gamma}$ is replaced by D . It follows that $k = (1 - \rho)^{-\gamma}$ and $J(x) = (1 - \rho)^{-\gamma} x^{1-\gamma}$, with $w = 1 - \rho$ as the optimal policy choice.

This example does not satisfy the boundedness condition (3) for $x \in \bigcup_i \mathcal{X}_i(x_0)$. But $f(x, w) = (wx)^{1-\gamma} \geq 0$ for all $x \geq 0$ and $w \geq 0$, so we invoke boundedness condition A in Note 3 above. We need to check that $J^w(x) = J(x)$ when $w = 1 - \rho$. It would be fairly easy to calculate $J^w(x)$ directly by taking the expectation inside the sum in the objective and summing the resulting geometric series. But there is no need to do this. Instead, it is evident that $X_t = x_0 \rho^t Z_1 \cdots Z_t$, so we must have $J^w(x_0) = k x_0^{1-\gamma}$ for some constant $k > 0$. Now $J^w(x_0)$ must also satisfy the Bellman equation in the problem where the set U is reduced to the single point $\{w\}$. But the only value of k that satisfies this equation is $k = (1 - \rho)^{-\gamma}$, as found above. Thus $J^w(x) = J(x)$ when $w = 1 - \rho$, so the specified policy w really is optimal. $\blacksquare$

Counterexamples

Two examples will be given. In the first, even though boundedness condition (3) is satisfied, the Bellman equation may still have a “false” solution (J, u) that fails to satisfy $J = J^u$.

EXAMPLE 2 Suppose that $\beta \in (0, 1)$, and consider the problem

$$\max \sum_{t=0}^{\infty} \beta^t (1 - u_t) \quad \text{subject to} \quad x_{t+1} = (1/\beta)(x_t + u_t), \quad u_t \in [0, 1] \quad x_0 > 0 \text{ given}$$

We show that the Bellman equation is satisfied by $J(x) = \gamma + x$, where $\gamma = 1/(1 - \beta)$, and that any $u = \hat{u} \in [0, 1]$ yields the maximum. Indeed, with $J(x) = \gamma + x$ the right-hand side of the Bellman equation becomes

$$\max_u [1 - u + \beta(\gamma + (1/\beta)(x + u))] = 1 - \beta\gamma + x$$

independent of u . Hence, $J(x) = \gamma + x$ solves the Bellman equation provided that $\gamma = 1 + \beta\gamma$, so $J(x) = (1 - \beta)^{-1} + x$.

But then is $u_t \equiv \hat{u} \equiv 1/2$, for instance, really the optimal control, and is $J(x) = (1 - \beta)^{-1} + x$ the optimal value function? In fact they are not. The first thing to note is that $J^u(x)$ is independent of x , so $J^u(x) \neq (1 - \beta)^{-1} + x$. It is evident that $u_t \equiv 0$ is optimal, with a criterion value $1/(1 - \beta)$ that is independent of x_0 , and twice as large as the criterion value of $u_t \equiv 1/2$. $\blacksquare$

In the next example boundedness condition B in Note 3 is satisfied, but the Bellman equation may still have a "false" solution (J, u) even though $J = J^u$.

EXAMPLE 3 Consider the problem

$$\max \sum_{t=0}^{\infty} \beta^t x_t (u_t - \alpha) \quad \text{subject to} \quad x_{t+1} = x_t u_t, \quad u_t \in U = [0, \alpha], \quad x_0 > 0 \text{ given}$$

where α, β are positive constants satisfying $\alpha\beta = 1$ and $\beta \in (0, 1]$. Note first that, regardless of which $u_t \in [0, \alpha]$ is chosen in each period, one has $x_t \geq 0$ for all t , so $\mathcal{X}(x_0) \subseteq \mathbb{R}_+$. Also $f(x, u) = x(u - \alpha) \leq 0$ for all $(x, u) \in \mathbb{R}_+ \times U$, so boundedness condition B in Note 3 is satisfied.

Evidently $J^u(x) \leq 0$ for all policies $u(x)$ and all $x \geq 0$. But $J^u(x) = 0$ for all $x \geq 0$ if we choose $u(x) = \alpha$. This must therefore be the optimal policy, with $J = 0$ as the corresponding solution of the Bellman equation

$$J(x) = \max_{u \in [0, \alpha]} \{x(u - \alpha) + \beta J(xu)\}$$

Nevertheless, this equation has an alternative "false" solution of the form $J(x) = \gamma x$, where γ is a constant. The condition for this to be a solution when $x > 0$ is that

$$\gamma = \max_{u \in [0, \alpha]} \{u - \alpha + \beta \gamma u\}$$

and we see that $\gamma = -1/\beta = -\alpha$ works. In this case any $u \in [0, \alpha]$ maximizes the right-hand side of the Bellman equation. If we choose the same $u \in [0, \alpha]$ in each period, then $J^u = \sum_{t=0}^{\infty} \beta^t x_t (u - \alpha)$ where $x_t = u^t x_0$. Because $\alpha\beta = 1$, we have

$$J^u(x_0) = x_0(u - \alpha) \sum_{t=0}^{\infty} (\beta u)^t = \frac{x_0(u - \alpha)}{1 - \beta u} = -\frac{x_0}{\beta} = \gamma x_0 = J(x_0)$$

which is independent of u . So the function $J(x) = -x/\beta = \gamma x$ solves the Bellman equation, and is the criterion value $J^u(x)$ of any corresponding stationary policy $u_t = \text{constant} \in [0, \alpha]$. However, $J(x) = \gamma x \neq J^u(x)$ for the optimal policy $u_t = \alpha$.

Iterative Methods

One way of finding an approximate solution of an infinite horizon dynamic programming problem has already been mentioned: under certain conditions, $J(0, \mathbf{x}, \mathbf{v}, T) \rightarrow J(\mathbf{x}, \mathbf{v})$ as the finite horizon $T \rightarrow \infty$. In this subsection, we describe two different approximation methods. Both work within the set $\mathcal{B}$ of all real-valued bounded functions $J(\mathbf{x}, \mathbf{v})$ defined on $\mathbb{R}^n \times U$. Given any policy function $\mathbf{u} = \mathbf{u}(\mathbf{x}, \mathbf{v})$, define the operator $T^{\mathbf{u}}: \mathcal{B} \rightarrow \mathcal{B}$ (as in Note 12.3.4) so that, for any real-valued bounded function $I(\mathbf{x}, \mathbf{v})$ in $\mathcal{B}$, the transformed function $T^{\mathbf{u}}(I)$ of $(\mathbf{x}, \mathbf{v})$ satisfies

$$T^{\mathbf{u}}(I)(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}, \mathbf{u}(\mathbf{x}, \mathbf{v})) + \beta E[I(\mathbf{g}(\mathbf{x}, \mathbf{u}(\mathbf{x}, \mathbf{v}), \mathbf{V}), \mathbf{V}) | \mathbf{v}]$$

Also, let $T: \mathcal{B} \rightarrow \mathcal{B}$ be the operator defined so that, for any real-valued bounded function $I(\mathbf{x}, \mathbf{v})$ in $\mathcal{B}$, the transformed function $T^{\mathbf{u}}(I)$ of $(\mathbf{x}, \mathbf{v})$ satisfies

$$T(I)(\mathbf{x}, \mathbf{v}) = \max_{\mathbf{u}} T^{\mathbf{u}}(I)(\mathbf{x}, \mathbf{v}) = \max_{\mathbf{u} \in U} \{f(\mathbf{x}, \mathbf{u}) + \beta E[I(\mathbf{g}(\mathbf{x}, \mathbf{u}, \mathbf{V}), \mathbf{V}) | \mathbf{v}]\}$$

The first **successive approximation method** can be formulated as follows. Starting with an arbitrary function $I_0 \in \mathcal{B}$ such as $I_0 = 0$, calculate successively $I_1 = T(I_0)$, $I_2 = T(I_1)$, and so on. For $k = 1, 2, \dots$ let the control $\mathbf{u}_k(\mathbf{x}, \mathbf{v})$ be one that yields a maximum at step k because it satisfies $I_k = T(I_{k-1}) = T^{\mathbf{u}_k}(I_{k-1})$. (We assume that all the maxima are attained.)

Provided that the boundedness condition (3) is satisfied, then as shown in Note 12.3.4 for the deterministic case, the operator T will be a contraction mapping (see Section 14.3). The constructed sequence of functions I_k will therefore converge to the unique solution of the Bellman equation $J = T(J)$, which is the maximum value function. Moreover, it follows that the controls $\mathbf{u}_k(\mathbf{x}, \mathbf{v})$ will be approximately optimal for k large.

The second approximation method is called **policy improvement**. As in Note 3, given any stationary policy $\mathbf{u}(\mathbf{x}, \mathbf{v})$, let $J^{\mathbf{u}}(\mathbf{x}, \mathbf{v})$ denote the expected value of the objective when starting from $(\mathbf{x}, \mathbf{v})$ at time 0, then using $\mathbf{u}(\mathbf{x}, \mathbf{v})$ all the time. Clearly, the boundedness condition (3) implies that $J^{\mathbf{u}} \in \mathcal{B}$. Now, instead of starting with an arbitrary function I_0 , the second method begins with an arbitrary initial stationary policy $\mathbf{u}_0$ whose value is $J^{\mathbf{u}_0}$. Also, let $\mathbf{u}_1(\mathbf{x}, \mathbf{v})$ be a control that yields the maximum when calculating $T(J^{\mathbf{u}_0})(\mathbf{x}, \mathbf{v}) = T^{\mathbf{u}_1}(J^{\mathbf{u}_0})(\mathbf{x}, \mathbf{v})$ for each $(\mathbf{x}, \mathbf{v})$. Next calculate $J^{\mathbf{u}_1}(\mathbf{x}, \mathbf{v})$, and find a control $\mathbf{u}_2(\mathbf{x}, \mathbf{v})$ that yields the maximum when calculating $T(J^{\mathbf{u}_1})(\mathbf{x}, \mathbf{v}) = T^{\mathbf{u}_2}(J^{\mathbf{u}_1})(\mathbf{x}, \mathbf{v})$ for each $(\mathbf{x}, \mathbf{v})$. Continuing in this way, we construct each control $\mathbf{u}_k(\mathbf{x}, \mathbf{v})$ recursively so that $T(J^{\mathbf{u}_k}) = T^{\mathbf{u}_{k+1}}(J^{\mathbf{u}_k})$ for $k = 0, 1, 2, \dots$

Let us now define the inequality relation $\geq$ on $\mathcal{B}$ so that, given any I, I' in $\mathcal{B}$, one has $I \geq I'$ if and only if $I(\mathbf{x}, \mathbf{v}) \geq I'(\mathbf{x}, \mathbf{v})$ for all $(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^n \times U$. We note that both operators T and $T^{\mathbf{u}}$ are **monotone** in the sense that if $I \geq I'$, then $T(I) \geq T(I')$ and $T^{\mathbf{u}}(I) \geq T^{\mathbf{u}}(I')$.

Arguing as in Note 12.3.4, we observe that each $J^{\mathbf{u}_k}(\mathbf{x}, \mathbf{v})$ satisfies the Bellman equation when the only possible choice of policy is $\mathbf{u}_k(\mathbf{x}, \mathbf{v})$. Therefore $J^{\mathbf{u}_k} = T^{\mathbf{u}_k}(J^{\mathbf{u}_k})$. But then the definition of T implies that $T(J^{\mathbf{u}_k}) \geq T^{\mathbf{u}_k}(J^{\mathbf{u}_k}) = J^{\mathbf{u}_k}$. Because the operator $T^{\mathbf{u}_k}$ is monotone, one has $T^{\mathbf{u}_{k+1}}(T(J^{\mathbf{u}_k})) \geq T^{\mathbf{u}_{k+1}}(J^{\mathbf{u}_k}) = T(J^{\mathbf{u}_k}) \geq J^{\mathbf{u}_k}$, then $T^{\mathbf{u}_{k+1}}(T^{\mathbf{u}_{k+1}}(T(J^{\mathbf{u}_k}))) \geq T^{\mathbf{u}_{k+1}}(J^{\mathbf{u}_k}) = T(J^{\mathbf{u}_k}) \geq J^{\mathbf{u}_k}$, and generally $(T^{\mathbf{u}_{k+1}})^n(T(J^{\mathbf{u}_k})) \geq J^{\mathbf{u}_k}$, where $(T^{\mathbf{u}_{k+1}})^n$ denotes the iterated operator that results from applying $T^{\mathbf{u}_{k+1}}$ iteratively n times.

Again, provided the boundedness condition (3) is satisfied, then as shown in Note 12.3.4 for the deterministic case, the operator $T^{\mathbf{u}_{k+1}}$, like T , will be a contraction mapping. Hence, the sequence $(T^{\mathbf{u}_{k+1}})^n(T(J^{\mathbf{u}_k}))$ must converge as $n \rightarrow \infty$ to a limit function $I^* \in \mathcal{B}$ which is the unique solution of $T^{\mathbf{u}_{k+1}}(I) = I$. But this unique limit must be $J^{\mathbf{u}_{k+1}}$. So the previous inequality implies that $J^{\mathbf{u}_{k+1}} \geq J^{\mathbf{u}_k}$. Therefore, $J^{\mathbf{u}_k}(\mathbf{x}, \mathbf{v})$ increases monotonically for each $(\mathbf{x}, \mathbf{v})$ when (3) holds. That is why it is called the policy improvement method, of course. Note that at each step $J^{\mathbf{u}_{k+1}}$ can be calculated approximately using the fact that $(T^{\mathbf{u}_{k+1}})^n(T(J^{\mathbf{u}_k})) \rightarrow J^{\mathbf{u}_{k+1}}$ as $n \rightarrow \infty$.

Finally, we show that $J^{\mathbf{u}_k}(\mathbf{x}, \mathbf{v}) \rightarrow J(\mathbf{x}, \mathbf{v})$ as $k \rightarrow \infty$, for all $(\mathbf{x}, \mathbf{v})$. Indeed, define $J^* = \sup_k J^{\mathbf{u}_k}(\mathbf{x}, \mathbf{v})$. Because $J^{\mathbf{u}_{k+1}} \geq J^{\mathbf{u}_k}$ and $T^{\mathbf{u}_{k+1}}$ is monotone, one has $J^{\mathbf{u}_{k+1}} = T^{\mathbf{u}_{k+1}}(J^{\mathbf{u}_{k+1}}) \geq T^{\mathbf{u}_{k+1}}(J^{\mathbf{u}_k}) = T(J^{\mathbf{u}_k})$. Monotonicity of T implies that $J^* = \sup_k J^{\mathbf{u}_{k+1}} \geq \sup_k T(J^{\mathbf{u}_k})$. By a similar argument, $J^* = \sup_k J^{\mathbf{u}_k} = \sup_k T^{\mathbf{u}_k}(J^{\mathbf{u}_k}) \leq \sup_k T(J^{\mathbf{u}_k})$. The two inequalities together imply that $J^* = \sup_k T(J^{\mathbf{u}_k})$. But monotonicity of T also gives $\sup_k T(J^{\mathbf{u}_k}) = T(\sup_k J^{\mathbf{u}_k}) = T(J^*)$. Therefore, J^* solves the Bellman equation

$J^* = T(J^*)$. But boundedness condition (3) implies that the unique solution of the Bellman equation is $J = J^0$, the value of an optimal policy $\hat{u}$. Hence, $J^* = J = J^0$, so $J^k \rightarrow J$ as $k \rightarrow \infty$.

PROBLEMS FOR SECTION 12.7

SM 1. Consider the problem

$$\max E \left[\sum_{t=0}^{\infty} \beta^t (-u_t^2 - X_t^2) \right], \quad \beta \in (0, 1), \quad u_t \in \mathbb{R}$$

$$X_{t+1} = X_t + u_t + V_t, \quad E[V_t] = 0, \quad E[V_t^2] = d$$

- Guess that $J(x)$ is of the form $ax^2 + b$, and insert it into the Bellman equation (6) to determine a and b .
- Solve the corresponding finite horizon problem assuming $J(t, x) = \beta^t (a_t x^2 + b_t)$. (We now sum only up to time T .) Find $J(0, x_0, T)$, let $T \rightarrow \infty$, and prove that $J(0, x_0, \infty) = J(x)$ (we are in case B of Note 3).

SM 2. Solve the problem

$$\max E \left[\sum_{t=0}^{\infty} \alpha^t (\ln u_t + \ln X_t) \right], \quad X_{t+1} = (X_t - u_t)V_{t+1}, \quad x_0 > 0, \quad u_t \in (0, x_t)$$

where $\alpha \in (0, 1)$, $V_t > 0$, and all the V_t are independent and identically distributed with $E[\ln V_t] < \infty$.

13

TOPOLOGY AND SEPARATION

We could, of course, dismiss the rigorous proof as being superfluous: if a theorem is geometrically obvious why prove it? This was exactly the attitude taken in the eighteenth century. The result, in the nineteenth century, was chaos and confusion: for intuition, unsupported by logic, habitually assumes that everything is much nicer behaved than it really is.

—I. Stewart (1975)

This chapter concentrates on a few theoretical topics that turn out to be useful in some parts of economics, notably general equilibrium theory and its applications to modern macroeconomics.

Section 13.1 takes a closer look at open and closed sets in $\mathbb{R}^n$, together with a number of closely associated concepts. Next, Sections 13.2 and 13.3 cover convergence, compactness, and continuity in $\mathbb{R}^n$. These concepts play an important part in mathematical analysis. Their systematic study belongs to *general or analytic topology*, an important branch of mathematics that saw a period of rapid development early in the 20th century. The precise definitions and carefully formulated arguments we provide may strike many readers as rather formal. Their primary purpose is less to provide methods of solving concrete problems than to equip the reader with the theoretical basis needed to understand why solutions may not even exist, as well as their regularity properties when they do exist. In the case of optimization problems, these ideas lead to the versions of the maximum theorem that are the subject of Section 13.4.

Section 13.5 introduces some concepts and results on convex sets, supplementing the material in Section 2.2. Separation theorems, which are useful in both general equilibrium theory and optimization theory, are discussed in Section 13.6. Section 13.7 on “productive economies” and the Perron-Frobenius root of a nonnegative square matrix concludes the chapter.

13.1 Point Set Topology in $\mathbb{R}^n$

This section begins by reviewing some basic facts concerning the n -dimensional Euclidean space $\mathbb{R}^n$, whose elements, or points, are n -vectors $\mathbf{x} = (x_1, \dots, x_n)$. The **Euclidean distance** $d(\mathbf{x}, \mathbf{y})$ between any two points $\mathbf{x} = (x_1, \dots, x_n)$ and $\mathbf{y} = (y_1, \dots, y_n)$ in $\mathbb{R}^n$ is the norm $\|\mathbf{x} - \mathbf{y}\|$ of the vector difference between $\mathbf{x}$ and $\mathbf{y}$. (See (1.1.37).) Thus,

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2} \quad (1)$$

Note that $d(\mathbf{x}, \mathbf{y}) \geq \sqrt{(x_j - y_j)^2} = |x_j - y_j|$ for each j and that $d(\mathbf{x}, \mathbf{y}) \leq \sum_{j=1}^n |x_j - y_j|$. (See Problem 3.) Moreover, if $\mathbf{x}, \mathbf{y}$, and $\mathbf{z}$ are points in $\mathbb{R}^n$, then

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \quad (\text{triangle inequality}) \quad (2)$$

which follows immediately from (1.1.39).

Recall that if $\mathbf{a}$ is a point in $\mathbb{R}^n$ and r is a positive number, then the set of all points $\mathbf{x}$ in $\mathbb{R}^n$ whose distance from $\mathbf{a}$ is less than r is called the **open ball** around $\mathbf{a}$ with radius r . This open ball is denoted by $B_r(\mathbf{a})$ or $B(\mathbf{a}; r)$. Thus,

$$B_r(\mathbf{a}) = B(\mathbf{a}; r) = \{\mathbf{x} \in \mathbb{R}^n : d(\mathbf{x}, \mathbf{a}) < r\} \quad (3)$$

On the real line $\mathbb{R} = \mathbb{R}^1$, with $\mathbf{a} = a_1$, the set $B_r(\mathbf{a})$ is the open interval $(a_1 - r, a_1 + r)$. If $n = 2$, then $B_r(\mathbf{a})$ is an open disk in the plane. In three-dimensional space $\mathbb{R}^3$, $B_r(\mathbf{a})$ is the set of all points strictly inside the surface of a sphere with centre $\mathbf{a}$ and radius r , as indicated in Fig. 1. (Points on a dashed curve do not belong to the set.) For $n > 3$, the open ball $B_r(\mathbf{a})$ is the set of all points strictly inside the surface of the *hypersphere* of points $\mathbf{x}$ satisfying $d(\mathbf{x}, \mathbf{a}) = r$.

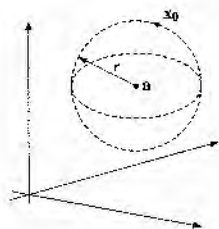


Figure 1 The open ball around $\mathbf{a}$ with radius r .

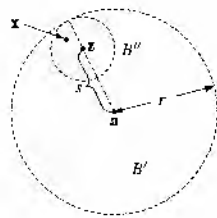


Figure 2 $B_r(\mathbf{a})$ is an open set.

Let S be any subset of $\mathbb{R}^n$. A point $\mathbf{a}$ in S is called an **interior point** of S if there is an open ball $B_r(\mathbf{a})$ centred at $\mathbf{a}$ that lies entirely within S . Thus, an interior point of S is immediately surrounded only by points of S . The set of all interior points of S is called the **interior** of S , and is denoted by $\text{int}(S)$ or S° .

A set S is called a **neighbourhood** of $\mathbf{a}$ if $\mathbf{a}$ is an interior point of S —that is, if S contains some open ball $B_r(\mathbf{a})$ around $\mathbf{a}$.

A set S in $\mathbb{R}^n$ is called **open** if all its members are interior points. On the real line $\mathbb{R}$, the simplest type of open set is an open interval.

EXAMPLE 1 Prove that any open ball $B' = B_r(\mathbf{a})$ is an open set.

Solution: Take any point $\mathbf{z}$ in B' and let $s = d(\mathbf{a}, \mathbf{z})$. Then $s < r$. Consider the open ball $B'' = B_{r-s}(\mathbf{z})$ with centre $\mathbf{z}$ and radius $r - s$. (See Fig. 2, which illustrates the proof when $n = 2$.) For any point $\mathbf{x} \in B''$, the triangle inequality implies that

$$d(\mathbf{a}, \mathbf{x}) \leq d(\mathbf{a}, \mathbf{z}) + d(\mathbf{z}, \mathbf{x}) < s + (r - s) = r$$

Hence, $\mathbf{x} \in B'$. We have proved that $B'' \subseteq B'$, which shows that B' is open. $\square$

EXAMPLE 2 Show that $A = \{(x, y) : x > y\}$ is an open set in $\mathbb{R}^2$.

Solution: Take any point (x_0, y_0) in A . Define $r = x_0 - y_0 > 0$. We claim that the open disk $B = B((x_0, y_0); r/2)$ with centre (x_0, y_0) and radius $r/2$ is contained in A , which will show that A is open.

To see this, take any point (x, y) in B . Then both $|x - x_0| < r/2$ and $|y - y_0| < r/2$. Hence $x > x_0 - r/2$ and $y < y_0 + r/2$. It follows that $x - y > x_0 - y_0 - r = 0$, and so $(x, y) \in A$ as claimed. $\square$

The interior of any set S is open. Indeed, take any $\mathbf{x}$ in $\text{int}(S)$. Then for some positive number r , the ball $B_r(\mathbf{x})$ is contained in S . Take any point $\mathbf{z}$ in $B_r(\mathbf{x})$ and choose r' so small that $B_{r'}(\mathbf{z}) \subseteq B_r(\mathbf{x})$ (see Example 1). But then $B_{r'}(\mathbf{z}) \subseteq S$, and so $\mathbf{z} \in \text{int}(S)$.

In fact, the interior of a set is its largest open subset (see Problem 10(a)). Hence, $S = \text{int}(S)$ if and only if S is open.

Some important properties of open sets are summarized in the following theorem:

THEOREM 13.1.1 (PROPERTIES OF OPEN SETS)

- (a) The whole space $\mathbb{R}^n$ and the empty set $\emptyset$ are both open.
- (b) Arbitrary unions of open sets are open.
- (c) The intersection of finitely many open sets is open.

Proof: (a) It is clear that $B_1(\mathbf{a}) \subseteq \mathbb{R}^n$ for all $\mathbf{a}$ in $\mathbb{R}^n$, so $\mathbb{R}^n$ is open. The empty set $\emptyset$ is open because there is no member of $\emptyset$ that fails to be an interior point.

(b) Let $\{U_i\}_{i \in I}$ be an arbitrary family of open sets in $\mathbb{R}^n$, and let $U^* = \bigcup_{i \in I} U_i$ be the union of the whole family. For each $\mathbf{x}$ in U^* there is at least one i in I such that $\mathbf{x} \in U_i$. Since U_i is open, there exists an open ball $B_r(\mathbf{x})$ with centre $\mathbf{x}$ such that $B_r(\mathbf{x}) \subseteq U_i \subseteq U^*$. Hence, $\mathbf{x}$ is an interior point of U^* . This shows that U^* is open.

(c) Let $\{U_i\}_{i=1}^m$ be a finite collection of open sets in $\mathbb{R}^n$, and let $U_* = \bigcap_{i=1}^m U_i$ be the intersection of all these sets. Let $\mathbf{x}$ be any point of U_* . Then for each $i = 1, \dots, m$, the point $\mathbf{x}$ belongs to U_i , and because U_i is open, there exists an open ball $B_{r_i}(\mathbf{x}) \subseteq U_i$ with centre $\mathbf{x}$ and radius $r_i > 0$ such that $B_{r_i}(\mathbf{x}) \subseteq U_i$. Let $B_* = B_r(\mathbf{x})$, where r is the smallest of the numbers $r_1, \dots, r_m$. Then $B_* \subseteq \bigcap_{i=1}^m B_{r_i}(\mathbf{x}) \subseteq \bigcap_{i=1}^m U_i$, so $\mathbf{x} \in B_*$ implies $\mathbf{x} \in U_*$. It follows that U_* is open. $\square$

NOTE 1 The intersection of an infinite number of open sets need not be open. For instance, the intersection of the infinite family $B_{1/k}(\mathbf{0})$, $k = 1, 2, \dots$, of open balls centred at the origin $\mathbf{0}$ is the one-element set $\{\mathbf{0}\}$. Yet the set $\{\mathbf{0}\}$ is not open, because $B_r(\mathbf{0})$ is not a subset of $\{\mathbf{0}\}$ for any positive r .

Recall that the **complement** of a set $S \subseteq \mathbb{R}^n$ is the set $\complement S = \mathbb{R}^n \setminus S$ of all points in $\mathbb{R}^n$ that do not belong to S . A point $\mathbf{x}_0$ in $\mathbb{R}^n$ is called a **boundary point** of the set $S \subseteq \mathbb{R}^n$

Note that $d(\mathbf{x}, y) \geq \sqrt{(x_j - y_j)^2} = |x_j - y_j|$ for each j and that $d(\mathbf{x}, y) \leq \sum_{j=1}^n |x_j - y_j|$. (See Problem 3.) Moreover, if $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are points in $\mathbb{R}^n$, then

$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \quad (\text{triangle inequality}) \quad (2)$$

which follows immediately from (1.1.39).

Recall that if $\mathbf{a}$ is a point in $\mathbb{R}^n$ and r is a positive number, then the set of all points $\mathbf{x}$ in $\mathbb{R}^n$ whose distance from $\mathbf{a}$ is less than r is called the **open ball** around $\mathbf{a}$ with radius r . This open ball is denoted by $B_r(\mathbf{a})$ or $B(\mathbf{a}; r)$. Thus,

$$B_r(\mathbf{a}) = B(\mathbf{a}; r) = \{\mathbf{x} \in \mathbb{R}^n : d(\mathbf{x}, \mathbf{a}) < r\} \quad (3)$$

On the real line $\mathbb{R} = \mathbb{R}^1$, with $\mathbf{a} = a_1$, the set $B_r(\mathbf{a})$ is the open interval $(a_1 - r, a_1 + r)$. If $n = 2$, then $B_r(\mathbf{a})$ is an open disk in the plane. In three-dimensional space $\mathbb{R}^3$, $B_r(\mathbf{a})$ is the set of all points strictly inside the surface of a sphere with centre $\mathbf{a}$ and radius r , as indicated in Fig. 1. (Points on a dashed curve do not belong to the set.) For $n > 3$, the open ball $B_r(\mathbf{a})$ is the set of all points strictly inside the surface of the *hypersphere* of points $\mathbf{x}$ satisfying $d(\mathbf{x}, \mathbf{a}) = r$.

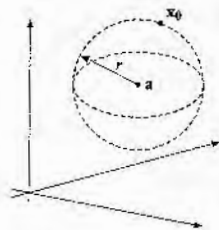


Figure 1 The open ball around $\mathbf{a}$ with radius r .

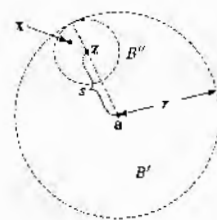


Figure 2 $B_r(\mathbf{a})$ is an open set.

Let S be any subset of $\mathbb{R}^n$. A point $\mathbf{a}$ in S is called an **interior point** of S if there is an open ball $B_r(\mathbf{a})$ centred at $\mathbf{a}$ that lies entirely within S . Thus, an interior point of S is immediately surrounded only by points of S . The set of all interior points of S is called the **interior** of S , and is denoted by $\text{int}(S)$ or S° .

A set S is called a **neighbourhood** of $\mathbf{a}$ if $\mathbf{a}$ is an interior point of S —that is, if S contains some open ball $B_r(\mathbf{a})$ around $\mathbf{a}$.

A set S in $\mathbb{R}^n$ is called **open** if all its members are interior points. On the real line $\mathbb{R}$, the simplest type of open set is an open interval.

EXAMPLE 1 Prove that any open ball $B' = B_r(\mathbf{a})$ is an open set.

Solution: Take any point $\mathbf{z}$ in B' and let $s = d(\mathbf{a}, \mathbf{z})$. Then $s < r$. Consider the open ball $B'' = B_{r-s}(\mathbf{z})$ with centre $\mathbf{z}$ and radius $r - s$. (See Fig. 2, which illustrates the proof when $n = 2$.) For any point $\mathbf{x} \in B''$, the triangle inequality implies that

$$d(\mathbf{a}, \mathbf{x}) \leq d(\mathbf{a}, \mathbf{z}) + d(\mathbf{z}, \mathbf{x}) < s + (r - s) = r$$

Hence, $\mathbf{x} \in B'$. We have proved that $B'' \subseteq B'$, which shows that B' is open. $\square$

EXAMPLE 2 Show that $A = \{(x, y) : x > y\}$ is an open set in $\mathbb{R}^2$.

Solution: Take any point (x_0, y_0) in A . Define $r = x_0 - y_0 > 0$. We claim that the open disk $B = B((x_0, y_0); r/2)$ with centre (x_0, y_0) and radius $r/2$ is contained in A , which will show that A is open.

To see this, take any point (x, y) in B . Then both $|x - x_0| < r/2$ and $|y - y_0| < r/2$. Hence $x > x_0 - r/2$ and $y < y_0 + r/2$. It follows that $x - y > x_0 - y_0 - r = 0$, and so $(x, y) \in A$ as claimed. $\square$

The interior of any set S is open. Indeed, take any $\mathbf{x}$ in $\text{int}(S)$. Then for some positive number r , the ball $B_r(\mathbf{x})$ is contained in S . Take any point $\mathbf{z}$ in $B_r(\mathbf{x})$ and choose r' so small that $B_{r'}(\mathbf{z}) \subseteq B_r(\mathbf{x})$ (see Example 1). But then $B_{r'}(\mathbf{z}) \subseteq S$, and so $\mathbf{z} \in \text{int } S$.

In fact, the interior of a set is its largest open subset (see Problem 10(a)). Hence, $S = \text{int}(S)$ if and only if S is open.

Some important properties of open sets are summarized in the following theorem:

THEOREM 13.1.1 (PROPERTIES OF OPEN SETS)

- (a) The whole space $\mathbb{R}^n$ and the empty set $\emptyset$ are both open.
- (b) Arbitrary unions of open sets are open.
- (c) The intersection of finitely many open sets is open.

Proof: (a) It is clear that $B_1(\mathbf{a}) \subseteq \mathbb{R}^n$ for all $\mathbf{a}$ in $\mathbb{R}^n$, so $\mathbb{R}^n$ is open. The empty set $\emptyset$ is open because there is no member of $\emptyset$ that fails to be an interior point.

(b) Let $\{U_i\}_{i \in I}$ be an arbitrary family of open sets in $\mathbb{R}^n$, and let $U^* = \bigcup_{i \in I} U_i$ be the union of the whole family. For each $\mathbf{x}$ in U^* there is at least one i in I such that $\mathbf{x} \in U_i$. Since U_i is open, there exists an open ball $B_r(\mathbf{x})$ with centre $\mathbf{x}$ such that $B_r(\mathbf{x}) \subseteq U_i \subseteq U^*$. Hence, $\mathbf{x}$ is an interior point of U^* . This shows that U^* is open.

(c) Let $\{U_i\}_{i=1}^m$ be a finite collection of open sets in $\mathbb{R}^n$, and let $U_* = \bigcap_{i=1}^m U_i$ be the intersection of all these sets. Let $\mathbf{x}$ be any point of U_* . Then for each $i = 1, \dots, m$, the point $\mathbf{x}$ belongs to U_i , and because U_i is open, there exists an open ball $B_{r_i}(\mathbf{x})$ with centre $\mathbf{x}$ and radius $r_i > 0$ such that $B_{r_i}(\mathbf{x}) \subseteq U_i$. Let $B_* = B_r(\mathbf{x})$, where r is the smallest of the numbers $r_1, \dots, r_m$. Then $B_* = \bigcap_{i=1}^m B_{r_i}(\mathbf{x}) \subseteq \bigcap_{i=1}^m U_i$, so $\mathbf{x} \in B_*$ implies $\mathbf{x} \in U_*$. It follows that U_* is open. $\square$

NOTE 1 The intersection of an infinite number of open sets need not be open. For instance, the intersection of the infinite family $B_{1/k}(\mathbf{0})$, $k = 1, 2, \dots$, of open balls centred at the origin $\mathbf{0}$ is the one-element set $\{\mathbf{0}\}$. Yet the set $\{\mathbf{0}\}$ is not open, because $B_r(\mathbf{0})$ is not a subset of $\{\mathbf{0}\}$ for any positive r .

Recall that the **complement** of a set $S \subseteq \mathbb{R}^n$ is the set $\complement S = \mathbb{R}^n \setminus S$ of all points in $\mathbb{R}^n$ that do not belong to S . A point $\mathbf{x}_0$ in $\mathbb{R}^n$ is called a **boundary point** of the set $S \subseteq \mathbb{R}^n$

if every open ball centred at x_0 contains at least one point in S and at least one point in $\mathbb{C}S$. Note that a boundary point of S is also a boundary point of $\mathbb{C}S$, and vice versa. For instance, the point x_0 in Fig. 1 is a boundary point of $B_r(a)$, as well as a boundary point of the complement of $B_r(a)$. In this particular case x_0 does not belong to the set. In general, a set may include none, some, or all of its boundary points. An open set, however, contains none of its boundary points.

If a point x belongs to a set S but is not an interior point of S , then every open ball centred at x intersects $\mathbb{C}S$, so x is a boundary point of S . The set of all boundary points of a set S is called the **boundary** of S and is denoted by ∂S or $\text{bd}(S)$. In Fig. 1, the boundary $\partial B_r(a)$ of the open ball $B_r(a)$ is the sphere consisting of all x with $d(x, a) = r$. The boundary of a nonempty open interval (a, b) on the real line consists of the two distinct points a and b .

These simple results imply that, given any set $S \subseteq \mathbb{R}^n$, there is a corresponding partition of $\mathbb{R}^n$ into three mutually disjoint sets (some of which may be empty), namely:

- the interior of S , which consists of all points x in $\mathbb{R}^n$ such that $B \subseteq S$ for some open ball B around x ;
- the exterior of S , which consists of all points x in $\mathbb{R}^n$ for which there exists some open ball B around x such that $B \subseteq \mathbb{R}^n \setminus S$;
- the boundary of S , which consists of all points x in $\mathbb{R}^n$ with the property that every open ball B around x intersects both S and its complement $\mathbb{R}^n \setminus S$.

A set S in $\mathbb{R}^n$ is said to be **closed** if it contains all its boundary points. The union $S \cup \partial S$ of S and its boundary is called the **closure** of S , denoted by $\bar{S}$ or $\text{cl}(S)$. A point a belongs to $\bar{S}$ if and only if every open ball $B_r(a)$ around a intersects S . The closure $\bar{S}$ of any set S is indeed closed (see Problem 9(b)). In fact, $\bar{S}$ is the smallest closed set containing S (see Problem 10(b)). It follows that S is closed if and only if $S = \bar{S}$.

We noted above that S and $\mathbb{C}S$ have the same boundary points. Furthermore, a set is open if and only if every point in the set is an interior point, i.e. if and only if it contains none of its boundary points. On the other hand, a set is closed if and only if it contains all its boundary points. It is easy to see that the following is true:

A set in $\mathbb{R}^n$ is closed if and only if its complement is open. (4)

Here are the most important properties of closed sets:

THEOREM 13.1.2 (PROPERTIES OF CLOSED SETS)

- The whole space $\mathbb{R}^n$ and the empty set $\emptyset$ are both closed.
- Arbitrary intersections of closed sets are closed.
- The union of finitely many closed sets is closed.

Proof: Part (a) is obvious. To prove (b) and (c), see Problem 12. ■

NOTE 2 Infinite unions of closed sets need not be closed. (See Problem 11.)

One should be careful to note the technical meaning of the words open and closed. In everyday usage these words are opposites. (A café is either open or closed!) In topology, however, any set containing some but not all boundary points is neither open nor closed. The half-open intervals $[a, b)$ and $(a, b]$ in $\mathbb{R}$, for example, are neither open nor closed.

Another example is indicated in Fig. 3. (We follow the usual convention that the dashed curves represent points that do not belong to S , whereas the solid curve consists of points that belong to S .) Here, a is a boundary point that belongs to S , whereas b is a boundary point that does not belong to S . The set S is neither open nor closed in $\mathbb{R}^2$. By contrast, the empty set, $\emptyset$, and the whole space, $\mathbb{R}^n$, are both open and closed. These are the only two sets in $\mathbb{R}^n$ that are both open and closed. (See Problem 14.)

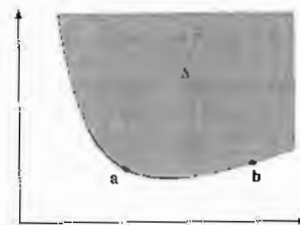


Figure 3

Economic analysis often involves sets defined in quite complicated ways. It may be difficult to see the practical relevance of knowing whether a given set includes or excludes a particular boundary point, yet such knowledge can determine which mathematical tools are applicable.

PROBLEMS FOR SECTION 13.1

- Show that if x and y are points in $\mathbb{R}^n$ such that $d(x, y) < r$, then $-r < x_j - y_j < r$ for all $j = 1, 2, \dots, n$.
- Show that if x, y , and z are points in $\mathbb{R}^n$, then $|d(z, x) - d(z, y)| \leq d(x, y)$.
- Show that if $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$, then $d(x, y) \leq \sum_{j=1}^n |x_j - y_j|$.
- The shaded areas and the curves in the figures below suggest five different sets in the plane. Which of them are open and/or closed? (Since the sets are imprecisely defined, the answer can only be rough.)



5. Sketch the set $S = \{(x, y) \in \mathbb{R}^2 : x > 0, y \geq 1/x\}$ in the plane. Is S closed?
6. (a) Let E be the subset in $\mathbb{R}^2$ consisting of the point $(0, 0)$ and all points of the form $(1/n, 1/m)$ for $n = 1, 2, \dots$ and $m = 1, 2, \dots$. Is E closed?
- (b) Let F be the subset in $\mathbb{R}^2$ defined by $F = \{(0, 0)\} \cup \{(1/n, 1/n) : n = 1, 2, \dots\}$. Is F closed?
7. Consider the following three subsets of $\mathbb{R}^2$:

$$\begin{aligned} A &= \{(x, y) : y = 1, x \in \bigcup_{n=1}^{\infty} (2n, 2n+1)\} \\ B &= \{(x, y) : y \in (0, 1), x \in \bigcup_{n=1}^{\infty} (2n, 2n+1)\} \\ C &= \{(x, y) : y = 1, x \in \bigcup_{n=1}^{\infty} [2n, 2n+1]\} \end{aligned}$$

For each of these sets determine whether it is open, closed, or neither.

8. Show that the boundary ∂S of any set S in $\mathbb{R}^n$ is closed.
9. (a) Show that, if S and T are subsets of $\mathbb{R}^n$ such that $S \subseteq T$, then
- $$\text{int}(S) \subseteq \text{int}(T) \quad \text{and} \quad \text{cl}(S) \subseteq \text{cl}(T)$$
- (b) Show that for any set S in $\mathbb{R}^n$, the closure $\text{cl}(S)$ is a closed set.
10. Let S be a subset of $\mathbb{R}^n$, and let $\mathcal{U} = \{U \subseteq \mathbb{R}^n : U \subseteq S \text{ and } U \text{ is open}\}$ be the family of all open subsets of S . Similarly, let $\mathcal{F} = \{F \subseteq \mathbb{R}^n : F \supseteq S \text{ and } F \text{ is closed}\}$ be the family of all closed supersets of S .
- (a) Show that $\text{int}(S) = \bigcup_{U \in \mathcal{U}} U$. Thus $\text{int}(S)$ is the largest open subset of S .
- (b) Show that $\text{cl}(S) = \bigcap_{F \in \mathcal{F}} F$. Thus $\text{cl}(S)$ is the smallest closed set containing S .
11. Show by an example that the union of infinitely many closed sets need not be closed. (Hint: Apply De Morgan's laws to the example in Note 1.)
12. Use De Morgan's laws (A.1.10) and the results of Theorem 13.1.1 to prove properties (b) and (c) in Theorem 13.1.2.

HARDER PROBLEMS

13. Let $\mathbb{Q}$ be the set of rational numbers. Prove that $\overline{\mathbb{Q}} = \mathbb{R}$ and $\partial \mathbb{Q} = \mathbb{R}$. What is the interior of $\mathbb{Q}$?
14. Prove that the empty set $\emptyset$ and the whole space $\mathbb{R}^n$ are the only sets in $\mathbb{R}^n$ that are both open and closed.
15. Which of the following statements are true for all subsets S and T of $\mathbb{R}^n$?
- (a) $\text{int}(\overline{S}) = \overline{\text{int}(S)}$ (b) $\overline{S \cup T} = \overline{S} \cup \overline{T}$
- (c) $\partial S \subseteq S$ (d) $S \text{ is open} \iff S \cap T \subseteq \overline{S \cap T}$

13.2 Topology and Convergence

A sequence $\{\mathbf{x}_k\} = \{\mathbf{x}_k\}_{k=1}^{\infty} = \{\mathbf{x}_k\}_k$ in $\mathbb{R}^n$ is a function that for each natural number k yields a corresponding point $\mathbf{x}_k$ in $\mathbb{R}^n$. (See Section A.3.) The point $\mathbf{x}_k$ is called the *kth term* or *kth element* of the sequence. Note that the terms of a sequence need not all be distinct.

CONVERGENCE OF A SEQUENCE

A sequence $\{\mathbf{x}_k\}$ in $\mathbb{R}^n$ **converges** to a point $\mathbf{x}$ if for each $\varepsilon > 0$ there exists a natural number N such that $\mathbf{x}_k \in B_{\varepsilon}(\mathbf{x})$ for all $k > N$, or, equivalently, if $d(\mathbf{x}_k, \mathbf{x}) \rightarrow 0$ as $k \rightarrow \infty$. (1)

In other words, each open ball around $\mathbf{x}$, however small its radius ε , must contain $\mathbf{x}_k$ for all sufficiently large k . Geometrically speaking, as k increases, the points $\mathbf{x}_k$ must eventually all become concentrated around $\mathbf{x}$. Note that $\mathbf{x}_k$ need not approach $\mathbf{x}$ from any fixed direction, and the distance $d(\mathbf{x}_k, \mathbf{x})$ need not decrease monotonically as k increases.

If $\{\mathbf{x}_k\}$ converges to $\mathbf{x}$ we write

$$\mathbf{x}_k \rightarrow \mathbf{x} \quad \text{as } k \rightarrow \infty, \quad \text{or} \quad \lim_{k \rightarrow \infty} \mathbf{x}_k = \mathbf{x}$$

and call $\mathbf{x}$ the **limit** of the sequence.

It follows from the definition of convergence that a sequence can have at most one limit (Problem 2). If a sequence is not convergent, it is **divergent**.

The definitions of limits and convergence generalize the corresponding definitions in Section A.3 for sequences of real numbers. The following result states that a sequence $\{\mathbf{x}_k\}$ in $\mathbb{R}^n$ will converge to a vector $\mathbf{x}$ if and only if each of its n component sequences converges (in $\mathbb{R}$) to the corresponding component of $\mathbf{x}$:

THEOREM 13.2.1 (CONVERGENCE OF EACH COMPONENT)

Let $\{\mathbf{x}_k\}$ be a sequence in $\mathbb{R}^n$. Then $\{\mathbf{x}_k\}$ converges to the vector $\mathbf{x}$ in $\mathbb{R}^n$ if and only if for each $j = 1, \dots, n$, the real number sequence $\{x_k^{(j)}\}_{k=1}^{\infty}$, consisting of the j th component of each vector $\mathbf{x}_k$, converges to $x^{(j)}$, the j th component of $\mathbf{x}$.

Proof: For every k and every j one has $d(\mathbf{x}_k, \mathbf{x}) = \|\mathbf{x}_k - \mathbf{x}\| \geq |x_k^{(j)} - x^{(j)}|$ by definition (13.1.1). So if $\mathbf{x}_k \rightarrow \mathbf{x}$, then $x_k^{(j)} \rightarrow x^{(j)} = x^{(j)}$.

Suppose on the other hand that $x_k^{(j)} \rightarrow x^{(j)}$ for $j = 1, 2, \dots, n$. Then, given any $\varepsilon > 0$, for each $j = 1, \dots, n$ there exists a number N_j such that $|x_k^{(j)} - x^{(j)}| < \varepsilon/\sqrt{n}$ for all $k > N_j$. It follows that

$$d(\mathbf{x}_k, \mathbf{x}) = \sqrt{|x_k^{(1)} - x^{(1)}|^2 + \dots + |x_k^{(n)} - x^{(n)}|^2} < \sqrt{\varepsilon^2/n + \dots + \varepsilon^2/n} = \sqrt{\varepsilon^2} = \varepsilon$$

for all $k > \max\{N_1, \dots, N_n\}$. Therefore $\mathbf{x}_k \rightarrow \mathbf{x}$ as $k \rightarrow \infty$. ■

This characterization makes it easy to translate theorems about sequences of numbers into theorems about sequences in $\mathbb{R}^n$.

Let $\{x_k\}$ be a sequence in $\mathbb{R}^n$. Consider a strictly increasing sequence $k_1 < k_2 < k_3 < \dots$ of natural numbers, and let $y_j = x_{k_j}$ for $j = 1, 2, \dots$. The sequence $\{y_j\}_{j=1}^{\infty}$ is called a **subsequence** of $\{x_k\}$, and is often denoted by $\{x_{k_j}\}_{j=1}^{\infty}$. All terms of the subsequence $\{x_{k_j}\}$ are present in the original sequence $\{x_k\}_k$. (See Section A.3.)

Cauchy Sequences

Cauchy sequences of real numbers are studied in Section A.3. There is a natural generalization to $\mathbb{R}^n$.

CAUCHY SEQUENCES

A sequence $\{x_k\}$ in $\mathbb{R}^n$ is called a **Cauchy sequence** if for every $\varepsilon > 0$ there exists a number N such that $d(x_k, x_m) < \varepsilon$ for all $k > N$ and all $m > N$. (2)

The main results in Section A.3 on Cauchy sequences in $\mathbb{R}$ carry over without difficulty to sequences in $\mathbb{R}^n$. In particular:

THEOREM 13.2.2 (CAUCHY'S CONVERGENCE CRITERION)

A sequence $\{x_k\}$ in $\mathbb{R}^n$ is convergent if and only if it is a Cauchy sequence.

Proof: The proof that a convergent sequence is a Cauchy sequence is left to the reader as Problem 3. As for the converse, let $\{x_k\}$ be a Cauchy sequence in $\mathbb{R}^n$. For each $j = 1, \dots, n$, the j th component sequence $\{x_k^{(j)}\}_k$ satisfies $|x_k^{(j)} - x_m^{(j)}| \leq \|x_k - x_m\|$, and so it is a Cauchy sequence in $\mathbb{R}$. Thus, according to Theorem A.3.5, for each j the component sequence $\{x_k^{(j)}\}_k$ must converge to a limit $\bar{x}^{(j)}$ in $\mathbb{R}$. But then Theorem 13.2.1 implies that $\{x_k\}$ converges to the point $\bar{x} = (\bar{x}^{(1)}, \dots, \bar{x}^{(n)})$. ■

Convergent sequences can be used to characterize very simply the closure of any set in $\mathbb{R}^n$.

THEOREM 13.2.3 (CLOSURE AND CONVERGENCE)

- (a) For any set $S \subseteq \mathbb{R}^n$, a point a in $\mathbb{R}^n$ belongs to $\bar{S}$ if and only if a is the limit of a sequence $\{x_k\}$ in S .
- (b) A set $S \subseteq \mathbb{R}^n$ is closed if and only if every convergent sequence of points in S has its limit in S .

Proof: (a) Let $a \in \bar{S}$. For each natural number k , the open ball $B(a; 1/k)$ must intersect S , so we can choose an x_k in $B(a; 1/k) \cap S$. Then $x_k \rightarrow a$ as $k \rightarrow \infty$.

On the other hand, assume that $a = \lim_{k \rightarrow \infty} x_k$ for some sequence $\{x_k\}$ in S . We claim that $a \in \bar{S}$. Indeed, for any $r > 0$, we know that $x_k \in B(a; r)$ for all large enough k . Since x_k also belongs to S , it follows that $B(a; r) \cap S \neq \emptyset$. Hence, $a \in \bar{S}$.

(b) Assume that S is closed, and let $\{x_k\}$ be a convergent sequence such that $x_k \in S$ for all k . By part (a), $x = \lim_{k \rightarrow \infty} x_k$ belongs to $\bar{S} = S$.

Conversely, suppose that every convergent sequence of points from S has its limit in S . Let a be a point in $\bar{S}$. By (a), $a = \lim_{k \rightarrow \infty} x_k$ for some sequence x_k in S , and therefore $a \in S$, by hypothesis. This shows that $\bar{S} \subseteq S$, hence S is closed. ■

Boundedness in $\mathbb{R}^n$

A set S in $\mathbb{R}^n$ is **bounded** if there exists a number M such that $\|x\| \leq M$ for all x in S . In other words, no point of S is at a distance greater than M from the origin. A set that is not bounded is called **unbounded**. Similarly, a sequence $\{x_k\}$ in $\mathbb{R}^n$ is **bounded** if the set $\{x_k : k = 1, 2, \dots\}$ is bounded.

It is easy to see that any convergent sequence is bounded. For if $x_k \rightarrow x$, then only finitely many terms of the sequence can lie outside the ball $B(x; 1)$. This ball is bounded and any finite set of points is bounded, so the sequence $\{x_k\}$ must be bounded. On the other hand, a bounded sequence $\{x_k\}$ in $\mathbb{R}^n$ is not necessarily convergent. In fact, a bounded sequence $\{x_k\}$ may well "jump around" and not converge to any point. A one-dimensional example in the line $\mathbb{R}$ is the bounded sequence $x_k = (-1)^k$, which has no limit.

Suppose $\{x_k\}$ is an arbitrary sequence in a bounded subset S of $\mathbb{R}^n$. Even though $\{x_k\}$ is not necessarily convergent, we now show that it must contain a convergent subsequence.

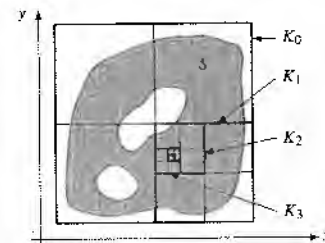


Figure 1

Consider first the case where $n = 2$, so that S is a bounded set in the plane, as illustrated in Fig. 1. Then there exists a square K_0 so large that S is contained in K_0 . Let L denote the length of each side of K_0 . The infinitely many terms of $\{x_k\}$ then all lie in K_0 . Divide K_0 into four equal squares, each of which has sides of length $L/2$. At least one of these four squares, say K_1 , must contain x_k for infinitely many k . Pick one of these terms, say x_{k_1} . Next, divide K_1 into four equal squares, each of which has sides of length $L/4$. In at least one of them, say K_2 , there will still be an infinite number of terms from the sequence. Take one of them, x_{k_2} , with $k_2 > k_1$. Continue in this way, dividing each successive square into smaller and smaller "subsquares" $K_2, K_3, \dots$, as indicated in Fig. 1, while also obtaining a subsequence $\{x_{k_j}\}$ of $\{x_k\}$ with x_{k_j} in K_j for $j = 1, 2, \dots$. It seems intuitively obvious that this subsequence converges to a unique point which is the intersection of all the squares K_j .

For the general case where n may not equal 2, suppose S is a bounded set in $\mathbb{R}^n$ and let K_0 be an n -dimensional cube containing S whose sides are all of length L . We can then divide each successive cube K_{j-1} , $j = 1, 2, \dots$, into 2^n equal parts, each having sides of length $L/2^j$, just as we divided the square into four equal parts when $n = 2$. At least one of these smaller cubes, call it K_j , will contain infinitely many terms of the sequence $\{x_k\}$. This gives us a sequence $\{K_j\}$ of cubes, each lying inside its predecessor and having sides of length $L/2^j$ and a diagonal of length $\sqrt{n}L/2^j$, $j = 1, 2, \dots$, and each containing infinitely many x_k . As in the two-dimensional case, we can find a subsequence $\{x_{k_j}\}$ of $\{x_k\}$ such that $x_{k_j} \in K_j$ for $j = 1, 2, \dots$. Then whenever $i, j \geq m$, the points x_{k_i} and x_{k_j} both belong to K_m and therefore $d(x_{k_i}, x_{k_j}) \leq \sqrt{n}L/2^m$. Hence the subsequence $\{x_{k_j}\}$ is a Cauchy sequence in $\mathbb{R}^n$, and is therefore convergent.

It is both unsurprising and easy to prove that, if S is unbounded, then there is a sequence in S without any convergent subsequence (see Problem 4). Hence, we have proved that:

THEOREM 13.2.4

A subset S of $\mathbb{R}^n$ is bounded if and only if every sequence of points in S has a convergent subsequence.

Compactness

A set S in $\mathbb{R}^n$ is called **compact** if it is both closed and bounded. An important example of a compact set in $\mathbb{R}^n$ is the **closed ball** $\bar{B}(a; r) = \{x : d(x, a) \leq r\}$ (with $r > 0$). Of course, this is the closure of the open ball $B(a; r) = \{x : d(x, a) < r\}$.

Compactness is a central concept in mathematical analysis. It also plays an important role in mathematical economics, for example when proving existence of solutions to maximization problems. Compact sets in $\mathbb{R}^n$ can be given the following very useful characterization:

THEOREM 13.2.5 (BOLZANO-WEIERSTRASS)

A subset S of $\mathbb{R}^n$ is compact (i.e. closed and bounded) if and only if every sequence of points in S has a subsequence that converges to a point in S .

Proof: Suppose S is compact and let $\{x_k\}$ be a sequence in S . By Theorem 13.2.4, $\{x_k\}$ contains a convergent subsequence. Since S is closed, it follows from Theorem 13.2.3 that the limit of the subsequence must be in S .

On the other hand, suppose that every sequence of points in S has a subsequence converging to a point of S . We must prove that S is closed and bounded. Boundedness follows from Theorem 13.2.4. To prove that S is closed, let x be any point in its closure $\bar{S}$. By Theorem 13.2.3 there is a sequence $\{x_k\}$ in S with $\lim_{k \rightarrow \infty} x_k = x$. By assumption $\{x_k\}$ has a subsequence $\{x_{k_j}\}$ that converges to a limit x' in S . But $\{x_{k_j}\}$ also converges also to x . Using the answer to Problem 2, it follows that $x = x' \in S$. ■

PROBLEMS FOR SECTION 13.2

- Find the limits of the following sequences in $\mathbb{R}^2$ if the limits exist.
 - $x_k = (1/k, 1 + 1/k)$
 - $x_k = (k, 1 + 3/k)$
 - $x_k = ((k+2)/3k, (-1)^k/2k)$
 - $x_k = (1 + 1/k, (1 + 1/k)^k)$
- Prove that a sequence in $\mathbb{R}^n$ cannot converge to more than one point.
- Prove that every convergent sequence in $\mathbb{R}^n$ is a Cauchy sequence. (Hint: See the proof of Theorem A.3.5.)
- Prove that if every sequence of points in a set S in $\mathbb{R}^n$ contains a convergent subsequence, then S is bounded. (Hint: If S is unbounded, then for each natural number k there exists an x_k in S with $\|x_k\| > k$.)
- Let $\{x_k\}$ be a sequence of points in a compact subset X of $\mathbb{R}^n$. Prove that if every convergent subsequence of $\{x_k\}$ has the same limit x^0 , then $\{x_k\}$ converges to x^0 .
- Show that if A and B are compact subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, then the Cartesian product $A \times B$ is a compact subset of $\mathbb{R}^{m+n}$ (when we identify $\mathbb{R}^m \times \mathbb{R}^n$ with $\mathbb{R}^{m+n}$ in the obvious way).

13.3 Continuous Functions

Section 2.9 dealt with some properties of transformations (i.e. vector-valued functions) from $\mathbb{R}^n$ to $\mathbb{R}^m$. In particular, the notion of differentiability was introduced. This section takes a closer look at continuous transformations. (Logically, it should really precede Section 2.9.)

Consider first a real-valued function $z = f(x) = f(x_1, \dots, x_n)$ of n variables. Roughly speaking, f is continuous if small changes in the independent variables cause only small changes in the function value. The precise “ ϵ - δ ” definition is as follows:

CONTINUITY OF REAL-VALUED FUNCTIONS

A function f with domain $S \subseteq \mathbb{R}^n$ is **continuous** at a point a in S if for every $\epsilon > 0$ there exists a $\delta > 0$ such that

$$|f(x) - f(a)| < \epsilon \text{ for all } x \text{ in } S \text{ with } \|x - a\| < \delta \quad (1)$$

If f is continuous at every point a in a set S , we say that f is continuous on S .

As in the one-variable case, we have the following useful rule:

Any function of n variables that can be constructed from continuous functions by combining the operations of addition, subtraction, multiplication, division, and composition of functions, is continuous wherever it is defined. (2)

Note that if $f(x_1, \dots, x_n) = g(x_i)$, so that f depends on x_i alone, then continuity of the function g with respect to x_i implies continuity of f with respect to $(x_1, \dots, x_n)$.

Consider next the case of transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ introduced in Section 2.7.

CONTINUITY OF TRANSFORMATIONS

A transformation (function) $\mathbf{f} = (f_1, \dots, f_m)$ from a subset S of $\mathbb{R}^n$ to $\mathbb{R}^m$ is said to be **continuous** at $\mathbf{x}^0$ in S if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}^0)) < \varepsilon$ for all $\mathbf{x}$ in S with $d(\mathbf{x}, \mathbf{x}^0) < \delta$, or equivalently, such that $\mathbf{f}(B_\delta(\mathbf{x}^0) \cap S) \subseteq B_\varepsilon(\mathbf{f}(\mathbf{x}^0))$. (3)

Intuitively, continuity of $\mathbf{f}$ at $\mathbf{x}^0$ means that $\mathbf{f}(\mathbf{x})$ is close to $\mathbf{f}(\mathbf{x}^0)$ whenever $\mathbf{x}$ is sufficiently close to $\mathbf{x}^0$. We call $\mathbf{f}$ **continuous** if it is continuous at every point in its domain.

Frequently, the easiest way to show that a transformation is continuous is to use the following condition. It can often be verified by applying (2) to each separate component.

THEOREM 13.3.1 (CONTINUITY OF EACH COMPONENT)

A function $\mathbf{f}$ from $S \subseteq \mathbb{R}^n$ to $\mathbb{R}^m$ is continuous at a point $\mathbf{x}^0$ in S if and only if each component function $f_j : S \rightarrow \mathbb{R}$, $j = 1, \dots, m$, is continuous at $\mathbf{x}^0$.

Proof: Suppose $\mathbf{f}$ is continuous at $\mathbf{x}^0$. Because $d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}^0)) \geq |f_j(\mathbf{x}) - f_j(\mathbf{x}^0)|$ for $j = 1, \dots, m$, it follows from (3) that for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $|f_j(\mathbf{x}) - f_j(\mathbf{x}^0)| < \varepsilon$ for every $\mathbf{x}$ in S with $d(\mathbf{x}, \mathbf{x}^0) < \delta$. Hence f_j is continuous at $\mathbf{x}^0$ for $j = 1, \dots, m$.

Suppose, on the other hand, that each component f_j is continuous at $\mathbf{x}^0$. Then, for every $\varepsilon > 0$ and every $j = 1, \dots, m$, there exists a $\delta_j > 0$ such that $|f_j(\mathbf{x}) - f_j(\mathbf{x}^0)| < \varepsilon/\sqrt{m}$ for every point $\mathbf{x}$ in S with $d(\mathbf{x}, \mathbf{x}^0) < \delta_j$. Let $\delta = \min\{\delta_1, \dots, \delta_m\}$. Then $\mathbf{x} \in B_\delta(\mathbf{x}^0) \cap S$ implies that

$$d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}^0)) = \sqrt{|f_1(\mathbf{x}) - f_1(\mathbf{x}^0)|^2 + \dots + |f_m(\mathbf{x}) - f_m(\mathbf{x}^0)|^2} < \sqrt{\frac{\varepsilon^2}{m} + \dots + \frac{\varepsilon^2}{m}} = \varepsilon$$

This proves that $\mathbf{f}$ is continuous at $\mathbf{x}^0$. ■

Continuity and Sequences

Continuity of a function can be characterized by means of convergent sequences. This is often the easiest way to check continuity.

THEOREM 13.3.2

A function $\mathbf{f}$ from $S \subseteq \mathbb{R}^n$ into $\mathbb{R}^m$ is continuous at a point $\mathbf{x}^0$ in S if and only if $\mathbf{f}(\mathbf{x}_k) \rightarrow \mathbf{f}(\mathbf{x}^0)$ for every sequence $\{\mathbf{x}_k\}$ of points in S that converges to $\mathbf{x}^0$.

Proof of "only if": Suppose that $\mathbf{f}$ is continuous at $\mathbf{x}^0$, and let $\{\mathbf{x}_k\}$ be any sequence in S that converges to $\mathbf{x}^0$. Let $\varepsilon > 0$ be given. Then there exists a $\delta > 0$ such that $d(\mathbf{f}(\mathbf{x}), \mathbf{f}(\mathbf{x}^0)) < \varepsilon$ whenever $\mathbf{x} \in B_\delta(\mathbf{x}^0) \cap S$. Because $\mathbf{x}_k \rightarrow \mathbf{x}^0$, there exists a number N such that $d(\mathbf{x}_k, \mathbf{x}^0) < \delta$ for all $k > N$. But then for all $k > N$ one has $\mathbf{x}_k \in B_\delta(\mathbf{x}^0) \cap S$, and so $d(\mathbf{f}(\mathbf{x}_k), \mathbf{f}(\mathbf{x}^0)) < \varepsilon$. This implies that $\{\mathbf{f}(\mathbf{x}_k)\}$ converges to $\mathbf{f}(\mathbf{x}^0)$.

The proof of the reverse implication is left to the reader as Problem 6. ■

The following property of continuous functions is much used:

THEOREM 13.3.3 (CONTINUOUS FUNCTIONS PRESERVE COMPACTNESS)

Let $S \subseteq \mathbb{R}^n$ and let $\mathbf{f} : S \rightarrow \mathbb{R}^m$ be continuous. Then $\mathbf{f}(K) = \{\mathbf{f}(\mathbf{x}) : \mathbf{x} \in K\}$ is compact for every compact subset K of S .

Proof: Let $\{y_k\}$ be any sequence in $\mathbf{f}(K)$. By definition, for each k there is a point $\mathbf{x}_k$ in K such that $y_k = \mathbf{f}(\mathbf{x}_k)$. Because K is compact, the sequence $\{\mathbf{x}_k\}$ has a subsequence $\{\mathbf{x}_{k_j}\}$ converging to a point $\mathbf{x}_0$ in K (by the Bolzano–Weierstrass theorem, Theorem 13.2.5). Because $\mathbf{f}$ is continuous, $\mathbf{f}(\mathbf{x}_{k_j}) \rightarrow \mathbf{f}(\mathbf{x}_0)$ as $j \rightarrow \infty$, where $\mathbf{f}(\mathbf{x}_0) \in \mathbf{f}(K)$ because $\mathbf{x}_0 \in K$. But then $\{y_{k_j}\}$ is a subsequence of $\{y_k\}$ that converges to a point $\mathbf{f}(\mathbf{x}_0)$ in $\mathbf{f}(K)$. So we have just proved that any sequence in $\mathbf{f}(K)$ has a subsequence converging to a point of $\mathbf{f}(K)$. By Theorem 13.2.5, it follows that $\mathbf{f}(K)$ is compact. ■

Theorem 13.3.3 can be used to prove the extreme value theorem, Theorem 3.1.3:

Proof of Theorem 3.1.3: By Theorem 13.3.3, $f(S)$ is compact. In particular, $f(S)$ is bounded, so $-\infty < a = \inf f(S)$ and $b = \sup f(S) < \infty$. Clearly, a and b are boundary points of $f(S)$. Because $f(S)$ is closed, both a and b belong to $f(S)$. Hence, there must exist points c and d in S such that $f(c) = a$ and $f(d) = b$. Obviously c is a minimum point and d is a maximum point. ■

A Characterization of Continuity

Suppose that $\mathbf{f}$ is a continuous function from (all of) $\mathbb{R}^n$ to $\mathbb{R}^m$. If V is an open set in $\mathbb{R}^n$, the image $\mathbf{f}(V) = \{\mathbf{f}(\mathbf{x}) : \mathbf{x} \in V\}$ of V need not be open in $\mathbb{R}^m$. Nor need $\mathbf{f}(C)$ be closed if C is closed. (See Problem 3.) Nevertheless, the **inverse image** (or **preimage**) $\mathbf{f}^{-1}(U) = \{\mathbf{x} : \mathbf{f}(\mathbf{x}) \in U\}$ of an open set U under a continuous function $\mathbf{f}$ is always open. Similarly, the inverse image of any closed set is closed. In fact, we have the following result:

THEOREM 13.3.4 (CHARACTERIZATION OF CONTINUITY)

Let $\mathbf{f}$ be any function from (all of) $\mathbb{R}^n$ to $\mathbb{R}^m$. Then $\mathbf{f}$ is continuous if and only if either of the following equivalent conditions is satisfied:

- (a) $\mathbf{f}^{-1}(U)$ is open for each open set U in $\mathbb{R}^m$.
- (b) $\mathbf{f}^{-1}(F)$ is closed for each closed set F in $\mathbb{R}^m$.

This theorem is a straightforward consequence of Theorem 13.3.5 below, which deals with the more general case in which f is not necessarily defined on all of $\mathbb{R}^n$.

EXAMPLE 1 In Examples 13.1.1 and 13.1.2 we used the definition of openness directly in order to prove that two particular sets were open. Such proofs become much easier once we understand how the above test can be applied.

For example, to show that the set $B' = B(x; r) = \{x : d(x, a) < r\}$ in Example 13.1.1 is open, define the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ by $f(x) = d(x, a)$. Then f is continuous (see Problem 4), $(-\infty, r)$ is open, and $B' = f^{-1}((-\infty, r))$. By Theorem 13.3.4, B' is open.

To prove that the set $A = \{(x, y) : x > y\}$ in Example 13.1.2 is open, define the continuous function g from $\mathbb{R}^2$ into $\mathbb{R}$ by $g(x, y) = x - y$. Note that $(x, y) \in A$ if and only if $g(x, y) > 0$. Hence $A = g^{-1}((0, \infty))$, and so A is open. $\square$

EXAMPLE 2 Let $U(x) = U(x_1, \dots, x_n)$ be a household's real-valued utility function, where x denotes its commodity vector and U is defined on the whole of $\mathbb{R}^n$. Recall from Example 2.2.2 that, for any real number a , the set $\Gamma_a = \{x \in \mathbb{R}^n : U(x) \geq a\}$ is an upper level set (or upper contour set) for U . If $U(x^0) = a$, then Γ_a consists of all vectors that the household values at least as much as x^0 . (See Fig. 2.2.5.)

Let F be the closed interval $[a, \infty)$. Then

$$\Gamma_a = \{x \in \mathbb{R}^n : U(x) \geq a\} = \{x \in \mathbb{R}^n : U(x) \in F\} = U^{-1}(F)$$

According to Theorem 13.3.4, if U is continuous, then the set Γ_a is closed for each value of a . Hence, *continuous functions generate closed upper level sets*. They also generate closed lower level sets, which are sets of the form $\{x \in \mathbb{R}^n : U(x) \leq a\}$.

In standard microeconomic theory the set $\{x \in \mathbb{R}^n : U(x) = a\}$ is called an **indifference surface** for U . The subset of $\mathbb{R}$ consisting of the single point $\{a\}$ is a closed set. The indifference surface corresponding to a is the set $U^{-1}(\{a\})$. We conclude that if U is continuous, then the indifference surfaces are all closed sets. $\square$

Relative Topology

Sometimes we are concerned only with a given subset S of $\mathbb{R}^n$. For example, S might be the domain of a function that, like a Cobb–Douglas production function, is not defined on the whole of $\mathbb{R}^n$ (see Example 2.5.5). Subsets of S may be open or closed relative to S in a sense that we shall now define. These definitions will be useful in giving a characterization of continuity that applies to functions whose domain is not the whole of $\mathbb{R}^n$.

Given a set S in $\mathbb{R}^n$, we define the **relative open ball** with radius r around a point a in S as $B_r^S(a) = B^S(a; r) = B(a; r) \cap S$. Once we have defined relative open balls, concepts like relative interior point, relative boundary point, relatively open set, and relatively closed set are defined in the same way as the ordinary versions of these concepts, except for the fact that $\mathbb{R}^n$ is replaced by S and balls by relative balls. Thus, given a subset A of S , a relative interior point of A is a point a in A such that $B^S(a; r) \subseteq A$ for some $r > 0$. A point a is a relative boundary point of A if $a \in S$ and all relative balls around a intersect both A and $S \setminus A$. By definition, a relatively open set consists only of relative interior points, and a relatively closed set contains all its relative boundary points. Note that $A \subseteq S$ is relatively closed in S if and only if $S \setminus A$ is relatively open in S .

Sometimes the word “relative(y)” is replaced by the expression “in the relative topology of”, e.g. “ A is open in the relative topology of S ”. Note the following result:

- (a) A is relatively open in $S \iff A = U \cap S$ for some open set U in $\mathbb{R}^n$.
 (b) A is relatively closed in $S \iff A = F \cap S$ for some closed set F in $\mathbb{R}^n$.
- (4)

Proof: To prove (a), suppose first that A is relatively open in S . By definition, for each a in A there exists a ball $B(a; r_a)$ such that $B(a; r_a) \cap S \subseteq A$. It follows that

$$A \subseteq \bigcup_{a \in A} (B(a; r_a) \cap S) \subseteq A \quad \text{and so} \quad A = \bigcup_{a \in A} (B(a; r_a) \cap S) = \left(\bigcup_{a \in A} B(a; r_a) \right) \cap S$$

Let $U = \bigcup_{a \in A} B(a; r_a)$. Then U is an open set by Theorem 13.1.1 (b), and we have just shown that $A = U \cap S$.

On the other hand, suppose U is open in $\mathbb{R}^n$ and that $A = U \cap S$. If a is an arbitrary point in A , then $a \in U$, so there exists an open ball $B(a; r) \subseteq U$. It follows that $a \in B(a; r) \cap S \subseteq U \cap S = A$, so a is a relative interior point of A . This shows that A is relatively open in S , so completes the proof of statement (a).

To prove (b), suppose A is relatively closed in S . Then $S \setminus A$ is relatively open, so $S \setminus A = U \cap S$ for some open set U in $\mathbb{R}^n$. But then

$$A = S \setminus (S \setminus A) = S \setminus (U \cap S) = S \setminus U = (\mathbb{R}^n \setminus U) \cap S$$

where $F = \mathbb{R}^n \setminus U$ is closed in $\mathbb{R}^n$.

Conversely, if $A = F \cap S$ for some closed F in $\mathbb{R}^n$, then

$$S \setminus A = S \setminus F = (\mathbb{R}^n \setminus F) \cap S$$

where $U = \mathbb{R}^n \setminus F$ is open in $\mathbb{R}^n$. It follows from (a) that $S \setminus A$ is relatively open in S , and so A is relatively closed. $\square$

Note that we can choose $F = \bar{A}$ in (4) (b), i.e. A is relatively closed in S if and only if $\bar{A} \cap S = A$. The following characterization of a relatively closed set is often useful (see Problem 7).

- A subset $A \subseteq S$ is **relatively closed** in S if and only if whenever a sequence $\{x_k\}$ in A converges to a limit in S , this limit belongs to A . (5)

Here is the promised characterization of continuous functions in terms of open or closed sets:

THEOREM 13.3.5

Let f be any function from $S \subseteq \mathbb{R}^n$ to $\mathbb{R}^m$. Then f is continuous if and only if either of the following conditions is satisfied:

- (a) $f^{-1}(U)$ is relatively open in S for each open set U in $\mathbb{R}^m$.
 (b) $f^{-1}(F)$ is relatively closed in S for each closed set F in $\mathbb{R}^m$.

Proof: (a) Let us first prove the “only if” part. Suppose f is continuous and U is an open set in $\mathbb{R}^m$. We want to show that $f^{-1}(U)$ is open in the relative topology of S . Let x be any point in $f^{-1}(U)$. Then $f(x) \in U$, and since U is open, there is an $\varepsilon > 0$ such that $B(f(x); \varepsilon) \subseteq U$. Since f is continuous at x , there exists a $\delta > 0$ such that $f(x') \in B(f(x); \varepsilon) \subseteq U$ for all $x' \in B(x; \delta) \cap S = B^S(x; \delta)$. Then $B^S(x; \delta) \subseteq f^{-1}(U)$, and so x is a relative interior point of $f^{-1}(U)$. It follows that $f^{-1}(U)$ is open.

To prove the “if” part of (a), suppose that the inverse image of every open set in $\mathbb{R}^m$ is relatively open in S . Let $\mathbf{x}$ be any point in S . We shall show that f is continuous at $\mathbf{x}$. Let ε be an arbitrary positive number. Then $U = B(f(\mathbf{x}); \varepsilon)$ is an open set in $\mathbb{R}^m$, and $f^{-1}(U)$ is a relatively open set in S . Since $\mathbf{x} \in f^{-1}(U)$, there is a relatively open ball $B^S(\mathbf{x}; \delta)$ around $\mathbf{x}$ such that $B^S(\mathbf{x}; \delta) \subseteq f^{-1}(U)$. Then for every $\mathbf{x}'$ in $B^S(\mathbf{x}; \delta) = B(\mathbf{x}; \delta) \cap S$, we have $f(\mathbf{x}') \in U = B(f(\mathbf{x}); \varepsilon)$. It follows that f is continuous at $\mathbf{x}$.

(b) Recall that a set F in $\mathbb{R}^m$ is closed if and only if its complement $\mathbb{R}^m \setminus F$ in $\mathbb{R}^m$ is open. Because

$$f^{-1}(\mathbb{R}^m \setminus F) = \{\mathbf{x} \in S : f(\mathbf{x}) \notin F\} = S \setminus f^{-1}(F)$$

the result for closed sets follows from that for open sets (and conversely). ■

PROBLEMS FOR SECTION 13.3

1. Prove that the set $S = \{(x, y) : 2x - y < 2 \text{ and } x - 3y < 5\}$ is open in $\mathbb{R}^2$.
2. Prove that the set $S = \{\mathbf{x} \in \mathbb{R}^n : g_j(\mathbf{x}) \leq 0, j = 1, \dots, m\}$ is closed if the functions g_j are all continuous.
3. Give examples of subsets S of $\mathbb{R}$ and continuous functions $f : \mathbb{R} \rightarrow \mathbb{R}$ such that
 - (a) S is closed, but $f(S)$ is not closed.
 - (b) S is open, but $f(S)$ is not open.
 - (c) S is bounded, but $f(S)$ is not bounded.
4. For a fixed $\mathbf{a}$ in $\mathbb{R}^n$, prove that the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ defined by $f(\mathbf{x}) = d(\mathbf{x}, \mathbf{a})$ is continuous. (Hint: See Problem 13.1.2.)
5. Let S be a closed set in $\mathbb{R}^n$ and $\mathbf{y}$ a fixed point in $\mathbb{R}^n$. Let $h(\mathbf{x}) = d(\mathbf{x}, \mathbf{y})$ for all $\mathbf{x}$ in S . Then h is continuous by Problem 4. Use the extreme value theorem to show that h attains a minimum at some point of S . (Hint: If $\mathbf{x}'$ is an arbitrary point in S , then any possible minimum point for $h(\mathbf{x})$ must lie in the intersection of S and the closed ball $\bar{B}(\mathbf{y}; r)$ with radius $r = d(\mathbf{x}', \mathbf{y})$.)
6. Prove the “if” part of Theorem 13.3.2.
7. Prove (5). Then use this characterization of relatively closed sets to offer an alternative proof of Theorem 13.3.5, part (b).
8. In a game between two players, the challenger and the defender, the defender tries to prove that a function $f : S \rightarrow \mathbb{R}^m$ with $S \subseteq \mathbb{R}^n$ is continuous at $\mathbf{x}^0 \in S$. The challenger tries to disprove this. The challenger makes the first move in the game by choosing a real $\varepsilon > 0$, which is observed by the defender. Then the defender responds by choosing a real $\delta > 0$ as a function $\delta(\varepsilon)$ of the observed ε . Finally, knowing both δ and ε , the challenger chooses an $\mathbf{x} \in S$. The rules are that the challenger wins if $\|\mathbf{x} - \mathbf{x}^0\| < \delta$ but $\|f(\mathbf{x}) - f(\mathbf{x}^0)\| \geq \varepsilon$; otherwise the defender wins. Explain why, with best play on both sides, the defender wins if and only if f is continuous at $\mathbf{x}^0$; otherwise the challenger wins.

13.4 Maximum Theorems

Economic theory abounds with “comparative statics” results. These describe what happens to an optimal solution in response to changes in exogenous parameters such as prices. In particular, will small changes in these parameters lead to only small changes in the criterion function? And to small changes in the optimal solution? The purpose of this section is to give some such results. Other results of this kind are presented in Chapter 14.

Suppose that $f(\mathbf{x}, \mathbf{y}) = f(x_1, \dots, x_n, y_1, \dots, y_m)$ is a continuous function defined for all $\mathbf{x}$ in X and all $\mathbf{y}$ in Y , where $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$. Suppose too that Y is compact. Then the extreme value theorem (Theorem 3.1.3) implies that for every $\mathbf{x}$ in X the problem of maximizing $f(\mathbf{x}, \mathbf{y})$ subject to $\mathbf{y} \in Y$ has a solution. The maximum value of $f(\mathbf{x}, \mathbf{y})$ will depend on $\mathbf{x}$. Define the (optimal) value function $V : X \rightarrow \mathbb{R}$ for the problem by

$$V(\mathbf{x}) = \max_{\mathbf{y} \in Y} f(\mathbf{x}, \mathbf{y}) \quad (1)$$

The next theorem tells us that $V(\mathbf{x})$ is in fact continuous. For a given $\mathbf{x}$ in X there may be several $\mathbf{y}$ in Y that maximize $f(\mathbf{x}, \mathbf{y})$. However, if for every $\mathbf{x}$ there is a *unique* $\mathbf{y} = \mathbf{y}(\mathbf{x})$ that solves problem (1), then $\mathbf{y}(\mathbf{x})$ varies continuously with $\mathbf{x}$.

THEOREM 13.4.1 (THE MAXIMUM THEOREM: THE SIMPLEST CASE)

Suppose that f is a continuous function from $X \times Y$ to $\mathbb{R}$, where $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$, and Y is compact, with $X, Y \neq \emptyset$. Then:

- (a) The value function $V(\mathbf{x}) = \max_{\mathbf{y} \in Y} f(\mathbf{x}, \mathbf{y})$ is a continuous function of $\mathbf{x}$.
- (b) If the maximization problem has a unique solution $\mathbf{y} = \mathbf{y}(\mathbf{x})$ for every $\mathbf{x}$, then $\mathbf{y}(\mathbf{x})$ is a continuous function of $\mathbf{x}$.

Proof: (a) We argue by contradiction. By the extreme value theorem, $V(\mathbf{x})$ is defined for every $\mathbf{x}$ in X . Suppose V happens to be discontinuous at some $\mathbf{x}^0$ in X . By Theorem 13.3.2, there exists a sequence $\{\mathbf{x}_k\}$ converging to $\mathbf{x}^0$ such that $\{V(\mathbf{x}_k)\}$ does not converge to $V(\mathbf{x}^0)$. So there is an $\varepsilon > 0$ such that

$$|V(\mathbf{x}_k) - V(\mathbf{x}^0)| \geq \varepsilon \quad (*)$$

for infinitely many k . Hence there is a subsequence of $\{\mathbf{x}_k\}$ such that $(*)$ holds for every term of that subsequence. This subsequence also converges to $\mathbf{x}^0$. By changing notation, denote this subsequence by $\{\mathbf{x}_k\}$.

By the extreme value theorem, for each k there is a $\mathbf{y}_k$ in Y such that $V(\mathbf{x}_k) = f(\mathbf{x}_k, \mathbf{y}_k)$. Use the Bolzano–Weierstrass theorem to choose a subsequence $\{\mathbf{y}_{k_j}\}$ that converges to some $\mathbf{y}^0$ in Y . For arbitrary $\mathbf{y}$ in Y , we have $f(\mathbf{x}_{k_j}, \mathbf{y}) \leq f(\mathbf{x}_{k_j}, \mathbf{y}_{k_j})$, so taking limits, we get $f(\mathbf{x}^0, \mathbf{y}) \leq f(\mathbf{x}^0, \mathbf{y}^0)$. Hence, $V(\mathbf{x}^0) = \max_{\mathbf{y} \in Y} f(\mathbf{x}^0, \mathbf{y}) \leq f(\mathbf{x}^0, \mathbf{y}^0)$. But by definition, $V(\mathbf{x}^0) \geq f(\mathbf{x}^0, \mathbf{y}^0)$, so $V(\mathbf{x}^0) = f(\mathbf{x}^0, \mathbf{y}^0)$. Then $V(\mathbf{x}^0) = \lim_j f(\mathbf{x}_{k_j}, \mathbf{y}_{k_j}) = \lim_j V(\mathbf{x}_{k_j})$, contradicting $(*)$.

(b) Suppose that $y(x)$ is not continuous at x^0 in X . Then there exists a sequence $\{x_k\}$ in X converging towards x^0 , such that $\{y(x_k)\}$ does not converge to $y(x^0)$. For some $\varepsilon > 0$ there exists a subsequence $\{x_{k_j}\}$ such that $\|y(x_{k_j}) - y(x^0)\| \geq \varepsilon$ for all j . By compactness of Y , this sequence again has a subsequence $\{\tilde{x}_i\}$ such that $\{\tilde{y}_i\} = \{y(\tilde{x}_i)\}$ converges to some $y' \neq y(x^0)$. Then $V(x^0) = \lim_i V(\tilde{x}_i) = \lim_i f(\tilde{x}_i, y(\tilde{x}_i)) = f(x^0, y')$, so y' also solves the maximization problem for $x = x^0$. This contradicts the hypothesis that the solution $y(x)$ is unique. ■

Note that the value function $V(x)$ in Theorem 13.4.1 is continuous even if $y(x)$ is not unique. The theorem is illustrated (for the case $n = m = 1$) in Figs. 1 and 2. Figure 1 shows the graph of a function $f(x, y)$. For each x in X the function has a maximum value $V(x)$ w.r.t. y . The figure suggests that if the function $f(x, y)$ is continuous, then $V(x)$ is also likely to be a continuous function of x . The graph of V is shown in Fig. 2. Furthermore, if for each x there is only one value of y that maximizes $f(x, y)$, it seems plausible that this maximizing y will also vary continuously with x .

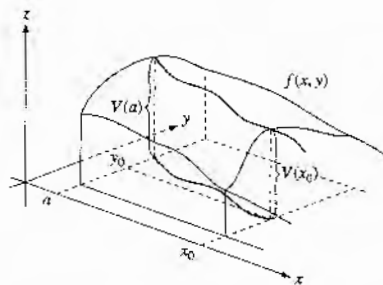


Figure 1

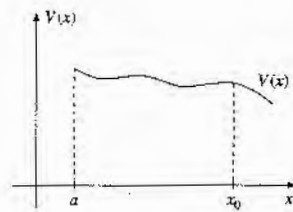


Figure 2

EXAMPLE 1 Let $X = \mathbb{R}$ and $Y = [-1, 2]$. Suppose that $f : X \times Y \rightarrow \mathbb{R}$ is defined by $f(x, y) = xy^2$. Consider the problem

$$\max f(x, y) \quad \text{subject to} \quad -1 \leq y \leq 2$$

For a fixed $x < 0$, $f(x, y) = xy^2$ is clearly maximized at $y = 0$, and the maximum value is 0. For a fixed $x > 0$, the function xy^2 is maximized at $y = 2$, and the maximum value is $4x$. Finally, for $x = 0$, all values of y in $[-1, 2]$ maximize xy^2 , and the maximum value is 0. Thus, the value function $V(x)$ for the problem is

$$V(x) = \begin{cases} 0 & \text{if } x < 0 \\ 4x & \text{if } x \geq 0 \end{cases}$$

Hence, V is continuous for all x , and differentiable for $x \neq 0$. Note also that the maximum point $y(x)$ is unique for all $x \neq 0$, and is a continuous function of x in each of the two intervals $(-\infty, 0)$ and $(0, \infty)$. ■

EXAMPLE 2 Let $X = Y = \mathbb{R}$, and define $f : X \times Y \rightarrow \mathbb{R}$ by $f(x, y) = e^{-(xy-1)^2}$. It is easy to see that when $x \neq 0$, then $f(x, y)$ is maximized w.r.t. y at $y = 1/x$, whereas when $x = 0$, any value of y is a maximizer. Hence,

$$V(x) = \max_{y \in \mathbb{R}} f(x, y) = \begin{cases} e^{-1} & \text{if } x = 0 \\ 1 & \text{if } x \neq 0 \end{cases}$$

Thus the value function is discontinuous at $x = 0$. In this example f is continuous, but Y is not compact. ■

EXAMPLE 3 (Maximum profit as a function of prices) Suppose the production of a commodity requires n input factors. If $\mathbf{v} = (v_1, \dots, v_n)$ is the vector of inputs, the number of units produced is $f(\mathbf{v})$. Assume that the production function f is defined and continuous on $\mathbb{R}_+^n = \{\mathbf{v} : \mathbf{v} \in \mathbb{R}^n, \mathbf{v} \geq 0\}$,¹ with $f(\mathbf{0}) = 0$ and $f(\mathbf{v}) \geq 0$ for all $\mathbf{v} \geq 0$. Assume further that, for each positive number a there exists a number K_a such that if $\|\mathbf{u}\| = 1$ and $\lambda \geq K_a$, then $f(\lambda\mathbf{u})/\lambda < a$. In particular, this implies that for each fixed $\mathbf{u}$, one has $f(\lambda\mathbf{u})/\lambda \rightarrow 0$ as $\lambda \rightarrow \infty$.

If the selling price per unit of the product is p and the unit prices of the input factors are given by the vector $\mathbf{q} = (q_1, \dots, q_n)$, then profit as a function of the input vector $\mathbf{v}$ is $\pi(\mathbf{v}) = pf(\mathbf{v}) - \mathbf{q} \cdot \mathbf{v}$. The following facts can be established:

- For given prices $p > 0$ and $\mathbf{q} \gg 0$, the profit function $\pi(\mathbf{v})$ attains a maximum value $V(p, \mathbf{q})$ as $\mathbf{v}$ runs through $\mathbb{R}_+^n$.
- $V(p, \mathbf{q})$ is a continuous function of $(p, \mathbf{q})$ over $\mathbb{R}_{++}^{n+1} = \{(p, \mathbf{q}) : p > 0, \mathbf{q} \gg 0\}$.

We would like to apply Theorem 13.4.1. The only difficulty in doing so arises because the vector $\mathbf{v}$ of chosen inputs can range over the whole of the unbounded set $\mathbb{R}_+^n$. The assumptions on f , however, allow us to prove that, for each fixed $(p^0, \mathbf{q}^0) \in \mathbb{R}_{++}^{n+1}$, there exist a neighbourhood N of $(p^0, \mathbf{q}^0)$ and a constant $a > 0$ such that the compact subset $K^0 = \{\mathbf{u} \in \mathbb{R}_+^n : \|\mathbf{u}\| \leq K_a\}$ contains any optimal choice of $\mathbf{v}$. Then the problem of maximizing $\pi(\mathbf{v})$ over $\mathbb{R}_+^n$ is unaltered if one imposes the additional constraint $\mathbf{v} \in K^0$. But because K^0 is compact, Theorem 13.4.1 does apply to this constrained problem, for all $(p, \mathbf{q}) \in N$. Finally, because this works for every fixed $(p^0, \mathbf{q}^0) \in \mathbb{R}_{++}^{n+1}$, the conclusions of Theorem 13.4.1 applies for all $(p, \mathbf{q}) \in \mathbb{R}_{++}^{n+1}$.

To show that $\max_{\mathbf{v} \geq 0} \pi(\mathbf{v}) = \max_{\mathbf{v} \in K^0} \pi(\mathbf{v})$, choose fixed positive numbers $k > p^0$ and $c < \min_i \{q_i^0\}$, so that $(c, \dots, c) \ll \mathbf{q}^0$. Next, let

$$N = \{(p, \mathbf{q}) \in \mathbb{R}_{++}^{n+1} : 0 < p < k, \mathbf{q} \gg (c, \dots, c)\}$$

which is a neighbourhood of $(p^0, \mathbf{q}^0)$. Define $a = c/(k\sqrt{n})$ and let K_a be defined as above. Given any $\mathbf{u}$ in $\mathbb{R}_+^n$ with $\|\mathbf{u}\| = 1$, notice that $\mathbf{u}$ has at least one component $\geq 1/\sqrt{n}$, so $\mathbf{q} \cdot \mathbf{u} \geq c/\sqrt{n} = ka > pa$ by definition of k and a . Let us show that K^0 contains all optimal $\mathbf{v}$. Take any $\mathbf{v} \in K^0$, i.e. $\lambda = \|\mathbf{v}\| > K_a$. Let $\mathbf{u} = \mathbf{v}/\lambda$, so that $\|\mathbf{u}\| = 1$. Then $pf(\mathbf{v}) = pf(\lambda\mathbf{u}) < p\lambda a$, by definition of K_a . Hence, $pf(\mathbf{v}) < p\lambda a < k\lambda a \leq \lambda \mathbf{q} \cdot \mathbf{u} = \mathbf{q} \cdot \mathbf{v}$, so $pf(\mathbf{v}) - \mathbf{q} \cdot \mathbf{v} < 0 = \pi(\mathbf{0})$. This shows that $\max_{\mathbf{v} \geq 0} \pi(\mathbf{v}) = \max_{\mathbf{v} \in K^0} \pi(\mathbf{v})$. ■

¹ The inequality $\mathbf{v} \geq \mathbf{u}$ means $v_i \geq u_i$ for all $i = 1, \dots, n$. If $v_i > u_i$ for all i , we write $\mathbf{v} \gg \mathbf{u}$.

Let us now extend the scope of the maximum theorem by allowing the fixed set Y to be replaced with a *constraint set* of the form

$$F(x) = \{y \in Y : g_i(x, y) \leq a_i, i = 1, \dots, l\} \quad (2)$$

that varies with x . Here the functions g_i and the numbers a_i are given. The maximization problem becomes

$$\text{maximize } f(x, y) \quad \text{subject to } y \in F(x)$$

Define the corresponding **value function**,

$$V(x) = \max_{y \in F(x)} f(x, y) \quad (3)$$

Then the following theorem holds:

THEOREM 13.4.2 (THE MAXIMUM THEOREM: A MORE GENERAL CASE)

Suppose that $f(x, y)$ and $g_i(x, y)$, $i = 1, \dots, l$, are continuous functions from $X \times Y$ into $\mathbb{R}$, where $X \subseteq \mathbb{R}^n$, $Y \subseteq \mathbb{R}^m$, and Y is compact. Suppose further that for every x in X , the constraint set (2) is nonempty and equal to the closure of $F^\circ(x) = \{y \in Y : g_i(x, y) < a_i, i = 1, \dots, l\}$. Then the value function $V(x)$ is continuous over X . Moreover, if the maximization problem has a unique maximum $y = y(x)$ for each x in X , then $y(x)$ is continuous.

The result follows from combining Theorem 14.2.1 with Example 14.1.5 in the next chapter.

EXAMPLE 4 Let $f(x_1, y_1)$ and $g(x_2, y_2)$ be two continuous production functions that give the quantities produced of two commodities as functions of the input factors $x_1 \geq 0$, $y_1 \geq 0$, $x_2 \geq 0$, and $y_2 \geq 0$. Say x_i denotes labour and y_i energy. The sale prices are p and q , respectively. Let $r > 0$ and $s > 0$ be the prices of the two inputs.

An entrepreneur wishes to choose x_1, x_2, y_1 , and y_2 such that total revenue $pf(x_1, y_1) + qg(x_2, y_2)$ is maximized subject to the total outlay for the input factors not exceeding budget allowance $m > 0$. Thus, the entrepreneur's problem is

$$\max pf(x_1, y_1) + qg(x_2, y_2) \quad \text{subject to } r(x_1 + x_2) + s(y_1 + y_2) \leq m \quad (*)$$

The constraint set

$$F(r, s, m) = \{(x_1, x_2, y_1, y_2) \in \mathbb{R}_+^4 : r(x_1 + x_2) + s(y_1 + y_2) \leq m\}$$

is obviously closed. It is also bounded, because if $(x_1, x_2, y_1, y_2) \in F(r, s, m)$ then $rx_i \leq m$ and $sy_i \leq m$, i.e. $x_i \in [0, m/r]$, $y_i \in [0, m/s]$. So $F(r, s, m)$ is compact, and therefore Theorem 13.4.2 already implies that the maximum revenue is a continuous function of

(p, q) . If f and g are strictly concave, then $pf(x_1, y_1) + qg(x_2, y_2)$ is strictly concave. Since Y is convex and any maximum of a strictly concave function over a convex set is unique (if it exists), in this case there is a unique maximum point $(x_1^*, x_2^*, y_1^*, y_2^*)$, which must be a continuous function of (p, q) .

But in this example it is easy to see that $\overline{F^\circ(r, s, m)} = F(r, s, m)$, so Theorem 13.4.2 implies that maximum profit is a continuous function of (p, q, r, s, m) wherever all five of these variables are positive.

PROBLEMS FOR SECTION 13.4

1. Let $f(x, y, z) = \ln(4 + y + z) + x^2 + e^2x^2 + e^2yz$ and define

$$V(x) = \max_{(y, z) \in S} f(x, y, z)$$

- (a) If $S = \{(y, z) : y \geq 1, z \geq 1, y^2 + z^2 \leq 4\} \subseteq \mathbb{R}^2$, is V continuous?
 (b) If $S = \{(y, z) : y > 0, z > 0, y^2 + z^2 \leq 4\} \subseteq \mathbb{R}^2$, is V continuous? (Hint: With y and z positive, $f(x, y, z)$ is strictly increasing in y and in z .)

2. Use the theorems in this section to determine whether each of the following functions is continuous:

$$(a) V_1(x) = \max_{u \in [0, 1]} (e^{-xu^2} - (u - x)^2) \quad (b) V_2(x) = \max_{u \in (-\infty, \infty)} (e^{-xu^2} - (u - x)^2)$$

3. Let $f(x, y) = -3xy^4 - 4(x - 1)y^3 + 6y^2$ for x in $X = (0, \infty)$ and y in $Y = [-3, 3]$. For each x in X , consider the problem of maximizing $f(x, y)$ subject to $y \in Y$. Let $M(x)$ be the corresponding set of (global) maximum points y in Y , and let

$$V(x) = \max_{y \in Y} f(x, y).$$

Verify that $V(x)$ is continuous, and draw a graph in the xy -plane showing the set $M(x)$ (one or more points) for each x . Are the maximizers given by a continuous function of x ?

4. Let $X = (0, \infty)$ and $Y = (-\infty, 1]$, and define the function $f : X \times Y \rightarrow \mathbb{R}$ by

$$f(x, y) = \ln(1 + xe^y) - y^2$$

Show that $V(x) = \max_{y \in Y} f(x, y)$ is a continuous function of x for $x > 0$.

13.5 Convex Sets

Section 2.2 gave the definition of convex sets and some of their basic properties. This section gives some further definitions and results that are occasionally useful in economics.

Let S and T be two arbitrary sets in $\mathbb{R}^n$. The (vector) **sum** of S and T is defined as

$$S + T = \{x + y : x \in S \text{ and } y \in T\} \quad (1)$$

Thus $S + T$ is the set of all possible sums $x + y$, when $x \in S$ and $y \in T$. For the case where S and T are subsets of $\mathbb{R}^2$, the construction is illustrated in Fig. 1.

More generally, if a and b are any two scalars in $\mathbb{R}$, define the **linear combination** $aS + bT$ of the two sets S and T as the set $\{ax + by : x \in S, y \in T\}$. This definition does not require a and b to be nonnegative: indeed, taking $a = 1$ and $b = -1$ gives the **vector difference** $S - T$ of the two sets. Note that this is entirely different from the set-theoretic difference $S \setminus T = \{x : x \in S \text{ and } x \notin T\}$.

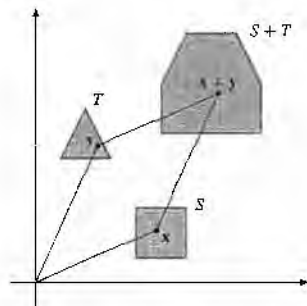


Figure 1 The sum of two sets.

EXAMPLE 1

Suppose that S is a firm's *production possibility set*, i.e. the set of all net output vectors that the firm can supply. If T is the corresponding set of a second firm, the set $S + T$ represents the *aggregate* net output vectors that the two firms can supply together. ■

If S and T are convex sets in $\mathbb{R}^n$, then $S + T$ is also convex, as Fig. 1 suggests. In fact, one has the following result:

$$S, T \text{ convex and } a, b \text{ real numbers} \implies aS + bT \text{ convex} \quad (2)$$

Proof: Let z and w belong to $Q = aS + bT$. Then $z = ax_1 + by_1$ and $w = ax_2 + by_2$, where $x_1, x_2 \in S$ and $y_1, y_2 \in T$. Let $\lambda \in [0, 1]$. We must prove that $\lambda z + (1 - \lambda)w \in Q$. In fact, $\lambda z + (1 - \lambda)w = \lambda(ax_1 + by_1) + (1 - \lambda)(ax_2 + by_2) = a(\lambda x_1 + (1 - \lambda)x_2) + b(\lambda y_1 + (1 - \lambda)y_2)$. This belongs to $Q = aS + bT$ because $\lambda x_1 + (1 - \lambda)x_2 \in S$ and $\lambda y_1 + (1 - \lambda)y_2 \in T$ due to the convexity of S and T . ■

So any linear combination of two convex sets is convex. It is easy to extend this result to linear combinations of an arbitrary finite number of convex sets.

Convex Hulls

Suppose that $x_1, \dots, x_m$ are vectors in $\mathbb{R}^n$. A point x that can be expressed in the form

$$x = \lambda_1 x_1 + \dots + \lambda_m x_m, \quad \text{with } \lambda_i \geq 0 \text{ for each } i \text{ and } \sum_{i=1}^m \lambda_i = 1 \quad (3)$$

is called a **convex combination** of the points $x_1, \dots, x_m$. It is a linear combination where the scalar weights λ_i are restricted to be nonnegative and sum to 1. Accordingly, these scalars are called **convex weights**.

In particular, a convex combination of two points x_1 and x_2 takes the form $\lambda_1 x_1 + \lambda_2 x_2$ with $\lambda_1 \geq 0, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$. Hence $\lambda_2 = 1 - \lambda_1$ and $\lambda_1 x_1 + \lambda_2 x_2 = \lambda_1 x_1 + (1 - \lambda_1)x_2$, with $\lambda_1 \in [0, 1]$. Thus a set S in $\mathbb{R}^n$ is convex if and only if it contains all convex combinations of each pair of points in S . Problem 6 indicates a proof of the following result:

$$\text{A convex set } S \text{ in } \mathbb{R}^n \text{ contains all convex combinations of points from } S. \quad (4)$$

If S is an arbitrary set in $\mathbb{R}^n$, the **convex hull** of S , denoted by $\text{co}(S)$, is defined as

$$\text{co}(S) = \text{the set of all convex combinations of points from } S \quad (5)$$

A point in S is clearly a convex combination of itself, because $x = 1 \cdot x$. Hence $S \subseteq \text{co}(S)$. In Problem 4 you are asked to prove that $\text{co}(S)$ is always convex. Because (4) and (5) imply that any convex set containing S also contains $\text{co}(S)$, the following must be true:

$$\text{co}(S) \text{ is the smallest convex set containing } S \quad (6)$$

The convex hulls of the two sets are illustrated in Figs. 2 and 3.

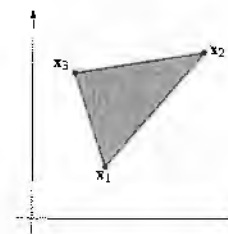


Figure 2 All convex combinations of x_1, x_2, x_3 .

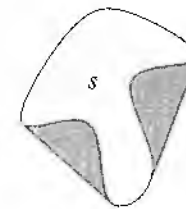


Figure 3 If S is the unshaded set, then $\text{co}(S)$ includes the shaded parts in addition.

Carathéodory's Theorem

An n -dimensional **simplex** in $\mathbb{R}^n$ is a set $S = \text{co}(V)$, where V is a set consisting of $n + 1$ points of $\mathbb{R}^n$, called the **vertices** of S , such that S has a nonempty interior. A particular example is the simplex $T = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_1 \geq 0, \dots, x_n \geq 0, x_1 + \dots + x_n \leq 1\}$, whose vertices are 0 and the n standard unit vectors $e^1, \dots, e^n$, where e^i has its i th component equal to 1 and the other $n - 1$ components equal to 0. Obviously, any point in the simplex $S = \text{co}(V)$ can be expressed as a convex combination of at most $n + 1$ vertices. The following theorem shows that a similar result holds for the convex hull of any set in $\mathbb{R}^n$:

THEOREM 13.5.1 (CARATHÉODORY)

If $S \subseteq \mathbb{R}^n$ and $\mathbf{x} \in \text{co}(S)$, then $\mathbf{x}$ can be expressed as a convex combination of at most $n + 1$ points in S .

Proof: Suppose $\mathbf{x}$ equals the convex combination $\lambda_1 \mathbf{x}_1 + \cdots + \lambda_k \mathbf{x}_k$ of k points in $\mathbb{R}^n$, where $k > n + 1$. For $i = 1, \dots, k$, let $\mathbf{y}_i = (\mathbf{x}_i, 1) \in \mathbb{R}^{n+1}$ denote the vector $\mathbf{x}_i$ augmented by an extra component equal to 1. The $k > n + 1$ vectors $\mathbf{y}_1, \dots, \mathbf{y}_k$ in $\mathbb{R}^{n+1}$ are linearly dependent, so there exist scalars $\alpha_1, \dots, \alpha_k$, not all 0, such that $\alpha_1 \mathbf{y}_1 + \cdots + \alpha_k \mathbf{y}_k = \mathbf{0} \in \mathbb{R}^{n+1}$, that is,

$$\alpha_1 \mathbf{x}_1 + \cdots + \alpha_k \mathbf{x}_k = \mathbf{0} \in \mathbb{R}^n \quad \text{and} \quad \alpha_1 + \cdots + \alpha_k = 0$$

Obviously, at least one α_j must be positive. Let $r = \max\{-\lambda_i/\alpha_j : \alpha_j > 0\}$. Then $r \leq 0$ and $\lambda_i + r\alpha_i \geq 0$ for all $i = 1, 2, \dots, k$. This latter inequality is satisfied with equality for at least one index i . Hence,

$$\begin{aligned} \mathbf{x} = \mathbf{x} + r\mathbf{0} &= \lambda_1 \mathbf{x}_1 + \cdots + \lambda_k \mathbf{x}_k + r(\alpha_1 \mathbf{x}_1 + \cdots + \alpha_k \mathbf{x}_k) \\ &= (\lambda_1 + r\alpha_1) \mathbf{x}_1 + \cdots + (\lambda_k + r\alpha_k) \mathbf{x}_k \end{aligned}$$

where $\lambda_i + r\alpha_i \geq 0$ for $i = 1, 2, \dots, k$, and $(\lambda_1 + r\alpha_1) + \cdots + (\lambda_k + r\alpha_k) = 1$. Because $\lambda_i + r\alpha_i = 0$ for at least one i , $\mathbf{x}$ must be a convex combination of at most $k - 1$ of the points $\mathbf{x}_1, \dots, \mathbf{x}_k$ in $\mathbb{R}^n$. Clearly, this process of eliminating points $\mathbf{x}_i$ one at a time can be repeated until $\mathbf{x}$ is expressed as a convex combination of at most $n + 1$ points. ■

Extreme Points of Convex Sets

An **extreme point** of a convex set S in $\mathbb{R}^n$ is a point in S that does not lie “properly inside” any line segment in S . More precisely, $\mathbf{z}$ is an extreme point of S if $\mathbf{z} \in S$ and there are no $\mathbf{x}$ and $\mathbf{y}$ in S and λ in $(0, 1)$ such that $\mathbf{z} \neq \mathbf{y}$ and $\mathbf{z} = \lambda \mathbf{x} + (1 - \lambda) \mathbf{y}$. Equivalently, $\mathbf{z}$ is a point of S that cannot be expressed as a convex combination of other points of S . In $\mathbb{R}$, a compact interval has two extreme points, an open interval has no extreme points, and a half-open interval has one extreme point.



Figure 4 A , B , and C are extreme points. D is not.

Any extreme point of a convex set must be a boundary point (see Problem 5). Thus, an open ball $B_r(\mathbf{a}) \subseteq \mathbb{R}^n$ has no extreme points. For a closed ball $\overline{B_r(\mathbf{a})}$, however, every boundary point is an extreme point. But not all boundary points of every convex set are extreme points. To see why, look at Fig. 4, where A , B , and C are the only extreme points. The point D is a boundary point that is not an extreme point. (Why? Because D is a convex combination of the two other points B and C .)

For a proof of the following theorem see e.g. Corollary 18.5.1 in Rockafellar (1970).

THEOREM 13.5.2 (KREIN–MILMAN)

Every compact convex set in $\mathbb{R}^n$ is the convex hull of its extreme points.

This finite-dimensional result is actually due to Minkowski, and is therefore also known as Minkowski’s theorem. Krein and Milman extended it to certain infinite-dimensional spaces.

Strictly Convex Sets

A **convex body** in $\mathbb{R}^n$ is a convex set with a nonempty interior. Suppose S is a convex body in $\mathbb{R}^n$ such that $\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}$ is an interior point of S whenever $\mathbf{x}$ and $\mathbf{y}$ are distinct points in S and $\lambda \in (0, 1)$. In this case S is called a **strictly convex body**.

For example, a ball in n -space is a strictly convex body, whether it is open or closed. On the other hand, a closed pyramid in 3-space is convex but not strictly convex.

It can be shown that a closed convex body S is strictly convex if and only if every boundary point of S is an extreme point. Generally, a convex body S is strictly convex if and only if every boundary point is an extreme point of the closure $\overline{S}$.

NOTE 1 Strict convexity can also be defined for sets S in $\mathbb{R}^n$ that contain no interior points. Call a point $\mathbf{z}$ in S a **relative interior point** if for every $\mathbf{c}$ in S there is a number $\mu > 1$ such that the point $\mu \mathbf{z} + (1 - \mu) \mathbf{c} = \mathbf{c} + \mu(\mathbf{z} - \mathbf{c}) \in S$, i.e. the line segment from $\mathbf{c}$ to $\mathbf{z}$ can be extended a little bit beyond $\mathbf{z}$ without leaving the set S .

The usual definition of strict convexity is this: S is **strictly convex** if for each pair of distinct points $\mathbf{x}$ and $\mathbf{y}$ in S , every point of the **open line segment** $(\mathbf{x}, \mathbf{y}) = \{\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} : 0 < \lambda < 1\}$ is a relative interior point of S . For example, a circular disk (like a coin whose thickness is zero) lying in $\mathbb{R}^3$ is strictly convex according to this definition. So is a line segment, or even a set consisting of a single point. When S does have interior points, the two definitions are equivalent.

PROBLEMS FOR SECTION 13.5

- Construct the set $S + T$ in the cases shown in Figs. (a) and (b).

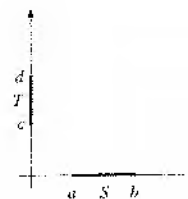


Figure (a)

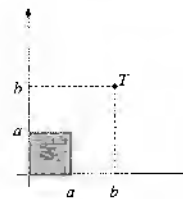


Figure (b)

2. Determine $\text{co}(S)$ in the cases shown in Figs. (c) and (d). (In (d), S consists of the four dots.)

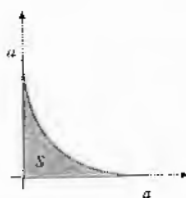


Figure (c)

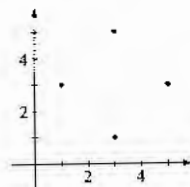


Figure (d)

3. Suppose that N units of a commodity (50,000 barrels of oil, for example) are spread out over points represented by a two-dimensional coordinate system so that n_1 units are to be found at the point $\mathbf{x}_1$, n_2 units are at $\mathbf{x}_2$, ..., n_m units are at $\mathbf{x}_m$, where $\sum_{i=1}^m n_i = N$. Explain why $\mathbf{z} = (1/N)(n_1\mathbf{x}_1 + n_2\mathbf{x}_2 + \dots + n_m\mathbf{x}_m)$ is a convex combination of $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m$. What is a common name for the point $\mathbf{z}$?
4. If S is an arbitrary set in $\mathbb{R}^n$, prove that the set $\text{co}(S)$ in (5) is convex. (Hint: Let $\mathbf{x} = \lambda_1\mathbf{u}_1 + \dots + \lambda_p\mathbf{u}_p$ and $\mathbf{y} = \mu_1\mathbf{v}_1 + \dots + \mu_q\mathbf{v}_q$ be arbitrary points in $\text{co}(S)$ with $\mathbf{u}_1, \dots, \mathbf{u}_p$ and $\mathbf{v}_1, \dots, \mathbf{v}_q$ all in S . Let $\lambda \in [0, 1]$ and prove by a direct argument that $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ is a convex combination of the points $\mathbf{u}_1, \dots, \mathbf{u}_p, \mathbf{v}_1, \dots, \mathbf{v}_q$.)
5. Show that an extreme point of a convex set must be a boundary point of the set. (Hint: Show that an interior point cannot be an extreme point.)

HARDER PROBLEMS

6. Prove (4). (Hint: The statement is true for $k = 2$ (and for $k = 1$). Suppose that it is true for $k = m$, where m is a positive integer, and let $\mathbf{x}_1, \dots, \mathbf{x}_{m+1}$ be $m+1$ points in S . Define $\mathbf{x} = \sum_{i=1}^m \lambda_i \mathbf{x}_i + \lambda_{m+1} \mathbf{x}_{m+1}$ with all $\lambda_i \geq 0$ and $\sum_{i=1}^{m+1} \lambda_i = 1$. If $\lambda_{m+1} = 1$, then $\mathbf{x} \in S$. Suppose next that $\lambda_{m+1} \neq 1$. Then

$$\mathbf{x} = (\lambda_1 + \dots + \lambda_m) \left[\sum_{i=1}^m \frac{\lambda_i}{\lambda_1 + \dots + \lambda_m} \mathbf{x}_i \right] + \lambda_{m+1} \mathbf{x}_{m+1}$$

is a convex combination of two points in S .

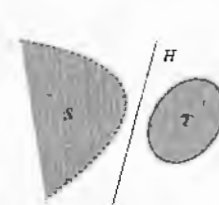
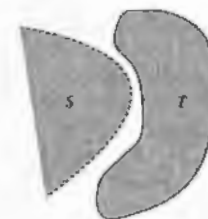
7. Use Carathéodory's theorem to show that the convex hull of any compact set in $\mathbb{R}^n$ is compact. Give an example to show that $\text{co}(S)$ need not be closed if S is closed but unbounded.

13.6 Separation Theorems

This section considers some theorems of a geometric nature with many applications in economic theory. The main result states that two disjoint convex sets in $\mathbb{R}^n$ can be separated by a hyperplane. In two dimensions, hyperplanes are straight lines, and the geometric content of the theorem in $\mathbb{R}^2$ is shown in Fig. 1.

Figure 2 shows an example of two disjoint sets in $\mathbb{R}^2$ that *cannot* be separated by a hyperplane; S is convex, but T is not. Of course, it may be possible to separate two sets even if either or both are not convex.

With its simple geometrical interpretation, the separation theorem in $\mathbb{R}^n$ is one of the most fundamental tools in modern optimization theory. In particular, the theorem makes it possible to state optimality conditions without differentiability requirements in cases where the functions involved are either concave or convex, as appropriate.

Figure 1 S and T are (strictly) separated by H .Figure 2 S and T cannot be separated by a hyperplane.

Separation theorems are also useful in many other areas. An early economic application of separation theorems was to welfare economics, where they were used to prove, under suitable hypotheses, that each non-extreme Pareto efficient allocation can be "decentralized" as a competitive equilibrium. (See Arrow (1951), or Mas-Colell et al. (1995).)

Recall from Example 2.2.1 (and (1.1.43)) that if $\mathbf{a}$ is a nonzero vector in $\mathbb{R}^n$ and α is a real number, then the set

$$H = \{\mathbf{x} : \mathbf{a} \cdot \mathbf{x} = \alpha\} \quad (1)$$

is a hyperplane in $\mathbb{R}^n$, with $\mathbf{a}$ as a normal vector. Moreover, the hyperplane H separates $\mathbb{R}^n$ into two closed half-spaces (see Example 2.2.1).

If S and T are subsets of $\mathbb{R}^n$, then H is said to **separate** S and T if S is contained in one of the closed half-spaces determined by H and T is contained in the other. In other words, S and T can be separated by a hyperplane if there exist a vector $\mathbf{a} \neq \mathbf{0}$ and a scalar α such that

$$\mathbf{a} \cdot \mathbf{x} \leq \alpha \leq \mathbf{a} \cdot \mathbf{y} \quad \text{for all } \mathbf{x} \text{ in } S \text{ and all } \mathbf{y} \text{ in } T \quad (*)$$

If both inequalities are strict, then the hyperplane $H = \{\mathbf{x} : \mathbf{a} \cdot \mathbf{x} = \alpha\}$ **strictly separates** S and T .

The first separation theorem we prove deals with the case where S is closed and convex, and T consists of only one point, $T = \{\mathbf{y}\}$. When a hyperplane separates a one-point set

from another set, one often says (a little imprecisely) that the hyperplane separates the point from the other set, as in the theorem below.

THEOREM 13.6.1 (A SPECIAL SEPARATION THEOREM)

Let S be a closed, convex set in $\mathbb{R}^n$, and let y be a point in $\mathbb{R}^n$ that does not belong to S . Then there exists a nonzero vector a in $\mathbb{R}^n$ and a number α such that

$$a \cdot x < \alpha < a \cdot y \quad \text{for all } x \text{ in } S \quad (2)$$

For every such α the hyperplane $H = \{x : a \cdot x = \alpha\}$ strictly separates S and y .

The geometric idea of the following proof is quite simple: Drop the “perpendicular” from y to the nearest point w of the set S . Let H' be the hyperplane through w with the vector $a = y - w$ as a normal. Then H' will separate y and S because S is convex. Figure 3 illustrates the construction in the case $n = 2$. The desired hyperplane H is obtained by choosing α as any number strictly between $a \cdot w$ and $a \cdot y$, thus shifting the hyperplane H' part way from w toward y .

This argument sounds quite convincing in 2 or even 3 dimensions, but what about the general case? What is the perpendicular from a point to a convex set in $\mathbb{R}^n$? A rigorous proof is needed.

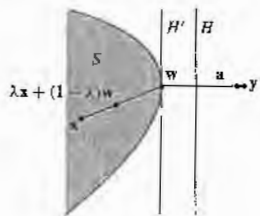


Figure 3 H' separates y from S , and H strictly separates y from S .

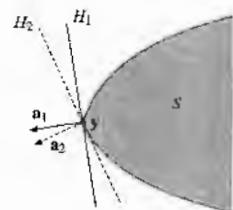


Figure 4 Two supporting hyperplanes to S at y are shown.

Proof: Because S is a closed set, among all the points of S there is one $w = (w_1, \dots, w_n)$ that is closest to y . (For a precise argument, see Problem 13.3.5. Because S is convex, the point w is actually unique, but we do not need this fact in the proof.)

Let $a = y - w$, the vector from w to y . (See Fig. 3.) Since $w \in S$ and $y \notin S$, it follows that $a \neq 0$. Note that $a \cdot (y - w) = a \cdot a > 0$, and so $a \cdot w < a \cdot y$. Suppose we prove that

$$a \cdot x \leq a \cdot w \quad \text{for all } x \text{ in } S \quad (i)$$

Then (2) will hold for every number α in the open interval $(a \cdot w, a \cdot y)$.

To prove (i), let x be any point in S , and let b denote $x - w$. Since S is convex, the point $w + \lambda b = \lambda x + (1 - \lambda)w$ belongs to S for each λ in $[0, 1]$. Now define $g(\lambda)$ as the square

of the distance from $w + \lambda b$ to the point y . Thus

$$g(\lambda) = \|w + \lambda b - y\|^2 = \|a - \lambda b\|^2$$

Differentiating $g(\lambda) = (a_1 - \lambda b_1)^2 + \dots + (a_n - \lambda b_n)^2$ w.r.t. λ gives

$$g'(\lambda) = -2(a_1 - \lambda b_1)b_1 - \dots - 2(a_n - \lambda b_n)b_n, \quad \text{or} \quad g'(\lambda) = -2(a - \lambda b) \cdot b$$

Also $g(0) = \|a\|^2 = \|y - w\|^2$, the square of the distance between y and w . It follows that $0 \leq g'(0) = -2a \cdot b$. This proves (i) because $b = x - w$. ■

In the proof of Theorem 13.6.1 it was essential that y did not belong to S , and this gave the strict inequality in (2). If S is an arbitrary convex set (not necessarily closed), and if y is not an interior point of S , then it seems plausible that y can still be separated from S by a hyperplane. If y is a boundary point of S , such a hyperplane is called a **supporting hyperplane** to S at y . It passes through y and has the property that, for a suitable normal $a = (a_1, \dots, a_n) \neq 0$ to it, $a \cdot x \leq a \cdot y$ for all $x = (x_1, \dots, x_n)$ in S (the vector a points away from S). Figure 4 shows two supporting hyperplanes to a set S at the same point y , together with normals to these hyperplanes.

THEOREM 13.6.2 (SEPARATING HYPERPLANE)

Let S be a convex set in $\mathbb{R}^n$ and suppose y is not an interior point of S . Then there exists a nonzero vector a in $\mathbb{R}^n$ such that

$$a \cdot x \leq a \cdot y \quad \text{for every } x \text{ in } S \quad (3)$$

Proof: Let $\bar{S}$ denote the closure of S . Because S is convex, so is $\bar{S}$ (see Problem 2). Because y is not an interior point of S and S is convex, y is not an interior point of $\bar{S}$ (see Problem 3). Hence there is a sequence $\{y_k\}$ of points outside $\bar{S}$ that converges to y . Now $y_k \notin \bar{S}$ and $\bar{S}$ is closed and convex, so according to the preceding separation theorem, for each $k = 1, 2, \dots$ there exists a vector $a_k \neq 0$ such that $a_k \cdot x < a_k \cdot y_k$ for all x in $\bar{S}$. Without loss of generality, after dividing by $\|a_k\|$ if necessary, we can assume that $\|a_k\| = 1$ for each k . Then $\{a_k\}$ is a sequence of vectors in the unit sphere of $\mathbb{R}^n$. Because this sphere is compact, the Bolzano–Weierstrass theorem (Theorem 13.2.5) shows that $\{a_k\}$ has a convergent subsequence $\{a_{k_j}\}$. Let $a = \lim_{j \rightarrow \infty} a_{k_j}$. Then $a \cdot x = \lim_{j \rightarrow \infty} (a_{k_j} \cdot x) \leq \lim_{j \rightarrow \infty} (a_{k_j} \cdot y_{k_j}) = a \cdot y$ for every x in $\bar{S}$, as required. Moreover, $a \neq 0$ because $\|a\| = \lim_{j \rightarrow \infty} \|a_{k_j}\| = 1$. ■

The general separation property illustrated in Fig. 1 turns out to be a rather simple consequence of the last two theorems.

THEOREM 13.6.3 (MINKOWSKI'S SEPARATING HYPERPLANE THEOREM)

Let S and T be two disjoint nonempty convex sets in $\mathbb{R}^n$. Then there exists a nonzero vector $\mathbf{a}$ in $\mathbb{R}^n$ and a scalar α such that

$$\mathbf{a} \cdot \mathbf{x} \leq \alpha \leq \mathbf{a} \cdot \mathbf{y} \quad \text{for all } \mathbf{x} \text{ in } S \text{ and all } \mathbf{y} \text{ in } T \quad (4)$$

Thus S and T are separated by the hyperplane $H = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{a} \cdot \mathbf{z} = \alpha\}$.

Proof: Let $\mathbf{W} = S - T$ be the vector difference of the two convex sets S and T . Since S and T are disjoint, $0 \notin \mathbf{W}$.

The set $\mathbf{W}$ is convex according to (13.5.2), so by Theorem 13.6.2 there exists an $\mathbf{a} \neq 0$ such that $\mathbf{a} \cdot \mathbf{w} \leq \mathbf{a} \cdot 0 = 0$ for all $\mathbf{w}$ in $\mathbf{W}$. Let $\mathbf{x}$ in S and $\mathbf{y}$ in T be any two points of these sets. Then $\mathbf{w} = \mathbf{x} - \mathbf{y} \in \mathbf{W}$, so $\mathbf{a} \cdot (\mathbf{x} - \mathbf{y}) \leq 0$. Hence

$$\mathbf{a} \cdot \mathbf{x} \leq \mathbf{a} \cdot \mathbf{y} \quad \text{for all } \mathbf{x} \text{ in } S \text{ and all } \mathbf{y} \text{ in } T \quad (*)$$

From (*) it follows, in particular, that the set $A = \{\mathbf{a} \cdot \mathbf{x} : \mathbf{x} \in S\}$ is bounded above by $\mathbf{a} \cdot \mathbf{y}$ for any $\mathbf{y}$ in T . Hence, A has a supremum α , say. Since α is the least of all the upper bounds of A , it follows that $\alpha \leq \mathbf{a} \cdot \mathbf{y}$ for every $\mathbf{y}$ in T . Therefore $\mathbf{a} \cdot \mathbf{x} \leq \alpha \leq \mathbf{a} \cdot \mathbf{y}$ for all $\mathbf{x}$ in S and all $\mathbf{y}$ in T . Thus S and T are separated by the hyperplane H . ■

An even more general separation theorem for convex sets in $\mathbb{R}^n$ is the following (for a proof see Rockafellar (1970), Theorem 11.3):

THEOREM 13.6.4 (GENERAL SEPARATING HYPERPLANE THEOREM)

Let S and T be two convex sets in $\mathbb{R}^n$ with no common relative interior point. Then S and T can be separated by a hyperplane, i.e. there exist a vector $\mathbf{a} \neq 0$ in $\mathbb{R}^n$ and a scalar α such that

$$\mathbf{a} \cdot \mathbf{x} \leq \alpha \leq \mathbf{a} \cdot \mathbf{y} \quad \text{for all } \mathbf{x} \text{ in } S \text{ and all } \mathbf{y} \text{ in } T \quad (5)$$

NOTE 1 Often in economics, a key price vector has components that are proportional to those of a normal to a suitable separating or supporting hyperplane.

PROBLEMS FOR SECTION 13.6

1. Let S be a nonempty, closed, convex set in $\mathbb{R}^n$ that does not contain the origin. Show that there exists a vector $\mathbf{a} = (a_1, \dots, a_n)$ and a positive real number α such that

$$\sum_{i=1}^n a_i x_i > \alpha \quad \text{for all } \mathbf{x} = (x_1, \dots, x_n) \text{ in } S$$

2. Prove that if S is a convex set in $\mathbb{R}^n$, then its closure, $\bar{S}$, is also convex. (Hint: Assume $\mathbf{x}, \mathbf{y} \in \bar{S}$ and let $\mathbf{x}_k \rightarrow \mathbf{x}, \mathbf{y}_k \rightarrow \mathbf{y}$, where $\mathbf{x}_k, \mathbf{y}_k \in S$.)

3. Prove that if S is a convex set in $\mathbb{R}^n$, and $\mathbf{x}$ is not an interior point of S , then $\mathbf{x}$ is not an interior point of $\bar{S}$.

4. If S is a set in $\mathbb{R}^n$ and $\mathbf{y}$ is a boundary point of S , is $\mathbf{y}$ necessarily a boundary point of $\bar{S}$? (Hint: The irrational number $\sqrt{2}$ is a boundary point of the set $\mathbb{Q}$ of rational numbers, but what is $\bar{\mathbb{Q}}$? If S is convex, then it is true that a boundary point of S is also a boundary point of $\bar{S}$ —see Problem 3.)

5. Some books in economics have suggested the following generalization of Theorem 13.6.3: Two convex sets in $\mathbb{R}^n$ with only one point in common can be separated by a hyperplane. Is this statement correct? What about the assertion that two convex sets in $\mathbb{R}^n$ with disjoint interiors can be separated by a hyperplane?

13.7 Productive Economies and Frobenius's Theorem

The final section of this chapter will indicate how some of the rather abstract concepts and results discussed lead to some interesting insights in economic models.

Consider an economy with n commodities. Producing the commodity vector $\mathbf{x} = (x_1, \dots, x_n)$ requires, in general, inputs of all goods. For $i = 1, \dots, n$, let $f_i(\mathbf{x})$ denote the amount of good i needed as an input to produce $\mathbf{x}$, so that $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_n(\mathbf{x}))$ is the commodity input vector needed to produce $\mathbf{x}$. It is reasonable to assume that the function $\mathbf{f}$ is increasing in the sense that²

$$\mathbf{u} \leq \mathbf{v} \Rightarrow \mathbf{f}(\mathbf{u}) \leq \mathbf{f}(\mathbf{v})$$

The vector of net outputs left for consumption and investment, the **final (net) supply**, is $\mathbf{y} = \mathbf{x} - \mathbf{f}(\mathbf{x})$.

Suppose that there exists a commodity vector $\mathbf{a} = (a^1, a^2, \dots, a^n) \geq (0, 0, \dots, 0)$ that can be produced, and for which the final supply $\mathbf{y} = \mathbf{a} - \mathbf{f}(\mathbf{a})$ is ≥ 0 . What other final supply vectors can be produced? A partial answer is given by the following theorem:

THEOREM 13.7.1

Let $\mathbf{f}$ be a continuous, increasing transformation from $\mathbb{R}_+^n$ into $\mathbb{R}_+^n$. Assume that there exists a vector $\mathbf{a} \geq 0$ such that $\mathbf{a} \geq \mathbf{f}(\mathbf{a})$. Then for every $\mathbf{y}$ such that $0 \leq \mathbf{y} \leq \mathbf{a} - \mathbf{f}(\mathbf{a})$, the equation $\mathbf{x} - \mathbf{f}(\mathbf{x}) = \mathbf{y}$ has a solution $\mathbf{x}$ with $\mathbf{y} \leq \mathbf{x} \leq \mathbf{a}$.

² Recall that $\mathbf{u} \leq \mathbf{v}$ means that $u_i \leq v_i$ for all $i = 1, \dots, n$. If the inequality is strict, with $u_i < v_i$ for all i , we write $\mathbf{u} < \mathbf{v}$.

Proof: Suppose y satisfies $0 \leq y \leq a - f(a)$. Let $x_0 = y$. Define $x_1 = f(x_0) + y$, and in general

$$x_m = f(x_{m-1}) + y, \quad m = 1, 2, \dots \quad (i)$$

We prove by induction on m that

$$x_{m-1} \leq x_m \quad \text{and} \quad x_{m-1} \leq a \quad \text{for } m = 1, 2, \dots \quad (ii)$$

Indeed, (ii) holds when $m = 1$ because $x_0 = y \leq a - f(a) \leq a$ and also $x_0 = y = x_1 - f(x_0) \leq x_1$. As the induction hypothesis, suppose that $k \geq 1$ and that (ii) holds when $m = k$. Then $x_k = f(x_{k-1}) + y \leq f(x_k) + y = x_{k+1}$ and also $x_k = f(x_{k-1}) + y \leq f(a) + y \leq a$, so (ii) is also true when $m = k + 1$. This completes the proof of (ii) by induction.

Define $x = \sup_m x_m$ (meaning that for $i = 1, \dots, n$, the i th component x^i of x satisfies $x^i = \sup_m x_m^i$). Because $x_{m-1} \leq x_m$ for $m = 1, 2, \dots$, it follows that $x_m^i \rightarrow x^i$ as $m \rightarrow \infty$ for every i , and therefore $x_m \rightarrow x$ as $m \rightarrow \infty$. Continuity of f and (i) together imply that $x = f(x) + y$, and so $x \geq y$. Also, (ii) implies that $x \leq a$. So x is the required solution. ■

NOTE 1 Suppose that the function f in Theorem 13.7.1 is *homogeneous of degree 1*, and that $a \gg f(a)$ for $a \gg 0$. Then $a - f(a) \gg 0$, so the theorem tells us that for every y with $0 \leq y \leq a - f(a)$, the equation $x - f(x) = y$ has a solution x . But replacing a with λa for any $\lambda > 0$ yields $\lambda a - f(\lambda a) = \lambda(a - f(a)) \gg 0$. Moreover, each component of $\lambda a - f(\lambda a)$ can be made as large as we please. Therefore, by Theorem 13.7.1, there is a solution x of $x - f(x) = y$ for all $y \geq 0$.

Now consider in particular the **Leontief case**, where f is linear, given by $f(x) = Ax$ for some $n \times n$ matrix $A = (a_{ij})_{n \times n}$ with nonnegative elements. If $v \geq u$, then $v - u \geq 0$ and $A(v - u) \geq 0$, or $Av \geq Au$. This shows that the function f is increasing, as well as obviously continuous, so Theorem 13.7.1 applies. If we want to get positive final supply of every good, as in Note 1, we are led to the concept of *productive matrices*:

PRODUCTIVE MATRICES

A nonnegative $n \times n$ matrix A is called **productive** if there exists a vector $a \gg 0$ such that $a \gg Aa$. (1)

For productive matrices, Theorem 13.7.1 can be significantly sharpened:

If A is productive, then for every $y \geq 0$ the equation $x - Ax = y$ has a unique solution, and this satisfies $x \geq y$. (2)

Existence, but not uniqueness, of x follows from the argument in Note 1 above. However, both existence and uniqueness follow from Theorem 13.7.2 below, which also enables us to find an explicit formula for x , namely $x = (I - A)^{-1}y$.

The next theorem gives us several ways to recognize productive matrices. Two of these involve a convergent sequence of square matrices, defined in the obvious way: If $\{B_k\}_k$ is a sequence of matrices, then $B_k \rightarrow B$ as $k \rightarrow \infty$ means that each element b_{ij}^k of B_k converges to the corresponding element b_{ij} of B .

THEOREM 13.7.2

For a nonnegative $n \times n$ matrix A the following statements are equivalent:

- (a) A is productive.
- (b) $A^m \rightarrow 0$ as $m \rightarrow \infty$.
- (c) $(I - A)^{-1} = I + A + A^2 + \dots$.
- (d) $(I - A)^{-1}$ exists and is nonnegative.

Proof: It suffices to prove that (a) $\Rightarrow$ (b) $\Rightarrow$ (c) $\Rightarrow$ (d) $\Rightarrow$ (a).

To prove (a) $\Rightarrow$ (b), choose a vector $a \gg 0$ such that $a \gg Aa$. Each component of a is then strictly larger than the corresponding component of Aa . Therefore, there exists a λ in $(0, 1)$ such that $\lambda a \gg Aa \gg 0$. Then $\lambda^2 a = \lambda(\lambda a) \gg \lambda Aa = A(\lambda a) \geq A(Aa) = A^2 a \gg 0$, and by induction we get $\lambda^m a \gg A^m a \gg 0$ for $m = 1, 2, \dots$. If we let $m \rightarrow \infty$, then $\lambda^m a \rightarrow 0$, because $\lambda \in (0, 1)$. Hence $A^m a \rightarrow 0$ as $m \rightarrow \infty$. But for each $j = 1, 2, \dots, n$, we have $A^m a = A^m (\sum_{i=1}^n a_i e_i) = \sum_{i=1}^n a_i A^m e_i \geq a_j A^m e_j$, and so the j th column $A^m e_j$ of A^m tends to 0 as a limit as $m \rightarrow \infty$. Therefore, $A^m \rightarrow 0$ as $m \rightarrow \infty$.

To prove (b) $\Rightarrow$ (c), note that, because $A^m \rightarrow 0$ and the determinant of a matrix is continuous in its elements, we have $|I - A^m| \rightarrow 1$ and so $|I - A^m| \neq 0$ for m sufficiently large. But $(I - A)(I + A + \dots + A^{m-1}) = I - A^m$, and so $|I - A| \neq 0$. It follows that $I - A$ is invertible and

$$I + A + \dots + A^{m-1} = (I - A)^{-1}(I - A^m)$$

Letting $m \rightarrow \infty$ yields the conclusion.

The implication (c) $\Rightarrow$ (d) is immediate.

Finally, to prove that (d) $\Rightarrow$ (a), choose any $y \gg 0$, and let $x = (I - A)^{-1}y$. Then (d) implies that $x \geq 0$, and $(I - A)x = y \gg 0$ implies $x \gg Ax$, so A is productive. ■

The last result of this section has several applications to economics. (See e.g. Nikaido (1970), Chapter 3.)

THEOREM 13.7.3 (PERRON-FROBENIUS)

Suppose that $A = (a_{ij})_{n \times n} \geq 0$, and define

$$\lambda_A = \inf \{ \mu : \mu > 0 \text{ and } \mu^{-1}A \text{ is productive} \}$$

Then $\lambda_A \geq 0$ is the largest real eigenvalue of A and has an associated nonnegative eigenvector. The eigenvalue λ_A is called the **Perron-Frobenius root** of A .

Proof: Let $\mathbf{1}$ denote the $n \times 1$ matrix with all elements equal to 1. For μ large enough, one has $\mu\mathbf{1} \gg \mathbf{A}\mathbf{1}$, so $\mathbf{I} \gg \mu^{-1}\mathbf{A}\mathbf{1}$. It follows that $\mu^{-1}\mathbf{A}$ is productive, so the set

$$J_A = \{ \mu : \mu > 0 \text{ and } \mu^{-1}\mathbf{A} \text{ is productive} \}$$

is nonempty. Moreover, if $\mu_0 > 0$ is such that $\mu_0^{-1}\mathbf{A}$ is productive, then $\mu^{-1}\mathbf{A}$ is productive for all $\mu > \mu_0$, and also for some μ slightly less than μ_0 . It follows that J_A is an open interval that is unbounded above; in fact, $J_A = (\lambda_A, \infty)$. It is clear that $\lambda_A \geq 0$.

Let $\mu > \lambda_A$, i.e. $\mu \in J_A$. Then $\mu > 0$ and $\mu^{-1}\mathbf{A}$ is productive, so by Theorem 13.7.2, $\mu\mathbf{I} - \mathbf{A} = \mu(\mathbf{I} - \mu^{-1}\mathbf{A})$ is invertible. Therefore, μ is not an eigenvalue of $\mathbf{A}$.

It remains to show that λ_A is an eigenvalue of $\mathbf{A}$ with an associated nonnegative eigenvector. Choose a sequence $\{\mu_k\}$ in J_A that converges to λ_A , and let $\mathbf{x}_k = (\mu_k\mathbf{I} - \mathbf{A})^{-1}\mathbf{1}$ for $k = 1, 2, \dots$. Then

$$(\mu_k\mathbf{I} - \mathbf{A})\mathbf{x}_k = \mathbf{1} \quad (i)$$

We demonstrate by contradiction that $\|\mathbf{x}_k\| \rightarrow \infty$ as $k \rightarrow \infty$. Indeed, if it did not, then the sequence $\{\|\mathbf{x}_k\|\}_k$ would have a bounded subsequence. But then, according to Theorem 13.2.4, $\{\mathbf{x}_k\}_k$ would have a convergent subsequence $\{\mathbf{x}_{k_j}\}_{j_j}$, with limit $\mathbf{x}^0 \geq \mathbf{0}$. Replace k by k_j in (i) and let $j \rightarrow \infty$. Then in the limit $(\lambda_A\mathbf{I} - \mathbf{A})\mathbf{x}^0 = \mathbf{1}$, so $\lambda_A\mathbf{x}^0 - \mathbf{A}\mathbf{x}^0 = \mathbf{1} \gg \mathbf{0}$. But that would imply that $\lambda_A > 0$ and that $\lambda_A^{-1}\mathbf{A}$ must be productive. In other words, λ_A would belong to J_A , but that is impossible. We conclude that $\|\mathbf{x}_k\| \rightarrow \infty$ as $k \rightarrow \infty$.

Put $\mathbf{y}_k = \mathbf{x}_k / \|\mathbf{x}_k\|$. Then $\|\mathbf{y}_k\| = 1$, and the sequence $\{\mathbf{y}_k\}$ has a convergent subsequence $\{\mathbf{y}_{k_i}\}_{i_i}$ converging as $i \rightarrow \infty$ to some $\mathbf{y}^0$ with $\|\mathbf{y}^0\| = 1$. Replacing k by k_i in (i) and dividing the equation by $\|\mathbf{x}_{k_i}\|$, we get

$$(\mu_{k_i}\mathbf{I} - \mathbf{A})\mathbf{y}_{k_i} = (1/\|\mathbf{x}_{k_i}\|)\mathbf{1}$$

Now let $i \rightarrow \infty$. It follows that $(\lambda_A\mathbf{I} - \mathbf{A})\mathbf{y}^0 = \mathbf{0}$, so $\mathbf{A}\mathbf{y}^0 = \lambda_A\mathbf{y}^0$. This equation shows that λ_A is an eigenvalue for $\mathbf{A}$, with an eigenvector $\mathbf{y}^0 \geq \mathbf{0}$. ■

NOTE 2 It can be shown that, if $\mathbf{A} \geq \mathbf{0}$, then $|\lambda_i| \leq \lambda_A$ for all eigenvalues λ of $\mathbf{A}$, whether real or complex. If $\mathbf{A} \gg \mathbf{0}$, then λ_A is a simple root of the eigenvalue equation and all other eigenvalues are strictly smaller in absolute value. See Berman and Plemmons (1994) or Horn and Johnson (1985) for more information.

PROBLEMS FOR SECTION 13.7

- (a) Show that $\mathbf{A} = \begin{pmatrix} 1/3 & 1/2 \\ 1/9 & 1/3 \end{pmatrix}$ is productive.
 (b) Show that if $\mathbf{A}$ is a nonnegative $n \times n$ matrix with all row sums less than 1, then $\mathbf{A}$ is productive.
- Find the Perron–Frobenius root and an associated nonnegative eigenvector for each of the following matrices.

$$(a) \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

$$(b) \begin{pmatrix} 1/3 & 1/2 \\ 1/9 & 1/3 \end{pmatrix}$$

$$(c) \begin{pmatrix} 2 & 1 & 1 \\ 4 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}$$

14

CORRESPONDENCES AND FIXED POINTS

Mathematics is like an addiction, or a disease; you can never truly shake it off, even if you want to.

—I. Stewart (1989)

A function f from a set A to a set B requires each element x of A to be mapped to exactly one element $f(x)$ in B . This is one of the most important concepts in mathematics and its applications. Many economic applications, however, use a generalization allowing any element in A to be mapped to a set consisting of several elements in B .

For example, suppose that $\mathbf{p}$ denotes a vector of parameters—for example, the list of prices at which an economic agent can buy or sell different commodities. Then $F(\mathbf{p})$ might be the set of the agent's optimal decisions given these parameters—for example, the set of a consumer's utility maximizing demands, or of a firm's profit maximizing production plans. Because optimal decisions may not always be unique, economists need to generalize the concept of a function in this way. Such “multi-valued” functions are usually called **correspondences**. Section 14.1 studies correspondences and some of their main properties.

The maximum theorems of Section 13.4 have a natural generalization to correspondences. Section 14.2 deals with such generalizations and also includes some economic applications.

Another focus of this chapter is on fixed point theorems. The brief Section 14.3 formulates and proves the existence of a fixed point for a special type of contraction mapping. This result has important applications to problems in the theory of differential equations, control theory and to infinite horizon dynamic programming.

Next, in Section 14.4, we study the fixed point theorems of Brouwer and Kakutani. In economics these results are widely used to prove that equilibrium exists in various models of perfectly competitive markets and in general strategic form games. These applications are regarded as major triumphs of mathematical economics. According to Scarf (1973), “This demonstration [of the existence of an equilibrium] has provided one of the rare instances in which abstract mathematical techniques are indispensable in order to solve a problem of central importance to economic theory”.

The final Section 14.5 proves the existence of an equilibrium in a pure exchange economy by using the Brouwer fixed point theorem.

14.1 Correspondences

This section considers correspondences and in particular introduces several continuity assumptions that have been found useful by economists. The rather intricate relationships between these different concepts are set out in some detail.

DEFINITION OF CORRESPONDENCES

A **correspondence** F from a set A into a set B is a rule that maps each x in A to a subset $F(x)$ of B . Then one writes $F : A \rightarrow B$ and $x \mapsto F(x)$ (with double arrows to distinguish a correspondence from a function).

(1)

The set $F(x)$ in the definition is allowed to be empty for some elements x in A . The **domain** of F is A , the set of all x for which $F(x)$ is defined; the **effective domain** is the set of all x in A at which $F(x)$ is nonempty.

Correspondences are also called **set-valued maps** or **multi-valued functions**. If the subset $F(x)$ always reduces to a single point, the correspondence is effectively the same as an ordinary function. The concept of a correspondence from $A \subseteq \mathbb{R}^3$ into $\mathbb{R}^2$ is illustrated in Fig. 1.

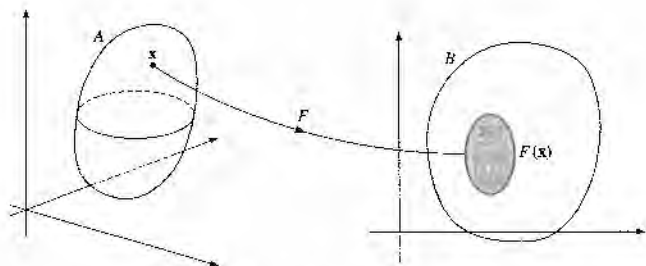


Figure 1 F is a correspondence from $A \subseteq \mathbb{R}^3$ to $\mathbb{R}^2$.

One familiar example of a correspondence is $y \mapsto f^{-1}(y)$, where f is an ordinary function from $\mathbb{R}$ to $\mathbb{R}$. Recall that $f^{-1}(y)$ is the pre-image set $\{x : f(x) = y\}$. It may well be empty for some x , or may contain more than one element unless f is one-to-one.

EXAMPLE 1 Example 2.2.2 defines a consumer's budget set

$$\mathcal{B}(\mathbf{p}, m) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{p} \cdot \mathbf{x} \leq m, \mathbf{x} \geq \mathbf{0}\} \quad (2)$$

for each price vector $\mathbf{p} \geq \mathbf{0}$ and income level $m \geq 0$. Thus, the budget set consists of all affordable nonnegative commodity vectors. Note that $(\mathbf{p}, m) \mapsto \mathcal{B}(\mathbf{p}, m)$ defines a correspondence from $\mathbb{R}_+^{n+1}$ into $\mathbb{R}_+^n$. See Fig. 2.

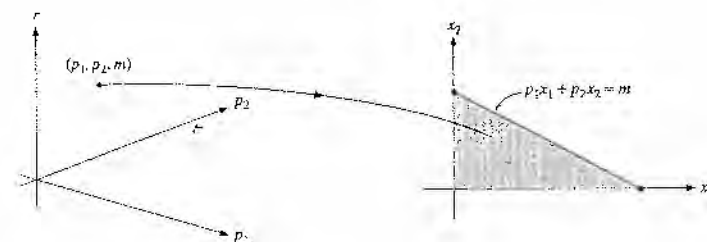


Figure 2 An individual's budget correspondence when there are two commodities.

EXAMPLE 2

Consider a firm producing a single commodity. Suppose that the total cost of production, as a function of output level $Q \geq 0$, is given by

$$C(Q) = \begin{cases} 0 & \text{if } Q = 0 \\ F + aQ + cQ^2 & \text{if } Q > 0 \end{cases}$$

where F , a , and c are positive constants with $P > a$. Note that $C(Q) > F$ whenever $Q > 0$. In this sense, F is a "fixed" or "setup" cost that must be incurred in order to produce any positive level of output.

Suppose the firm faces the output price P . Then its profit, as a function of Q , is given by

$$\pi(Q) = PQ - C(Q) = \begin{cases} 0 & \text{if } Q = 0 \\ -F + (P - a)Q - cQ^2 & \text{if } Q > 0 \end{cases}$$

For $Q > 0$, one has $\pi'(Q) = P - a - 2cQ$ and $\pi''(Q) = -2c < 0$. Note that $\pi'(Q) = 0$ for $Q = Q^* = (P - a)/2c$, with $\pi(Q^*) = (P - a)^2/4c - F$. We see that $\pi(Q^*) \geq 0$ if and only if $(P - a)^2 \geq (2\sqrt{cF})^2$, i.e. if and only if $P \geq a + 2\sqrt{cF}$. It follows that the profit maximizing choice of output is

$$Q(P) = \begin{cases} 0 & \text{if } P \leq a + 2\sqrt{cF} \\ (P - a)/2c & \text{if } P \geq a + 2\sqrt{cF} \end{cases}$$

yielding the profit level

$$\pi(Q(P)) = \begin{cases} 0 & \text{if } P \leq a + 2\sqrt{cF} \\ (P - a)^2/4c - F & \text{if } P \geq a + 2\sqrt{cF} \end{cases}$$

Note that the producer's behaviour is described by a **supply correspondence** rather than a function because, when $P = a + 2\sqrt{cF}$, both 0 and $\sqrt{F/c}$ are quantities giving the producer zero maximal profit. The supply correspondence is illustrated in Fig. 3 and also in Fig. 4, where the axes have been interchanged to conform with the standard convention in economics.

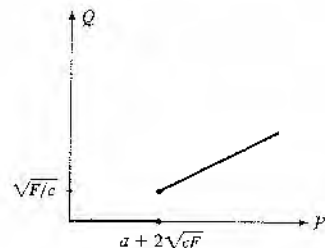


Figure 3

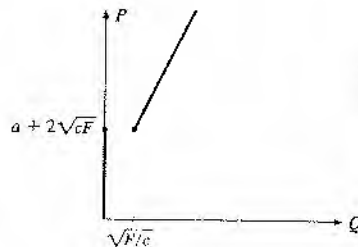


Figure 4

The graph of a correspondence $F : A \rightarrow B$ is defined as

$$\text{graph}(F) = \{(x, y) \in A \times B : x \in A \text{ and } y \in F(x)\} \quad (3)$$

If F is an ordinary function, its graph reduces to the familiar graph of a function.

Given a correspondence $F : A \rightarrow B$ and any set $S \subseteq A$, the range or image of S under F is defined as the set $F(S) = \bigcup_{x \in S} F(x)$.

EXAMPLE 3 Let F be the correspondence from $\mathbb{R}$ to $\mathbb{R}$ that maps every $x < 1$ to the interval $[1, 3]$, and every $x \geq 1$ to the set consisting only of the number 2. Hence,

$$F(x) = \begin{cases} [1, 3], & x < 1 \\ \{2\}, & x \geq 1 \end{cases}$$

Draw the graph of F .

Solution: The graph of F is shown in Fig. 5. The dashed line indicates boundary points that are not part of the graph.

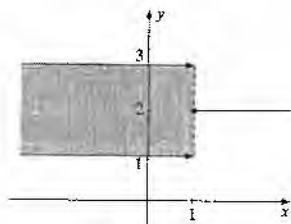


Figure 5 The correspondence F in Example 3 is lower hemicontinuous, but its graph is not closed.

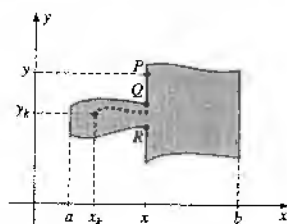


Figure 6 The correspondence has the closed graph property, but is not lower hemicontinuous.

Correspondences as relations

The graph of a correspondence from A to B is a subset of the Cartesian product $A \times B$. This subset can also be interpreted as a relation from A to B (see Appendix A). Indeed, for every correspondence $F : A \rightarrow B$ there is a unique relation R_F from A to B such that $a R_F b \iff b \in F(a)$. Similarly, to each relation R from A to B there is a correspondence $F_R : A \rightarrow B$ given by $F_R(a) = \{b \in B : a R b\}$.

The Closed Graph Property

The rest of this section will deal with correspondences that map points in $\mathbb{R}^n$ to subsets of $\mathbb{R}^m$.

Continuity is an important concept for ordinary functions. It ensures that small changes in the independent variable do not lead to large changes in the function value. For a correspondence F it is equally important to introduce certain continuity conditions ensuring that small changes in $\mathbf{x}$ do not change the image set $F(\mathbf{x})$ too drastically. However, unlike for (single-valued) functions, there are several different kinds of continuity for (multi-valued) correspondences. The distinctions between them are of some importance in economic theory. Of these different kinds of continuity, the simplest is the closed graph property:

CLOSED GRAPH PROPERTY OF A CORRESPONDENCE

A correspondence $F : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ has the **closed graph property at a point $\mathbf{x}^0$** in X if, whenever $\{x_k\}$ is a sequence in X that converges to $\mathbf{x}^0$, and $\{y_k\}$ is a sequence in $\mathbb{R}^m$ that satisfies $y_k \in F(x_k)$, $k = 1, 2, \dots$, and converges to $\mathbf{y}^0$, one has $\mathbf{y}^0 \in F(\mathbf{x}^0)$.

(4)

The correspondence F has the **closed graph property** if it has the closed graph property at every point in X , i.e. if for every convergent sequence $\{(x_k, y_k)\}$ in $\text{graph}(F)$ whose limit is a point $(\mathbf{x}^0, \mathbf{y}^0)$ in $X \times \mathbb{R}^m$, one has $\mathbf{y}^0 \in F(\mathbf{x}^0)$.

If a correspondence F has the closed graph property at $\mathbf{x}^0$, then in particular the set $F(\mathbf{x}^0)$ is closed. In the language of Section 13.3, $F : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ has the closed graph property if and only if $\text{graph}(F)$ is relatively closed in $X \times \mathbb{R}^m$, i.e. if and only if $\text{graph}(F)$ is the intersection of $X \times \mathbb{R}^m$ and a closed set in $\mathbb{R}^n \times \mathbb{R}^m$. In particular, if X is closed, then F has the closed graph property if and only if $\text{graph}(F)$ is closed.

Figure 6 shows the graph of a correspondence $F : [a, b] \rightarrow \mathbb{R}$ which *does* have the closed graph property. It is clear from Fig. 5 that the correspondence F in Example 3 *does not* have the closed graph property at $x = 1$.

EXAMPLE 4 Suppose that $g : \mathbb{R}^{n+m} \rightarrow \mathbb{R}$ is a continuous function. For all $\mathbf{x}$ in $\mathbb{R}^n$, define

$$\mathcal{P}(\mathbf{x}) = \{y \in \mathbb{R}^m : g(\mathbf{x}, y) \leq 0\}$$

Show that the correspondence $\mathbf{x} \mapsto \mathcal{P}(\mathbf{x})$ has the closed graph property.

Solution: The domain $\mathbb{R}^n$ of the correspondence $\mathcal{P}$ is closed. Its graph is the subset $g^{-1}(-\infty, 0] = \{(\mathbf{x}, y) : g(\mathbf{x}, y) \leq 0\}$ of $\mathbb{R}^{n+m}$. Because $(-\infty, 0]$ is closed and g is continuous, this inverse image is closed. Because $\mathcal{P}$ has a closed graph, it has the closed graph property.

Upper Hemicontinuity

A concept closely related to the closed graph property is upper hemicontinuity. Note 1 and Theorem 14.1.2 show the relationship between the two concepts.

UPPER HEMICONTINUOUS CORRESPONDENCES

A correspondence $F : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ is said to be **upper hemicontinuous** (or **u.h.c.**) at a point x^0 in X if for every open set U that contains $F(x^0)$ there exists a neighbourhood N of x^0 such that $F(x) \subseteq U$ for every x in $N \cap X$, i.e. such that $F(N \cap X) \subseteq U$.

F is **upper hemicontinuous** (or **u.h.c.**) in X if it is u.h.c. at every x in X .

(5)

The following result is an immediate implication of definition (5):

THEOREM 14.1.1 (CONTINUOUS FUNCTIONS ARE U.H.C. CORRESPONDENCES)

The function $x \mapsto f(x)$ is continuous at a point x^0 of its domain X if and only if the associated correspondence $x \mapsto \{f(x)\}$ is upper hemicontinuous at x^0 .

Proof: Let U be a open set containing $f(x^0)$ and let N be a neighbourhood of x^0 . Note that $f(N \cap X) \subseteq U \iff N \cap X \subseteq f^{-1}(U)$. Because of Theorem 13.3.5(a), the theorem follows from this equivalence. ■

NOTE 1 If $F : X \rightarrow \mathbb{R}^m$ is upper hemicontinuous at a point x^0 in X and $F(x^0)$ is a closed set, then $F(x)$ has the closed graph property at x^0 . (*Proof:* Let $x_k \rightarrow x^0$, $y_k \rightarrow y^0$, and $y_k \in F(x_k)$. Consider any $y \notin F(x^0)$. There is a closed ball B around y which is small enough not to intersect the closed set $F(x^0)$. Applying (5) to the open set $\mathbb{C}B = \mathbb{R}^m \setminus B$, there exists a neighbourhood N of x^0 such that $F(N \cap X) \subseteq \mathbb{C}B$. But for k large, $x_k \in N \cap X$ and so $y_k \in F(x_k) \subseteq \mathbb{C}B$. Therefore y_k does not converge to y . Hence $y^0 \in F(x^0)$.)

On the other hand, the following theorem also holds:

THEOREM 14.1.2

Suppose that the correspondence $F : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ has the closed graph property at x^0 and that F is locally bounded near x^0 in the sense that there exists a neighbourhood N of x^0 such that $F(N \cap X)$ is bounded. Then F is upper hemicontinuous at x^0 .

Proof: Let N be as in the theorem, and let $B(x^0; \alpha) \subseteq N$. Suppose that F is not u.h.c. at x^0 in X . Then there must exist an open set $U \supseteq F(x^0)$ such that, given any ball $B(x^0; \alpha/k)$, $k = 1, 2, \dots$, there exists an x_k in $B(x^0; \alpha/k) \cap X$ for which $F(x_k) \not\subseteq U$. Choose arbitrary vectors y_k in $F(x_k) \setminus U$. The boundedness property in the theorem implies that $\{y_k\}$ has a convergent subsequence $\{y_{k_r}\}$, which converges to some point y^0 in $F(x^0)$ as $r \rightarrow \infty$, by the closed graph property. Because $y_{k_r} \in \mathbb{R}^m \setminus U$ and $\mathbb{R}^m \setminus U$ is closed, $y^0 = \lim_{r \rightarrow \infty} y_{k_r} \in \mathbb{R}^m \setminus U$. Hence $y^0 \notin F(x^0)$, contradicting the closed graph property at x^0 . ■

The following important result is an immediate corollary:

THEOREM 14.1.3 (COMPACT GRAPH TEST FOR UPPER-HEMICONTINUITY)

If a correspondence F from $X \subseteq \mathbb{R}^n$ to $Y \subseteq \mathbb{R}^m$ has a compact graph, then it is upper hemicontinuous.

Problem 1 concerns an example of a correspondence, some of whose values are not compact, that has the closed graph property, but is not upper hemicontinuous. Theorem 14.1.1 shows that, if a correspondence is single-valued and so collapses to a function, then upper hemicontinuity implies that the function is continuous.

Lower Hemicontinuity

Another frequently encountered continuity condition for correspondences is the following:

LOWER HEMICONTINUOUS CORRESPONDENCE

A correspondence $F : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ is **lower hemicontinuous** (or **l.h.c.**) at a point x^0 in X if, whenever $y^0 \in F(x^0)$ and $\{x_k\}$ is a sequence in X that converges to x^0 , there exist a number k_0 and a sequence $\{y_k\}_{k=k_0}^\infty$ in $\mathbb{R}^m$ that converges to y^0 and satisfies $y_k \in F(x_k)$ for all $k \geq k_0$. (6)

F is **lower hemicontinuous** (or **l.h.c.**) in X if it is l.h.c. at every point x in X .

Lower hemicontinuity requires a correspondence to be continuous in a sense that is almost the opposite of upper hemicontinuity. For example, the correspondence shown in Fig. 5 is lower but not upper hemicontinuous. Yet the opposite holds for the correspondence in Fig. 6. (To see why the correspondence in Fig. 6 is not lower hemicontinuous, consider the point P on the graph, and let the sequence $\{x_k\}$ be as suggested by the dots in the figure. In particular, suppose $\{x_k\}$ converges to x . It is obviously impossible to choose a sequence $y_k \in F(x_k)$ that converges to y , because the corresponding sequence $\{(x_k, y_k)\}$ can only converge to a point on the line segment RQ , and not to P .)

Roughly speaking, if a correspondence F is upper hemicontinuous at a point x^0 of its domain, then $F(x)$ cannot "explode" as x moves slightly away from x^0 , as happens at $x = 1$ in Fig. 5, but it may "implode", as happens at x in Fig. 6. For lower hemicontinuous correspondences the opposite is true: $F(x)$ cannot implode as x moves slightly away from x^0 , but it may explode.

EXAMPLE 5 Let $g = (g_1, \dots, g_l) : \mathbb{R}^{n+m} \rightarrow \mathbb{R}^l$ be continuous, let b be a given l -vector, let A be a given closed set in $\mathbb{R}^n$, and let X be a given set in $\mathbb{R}^n$. For each x in X , define the set $\mathcal{P}(x) \subseteq \mathbb{R}^m$ by

$$\mathcal{P}(x) = \{y \in A : g(x, y) \leq b\} = \{y \in A : g_i(x, y) \leq b_i, i = 1, \dots, l\} \quad (7)$$

Show that the correspondence $x \mapsto \mathcal{P}(x)$ has the closed graph property and is also lower hemicontinuous if $\mathcal{P}(x) = \overline{\mathcal{P}^o(x)}$ for every x in X , where $\mathcal{P}^o(x) = \{y \in A : g(x, y) < b\}$.

Solution: Proof of the closed graph property: Assume that $x_k \rightarrow x^0 \in X$ and $y_k \rightarrow y^0$ when $k \rightarrow \infty$, where $y_k \in \mathcal{P}(x_k)$ for all k . Then $y_k \in A$ and $g(x_k, y_k) \leq b$. Because A is closed, $y^0 \in A$. By continuity of g , letting $k \rightarrow \infty$ in the inequality yields $g(x^0, y^0) \leq b$. It follows that $y^0 \in \mathcal{P}(x^0)$.

Proof of lower hemicontinuity: Let $y^0 \in \mathcal{P}(x^0)$ and let $x_k \rightarrow x^0 \in X$. Because $\mathcal{P}(x^0) = \overline{\mathcal{P}^o(x^0)}$, there exist vectors y^j in $\mathcal{P}^o(x^0)$ such that $\|y^j - y^0\| < 1/j$ for $j = 1, 2, \dots$. Because $g_i(x^0, y^j) < b_i$ for $i = 1, \dots, l$, there exists a strictly increasing sequence of numbers k_j such that for $k \geq k_j$, the inequality $g_i(x_k, y^j) < b_i$ holds for all i . Let $y_k = y^j$ for $k_j \leq k < k_{j+1}$. Then $\|y_k - y^0\| \leq 1/j$ for all $k \geq k_j$, so $y_k \rightarrow y^0$. Also, y_k belongs to $\mathcal{P}(x_k)$ for $k \geq k_j$, thus confirming lower hemicontinuity at x^0 . $\blacksquare$

EXAMPLE 6 Show that the budget correspondence $\mathcal{B}(p, m)$ defined in Example 1 has the closed graph property and is lower hemicontinuous at any point (p, m) where $m > 0$.

Solution: The closed graph property follows immediately from the previous example (with $A = \{x : x \geq 0\}$).

To prove that the correspondence is lower hemicontinuous when $m > 0$, because of Example 5 it is enough to show that $\mathcal{B}(p, m) \subseteq \overline{\mathcal{B}^o(p, m)}$, where $\mathcal{B}^o(p, m) = \{x \geq 0 : p \cdot x < m\}$. Given any x in $\mathcal{B}(p, m)$, let $\alpha_k = (1 - 1/k)$, $k = 1, 2, \dots$. Then $\alpha_k x \rightarrow x$ as $k \rightarrow \infty$. Moreover, because $p \cdot x \leq m$, $0 < \alpha_k < 1$, and $m > 0$, one has $p \cdot \alpha_k x \leq \alpha_k m < m$, so $\alpha_k x$ belongs to $\mathcal{B}^o(p, m)$, and $x = \lim_k \alpha_k x \in \overline{\mathcal{B}^o(p, m)}$, which proves the asserted inclusion. $\blacksquare$

Here is an alternative condition for lower hemicontinuity:

ALTERNATIVE CHARACTERIZATION OF LOWER HEMICONTINUITY

A correspondence $F : X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ is lower hemicontinuous at x^0 in X if and only if for each y^0 in $F(x^0)$ and each neighbourhood U of y^0 , there exists a neighbourhood N of x^0 such that $F(x) \cap U \neq \emptyset$ for all x in $N \cap X$. $\quad (8)$

Proof: Let us show first that F is lower hemicontinuous at a point x^0 in X if condition (8) holds. Let $y^0 \in F(x^0)$ and let $x_k \rightarrow x^0$ as $k \rightarrow \infty$. For each ball $B(y^0, 1/j)$, $j = 1, 2, \dots$, the condition in (8) implies that there exists a neighbourhood N_j of x^0 such that $F(x) \cap B(y^0, 1/j) \neq \emptyset$ whenever $x \in N_j \cap X$. There exists a k_j such that $k \geq k_j$ implies $x_k \in N_j \cap X$, so there exists a $y_k^j \in F(x_k)$ such that $y_k^j \in B(y^0, 1/j)$, $k = k_j, k_j + 1, \dots$. Obviously $\{k_j\}$ may be chosen to be increasing. Put $y_k = y_k^j$ when $k_j \leq k < k_{j+1}$. Clearly $y_k \rightarrow y^0$, so F satisfies (6) at x^0 .

To prove the reverse implication, suppose F does not satisfy (8) at a point x^0 in X . Then there exist a point $y^0 \in F(x^0)$ and a neighbourhood U of y^0 such that every ball $B(x^0, 1/j)$, $j = 1, 2, \dots$, includes at least one point x_j satisfying $F(x_j) \cap U = \emptyset$. Hence, there exists a sequence $\{x_j\}$ in X such that $x_j \rightarrow x^0$ as $j \rightarrow \infty$ and $F(x_j) \cap U = \emptyset$ for all $j \geq 1$. But then no sequence $\{y_j\}$ with $y_j \in F(x_j)$, $j = 1, 2, \dots$, can possibly converge to y^0 . So F cannot satisfy (6) at x^0 . $\blacksquare$

Comparing (8) with the corresponding topological condition for continuous functions leads immediately to the following important result:

THEOREM 14.1.4 (CONTINUOUS FUNCTIONS ARE L.H.C. CORRESPONDENCES)

A function $x \mapsto f(x)$ is continuous at a point x^0 of its domain X if and only if the associated correspondence $x \mapsto \{f(x)\}$ is lower hemicontinuous at x^0 .

Proof: Given any neighbourhood U of $\{f(x^0)\}$, one has $\{f(x)\} \cap U \neq \emptyset \iff f(x) \in U$. So $\{f(x)\} \cap U \neq \emptyset$ for all x in $N \cap X$ if and only if $f(N \cap X) \subseteq U$, i.e. if and only if $N \cap X \subseteq f^{-1}(U)$. The result follows from (8) because the topological condition of Theorem 13.3.5 for f to be continuous at x^0 is that $N \cap X \subseteq f^{-1}(U)$ for some neighbourhood N of x^0 whenever U is a neighbourhood of $\{f(x^0)\}$. $\blacksquare$

A correspondence that is both upper and lower hemicontinuous is called **continuous**. Of course, any constant-valued correspondence is continuous, as is any single-valued correspondence that collapses to a continuous function. But so are many others.

Composite Correspondences

If $f : A \rightarrow B$ and $g : B \rightarrow C$ are functions, then the composition $h = g \circ f$ is the function $h : A \rightarrow C$ given by $h(x) = g(f(x))$. Similarly, if $F : A \rightarrow B$ and $G : B \rightarrow C$ are correspondences, then the **composite correspondence** $H = G \circ F : A \rightarrow C$ is defined by

$$H(x) = G(F(x)) = \bigcup_{y \in F(x)} G(y)$$

Thus, $H(x)$ is the union of all the sets $G(y)$ that are obtained as y runs through $F(x)$. This means that $z \in H(x)$ if and only if $z \in G(y)$ for at least one y in $F(x)$.

Recall that the composition of two continuous functions is continuous. For correspondences between sets in Euclidean spaces we have the following results:

THEOREM 14.1.5

Let $F : X \subseteq \mathbb{R}^n \rightarrow Y \subseteq \mathbb{R}^m$ and $G : Y \rightarrow Z \subseteq \mathbb{R}^p$ be correspondences, and let $H = G \circ F : X \rightarrow Z$ be their composition. Then:

- If F and G have the closed graph property and Y is compact, then H has the closed graph property.
- If F is upper hemicontinuous at x^0 in X and G is upper hemicontinuous at every point of $F(x^0)$, then H is upper hemicontinuous at x^0 .
- If F is lower hemicontinuous at x^0 in X and G is lower hemicontinuous at every point of $F(x^0)$, then H is lower hemicontinuous at x^0 .

Proof: (a) Suppose that $\{x_k\}$ and $\{z_k\}$ are convergent sequences in X and Z respectively, such that $z_k \in H(x_k)$ for all k , $x^0 = \lim_k x_k \in X$, and $z^0 = \lim_k z_k$. We must show that $z^0 \in H(x^0)$. The

definition of a composite correspondence implies that for each k there exists a y_k in $F(x_k)$ such that $z_k \in G(y_k)$. Because Y is compact, $\{y_k\}$ has a convergent subsequence $\{y_{k_r}\}$. Let $y^0 = \lim y_{k_r}$. The corresponding subsequences $\{x_{k_r}\}$ and $\{z_{k_r}\}$ of $\{x_k\}$ and $\{z_k\}$ converge to x^0 and z^0 respectively. Because F and G have the closed graph property, and because $y_{k_r} \in F(x_{k_r})$ for all r , it follows that $y^0 \in F(x^0)$ and $z^0 \in G(y^0)$. Hence, $z^0 \in H(x^0)$.

(b) Let U be any open set containing $H(x^0)$. Since $H(x^0) = \bigcup_{y \in F(x^0)} G(y)$, the set U contains $G(y)$ for every $y \in F(x^0)$. Because G is u.h.c., for each such y there exists an open neighbourhood N_y of y such that $G(y') \subseteq U$ whenever $y' \in N_y \cap Y$. Define $N^* = \bigcup_{y \in F(x^0)} N_y$. As the union of open sets, this is an open set containing $F(x^0)$. Because F is u.h.c. at x^0 , there must exist a neighbourhood N of x such that $F(x) \subseteq N^* \cap Y$ whenever $x \in N \cap X$. But then, for all such x , one has

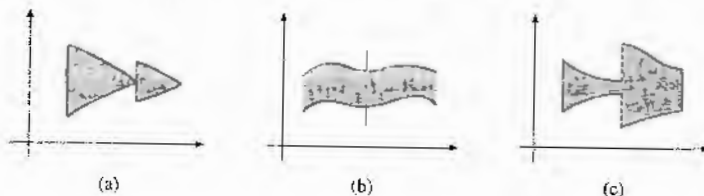
$$H(x) = \bigcup_{y \in F(x)} G(y) \subseteq \bigcup_{y' \in N^* \cap Y} G(y') = \bigcup_{z \in F(x^0)} \left[\bigcup_{y' \in N_y \cap Y} G(y') \right] \subseteq U$$

This confirms that H is u.h.c. at x^0 .

(c) This (much easier) proof is left to the reader—see the Student's Manual.

PROBLEMS FOR SECTION 14.1

1. Let $F(x) = \{1/x\}$ for $x \neq 0$, with $F(0) = \{0\}$. Prove that f has the closed graph property, but is not upper hemicontinuous.
2. Determine by a geometric argument whether or not the correspondences given by the following graphs have the closed graph property and/or are lower hemicontinuous.



3. Let the correspondence $F: \mathbb{R}_+ = [0, \infty) \rightarrow \mathbb{R}_+$ be given by $F(x) = [0, 1/x]$ for $x > 0$ and $F(0) = \mathbb{R}_+$. Prove that F has the closed graph property and is lower hemicontinuous.
4. If $F: X \rightarrow \mathbb{R}^l$ and $G: X \rightarrow \mathbb{R}^m$ are correspondences defined on a set $X \subseteq \mathbb{R}^n$, the product correspondence $H = F \times G: X \rightarrow \mathbb{R}^{l+m}$ is defined by $H(x) = F(x) \times G(x)$ for all x in X . Prove that if F and G are lower hemicontinuous at x^0 , then H is also lower hemicontinuous at x^0 . Similarly, prove that if F and G are compact-valued and also upper hemicontinuous at x^0 , then H is also upper hemicontinuous at x^0 .
5. Suppose that the two compact-valued correspondences $F, G: X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ are upper hemicontinuous at a point x^0 in X . Consider the summation correspondence $H: X \rightarrow \mathbb{R}^m$ defined by $H(x) = F(x) + G(x)$ for all x in X . Prove that H is upper hemicontinuous at x^0 . What may go wrong if F and G are not compact-valued?

SM 6. Suppose $P: X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ is lower hemicontinuous at a point x^0 in X . Let $G: X \rightarrow \mathbb{R}^m$ be the correspondence whose value at each x in X is the convex hull $\text{co}(F(x))$. Prove that G is l.h.c. at x^0 .

7. (Sequence test for upper hemicontinuity) Let $F: X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a compact-valued correspondence. Prove the following result: F is upper hemicontinuous at x^0 in X if, whenever $\{x_k, y_k\}$ is a sequence of points in $\text{graph}(F)$ for which $x_k \rightarrow x^0$ as $k \rightarrow \infty$, the corresponding sequence $\{y_k\}$ has a subsequence converging to a point of $F(x^0)$. (The converse is also true.)

HARDER PROBLEMS

8. Let the functions $g_i(x, y)$ in Example 5 be continuous in (x, y) and convex in y . Furthermore, suppose that $\{y: g_i(x, y) < b_i \text{ for } i = 1, \dots, l\}$ is nonempty for all x . Show that the correspondence $\mathcal{P}$ defined in Example 5 is then lower hemicontinuous. (Hint: Use the result in Example 5 and, for any y' in $\mathcal{P}(x)$, take y in $\mathcal{P}^o(x)$ and show that $y'' = \lambda y + (1 - \lambda)y' \in \mathcal{P}^o(x)$ for λ in $(0, 1)$, then let $\lambda \rightarrow 0$.)
9. Let the functions $g_i(x, y)$ in Example 5 be continuous, and have continuous partial derivatives w.r.t. $y_1, \dots, y_m$. Furthermore, suppose that for all pairs (x, y) with $g(x, y) \leq b$, the rank of the matrix with entries $\partial g_i(x, y) / \partial x_j$, where $j = 1, \dots, n$ and $i \in S(x, y) = \{i: g_i(x, y) = b_i\}$ is equal to the number of elements in the set $S(x, y)$. Prove that the correspondence $\mathcal{P}(x)$ is then lower hemicontinuous. (Hint: Use Example 5.)
- SM 10. Let $a(x)$ and $b(x)$ be two continuous functions mapping $\mathbb{R}$ into $\mathbb{R}$, with $a(x) \leq b(x)$ for all x . Of all the different possible correspondences $F: \mathbb{R} \rightarrow \mathbb{R}$ that satisfy

$$(a(x), b(x)) \subseteq F(x) \subseteq [a(x), b(x)]$$
 for all x , which are lower hemicontinuous, and which are upper hemicontinuous? (Hint: First examine the case when $a(x)$ and $b(x)$ are both constants.)
- SM 11. Let $F: X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^m$ be a compact-valued correspondence and let $G: X \rightarrow \mathbb{R}^m$ be the correspondence whose value at each x in X is the convex hull $\text{co}(F(x))$. Prove that if F is upper hemicontinuous at x^0 , then G is also upper hemicontinuous at x^0 . (Hint: Apply Carathéodory's theorem.)

14.2 A General Maximum Theorem

Assume that $F(x)$ is a correspondence from $X \subseteq \mathbb{R}^n$ into $Y \subseteq \mathbb{R}^m$, and let $f(x, y)$ be a function from $X \times Y$ into $\mathbb{R}$. Consider the maximization problem

$$\text{maximize } f(x, y) \text{ subject to } y \in F(x) \quad (1)$$

Define the choice or behaviour correspondence F^* from X into Y by¹

$$F^*(x) = \arg \max_{y \in F(x)} f(x, y) = \{y \in F(x) : f(x, z) \leq f(x, y) \text{ for all } z \text{ in } F(x)\} \quad (2)$$

¹ In general, for a function $\phi: S \rightarrow \mathbb{R}$ the notation $\arg \max_{s \in S} \phi(s)$ ($\arg \min_{s \in S} \phi(s)$) is used to denote the set of all values of the argument s in S that maximize (minimize) $\phi(s)$.

Thus, for each x in X , the set $Y^*(x)$ consists of all the values of the argument y that maximize $f(x, y)$ as y runs through the set $F(x)$. The problem is illustrated in Figure 1.

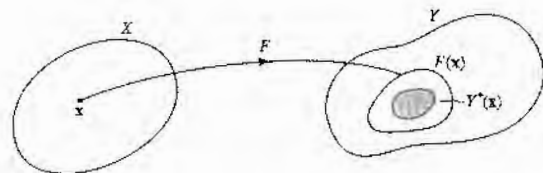


Figure 1

Also define the corresponding value function:

$$V(x) = \sup_{y \in F(x)} f(x, y) \quad (3)$$

The function V is well defined on the effective domain of F . If for some x the supremum is attained at some $\hat{y}$, then $V(x) = f(x, \hat{y}) = \max_{y \in F(x)} f(x, y)$. Note that if $x \in X$ and $y \in Y^*(x)$, then $V(x) = f(x, y)$. In fact, $Y^*(x) = \{y \in F(x) : f(x, y) = V(x)\}$.

Here is a general (but somewhat vague) economic interpretation. First, x is a vector of exogenous parameters that jointly describe the "environment" faced by some maximizing economic agent. Given this vector x , the feasible set $F(x)$ describes what options are available. Let $f(x, y)$ measure the benefit to the economic agent from choosing the point y in $F(x)$ in situation x . Then $Y^*(x)$ is the set of choices of y that maximize benefit.

In general, we are interested in finding the strongest possible continuity properties of $Y^*(x)$ and $V(x)$. Here is the main result:

THEOREM 14.2.1 (THE MAXIMUM THEOREM)

Suppose that $x \mapsto F(x)$ is a correspondence from $X \subseteq \mathbb{R}^n$ into $Y \subseteq \mathbb{R}^m$ that has nonempty compact values for all x in X and is continuous (i.e. both upper and lower hemicontinuous) at x^0 in X . Let $f(x, y)$ be a continuous function from $X \times Y$ into $\mathbb{R}$. Define

$$Y^*(x) = \{y \in F(x) : f(x, z) \leq f(x, y) \text{ for all } z \in F(x)\}$$

Then $Y^*(x)$ is nonempty for all x in X . Moreover, the choice correspondence $x \mapsto Y^*(x)$ is u.h.c. at x^0 , and the value function $V(x)$ is continuous at x^0 .

Proof: Because F has nonempty compact values for all x in X , the extreme-value theorem implies that $Y^*(x) \neq \emptyset$ for all x in X . Also, because $F(x^0)$ is compact, there exists an open and bounded set W containing $F(x^0)$, and since F is assumed to be upper hemicontinuous, there is an open ball U around x^0 such that $F(x) \subseteq W$ for all x in U . Then for every x in U and every y in $F(x)$, the point (x, y) belongs to the compact set $\bar{U} \times \bar{W}$.

Upper hemicontinuity of Y^ :* (a) Applying Theorem 14.1.2 (with $X = U$ and $F = Y^*$), upper hemicontinuity of $Y^*(x)$ at x^0 will follow provided we can show that Y^* has the closed graph property at x^0 . Let $\{x_k\}$ be any sequence of points in X that converges to x^0 as $k \rightarrow \infty$ and let $y_k \in Y^*(x_k)$ converge to y^0 . Because F has the closed graph property at x^0 and $y_k \in Y^*(x_k) \subseteq F(x_k)$ for each k , we have $y^0 \in F(x^0)$.

(b) Take an arbitrary z^0 in $F(x^0)$. Because F is lower hemicontinuous at x^0 and $x_k \rightarrow x^0$, there exists a sequence $\{z_k\}$ with $z_k \in F(x_k)$ and $z_k \rightarrow z^0$. Now $y_k \in Y^*(x_k)$ and $z_k \in F(x_k)$, so the definition of Y^* implies that $f(x_k, z_k) \leq f(x_k, y_k)$. Taking limits as $k \rightarrow \infty$, continuity of f implies that $f(x^0, z^0) \leq f(x^0, y^0)$. Because the choice of z^0 in $F(x^0)$ was arbitrary, it follows that $y^0 \in Y^*(x^0)$. Hence Y^* does have the closed graph property at x^0 .

Continuity of V : Let $\{x_k\}$ be a sequence in X that converges to x^0 , and assume for a contradiction that $V(x_k)$ does not converge to $V(x^0)$. (We can assume $x_k \in U$ for all k .) Then there exist an $\varepsilon > 0$ and a subsequence $\{x_{k_r}\}$ such that $|V(x_{k_r}) - V(x^0)| > \varepsilon$ for all r . Let $y_{k_r} \in Y^*(x_{k_r})$. The points (x_{k_r}, y_{k_r}) all belong to $\bar{U} \times \bar{W}$, so by Theorem 13.2.5 the sequence $\{(x_{k_r}, y_{k_r})\}$ has a convergent "subsubsequence" $\{(x'_m, y'_m)\}_m$, whose limit must be a pair (x^0, y') with $y' \in Y^*(x^0)$, and so $f(x^0, y') = V(x^0)$. Obviously, $y'_m \in Y^*(x'_{k_r})$ for each m , and so $V(x'_{k_r}) = f(x'_{k_r}, y'_m)$. Since f is continuous, $V(x'_{k_r}) = f(x'_{k_r}, y'_m) \rightarrow f(x^0, y') = V(x^0)$. But $|V(x'_{k_r}) - V(x^0)| > \varepsilon$, which yields a contradiction. ■

NOTE 1 Suppose there is a real-valued continuous function $\alpha(x)$ such that the correspondence $x \mapsto F^+(x) = F(x) \cap \{y \in Y : f(x, y) \geq \alpha(x)\}$ has nonempty and compact values for all x in X , and is upper hemicontinuous at x^0 . Then in Theorem 14.2.1, upper hemicontinuity of $F(x)$ at x^0 and compactness of $F(x)$ can be dropped. Lower hemicontinuity of $F(x)$ must be kept, however. Note that the set $Y^*(x)$ in the theorem equals $\{y \in F^+(x) : f(x, z) \leq f(x, y) \text{ for all } z \in F^+(x)\}$. To prove upper hemicontinuity of Y^* at x^0 , we now let W be an open and bounded set containing $F^+(x^0)$ and let U be an open ball around x^0 such that $F^+(x) \subseteq W$ for all x in U . Then replace F by F^+ in part (a) of the proof above.

EXAMPLE 1

In connection with Example 13.4.1 define the correspondence F for all x by $F(x) = [-1, 2]$. Then all the assumptions in Theorem 14.2.1 are satisfied and the set $Y^*(x)$ is $Y^*(x) = \{y \in [-1, 2] : xz^2 \leq xy^2 \text{ for all } z \in [-1, 2]\}$. It follows that

$$Y^*(x) = \begin{cases} \{0\} & \text{when } x < 0 \\ [-1, 2] & \text{when } x = 0 \\ \{2\} & \text{when } x > 0 \end{cases}$$

(For instance, if $x < 0$, then $Y^*(x)$ consists of all y in $[-1, 2]$ such that $z^2 \geq y^2$ for all z in $[-1, 2]$. In particular, $y^2 \leq 0$, so $y = 0$.) The correspondence $x \mapsto Y^*(x)$ is upper hemicontinuous. (Draw a figure!)

EXAMPLE 2

Let K be a nonempty compact convex set in $\mathbb{R}^n$. For all x in $\mathbb{R}^n$, define

$$\delta(x, K) = \min_{y \in K} d(x, y), \quad \psi(x) = \arg \min_{y \in K} d(x, y)$$

where $d(x, y)$ is the Euclidean distance. Thus $\delta(x, K)$ can be interpreted as the minimum distance from x to K , and $\psi(x)$ as the set of closest points. Since $[d(x, y)]^2$ is a strictly convex function of y , it has a unique minimum over the convex set K of possible values of y . Hence, $\psi(x)$ is single-valued, i.e. $\psi(x) = \{y^*(x)\}$ for some function $y^*(x)$. In addition, the correspondence $x \mapsto K$ is constant-valued, so continuous. Hence, the maximum theorem applies. The correspondence $x \mapsto \psi(x)$ is therefore u.h.c., implying that the function $x \mapsto y^*(x)$ is continuous. In addition, $x \mapsto \delta(x, K) = d(x, y^*(x))$ is also continuous. $\square$

EXAMPLE 3 (Profit maximization) Suppose a firm produces a single output commodity using n different factors of production as inputs. Let the vector of strictly positive unit prices for the inputs be $w = (w_1, \dots, w_n) \gg 0$. Suppose that the firm's minimum cost, when faced with input prices $w \gg 0$ and producing output $y \geq 0$, is given by the function $C(w, y)$. One expects C to be increasing in w in the sense that $C(w', y) \geq C(w, y)$ whenever $w' \geq w$, and that $C(w, 0) = 0$. This will be assumed, as well as that $C(w, y)$ is a continuous function. One also expects C to be increasing in y and homogeneous of degree one in w , in the sense that $C(\lambda w, y) = \lambda C(w, y)$ for all $\lambda > 0$. All these properties are simply assumed here; their plausibility is demonstrated in the next example.

Finally, in order to ensure that profits remain bounded, assume that for each fixed $w \gg 0$ the average cost $C(w, y)/y$ tends to ∞ as $y \rightarrow \infty$.

Let $p > 0$ denote the price per unit of output. Consider the problem of maximizing the firm's profit $py - C(w, y)$ by an appropriate choice of output y . We are interested in showing that the *supply correspondence* $(p, w) \mapsto \eta(p, w) = \arg \max_y \{py - C(w, y)\}$ is u.h.c., and that the *profit function* $(p, w) \mapsto \pi(p, w) = \max_y \{py - C(w, y)\}$ is continuous. It seems that the maximum theorem should be helpful, but there is a difficulty because the relevant feasible set $\{y : y \geq 0\}$ is not compact. However, define the set $F^+(p, w) = \{y \geq 0 : py - C(w, y) \geq 0\}$. It is nonempty because $0 \in F^+(p, w)$. Choose any $p^0 > 0$ and $w^0 \gg 0$. Because average cost $C(w^0/2, y)/y \rightarrow \infty$ as $y \rightarrow \infty$, by hypothesis, a number y^* can be so chosen that $C(w^0/2, y) > 2p^0 y$ for all $y > y^*$. We claim that $F^+(p, w) \subseteq [0, y^*]$ for $w \gg w^0/2$, $p < 2p^0$. To see this, note that $y > y^*$ gives negative profits because $py - C(w, y) < 2p^0 y - C(w^0/2, y) < 0$. Arguing as in Example 14.1.5, $F^+(p, w)$ has the closed graph property for $w \gg w^0$ and $p < p^0$. Because $[0, y^*]$ is bounded, Theorem 14.1.2 implies that $F^+(p, w)$ is upper hemicontinuous for all such (p, w) . Finally, applying Note 1 with $F(p, w) = [0, \infty)$ shows that $\pi(p, w)$ is continuous and $\eta(p, w)$ is upper hemicontinuous for $p < 2p^0$ and $w \gg w^0/2$. The same result extends to all $p > 0$ and $w \gg 0$, since p^0 and w^0 were arbitrary. $\square$

EXAMPLE 4 (Cost minimization) Consider the same firm as in the previous example. Suppose that the level of output is determined by the production function $f(x)$, where $x = (x_1, \dots, x_n)$ is the input vector. Suppose also that f is defined and continuous on the set $\mathbb{R}_+^n = \{x \in \mathbb{R}^n : x \geq 0\}$, that $f(0) = 0$, and that f is also *monotone* in the sense that $f(x') \geq f(x)$ whenever $x' \geq x$, with $f(x') > f(x)$ whenever $x' \gg x$.

Consider the set $Y = f(\mathbb{R}_+^n) \subseteq \mathbb{R}$ of all possible output levels that the firm can produce. It must be an interval of the form $[0, \bar{y})$, where $\bar{y}$ may be $+\infty$. (There can be no $\bar{x}$ such that $f(\bar{x}) = \bar{y}$ because $f(x) > f(\bar{x})$ whenever $x \gg \bar{x}$.) Given any input price vector $w \gg 0$, the firm's (total) cost is given by $w \cdot x$. We shall now study the firm's *cost function*, which specifies the minimum cost of producing a given output level y in Y when the input price vector is w . It is given by

$$C(w, y) = \min_x \{w \cdot x : f(x) \geq y\}$$

In particular, we would like to apply Theorem 14.2.1 in order to demonstrate that $C(w, y)$ is a continuous function, and that the *input demand correspondence*, defined by

$$\xi(w, y) = \arg \min_x \{w \cdot x : f(x) \geq y\}$$

is upper hemicontinuous. A difficulty here is that the constraint set $X(w, y) = \{x \in \mathbb{R}_+^n : f(x) \geq y\}$ is definitely unbounded. Thus, even though it is closed, it is not compact. So let us turn to Note 1. Given any $\bar{w} \gg 0$ and any $\bar{y}$ in $Y = [0, \bar{y})$, it is enough to prove continuity of C and upper hemicontinuity of ξ in a neighbourhood of $(\bar{w}, \bar{y})$ such as $W \times \bar{Y}$, where $W = \{w : \bar{w}/2 \leq w \leq 2\bar{w}\}$ and $\bar{Y} = [0, \bar{y})$ for some point $\bar{y}$ in $(\bar{y}, \bar{y})$. Let $\mathbf{1}$ denote the vector whose components are all 1. Monotonicity guarantees that $f(\beta \mathbf{1})$ is a strictly increasing function of $\beta \geq 0$, with $f(0\mathbf{1}) = 0$ and $f(\beta \mathbf{1}) \geq f(x)$ when $f(x) = \bar{y}$ and β is so large that $\beta \mathbf{1} \geq x$. Hence, there exists a $\hat{\beta}$ such that $f(\hat{\beta} \mathbf{1}) = \bar{y}$. Given this $\hat{\beta}$, define

$$\tilde{F}(w, y) = \{x \geq 0 : f(x) \geq y, w \cdot x \leq 2\bar{w} \cdot (\hat{\beta} \mathbf{1})\}$$

Consider any fixed (w, y) in $W \times \bar{Y}$. Note that $\hat{\beta} \mathbf{1} \in \tilde{F}(w, y)$. Also any x in $\tilde{F}(w, y)$ must satisfy $\frac{1}{2}\bar{w} \cdot x \leq w \cdot x \leq 2\bar{w} \cdot (\hat{\beta} \mathbf{1})$, and so $\tilde{F}(w, y) \subseteq A$, where $A = \{x \in \mathbb{R}_+^n : \frac{1}{2}\bar{w} \cdot x \leq 2\bar{w} \cdot (\hat{\beta} \mathbf{1})\}$, which is obviously a bounded set. Example 14.1.5 implies that $\tilde{F}$ has the closed graph property at every point of $W \times \bar{Y}$, so Theorem 14.1.2 implies that $\tilde{F}$ is u.h.c. throughout $W \times \bar{Y}$. So Note 1 applies and gives continuity of $C(w, y)$ and upper hemicontinuity of $\xi(w, y)$ in $W \times \bar{Y}$. Since $\bar{w}$ and $\bar{y}$ were arbitrary, these properties hold for all $w \gg 0, y \geq 0$. $\square$

PROBLEMS FOR SECTION 14.2

1. Let $f(x, y) = -y^4 + x(y^2 - 1)$ for all x and $-1 \leq y \leq 1$, and consider the maximization problem $\max_{-1 \leq y \leq 1} f(x, y)$. Determine the value function for this problem, and describe the correspondence $F^*(x) = \{y \in [-1, 1] : y \text{ maximizes } f(x, y) \text{ over } [-1, 1]\}$. Show that F^* has the closed graph property.
2. Suppose that a consumer has a continuous and strictly quasiconcave utility function $U(x)$ defined on the set $\mathbb{R}_+^n$, which is maximized subject to the constraint $x \in \mathcal{B}(p, m)$, where $(p, m) \mapsto \mathcal{B}(p, m)$ is the budget correspondence described in Example 14.1.1. Explain why the consumer's (single-valued) demand function $x(p, m)$ and the associated indirect utility function $V(p, m)$ are both continuous whenever $p \gg 0$ and $m \geq 0$. What can go wrong if $p_i = 0$ for some i ?
3. Suppose that the utility function of the consumer in Problem 2 is continuous but not even quasiconcave. What continuity properties can then be expected of the consumer's demand correspondence $(p, m) \mapsto \xi(p, m)$ and indirect utility function $V(p, m)$? What difference would quasiconcavity make to the demand correspondence?

14.3 Fixed Points for Contraction Mappings

This brief section presents a so-called fixed point theorem with important applications to economics. In particular it is used in Section 12.3 in connection with the Bellman equation in infinite horizon dynamic programming.

A function F from a set S into $\mathbb{R}^m$ is called **bounded** on S if there exists a positive number M such that $\|F(x)\| \leq M$ for all x in S .

Let S be a subset of $\mathbb{R}^n$, and let $\mathcal{B}$ denote the set of all bounded functions from S into $\mathbb{R}^m$. We define the distance between two functions ϕ and ψ in $\mathcal{B}$ as

$$d(\phi, \psi) = \sup_{x \in S} \|\phi(x) - \psi(x)\|$$

Let $T: \mathcal{B} \rightarrow \mathcal{B}$ be a function (or "operator") that maps each function φ in $\mathcal{B}$ to a function $T(\varphi)$ in $\mathcal{B}$. Thus $T(\varphi)$ is also a bounded function $S \rightarrow \mathbb{R}^m$. We will write $T(\varphi)(x)$ for the value of $T(\varphi)$ at a point x in S . The function T is called a **contraction mapping** if there exists a constant β in $(0, 1)$ such that for all φ and ψ in $\mathcal{B}$, one has

$$\|T(\varphi)(x) - T(\psi)(x)\| \leq \beta d(\varphi, \psi) \quad \text{for all } x \text{ in } S \quad (1)$$

or, equivalently,

$$d(T(\varphi), T(\psi)) \leq \beta d(\varphi, \psi) \quad (2)$$

For any two elements φ and ψ of $\mathcal{B}$, the distance between $T(\varphi)$ and $T(\psi)$ is then at most β times the distance between φ and ψ , hence the name contraction mapping.

THEOREM 14.3.1 (CONTRACTION MAPPING THEOREM)

Let S be a nonempty subset of $\mathbb{R}^n$ and let $\mathcal{B}$ be the set of all bounded functions from S into $\mathbb{R}^m$. Suppose that the operator $T: \mathcal{B} \rightarrow \mathcal{B}$ is a contraction mapping. Then there exists a unique function φ^* in $\mathcal{B}$ such that $\varphi^* = T(\varphi^*)$.

Proof: Since T is a contraction, there exists a β in $(0, 1)$ such that (2) is satisfied for all φ and ψ in $\mathcal{B}$. Choose an arbitrary function φ_0 in $\mathcal{B}$. Define $\varphi_1 = T(\varphi_0)$, and generally $\varphi_{n+1} = T(\varphi_n)$ for $n = 0, 1, 2, \dots$. Let $\gamma_n = d(\varphi_{n+1}, \varphi_n)$. Then (2) implies that

$$\gamma_{n+1} = d(\varphi_{n+2}, \varphi_{n+1}) = d(T(\varphi_{n+1}), T(\varphi_n)) \leq \beta d(\varphi_{n+1}, \varphi_n) = \beta \gamma_n, \quad n \geq 0 \quad (i)$$

An obvious induction argument shows that $\gamma_n \leq \beta^n \gamma_0$. We want to prove that for each point x in S , the sequence $\{\varphi_n(x)\}$ is a Cauchy sequence in $\mathbb{R}^m$. To this end note that

$$\varphi_{n+k} - \varphi_n = (\varphi_{n+k} - \varphi_{n+k-1}) + (\varphi_{n+k-1} - \varphi_{n+k-2}) + \dots + (\varphi_{n+1} - \varphi_n)$$

Therefore, for every x in S , whenever $m > n$ it follows from the triangle inequality that

$$\begin{aligned} \|\varphi_m(x) - \varphi_n(x)\| &= \left\| \sum_{r=n}^{m-1} (\varphi_{r+1}(x) - \varphi_r(x)) \right\| \leq \sum_{r=n}^{m-1} \|\varphi_{r+1}(x) - \varphi_r(x)\| \\ &\leq \sum_{r=n}^{m-1} \gamma_r \leq \sum_{r=n}^{m-1} \beta^r \gamma_0 = \beta^n \gamma_0 \frac{1 - \beta^{m-n}}{1 - \beta} \leq \frac{\beta^n \gamma_0}{1 - \beta} \end{aligned} \quad (ii)$$

The last expression is small when n is large. Hence, $\varphi_n(x)$ is indeed a Cauchy sequence, with a limit that we denote by $\varphi^*(x)$. Letting $m \rightarrow \infty$ in the inequalities (ii), with n fixed, we see that $\|\varphi^*(x) - \varphi_n(x)\| \leq \beta^n \gamma_0 / (1 - \beta)$ for all x in S . Now,

$$\|T(\varphi^*)(x) - \varphi_{n+1}(x)\| = \|T(\varphi^*)(x) - T(\varphi_n(x))\| \leq \beta^{n+1} \gamma_0 / (1 - \beta)$$

by (i). Letting $n \rightarrow \infty$ yields $\varphi^*(x) = T(\varphi^*)(x)$ for all x , and so $T(\varphi^*) = \varphi^*$.

If another function φ^{**} satisfies $T(\varphi^{**}) = \varphi^{**}$, then by (2),

$$d(\varphi^*, \varphi^{**}) = d(T(\varphi^*), T(\varphi^{**})) \leq \beta d(\varphi^*, \varphi^{**})$$

Because $0 < \beta < 1$ and $d(\varphi^*, \varphi^{**}) \geq 0$, it follows that $d(\varphi^*, \varphi^{**}) = 0$, hence $\varphi^{**} = \varphi^*$. ■

NOTE 1 The conclusion of the theorem remains true if $\mathcal{B}$ is restricted to only those bounded functions $\varphi: S \rightarrow \mathbb{R}^m$ that satisfy the inequality $\|\varphi(x) - \varphi_0\| \leq A$ for a given point φ_0 in $\mathbb{R}^m$ and a given number A , i.e. every φ in $\mathcal{B}$ maps S into the closed ball of radius A around the point φ_0 . It also remains true if we require that $\|\varphi(x) - \varphi(x')\| \leq M\|x - x'\|$ for all x and x' in S and a common given number M . It even remains true if both conditions are imposed on the elements of $\mathcal{B}$. To see this, note that if the relevant inequality or inequalities hold for each φ_n in the proof above, they still hold after passing to the limit (as $n \rightarrow \infty$).

EXAMPLE 1 As an example of how one can use the contraction mapping theorem, we prove Theorem 5.8.2 on the existence and uniqueness of solutions of differential equations.

Proof: We use the notation in Theorem 5.8.2. Further, let

$$K = \max_{(t,x) \in \Gamma} |F'_x(t, x)| \quad \text{and} \quad k = \min\{a, b/M, 1/(2K)\}$$

By the mean value theorem, $|F(t, x) - F(t, x')| \leq K|x - x'|$ for all t in $[t_0 - a, t_0 + a]$ and all x, x' in $[x_0 - b, x_0 + b]$.

We first prove the existence of a unique solution over the interval $I = [t_0 - k, t_0 + k]$ instead of $(t_0 - r, t_0 + r)$. A function x^* solves the initial value problem $\dot{x} = F(t, x)$, $x(t_0) = x_0$ if and only if

$$x^*(t) = x_0 + \int_{t_0}^t F(s, x^*(s)) ds$$

for all t in I . Suppose x^* is a solution. Then for t and t' in I we have

$$|x^*(t') - x^*(t)| = \left| \int_t^{t'} F(s, x^*(s)) ds \right| \leq M|t' - t|$$

since $|F(s, x)| \leq M$ for all (s, x) in Γ . Let $\mathcal{B}$ be the set of all functions $x: I \rightarrow \mathbb{R}$ that satisfy $x(t_0) = x_0$ and $|x(t') - x(t)| \leq M|t' - t|$ for all t, t' in I . The set $\mathcal{B}$ is nonempty, since it contains the constant function $x = x_0$, and all functions that belong to $\mathcal{B}$ are continuous. Then the operator $T: \mathcal{B} \rightarrow \mathcal{B}$ defined by $T(x)(t) = x_0 + \int_{t_0}^t F(s, x(s)) ds$ is well defined. (You should verify that $(s, x(s))$ lies in Γ .) For any two functions x and $\bar{x}$ in $\mathcal{B}$ and any s in I we have

$$|F(s, \bar{x}(s)) - F(s, x(s))| \leq K|\bar{x}(s) - x(s)| \leq Kd(\bar{x}, x)$$

and so

$$|T(\bar{x})(t) - T(x)(t)| = \left| \int_{t_0}^t (F(s, \bar{x}(s)) - F(s, x(s))) ds \right| \leq |t - t_0| K d(\bar{x}, x) \leq \frac{1}{2} d(\bar{x}, x)$$

since $|t - t_0| \leq k \leq 1/(2K)$. It follows that $d(T(\bar{x}), T(x)) = \sup_{t \in I} |T(\bar{x})(t) - T(x)(t)| \leq \frac{1}{2} d(\bar{x}, x)$. Hence, T is a contraction, and by Theorem 14.3.1 and the succeeding note, it has a unique fixed point x^* in the set $\mathcal{B}$.

So far we have proved the existence of a unique solution over $[t_0 - k, t_0 + k]$. To extend this solution to all of $(t_0 - r, t_0 + r)$, note that we can use the same construction to show that there is a unique solution in a neighbourhood of any point in Γ . One can then splice together two solutions that agree on some common subinterval of their domains, and obtain a solution over the union of the domains. In this way, one can obtain a solution over all of $(t_0 - r, t_0 + r)$. We refrain from going into the rather tedious details here. ■

14.4 Brouwer's and Kakutani's Fixed Point Theorems

Consider a function f that maps each point x of a set K in $\mathbb{R}^n$ to a point $f(x)$ of the same set K . We say that f maps the set K into itself. Usually, x and $f(x)$ will be different. If x^* is a point such that $f(x^*) = x^*$, that is, if the point x^* is mapped to itself, then x^* is called a **fixed point** of f .

We would like to find conditions ensuring that any continuous function mapping K into itself has a fixed point. Note that some restrictions must be placed on K . For instance, the continuous mapping $f(x) = x + 1$ of the real line into itself has no fixed point; this is because $f(x^*) = x^*$ would imply that $x^* + 1 = x^*$, which is absurd.

The following result by L. E. J. Brouwer yields sufficient conditions for the existence of a fixed point (see Ichiiishi (1983) for a proof):

THEOREM 14.4.1 (BROUWER'S FIXED POINT THEOREM)

Let K be a nonempty compact (closed and bounded) convex set in $\mathbb{R}^n$, and f a continuous function mapping K into itself. Then f has a fixed point x^* , i.e. a point x^* in K such that $f(x^*) = x^*$.

The function f in the theorem maps points x in $\mathbb{R}^n$ into points y in $\mathbb{R}^n$. It is therefore described by the system

$$\begin{aligned} y_1 &= f_1(x_1, \dots, x_n) \\ &\dots \dots \dots \\ y_n &= f_n(x_1, \dots, x_n) \end{aligned} \quad (1)$$

So a fixed point $x^* = (x_1^*, \dots, x_n^*)$ of f must satisfy the equation system

$$\begin{aligned} x_1 &= f_1(x_1, \dots, x_n) \\ &\dots \dots \dots \\ x_n &= f_n(x_1, \dots, x_n) \end{aligned} \quad (2)$$

This immediately shows how Brouwer's fixed point theorem can be used to establish the existence of a solution to a nonlinear system of equations. Note, however, that in order to apply the theorem one must establish the continuity of f and prove that f maps a suitable domain K into itself.

There are numerous applications of Brouwer's theorem in which the set K is the **standard unit simplex** Δ^{n-1} in $\mathbb{R}^n$ defined by²

$$\Delta^{n-1} = \{x = (x_1, \dots, x_n) : x_1 \geq 0, \dots, x_n \geq 0, \sum_{i=1}^n x_i = 1\} \quad (3)$$

² Note that Δ^{n-1} is $(n-1)$ -dimensional. For example, Δ^2 is contained in the (two-dimensional) plane $x_1 + x_2 + x_3 = 1$ in $\mathbb{R}^3$. See Fig. 1.

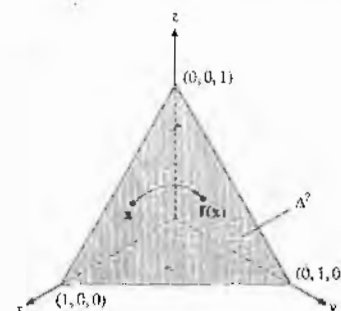


Figure 1 f maps Δ^2 into itself.

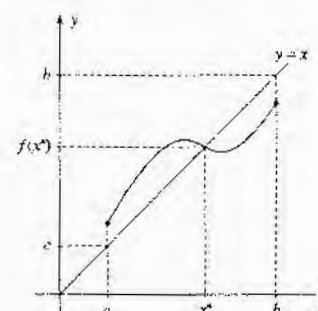


Figure 2 $f(x^*) = x^*$

For instance, $x_1, \dots, x_n$ might denote the nonnegative prices of n different commodities, these prices being normalized by the convention that $x_1 + \dots + x_n = 1$.

The set Δ^{n-1} is convex and compact. To see whether Brouwer's theorem applies to a given continuous function f defined on Δ^{n-1} , it is only necessary to check that f maps Δ^{n-1} into itself. If f is given by (1), then f will map Δ^{n-1} into itself provided that for all $(x_1, \dots, x_n)$ in Δ^n , one has

$$f_1(x_1, \dots, x_n) \geq 0, \dots, f_n(x_1, \dots, x_n) \geq 0, \sum_{i=1}^n f_i(x_1, \dots, x_n) = 1 \quad (4)$$

The case when $n = 3$ is illustrated in Fig. 1. Brouwer's theorem implies that if f maps Δ^2 continuously into Δ^2 , then there must be at least one point x^* in the simplex Δ^2 for which $f(x^*) = x^*$.

In $\mathbb{R}^1$ (the real line), a nonempty compact convex set must be a closed and bounded interval $[a, b]$ (or a single point). So Brouwer's theorem asserts that a continuous function $f : [a, b] \rightarrow [a, b]$ must have a fixed point. But this follows from the intermediate value theorem. (Indeed, $g(x) = f(x) - x$ satisfies $g(a) \geq 0$, and $g(b) \leq 0$, so for some x^* in $[a, b]$, $g(x^*) = 0$.) The geometric content of this proposition is illustrated in Fig. 2. The graph of f must cross the diagonal $y = x$.

An Illustration

For the two-dimensional case the following illustration of the theorem might aid your intuition. Do not take the illustration too seriously, however!

Imagine a flock of sheep crammed into a circular pen. Suppose that the flock suddenly starts moving and then stops after a certain time. At a given moment of time each sheep has a definite position in the pen, as shown in Fig. 3. Although each sheep can move, no sheep can move against or across the stream, so each sheep must stay close to its original neighbours.

Consider the mapping from the position originally occupied by each sheep to its final position after stopping. By the assumptions above, this is a continuous mapping (a dubious

claim) of the pen into itself. Because the pen is a compact convex set, Brouwer's theorem applies. So there must be at least one "fixed sheep" which stops exactly where it started.

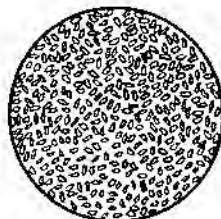


Figure 3

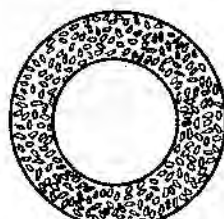


Figure 4

Now assume instead that the flock is enclosed in a circular ring, as indicated in Fig. 4. Suppose all the sheep move 90 degrees clockwise around the ring. Then each sheep will stop in an entirely new position. No "fixed sheep" exists in this case. As above, the movement indicated defines a continuous mapping of the pen into itself, but there is no fixed point. In fact, Brouwer's theorem does not apply since the ring of Fig. 4 is not a convex set.

The importance of this result for sheep farmers can hardly be underestimated, but it does indicate an important "topological difference" between a circular disc and a ring.

A Generalization

The convexity hypothesis in Theorem 14.4.1 can be relaxed. Let $L \subseteq \mathbb{R}^n$ be a **homeomorphic image** of K in the sense that there exists a one-to-one continuous mapping g of K onto L (i.e. $g(K) = L$) whose inverse mapping g^{-1} is also continuous. Intuitively, homeomorphic images of a rubber ball are obtained by squashing or stretching, as long as we do not tear it apart, make any holes, or glue parts together. Note, in particular, that homeomorphic images of a convex set are not necessarily convex. The natural generalization of Brouwer's theorem is:

Any homeomorphic image L of a nonempty compact convex set $K \subseteq \mathbb{R}^n$ has the fixed point property, i.e. any continuous function f mapping L into L has a fixed point. (5)

Proof: Let f be the continuous function mapping L into L , and let g be a homeomorphism, i.e. a continuous mapping of K onto L with a continuous inverse g^{-1} . If $x \in K$, then $g(x) \in L$, so $f(g(x)) \in L$, which implies that $g^{-1}(f(g(x))) \in K$. So the mapping $g^{-1}fg$ must be a continuous function of K into itself. According to Theorem 14.4.1, there exists a fixed point x^* in K such that $g^{-1}(f(g(x^*))) = x^*$. But then $g(x^*) = f(g(x^*))$, and so $g(x^*)$ in L is a fixed point for f . ■

Brouwer's original motivation for his theorem was to examine the topological differences between various sets in $\mathbb{R}^n$. It was recognized only later that the theorem had interesting applications outside topology. The next section shows how Brouwer's theorem can be used to prove the existence of an equilibrium in a pure exchange economy. Although this type of

economy is very simple and unrealistic, the existence proof contains many of the essential features that arise in richer general equilibrium models.

Kakutani's Fixed Point Theorem

Brouwer's theorem deals with fixed points of continuous functions on appropriate domains. Kakutani's theorem generalizes the theorem to correspondences. It extends an existence theorem for saddle points proved by von Neumann in 1928 and used in his work on both game theory and growth theory. It also greatly simplified Nash's proof that a mixed strategy equilibrium always exists in an n -player game with finite strategy sets. (For a proof of Kakutani's theorem, see Aubin and Frankowska (1990).)

THEOREM 14.4.2 (KAKUTANI'S FIXED POINT THEOREM)

Let K be a nonempty compact convex set in $\mathbb{R}^n$ and F a correspondence $K \rightarrow K$. Suppose that:

- (a) $F(x)$ is a nonempty convex set in K for each x in K .
- (b) F is upper hemicontinuous.

Then F has a fixed point x^* in K , i.e. a point x^* such that $x^* \in F(x^*)$.

It is worth emphasizing that, because of Theorem 14.1.3, the correspondence F from the compact domain K into itself will be upper hemicontinuous if and only if its graph is compact. Notice that Brouwer's theorem is implied by Kakutani's because a continuous function f mapping the compact domain $K \subseteq \mathbb{R}^n$ into itself has an associated correspondence with nonempty convex values defined by $F(x) = \{f(x)\}$ for all x in K . This correspondence is upper hemicontinuous by Theorem 14.1.1, and so has the closed graph property.

In the one-dimensional case, Theorem 14.4.2 takes the form:

If the correspondence $F: [a, b] \rightarrow [a, b]$ has the closed graph property and $F(x)$ is a nonempty closed interval for each x in $[a, b]$, then F has a fixed point. (6)

Proof of (6): For each x in $[a, b]$, the image of F is a closed interval depending on x , say $F(x) = [f(x), g(x)]$. Define $x^* = \sup\{x \in [a, b] : f(x) \geq x\}$. We claim that $f(x^*) \leq x^* \leq g(x^*)$, which means that $x^* \in F(x^*)$ and thus x^* is a fixed point.

Suppose for a contradiction that $f(x^*) > x^*$. Then also $f(x^*) > x^* + \varepsilon$ for some $\varepsilon > 0$. By Theorem 14.1.2, F is upper hemicontinuous at x^* . According to definition (14.1.5), to the open set $U = (x^* + \varepsilon, \infty)$, which contains $F(x^*) = [f(x^*), g(x^*)]$, there exists a neighbourhood N of x^* such that for x in $N \cap [a, b]$, we have $F(x) = [f(x), g(x)] \subseteq U$. Then, in particular, for x in $N \cap [a, b]$, $f(x) > x^* + \varepsilon$. If we choose N so small that all x in N satisfy $x < x^* + \varepsilon$, then $f(x) > x^* + \varepsilon > x$, contradicting the definition of x^* . Hence, $f(x^*) \leq x^*$.

To prove that $g(x^*) \geq x^*$, suppose to the contrary that $g(x^*) < x^*$. Then also $g(x^*) < x^* - \varepsilon$ for some $\varepsilon > 0$. Upper hemicontinuity of F at x^* implies that $F(x) = [f(x), g(x)] \subseteq (-\infty, x^* - \varepsilon)$ for all x in some neighbourhood of x^* . In particular, $f(x) \leq g(x) < x^* - \varepsilon < x$ for all x in $(x^* - \varepsilon, x^*)$ that are close enough to x^* , yielding another contradiction to the definition of x^* . It follows that we must have $g(x^*) \geq x^*$. ■

The one-dimensional case is illustrated in Fig. 5. The fixed point is x^* . Figure 6 illustrates that in (6) one cannot drop the requirement that $F(x)$ be an interval (and so convex).

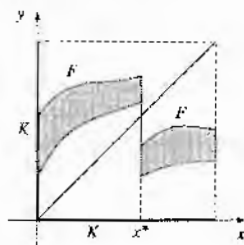


Figure 5 x^* is a fixed point for F .

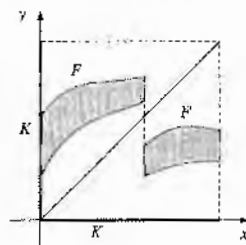


Figure 6 F has no fixed point.

PROBLEMS FOR SECTION 14.4

1. Consider the function f defined for all x in $(0, 1)$ by

$$f(x) = \frac{1}{2}(x+1)$$

Prove that f maps $(0, 1)$ into itself, but f has no fixed point. Why does Brouwer's theorem not apply?

2. Consider the continuous transformation $T: (x, y) \mapsto (-y, x)$ from the xy -plane into itself, consisting of a 90° rotation around the origin. Define the sets

$$E = \{(x, y) : x^2 + y^2 = 1\}, \quad B = \{(x, y) : x^2 + y^2 \leq 1\}$$

Are these sets compact? T induces continuous maps $T_E: E \rightarrow E$ and $T_B: B \rightarrow B$. Does either transformation have a fixed point? Explain the results in the light of Brouwer's theorem.

3. Let $A = (a_{ij})$ be an $n \times n$ matrix whose elements all satisfy $a_{ij} \geq 0$. Assume that all column sums are 1, so that $\sum_{i=1}^n a_{ij} = 1$ ($j = 1, \dots, n$). Prove that if $x \in \Delta^{n-1}$, then $Ax \in \Delta^{n-1}$, where Δ^{n-1} is the unit simplex defined by (3). Hence, $x \mapsto Ax$ is a (linear) transformation of Δ^{n-1} into itself. What does Brouwer's theorem say in this case?

4. Consider the correspondence $F: [0, 2] \rightrightarrows [0, 2]$ that maps each x in $(0, 1)$ to $\{2\}$, maps $x = 1$ to $\{0, 2\}$, and finally maps each x in $(1, 2]$ to $\{0\}$. Draw the graph of F and determine whether F has a closed graph. Does Kakutani's theorem apply?

HARDER PROBLEMS

5. Assume that $f: [0, 1] \rightarrow [0, 1]$ satisfies

$$\lim_{s \rightarrow x^+} f(s) \leq f(x) \leq \lim_{s \rightarrow x^-} f(s) \quad \text{for all } x \text{ in } [0, 1]$$

(Only the right-hand (left-hand) inequality is required to hold for $x = 0$ ($x = 1$).) Prove that f has a fixed point. (Hint: Consider $x^* = \sup A$, where $A = \{x \in [0, 1] : f(x) \geq x\}$.)

14.5 Equilibrium in a Pure Exchange Economy

Consider an economy with m consumers, each of whom is initially endowed with fixed quantities of n different commodities or goods. No production is possible, so the consumers merely engage in exchange. Trade takes place because each consumer wishes to acquire a bundle of commodities that is preferred to the initial endowment. This is described as a **pure exchange economy**.

As usual when discussing perfectly competitive markets, we assume that consumers can exchange their commodities at fixed price ratios. Specifically, we assume that a price vector $\mathbf{p} = (p_1, \dots, p_n)$ is announced, where p_i is the nonnegative price per unit of commodity number i . Thus, any consumer can sell one unit of commodity j for the amount p_j , and use that amount to buy any other commodity i at the price p_i . In this way, the one unit of commodity j has the same value as p_j/p_i units of commodity i (assuming that $p_i > 0$).

The price vector $\mathbf{p}$ determines the market value $\mathbf{p} \cdot \mathbf{c} = \sum_{i=1}^n p_i c_i$ of any commodity bundle $\mathbf{c} = (c_1, \dots, c_n)$, including any consumer's initial endowment. By exchanging at the fixed price ratios p_j/p_i , the consumer can achieve any commodity bundle whose market value equals that of the initial endowment. These are the **affordable** consumption bundles at which the consumer satisfies his or her **budget constraint**.

Among all affordable commodity bundles, each consumer selects one that is preferred to all the others. In other words, each consumer's demands represent the choice of commodity bundle that maximizes a utility function subject to that consumer's own budget constraint. (We assume that there is a unique utility maximizing consumption bundle.)

Next, add the demands of all consumers for each commodity i and subtract the total initial endowment of i . The result is called the **excess demand** for that commodity. Because it depends on the price vector $\mathbf{p}$, the excess demand will be denoted by $g_i(\mathbf{p})$. When the price vector is $\mathbf{p}$, the sign of $g_i(\mathbf{p})$ indicates whether the total demand for commodity i is greater or less than the total endowment of that good.

The following question arises naturally in the minds of most economists. Is it possible to find prices of all commodities which ensure that the aggregate demand for each does not exceed the corresponding aggregate endowment? Prices with this property are called **equilibrium prices**. This is because, if all consumers do face such prices, then all their demands for every good can be satisfied simultaneously. So there are no unfulfilled demands that can force consumers to change their plans.

Let $\mathbf{p}^* = (p_1^*, \dots, p_n^*)$ denote such an equilibrium price vector. By definition, it must satisfy the inequalities $g_i(\mathbf{p}^*) \leq 0$ for $i = 1, \dots, n$. It will now be shown how Brouwer's fixed point theorem can be used to prove existence of an equilibrium price vector, provided suitable continuity conditions are imposed.

To proceed further requires a little extra notation. For each consumer j and commodity i , let w_i^j denote j 's initial endowment of i , and $x_i^j(\mathbf{p})$ the same consumer's final demand when the price vector is $\mathbf{p}$. In addition, let

$$w_i = \sum_{j=1}^m w_i^j \quad \text{and} \quad x_i(\mathbf{p}) = \sum_{j=1}^m x_i^j(\mathbf{p}), \quad i = 1, \dots, n$$

denote respectively the **total endowment** and **aggregate demand** for each commodity i . The latter is equal to the total demand for commodity i by all consumers. The excess demand

functions referred to above are then given by

$$g_i(\mathbf{p}) = x_i(\mathbf{p}) - w_i$$

Now, the total value of consumer j 's initial endowment at the price vector $\mathbf{p}$ is $\sum_{i=1}^n p_i w_i^j$, so the budget constraint is

$$\sum_{i=1}^n p_i x_i^j(\mathbf{p}) = \sum_{i=1}^n p_i w_i^j \quad (1)$$

This is valid for each consumer, so summing (1) from $j = 1$ to $j = m$ and using the definitions of the aggregates w_i and $x_i(\mathbf{p})$, we obtain

$$\sum_{i=1}^n p_i x_i(\mathbf{p}) = \sum_{i=1}^n p_i w_i \quad (\text{Walras's Law}) \quad (2)$$

Thus, the value of the aggregate excess demand vector $g_i(\mathbf{p}) = \sum_{j=1}^m (x_i(\mathbf{p}) - w_i)$ at prices $\mathbf{p}$ is identically zero.

An equilibrium price vector $\mathbf{p}^*$ is characterized by the inequalities

$$x_i(\mathbf{p}^*) \leq w_i \quad \text{or} \quad g_i(\mathbf{p}^*) \leq 0 \quad \text{for all } i = 1, \dots, n \quad (3)$$

so that the equilibrium market demand for each commodity does not exceed the total endowment of that commodity. Also, observe that because $\mathbf{p}^* \geq 0$ and $x_i(\mathbf{p}^*) \leq w_i$, each product $p_i^*(x_i(\mathbf{p}^*) - w_i)$ is ≤ 0 . But the sum over $i = 1, \dots, n$ of all these products is 0 because of Walras's law (2). Consequently, it is impossible that $p_i^*(x_i(\mathbf{p}^*) - w_i) < 0$ for any i . Hence, we have proved that if $\mathbf{p}^*$ is an equilibrium price vector, then

$$x_i(\mathbf{p}^*) < w_i \Rightarrow p_i^* = 0, \quad i = 1, \dots, n \quad (4)$$

This is the *rule of free goods*: if any commodity is in excess supply in equilibrium, its price must be zero. In other words, if there is a commodity for which the market demand is strictly less than the total stock, then the equilibrium price for that commodity must be 0.

It is rather obvious that only price ratios, or relative prices, matter in this economy. For this reason, we can normalize by dividing the price vector $\mathbf{p}$ by the (positive) sum $p_1 + \dots + p_n$ of the prices to ensure that this sum is equal to 1 (see Problem 1). Then all normalized price vectors will lie in the simplex Δ^{n-1} defined by (14.4.3).

Any existence proof requires continuity assumptions. It will be enough to assume that the market demand functions $\mathbf{p} \mapsto x_i(\mathbf{p})$, $i = 1, \dots, n$, are continuous functions on the simplex Δ^{n-1} or, what amounts to the same thing, that the excess demand functions $g_1, \dots, g_n$ are continuous on Δ^{n-1} . Our problem can now be stated as follows:

Suppose that $g_1, \dots, g_n$ are continuous on Δ^{n-1} and assume that

$$\sum_{i=1}^n p_i g_i(\mathbf{p}) = 0 \quad \text{for all } \mathbf{p} \text{ in } \Delta^{n-1} \quad (5)$$

Is there a vector $\mathbf{p}^* = (p_1^*, \dots, p_n^*)$ in Δ^n such that

$$g_1(\mathbf{p}^*) \leq 0, \quad \dots, \quad g_n(\mathbf{p}^*) \leq 0?$$

(Note that (5) is a restatement of Walras's law.)

We shall use Brouwer's theorem to prove existence. To do so, we construct a continuous mapping of Δ^{n-1} into itself for which any fixed point gives equilibrium prices. Consider first the mapping $(p_1, \dots, p_n) \mapsto (p'_1, \dots, p'_n)$ defined by

$$p'_1 = p_1 + g_1(\mathbf{p}), \quad p'_2 = p_2 + g_2(\mathbf{p}), \quad \dots, \quad p'_n = p_n + g_n(\mathbf{p}) \quad (6)$$

This simple price adjustment mechanism has a certain economic appeal: it maps p_i , the "old" price of commodity i , to the new adjusted price $p'_i = p_i + g_i(\mathbf{p})$. If excess demand $g_i(\mathbf{p})$ is positive, so that the market demand exceeds the total available endowment, then the price is increased. The opposite is true if $g_i(\mathbf{p}) < 0$, when the price is lowered. So far, this is all very sensible. Note, however, that there is no guarantee that $p'_i \geq 0$. Moreover, the new prices p'_i usually will not sum to 1. Hence, the new price vector $(p'_1, \dots, p'_n)$ will not necessarily belong to the simplex Δ^{n-1} . As a consequence, Brouwer's theorem does not apply to the mapping defined in (6). The mapping must be altered somewhat in order to work.

Before we present an alternative mapping, recall that if x is a real number, then $\max\{0, x\}$ denotes the larger of the two numbers 0 and x . Hence, if $x > 0$ then $\max\{0, x\} = x$, whereas $\max\{0, x\} = 0$ if $x \leq 0$. Evidently the function $x \mapsto \max\{0, x\}$ is continuous.

With this in mind, instead of (6) we define a mapping $(p_1, \dots, p_n) \mapsto (p'_1, \dots, p'_n)$ by

$$p'_i = \frac{1}{d(\mathbf{p})} (p_i + \max\{0, g_i(\mathbf{p})\}), \quad i = 1, \dots, n \quad (7)$$

where $d(\mathbf{p}) = 1 + \sum_{i=1}^n \max\{0, g_i(\mathbf{p})\} \geq 1$. It is difficult to provide a good economic motivation for this particular mapping. Nevertheless, it does what is needed. Note first that $p'_i \geq 0$ for all i . Also, the new prices p'_i sum to unity whenever the old prices p_i do. Hence, (7) defines a mapping of Δ^{n-1} into itself. We see, moreover, that each p'_i is a continuous function of $(p_1, \dots, p_n)$. Thus Brouwer's theorem applies, and so there must exist a fixed point $\mathbf{p}^* = (p_1^*, \dots, p_n^*)$ in Δ^{n-1} . At $\mathbf{p}^*$, for any $i = 1, \dots, n$, one has

$$p_i^* = \frac{1}{d(\mathbf{p}^*)} (p_i^* + \max\{0, g_i(\mathbf{p}^*)\})$$

This is easily seen to be equivalent to

$$(d(\mathbf{p}^*) - 1)p_i^* = \max\{0, g_i(\mathbf{p}^*)\} \quad (8)$$

The definition of $d(\mathbf{p})$ implies that $d(\mathbf{p}^*) \geq 1$. Suppose that $d(\mathbf{p}^*) > 1$. Then (8) implies that for those i with $p_i^* > 0$ one has $\max\{0, g_i(\mathbf{p}^*)\} > 0$, and so $g_i(\mathbf{p}^*) > 0$. Because $p_1^* + \dots + p_n^* = 1$, however, at least one p_i^* is positive. It follows that $\sum_{i=1}^n p_i^* g_i(\mathbf{p}^*) > 0$, a contradiction of Walras's law. Hence, we conclude that $d(\mathbf{p}^*) = 1$. But then (8) implies that $\max\{0, g_i(\mathbf{p}^*)\} = 0$ for $i = 1, \dots, n$, and so $g_i(\mathbf{p}^*) \leq 0$ for $i = 1, \dots, n$. This proves that $\mathbf{p}^*$ is an equilibrium price vector. The existence of an equilibrium in the pure exchange economy is thereby established.

Brouwer's fixed point theorem can only be used to prove existence. For the last example in particular, it does not by itself indicate any practical method for finding equilibrium prices.

The economic model considered above was one of pure exchange in the sense that there was no production of commodities. Moreover, consumer demand functions were single-valued. More realistic equilibrium models include producers as well as consumers, and allow (multi-valued) demand (and supply) correspondences. Obviously, existence of an equilibrium is an important issue in these more general models as well. It turns out that existence can still be established under suitable assumptions, making use of Kakutani's fixed point theorem for correspondences.

PROBLEMS FOR SECTION 14.5

1. In the pure exchange model studied above, suppose that each consumer j 's demand functions $x_1^j(p), \dots, x_n^j(p)$ result from utility maximization subject to j 's own budget constraint. Explain why the demand functions $x_i^j(p)$ are then all homogeneous of degree 0, and why this entitles us to normalize prices by setting $p_1 + \dots + p_n = 1$.

A

APPENDIX

SETS, COMPLETENESS, AND CONVERGENCE

If we can't imagine how something might fail to happen, we are tempted to conclude that it must always happen. Of course, according to this principle of reasoning, the poorer our imagination the more facts we could establish!
—Loomis (1974)

This appendix considers a few selected topics from the foundations of mathematical analysis. Section A.1 introduces terminology used in almost all of mathematics, like the basic concepts of set theory and general functions. Section A.2 discusses the least upper bound principle, which is a crucial property of the real number system. Sequences, subsequences, and Cauchy sequences on the real line are the topics of Section A.3. The last Section A.4 introduces some results about the infimum and supremum of functions that are occasionally useful in economics.

A.1 Sets and Functions

A **set** is a "collection of objects". These objects are called the **elements** or **members** of the set. A finite set can be described by listing the objects: $\{a, b, c, \dots, t\}$. Some infinite sets can be written in a similar way, like the set $\mathbb{N} = \{1, 2, 3, \dots\}$ of all natural numbers, provided it is clear from the context precisely what the elements of the set are. We use the notation $x \in S$ to indicate that x is an element of S (or "belongs to S " or "is a member of S ").

Two sets A and B are equal ($A = B$) if and only if they have the same elements. This implies that repeating any listed element has no effect: $\{1, 3, 5, 1, 5, 2, 1\} = \{1, 2, 3, 5\}$. This example also illustrates that the order of the elements in the listing makes no difference.

If A and B are two sets such that every element of A is also an element of B , then A is a **subset** of B and one writes $A \subseteq B$ (read as " A is a subset of B " or " A is included in B ") or $B \supseteq A$ (" B includes A " or " B is a **superset** of A "). The set A is a **proper subset** of B if $A \subseteq B$ and $A \neq B$; sometimes one writes $A \subsetneq B$ in this case. The symbol $\subseteq$ is called the **inclusion symbol**.¹ It is clear that $A = B$ if and only if $A \subseteq B$ and $B \subseteq A$. It is also easy

¹ Some authors use $\subset$ as the inclusion symbol, and some use $\subseteq$ for inclusion and reserve $\subset$ for proper inclusion. In this book we use $\subseteq$ for inclusion, whether proper or not.

to see that if $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$.

The **empty set**, $\emptyset$, is the only set with no elements at all. It is a subset of every set.

There are several ways to build new sets out of old sets. One very common construction is the creation of a subset of a given set by selecting those elements that have a certain property: if S is a set and $\alpha(x)$ is a condition that an element x of S may or may not satisfy, then $A = \{x \in S : \alpha(x)\}$ is the set of all those elements of S that satisfy the condition. For example, the set $A = \{1, 2, 3, 4\}$ of all natural numbers between 1 and 4 can be written as $\{x \in \mathbb{N} : 1 \leq x \leq 4\}$. If it is clear from the context exactly what the set S is, one often simply writes $\{x : \alpha(x)\}$ for the set $\{x \in S : \alpha(x)\}$.

If A and B are sets, then $A \cup B$, the **union** of A and B , is the set of all elements that belong to A or B (or both). The **intersection** $A \cap B$ of A and B is the set of the elements that belong to both A and B . If $A \cap B = \emptyset$, the sets A and B are **disjoint**. The **set theoretic difference** $A \setminus B$ ("A minus B") is the set of all elements in A that do not belong to B . The **symmetric difference** $A \Delta B = (A \setminus B) \cup (B \setminus A)$ is the set of all elements that belong to exactly one of the sets A and B .

The following are some important identities involving the operations defined above.

$$A \cup B = B \cup A, \quad (A \cup B) \cup C = A \cup (B \cup C), \quad A \cup \emptyset = A \quad (1)$$

$$A \cap B = B \cap A, \quad (A \cap B) \cap C = A \cap (B \cap C), \quad A \cap \emptyset = \emptyset \quad (2)$$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C), \quad A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \quad (3)$$

$$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C), \quad A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C) \quad (4)$$

$$A \Delta B = B \Delta A, \quad (A \Delta B) \Delta C = A \Delta (B \Delta C), \quad A \Delta \emptyset = A \quad (5)$$

The formulas in (3) are called **distributive laws** and the formulas in (4) are known as **De Morgan's laws**.

In discussions involving sets, it is often the case that all the sets considered are subsets of some given "universal" set, Ω , say. When this is the case, the set difference $\Omega \setminus S$ is often written as A^c or $\complement S$, and called the **complement** of S . When we discuss subsets of $\mathbb{R}^n$, for instance, $\complement S = \mathbb{R}^n \setminus S$. With this notation, and regarding A as the universal set, De Morgan's laws can be written as

$$\complement(B \cup C) = \complement B \cap \complement C, \quad \complement(B \cap C) = \complement B \cup \complement C \quad (6)$$

The collection of all subsets of a set A is also a set, called the **power set** of A and denoted by $\mathcal{P}(A)$. Thus, $B \in \mathcal{P}(A) \iff B \subseteq A$.

We noted above that in a set specification such as $\{a, b, \dots, t\}$ the order in which the elements are listed does not matter. Thus, in particular $\{a, b\} = \{b, a\}$. However, on many occasions one is interested in distinguishing between the first and the second elements of a pair. One such example is the coordinates of a point in the xy -plane. These coordinates are given as an **ordered pair** (a, b) of real numbers. The important property of ordered pairs is that $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$. In particular, $(a, b) = (b, a)$ if and only if $a = b$. See Problem 1 for one possible way to define an ordered pair in set-theoretic terms. Once ordered pairs have been defined, one can go on to define ordered triples, quadruples, etc. as $(a, b, c) = ((a, b), c)$, $(a, b, c, d) = ((a, b, c), d)$, etc. Of course, there is a natural

one-to-one correspondence $((a, b), c) \leftrightarrow (a, (b, c))$, so it would not matter much if an ordered triple were defined as $(a, (b, c))$ instead of $((a, b), c)$. The important thing again is that $(a, b, c) = (d, e, f)$ if and only if $a = d$, $b = e$, and $c = f$.

If A and B are sets, their **Cartesian product** is the set $A \times B$ consisting of all ordered pairs (a, b) such that $a \in A$ and $b \in B$. Similarly, the Cartesian product of the sets A , B , and C is the set of all ordered triples (a, b, c) such that $a \in A$, $b \in B$, and $c \in C$. The natural one-to-one correspondence $((a, b), c) \leftrightarrow (a, (b, c))$ referred to above gives a one-to-one correspondence between $(A \times B) \times C$ and $A \times (B \times C)$, so one can well identify the two and write either product simply as $A \times B \times C$.

The Euclidean plane $\mathbb{R}^2$ is the Cartesian product $\mathbb{R} \times \mathbb{R}$. More generally, $\mathbb{R}^m = \mathbb{R} \times \dots \times \mathbb{R}$ (with m factors), and there is a natural one-to-one correspondence between the elements $((x_1, \dots, x_m), (y_1, \dots, y_n))$ of $\mathbb{R}^m \times \mathbb{R}^n$ and the elements $(x_1, \dots, x_m, y_1, \dots, y_n)$ of $\mathbb{R}^{m+n}$.

Indexed Sets

There is often a need to go beyond ordered pairs, or triples, quadruples, even n -tuples for any finite n . Suppose that, for each i in some set I , we specify an object a_i (which can be a number, a set, or any other entity). Then these objects form an **indexed set** $\{a_i\}_{i \in I}$ with I as its **index set**. In formal terms, an indexed set is a function whose domain is the index set (see below).

There is an important difference between the indexed set $\{a_i\}_{i \in I}$ and the set of all the values a_i . For example, an n -vector $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is an indexed set with $\{1, 2, \dots, n\}$ as its index set. Here the order of the elements does matter, and multiple occurrences of the same value will also matter. Thus the five-dimensional vector $(3, -1, 3, 3, -2)$ is different from the vector $(3, -2, -1, 3, -1)$, whereas the sets $\{3, -1, 3, 3, -2\}$ and $\{3, -2, -1, 3, -1\}$ are equal (and equal to the set $\{-1, -2, 3\}$). Indexed sets allow one to distinguish between sets whose elements appear in a different order, and also to talk about sets where some elements are repeated. A **sequence** is an indexed set $\{a_k\}_{k \in \mathbb{N}}$ with the set $\mathbb{N}$ of natural numbers as its index set. Instead of $\{a_k\}_{k \in \mathbb{N}}$ one often writes $\{a_k\}_{k=1}^\infty$.

A set whose elements are themselves sets is often called a **family** of sets, and so an indexed set of sets is also called an **indexed family** of sets.

Consider a nonempty indexed family $\{A_i\}_{i \in I}$ of sets (i.e. the index set I is nonempty). The **union** and the **intersection** of this family are the sets

$$\bigcup_{i \in I} A_i = \text{the set consisting of all } x \text{ that belong to } A_i \text{ for at least one } i \text{ in } I \quad (7)$$

$$\bigcap_{i \in I} A_i = \text{the set consisting of all } x \text{ that belong to } A_i \text{ for all } i \text{ in } I \quad (8)$$

The distributive laws in (3) can be generalized to

$$A \cup \left(\bigcap_{i \in I} B_i \right) = \bigcap_{i \in I} (A \cup B_i), \quad A \cap \left(\bigcup_{i \in I} B_i \right) = \bigcup_{i \in I} (A \cap B_i) \quad (9)$$

and De Morgan's laws (4) to

$$A \setminus \left(\bigcup_{i \in I} B_i \right) = \bigcap_{i \in I} (A \setminus B_i), \quad A \setminus \left(\bigcap_{i \in I} B_i \right) = \bigcup_{i \in I} (A \setminus B_i) \quad (10)$$

The union and the intersection of a sequence $\{A_n\}_{n \in \mathbb{N}} = \{A_n\}_{n=1}^{\infty}$ of sets is often written as $\bigcup_{n=1}^{\infty} A_n$ and $\bigcap_{n=1}^{\infty} A_n$. The meaning of notation like $\bigcup_{n=1}^k A_n$ should be obvious.

One can also form the Cartesian product of indexed families. If $\{A_i\}_{i \in I}$ is an indexed family, then $\prod_{i \in I} A_i$ is the set of all indexed sets $\{a_i\}_{i \in I}$ such that $a_i \in A_i$ for all $i \in I$. In particular,

$$\prod_{k=1}^n A_k = A_1 \times \cdots \times A_n = \{(a_1, \dots, a_n) : a_k \in A_k \text{ for } k = 1, \dots, n\}. \quad (11)$$

Relations

A **relation** from a set A to a set B is a subset of $A \times B$, that is, a set of ordered pairs (a, b) such that $a \in A$ and $b \in B$. If R is a relation, one often writes aRb instead of $(a, b) \in R$.

A relation from A to A is also called a **(binary) relation in A** . As an example of a relation in $\mathbb{R}$, consider the “less than” relation consisting of all ordered pairs of real numbers (x, y) with $x < y$.

The **domain** of a relation R from A to B is the set

$$\text{dom}(R) = \{a \in A : (a, b) \in R \text{ for some } b \in B\} = \{a \in A : aRb \text{ for some } b \in B\}$$

of possible first elements, whereas the **range** of R is the set

$$\text{range}(R) = \{b \in B : (a, b) \in R \text{ for some } a \in A\} = \{b \in B : aRb \text{ for some } a \in A\}$$

of possible second elements in the set R of ordered pairs. The **inverse** of a relation R from A to B is the relation R^{-1} from B to A given by $R^{-1} = \{(b, a) \in B \times A : (a, b) \in R\}$.

If R is a relation from A to B and S is a relation from B to C , we define the **composition** $S \circ R$ of R and S as the set of all (a, c) in $A \times C$ such that there is an element b in B with aRb and bSc . $S \circ R$ is then a relation from A to C .

A relation R in X is **reflexive** if xRx for all x in X . It is **transitive** if xRy and yRz imply xRz , it is **symmetric** if xRy implies yRx , it is **anti-symmetric** if xRy and yRx imply $x = y$, and it is **complete** if for all x and y in X at least one of xRy or yRx holds. A **partial ordering** in X is a relation in X that is reflexive, transitive, and anti-symmetric. If a partial ordering is complete, it is called a **linear (or total) ordering**. The relation $\leq$ in $\mathbb{R}$ is a linear ordering.

For $n \geq 2$, the less-than-or-equal-to relation $\leq$ in $\mathbb{R}^n$ is defined by $(x_1, \dots, x_n) \leq (y_1, \dots, y_n) \Leftrightarrow x_k \leq y_k$ for $k = 1, \dots, n$. The symbol $\leq$ is then usually taken to mean “ $\leq$, but not $=$ ”. (Some authors also use $<$ in this sense when comparing vectors. Caution is necessary!) There is also a strict inequality relation $\ll$, given by $(x_1, \dots, x_n) \ll (y_1, \dots, y_n) \Leftrightarrow x_k < y_k$ for all $k = 1, \dots, n$. The relation $\leq$ is a partial ordering in $\mathbb{R}^n$.

An **equivalence relation** in X is a relation that is reflexive, transitive, and symmetric. If R is an equivalence relation in X , then R induces a **partition** of X into pairwise disjoint **equivalence classes** $[x] = \{y \in X : yRx\}$. The union of all these equivalence classes is X .

Functions

A **function** (also called a **mapping**, **map**, or **transformation**) $f : X \rightarrow Y$ from a set X to a set Y is a rule that assigns exactly one element $y = f(x)$ in Y to each x in X . In set-theoretic terms f is a **relation** from X to Y such that:

$$(1) \text{ dom}(f) = X;$$

$$(2) \text{ for each } x \text{ in } X \text{ there is exactly one } y \text{ in } Y \text{ such that } xfy.$$

Thus f is “single-valued” and “operates on” every x in X . One usually writes $f(x) = y$ instead of xfy . The set X is the **domain** and Y is the **codomain** of f . The **range** of f is the same as the range of f considered as a relation:

$$\text{range}(f) = \{y \in Y : y = f(x) \text{ for at least one } x \text{ in } X\} = \{f(x) : x \in X\}$$

The last formulation is an example of a somewhat sloppy notation that is often used when the meaning is clear from the context. The **graph** of f is the set

$$\text{graph}(f) = \{(x, y) \in X \times Y : y = f(x)\}$$

This is of course the same as the relation f as defined in the previous subsection.

If $f(x) = y$, one also writes $f : x \mapsto y$. The squaring function $s : \mathbb{R} \rightarrow \mathbb{R}$, for example, can then be written as $s : x \mapsto x^2$. Thus, $\mapsto$ indicates the effect of the function on an element of the domain. If the domain A of $f : A \rightarrow B$ is a subset of a set X , it is sometimes convenient to write $f : A \subseteq X \rightarrow B$, and maybe even $f : A \subseteq X \rightarrow B \subseteq Y$ if B is a subset of Y .

If $f : A \rightarrow B$ is a function and $S \subseteq A$, the **restriction** of f to S is the function $f|_S$ defined by $f|_S(x) = f(x)$ for every x in S .

A function $f : A \rightarrow B$ is “**one-to-one**” or **injective** if $f(a) \neq f(a')$ whenever $a \neq a'$, i.e. if f always maps distinct points in A to distinct points in B .

If the range of $f : A \rightarrow B$ is all of B , then f is called **surjective** or **onto**, and f is said to map A **onto** B .

When $f : A \rightarrow B$ is both injective and surjective (i.e. when it is both one-to-one and onto), it is called **bijective**.

If $f : A \rightarrow B$ is injective, then f has an inverse function $f^{-1} : \text{range}(f) \rightarrow A$, defined by $f^{-1}(b) = a \Leftrightarrow b = f(a)$. (Considered as a relation, f always has an inverse with $\text{range}(f)$ as its domain, but this inverse relation is a function only when f is injective.)

The **composition** of a function $f : A \rightarrow B$ and a function $g : B \rightarrow C$ is the function $g \circ f : A \rightarrow C$ given by $(g \circ f)(a) = g(f(a))$ for all a in A . It is easy to check that this is the same as the composition of f and g as relations.

Direct and Inverse Images

Let $f : A \rightarrow B$ be a function. The **(direct) image** under f of a subset S of A is the set

$$f(S) = \{y \in B : y = f(x) \text{ for some } x \text{ in } S\}$$

and the **inverse image** or **preimage** under f of a set $T \subseteq B$ is

$$f^{-1}(T) = \{x \in A : f(x) \in T\}$$

Direct and inverse images satisfy a number of relations. Given indexed families $\{S_i\}_{i \in I}$ and

$\{T_i\}_{i \in I}$ of subsets of A and B , respectively, these include:

$$f(S_1 \cup S_2) = f(S_1) \cup f(S_2), \quad f\left(\bigcup_{i \in I} S_i\right) = \bigcup_{i \in I} f(S_i) \quad (12)$$

$$f(S_1 \cap S_2) \subseteq f(S_1) \cap f(S_2), \quad f\left(\bigcap_{i \in I} S_i\right) \subseteq \bigcap_{i \in I} f(S_i) \quad (13)$$

$$f^{-1}(T_1 \cup T_2) = f^{-1}(T_1) \cup f^{-1}(T_2), \quad f^{-1}\left(\bigcup_{i \in I} T_i\right) = \bigcup_{i \in I} f^{-1}(T_i) \quad (14)$$

$$f^{-1}(T_1 \cap T_2) = f^{-1}(T_1) \cap f^{-1}(T_2), \quad f^{-1}\left(\bigcap_{i \in I} T_i\right) = \bigcap_{i \in I} f^{-1}(T_i) \quad (15)$$

Note that inverse images preserve both unions and intersections, whereas direct images preserve only unions, not intersections (see Problem 3).

Some other properties of direct and inverse images are:

$$S \subseteq f^{-1}(f(S)), \quad f(f^{-1}(T)) \subseteq T, \quad f^{-1}(T) = f^{-1}(T \cap \text{range}(f)) \quad (16)$$

$$S_1 \subseteq S_2 \implies f(S_1) \subseteq f(S_2) \text{ and } f(S_2) \setminus f(S_1) \subseteq f(S_2 \setminus S_1) \quad (17)$$

$$T_1 \subseteq T_2 \implies f^{-1}(T_1) \subseteq f^{-1}(T_2) \text{ and } f^{-1}(T_2 \setminus T_1) = f^{-1}(T_2) \setminus f^{-1}(T_1) \quad (18)$$

PROBLEMS FOR SECTION A.1

1. The ordered pair (a, b) is most commonly defined as the set $(a, b) = \{\{a\}, \{a, b\}\}$. Show that with this definition, $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$.
2. Show the equalities $\text{dom}(R^{-1}) = \text{range}(R)$ and $\text{range}(R^{-1}) = \text{dom}(R)$ for a relation R from A to B .
3. Give an example to show that the inclusion signs in (13) cannot be replaced by equals signs.
4. Show that if R is a linear ordering in a set X , then the inverse relation R^{-1} is also a linear ordering.
5. Prove (16). Also, give examples to show that $\subseteq$ cannot always be replaced by $=$.

A.2 Least Upper Bound Principle

The real number system is fully characterized by a rather small number of axioms. The usual algebraic rules and the rules for inequalities are all well known. Here we consider briefly only the so-called *least upper bound principle*. An understanding of this principle is crucial for many of the arguments in this book. The need for this principle can be illustrated by the problem of determining the area of a circle. We know how to calculate the area of plane regions bounded by straight lines. Figure 1 shows a sequence of regular polygons inscribed in a circle. For $n \geq 3$, let A_n be the area of a regular n -sided polygon inscribed in the circle. Thus, A_3 is the area of an equilateral triangle, A_4 is the area of a square, A_5 is the area of a regular pentagon, and so on.

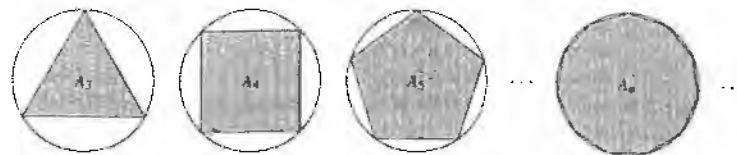


Figure 1

The area of the circle must be greater than each of the numbers $A_3, A_4, A_5, \dots$. On the other hand it seems clear that by choosing n sufficiently large, we can make the difference between the area of the circle and the area of an inscribed regular n -gon as small as we please. We now *define* the area of the circle as the smallest number greater than or equal to each of the numbers $A_3, A_4, A_5, \dots$. This definition makes sense only because the existence of such a number is a basic property of the set of real numbers, called the principle of least upper bound.

Recall that a set S of real numbers is **bounded above** if there exists a real number b such that $b \geq x$ for all x in S . Any such number b is called an **upper bound** for S . A set that is bounded above has many upper bounds. A **least upper bound** for the set S is a number b^* that is an upper bound for S and is such that $b^* \leq b$ for every upper bound b .

The existence of a least upper bound is a basic and non-trivial property of the real number system.

LEAST UPPER BOUND PRINCIPLE

Any nonempty set of real numbers that is bounded above has a least upper bound. (1)

A set S can have at most one least upper bound, because if b_1^* and b_2^* are both least upper bounds for S , then $b_1^* \leq b_2^*$ and $b_2^* \leq b_1^*$, and thus $b_1^* = b_2^*$. The least upper bound b^* of S is often called the **supremum** of S . We write $b^* = \sup S$ or $b^* = \sup_{x \in S} x$.

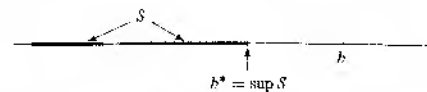


Figure 2 Any $b \geq b^*$ is an upper bound for S , but $b^* = \sup S$ is the unique least upper bound for S .

EXAMPLE 1

The set $S = (0, 5)$, consisting of all x such that $0 < x < 5$, has many upper bounds, including 100, 6.73, and 5. Clearly, no number smaller than 5 can be an upper bound, so 5 is the least upper bound. Thus $\sup S = 5$.

The set $T = \{x : x^2 < 2\} = (-\sqrt{2}, \sqrt{2})$ has many upper bounds, which include 9, 2, and $\sqrt{2}$. Clearly, no number smaller than $\sqrt{2}$ can be an upper bound of T , and so $\sqrt{2}$ is the least upper bound, $\sup T = \sqrt{2}$.

A set S is **bounded below** if there exists a real number a such that $x \geq a$ for all x in S . The number a is called a **lower bound** for S . A set S that is bounded below has a **greatest lower bound** a^* , with the property $a^* \leq x$ for all x in S , and $a^* \geq a$ for all lower bounds a . The number a^* is called the **infimum** of S and we write $a^* = \inf S$ or $a^* = \inf_{x \in S} x$. Thus,

$$\sup S = \text{the least number greater than or equal to all numbers in } S \quad (2)$$

$$\inf S = \text{the greatest number less than or equal to all numbers in } S \quad (3)$$

If S is not bounded below, we write $\inf S = -\infty$. If S is not bounded above, we write $\sup S = \infty$.

Every real number is an upper bound for $\emptyset$, the *empty set*. Therefore $\emptyset$ has no *least* upper bound, and by convention we write $\sup \emptyset = -\infty$. Similarly, $\inf \emptyset = \infty$.

The following characterization of the supremum is easy to prove:

THEOREM A.2.1

Let S be a set of real numbers and b^* a real number. Then $\sup S = b^*$ if and only if the following two conditions are satisfied:

- (a) $x \leq b^*$ for all x in S .
- (b) For each $\varepsilon > 0$ there exists an x in S such that $x > b^* - \varepsilon$.

NOTE 1 The existence of a supremum of a set bounded above may seem to be “intuitively evident”: Start with a nonempty set S of real numbers and an upper bound b for S , as in Fig. 2. Then move b to the left until it is “stopped” by the set S . The number b is still an upper bound for S , and it is the least of all the upper bounds.

The least upper bound principle is a non-trivial property of the real number system. To appreciate this, observe that the principle does not hold within the set $\mathbb{Q}$ of rational numbers. For instance, let S be the set of all rational numbers r such that $r^2 < 2$. Within the rational number system, the set S has no least upper bound. All rational numbers larger than or equal to $\sqrt{2}$ are upper bounds for S , but there is not a smallest one among these numbers because $\sqrt{2}$ is irrational. (See Problem 2.)

PROBLEMS FOR SECTION A.2

1. Determine $\sup$ and $\inf$ for each of these three sets:

$$A = (-3, 7], \quad B = \{1/n : n = 1, 2, 3, \dots\}, \quad C = \{x : x > 0 \text{ and } x^2 > 3\}$$

2. Suppose that r is a rational number with $r > \sqrt{2}$. Show that the rational number $s = (2+r^2)/2r$ satisfies $\sqrt{2} < s < r$.

3. Show that $\sup S = \infty$ iff for every b in $\mathbb{R}$ there exists an x in S such that $x > b$.

A.3 Sequences of Real Numbers

A **sequence** can be viewed as a function $k \mapsto x(k)$ with the set $\mathbb{N} = \{1, 2, 3, \dots\}$ of all positive integers as its domain. The **terms** $x(1), x(2), x(3), \dots, x(k), \dots$ of the sequence are usually denoted by using subscripts: $x_1, x_2, x_3, \dots, x_k, \dots$. We shall use the notation $\{x_k\}_{k=1}^\infty$, or simply $\{x_k\}_k$, or even just $\{x_k\}$, to indicate an arbitrary sequence of real numbers.

A sequence $\{x_k\}$ of real numbers is said to be

- (a) **increasing** (or **nondecreasing**) if $x_k \leq x_{k+1}$ for $k = 1, 2, \dots$;
- (b) **strictly increasing** if $x_k < x_{k+1}$ for $k = 1, 2, \dots$;
- (c) **decreasing** (or **nonincreasing**) if $x_k \geq x_{k+1}$ for $k = 1, 2, \dots$;
- (d) **strictly decreasing** if $x_k > x_{k+1}$ for $k = 1, 2, \dots$.

A sequence that is increasing or decreasing is called **monotone**.

EXAMPLE 1 Decide whether or not the three sequences of real numbers whose general terms are given below are monotone:

$$(a) \ x_k = 1 - 1/k \quad (b) \ y_k = (-1)^k \quad (c) \ z_k = \sqrt{k+1} - \sqrt{k}$$

Solution: The sequence $\{x_k\}$ is (strictly) increasing, because for $k = 1, 2, \dots$,

$$x_{k+1} > x_k \iff 1 - 1/(1+k) > 1 - 1/k \iff 1/(k+1) < 1/k$$

and the last inequality clearly holds for all $k \geq 1$.

The sequence $\{y_k\}$ is not monotone. It is clearly neither increasing nor decreasing, because its terms are $-1, 1, -1, 1, -1, \dots$.

The first three terms of the sequence $\{z_k\}$ are: $z_1 = \sqrt{2} - \sqrt{1} \approx 0.4142$, $z_2 = \sqrt{3} - \sqrt{2} \approx 0.3178$, and $z_3 = \sqrt{4} - \sqrt{3} \approx 0.2679$. Note that $z_1 > z_2 > z_3$. In fact, a standard trick shows that $\{z_k\}$ is indeed strictly decreasing:

$$z_k = \sqrt{k+1} - \sqrt{k} = \frac{(\sqrt{k+1} - \sqrt{k})(\sqrt{k+1} + \sqrt{k})}{\sqrt{k+1} + \sqrt{k}} = \frac{1}{\sqrt{k+1} + \sqrt{k}}$$

From the last fraction we see that z_k decreases when k increases.

A sequence $\{x_k\}$ is said to **converge** to a number x if x_k becomes arbitrarily close to x for all sufficiently large k . We write

$$\lim_{k \rightarrow \infty} x_k = x \quad \text{or} \quad x_k \rightarrow x \quad \text{as} \quad k \rightarrow \infty$$

The precise definition of convergence is as follows:

DEFINITION OF A CONVERGENT SEQUENCE

The sequence $\{x_k\}$ converges to x , and we write

$$\lim_{k \rightarrow \infty} x_k = x$$

(1)

if for every $\varepsilon > 0$ there exists a natural number N such that $|x_k - x| < \varepsilon$ for all $k > N$. The number x is called the **limit** of the sequence $\{x_k\}$. A **convergent** sequence is one that converges to some number.

Note that the limit of a convergent sequence is unique. (See Problem 7.) A sequence that does not converge to any real number is said to **diverge**.

In some cases we use the notation $\lim_{k \rightarrow \infty} x_k$ even if the sequence $\{x_k\}$ is divergent: if for each number M there exists a number N such that $x_k \geq M$ for all natural number $k \geq N$, then we say that x_k tends to ∞ , and write $\lim_{k \rightarrow \infty} x_k = \infty$. In the same way we write $\lim_{k \rightarrow \infty} x_k = -\infty$ if for every number M there exists a number N such that $x_k \leq -M$ for all $k \geq N$.

In Example 1, the sequence $\{x_k\}$ converges to 1 because $1/k$ tends to 0 as k tends to ∞ . (Using (1): Given $\varepsilon > 0$, we must find a number N such that for $k > N$ we have $|(1 - 1/k) - 1| < \varepsilon$, i.e. $1/k < \varepsilon$, or $k > 1/\varepsilon$. Clearly, this is accomplished by choosing an $N \geq 1/\varepsilon$.)

The sequence $\{y_k\}$ is divergent. If k is even, $y_k = 1$ and if k is odd, $y_k = -1$. So there is clearly no number y such that y_k tends to y as k tends to ∞ .

The sequence $\{z_k\}$ is convergent, with $\lim_{k \rightarrow \infty} z_k = \lim_{k \rightarrow \infty} 1/(\sqrt{k+1} + \sqrt{k}) = 0$.

A sequence $\{x_k\}$ is **bounded** if there exists a number M such that $|x_k| \leq M$ for all $k = 1, 2, \dots$. It is easy to see that every convergent sequence is bounded: If $x_k \rightarrow x$, then by the definition of convergence, only finitely many terms of the sequence can lie outside the interval $I = (x - 1, x + 1)$. The set I is bounded and the finite set of points from the sequence that are not in I is bounded, so $\{x_k\}$ must be bounded. On the other hand, is every bounded sequence convergent? No. For example, the sequence $\{y_k\} = \{(-1)^k\}$ in Example 1 is bounded but not convergent. Suppose, however, that the sequence is monotone as well as bounded. Then it is convergent.

THEOREM A.3.1

Every bounded monotone sequence is convergent.

Proof: Suppose that $\{x_k\}$ is increasing and bounded. Let b^* be the least upper bound of the set $X = \{x_k : k = 1, 2, \dots\}$, and let ε be an arbitrary positive number. Then $b^* - \varepsilon$ is not an upper bound of X , so there must be a term x_N of the sequence for which $x_N > b^* - \varepsilon$. Because the sequence is increasing, $b^* - \varepsilon < x_N \leq x_k$ for all $k > N$. But the x_k are all less than or equal to b^* , so $b^* - \varepsilon < x_k \leq b^*$. Thus, for any $\varepsilon > 0$ there exists a number N

such that $|x_k - b^*| < \varepsilon$ for all $k > N$. Hence, $\{x_k\}$ converges to b^* . If $\{x_k\}$ is decreasing and bounded, the argument is analogous. ■

EXAMPLE 2

Consider the sequence $\{x_k\}$ defined by

$$x_1 = \sqrt{2}, \quad x_{k+1} = \sqrt{x_k + 2}, \quad k = 1, 2, \dots \quad (*)$$

Use Theorem A.3.1 to prove that the sequence is convergent and find its limit. (Hint: Prove by induction that $x_k < 2$ for all k . Then prove that the sequence is (strictly) increasing.)

Solution: Note that $x_1 = \sqrt{2} < 2$, and if $x_k < 2$, then $x_{k+1} = \sqrt{x_k + 2} < \sqrt{2 + 2} = 2$, so by induction, $x_k < 2$ for all k . Moreover, because $x_k < 2$, one has

$$x_{k+1} = \sqrt{x_k + 2} > \sqrt{x_k + x_k} = \sqrt{2x_k} > \sqrt{x_k^2} = x_k$$

so $\{x_k\}$ is (strictly) increasing. By Theorem A.3.1 the sequence is convergent. If x is its limit, then letting $k \rightarrow \infty$ in (*) yields $x = \sqrt{x + 2}$, by the continuity of $\sqrt{x}$. This equation implies that $x^2 = x + 2$, which has the two solutions -1 and $x = 2$. Because -1 is obviously not a solution of $x = \sqrt{x + 2}$, the only solution is $x = 2$, and thus $\lim_{k \rightarrow \infty} x_k = 2$. ■

Rules for Handling Convergent Sequences

Suppose that $\{x_k\}$ converges to x and $\{y_k\}$ converges to y as $k \rightarrow \infty$. For k sufficiently large, x_k is close to x and y_k is close to y . Then $x_k + y_k$ must be close to $x + y$, and it is therefore reasonable to believe that $x_k + y_k \rightarrow x + y$ as $k \rightarrow \infty$. Corresponding results hold for subtraction, multiplication, and division. In fact, we have the following result:

THEOREM A.3.2 (RULES FOR SEQUENCES)

Suppose that the sequences $\{x_k\}$ and $\{y_k\}$ converge to x and y , respectively. Then:

- (a) $\lim_{k \rightarrow \infty} (x_k + y_k) = x + y$
- (b) $\lim_{k \rightarrow \infty} (x_k - y_k) = x - y$
- (c) $\lim_{k \rightarrow \infty} (x_k \cdot y_k) = x \cdot y$
- (d) $\lim_{k \rightarrow \infty} (x_k / y_k) = x / y$, assuming that $y_k \neq 0$ for all k and $y \neq 0$.

Proof: (a) Here is a formal proof: Let ε be an arbitrary positive number. Since $\{x_k\}$ is convergent, there exists a number N_1 such that $|x_k - x| < \varepsilon/2$ for all $k > N_1$. In the same way there exists a number N_2 such that $|y_k - y| < \varepsilon/2$ for all $k > N_2$. Let N be the greater of the two numbers N_1 and N_2 . Then for $k > N$,

$$|(x_k + y_k) - (x + y)| = |(x_k - x) + (y_k - y)| \leq |x_k - x| + |y_k - y| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

But this means that $\lim_{k \rightarrow \infty} (x_k + y_k) = x + y = \lim_{k \rightarrow \infty} x_k + \lim_{k \rightarrow \infty} y_k$.

The statement in (b) is proved in the same way. Proving (c) and (d) requires more complicated arguments to show that $|x_k y_k - xy|$ and $|x_k / y_k - x / y|$ are less than an arbitrary positive ε . (For a precise proof we refer to e.g. Marsden and Hoffman (1993).) ■

Subsequences

Let $\{x_k\}$ be a sequence. Consider a strictly increasing sequence of natural numbers $k_1 < k_2 < k_3 < \dots$, and form a new sequence $\{y_j\}_{j=1}^{\infty}$, where $y_j = x_{k_j}$ for $j = 1, 2, \dots$. The sequence $\{y_j\}_{j=1}^{\infty} = \{x_{k_j}\}_{j=1}^{\infty}$ is called a **subsequence** of $\{x_k\}$. Because the sequence $\{k_j\}$ is strictly increasing, $k_j \geq j$ for all j . The terms of the subsequence are all present in the original one. In fact, a subsequence can be viewed as the result of removing some (possibly none) of the terms of the original sequence. For example, $x_5, x_6, x_7, \dots$ is the subsequence obtained by striking out the first four terms of the original sequence, and $x_2, x_4, x_6, \dots$ is the subsequence obtained by removing all terms with an odd index. If $x_k = (-1)^k$ is the divergent sequence mentioned above, we may for example define the two subsequences $\{x_{2k}\}$ and $\{x_{2k-1}\}$. Here $x_{2k} = (-1)^{2k} = 1$, and $x_{2k-1} = (-1)^{2k-1} = -1$ for all k . Note that these two particular subsequences happen to be both convergent.

NOTE 1 Some proofs involve pairs of sequences $\{x_k\}_{k=1}^{\infty}$ and $\{x_{k_j}\}_{j=1}^{\infty}$ where $k_j \geq j$ for all j , but where the sequence $k_1, k_2, \dots$ is not necessarily strictly increasing. Thus $\{x_{k_j}\}_{j=1}^{\infty}$ is “not quite” a subsequence of $\{x_k\}_{k=1}^{\infty}$. However, it is always possible to select terms from $\{x_{k_j}\}_{j=1}^{\infty}$ in such a way that we get a subsequence $\{x_{k_i}\}_{i=1}^{\infty}$ of $\{x_k\}_{k=1}^{\infty}$: Let $k_1 = k_1$, and generally $k_{i+1} = k_{k_i+1}$. Then $k_{i+1} \geq k_i + 1 > k_i$.

The following important fact follows immediately from the definition of convergence:

THEOREM A.3.3

Every subsequence of a convergent sequence is itself convergent, and has the same limit as the original sequence.

The following result is less obvious but very useful:

THEOREM A.3.4

If the sequence $\{x_k\}$ is bounded, then it contains a convergent subsequence.

Proof: Suppose that $|x_k| \leq M$ for all $k = 1, 2, \dots$. Let $y_n = \sup\{x_k : k \geq n\}$ for $n = 1, 2, \dots$. Then $\{y_n\}$ is a decreasing sequence because the set $\{x_k : k \geq n\}$ shrinks as n increases. The sequence is also bounded because $y_n \in [-M, M]$. According to Theorem A.3.1, the sequence $\{y_n\}$ has a limit $x = \lim_{n \rightarrow \infty} y_n \in [-M, M]$. By Theorem A.2.1, the definition of y_n implies that there is a term x_{k_n} from the original sequence $\{x_k\}$ (with $k_n \geq n$) satisfying $|y_n - x_{k_n}| < 1/n$. But then

$$|x - x_{k_n}| = |x - y_n + y_n - x_{k_n}| \leq |x - y_n| + |y_n - x_{k_n}| < |x - y_n| + 1/n \quad (*)$$

This shows that $x_{k_n} \rightarrow x$ as $n \rightarrow \infty$. By using the construction in Note 1, we can extract from $\{x_{k_n}\}$ a subsequence of $\{x_k\}$ that converges to x . ■

Cauchy Sequences

The definition (1) of a convergent sequence involves the specific value of the limit. If this limit is unknown, or inconvenient to calculate, the definition is not very useful because one cannot test all numbers to see if they meet the criterion. An important alternative necessary and sufficient condition for convergence is based on the following concept:

DEFINITION OF A CAUCHY SEQUENCE

A sequence $\{x_k\}$ of real numbers is called a **Cauchy sequence** if for every $\varepsilon > 0$, there exists a natural number N such that

$$|x_n - x_m| < \varepsilon \quad \text{for all } n > N \text{ and all } m > N \quad (2)$$

Note that the terms of a Cauchy sequence will eventually be close together, which need not be the case for an arbitrary sequence. In particular, the sequence $\{x_k\} = \{(-1)^k\}$, whose terms are alternatively -1 and 1 , is clearly not a Cauchy sequence. But the sequence $\{y_k\} = \{(-\frac{1}{2})^k\}$ is.

All the terms of a convergent sequence eventually cluster around the limit, so the sequence is a Cauchy sequence. The converse is also true—that is, every Cauchy sequence is convergent:

THEOREM A.3.5

A sequence is convergent if and only if it is a Cauchy sequence.

Proof: To prove the “only if” part, suppose that $\{x_k\}$ converges to x . Given $\varepsilon > 0$, choose a natural number N such that $|x_k - x| < \varepsilon/2$ for all $k > N$. Then for $k > N$ and $m > N$,

$$|x_k - x_m| = |(x_k - x) + (x - x_m)| \leq |x_k - x| + |x - x_m| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

Therefore $\{x_k\}$ is a Cauchy sequence.

To prove the “if” part, suppose $\{x_k\}$ is a Cauchy sequence. We first show that the sequence is bounded. By the Cauchy property, there is a number M such that $|x_k - x_M| < 1$ for $k > M$. This means that all points x_k with $k > M$ have a distance from x_M that is less than 1. Moreover, the finite set $\{x_1, x_2, \dots, x_{M-1}\}$ is surely bounded. Hence, $\{x_k\}$ is bounded.

By Theorem A.3.4, therefore, the sequence $\{x_k\}$ has a convergent subsequence $\{x_{k_j}\}$. Let $x = \lim_{j \rightarrow \infty} x_{k_j}$. Because $\{x_k\}$ is a Cauchy sequence, for every $\varepsilon > 0$ there is a natural number N such that $|x_n - x_m| < \varepsilon/2$ for $n > N$ and $m > N$. Moreover, if J is sufficiently large, $|x_{k_j} - x| < \varepsilon/2$ for all $j > J$. Then for $k > N$ and $j > \max\{N, J\}$,

$$|x_k - x| \leq |x_k - x_{k_j}| + |x_{k_j} - x| < \varepsilon/2 + \varepsilon/2 = \varepsilon$$

Hence, $x_k \rightarrow x$ as $k \rightarrow \infty$. ■

EXAMPLE 3 Prove that the sequence $\{x_k\}$ with the general term $x_k = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{k^2}$ is a Cauchy sequence.

Solution: Let n and m be natural numbers with $m > n$, and define $p = m - n$. Then

$$\begin{aligned} |x_m - x_n| &= |x_{n+p} - x_n| = \frac{1}{(n+1)^2} + \frac{1}{(n+2)^2} + \cdots + \frac{1}{(n+p)^2} \\ &< \frac{1}{n(n+1)} + \frac{1}{(n+1)(n+2)} + \cdots + \frac{1}{(n+p-1)(n+p)} \\ &= \left(\frac{1}{n} - \frac{1}{n+1}\right) + \left(\frac{1}{n+1} - \frac{1}{n+2}\right) + \cdots + \left(\frac{1}{n+p-1} - \frac{1}{n+p}\right) \\ &= \frac{1}{n} - \frac{1}{n+p} < \frac{1}{n} \end{aligned}$$

Thus, for any $\varepsilon > 0$, if we choose $n > 1/\varepsilon$, then $|x_m - x_n| < \varepsilon$ for all $m > n$. This proves that $\{x_k\}$ is a Cauchy sequence. $\blacksquare$

NOTE 2 The infinite series $\sum_{n=1}^{\infty} x_n$ is said to converge if the sequence $\{s_k\}$ of partial sums $s_k = x_1 + x_2 + \cdots + x_k$ is convergent. It follows from the previous example and Theorem A.3.5 that the infinite series

$$\sum_{n=1}^{\infty} \frac{1}{n^2}$$

is convergent. In fact, one can prove that this infinite series converges to $\pi^2/6$.

Upper and Lower Limits

Let $\{x_k\}$ be a sequence that is bounded above, and define $y_n = \sup\{x_k : k \geq n\}$ for $n = 1, 2, \dots$. Each y_n is a finite number and $\{y_n\}_n$ is a decreasing sequence. Then $\lim_{n \rightarrow \infty} y_n$ is either finite or is $-\infty$. We call this limit the **upper limit** (or **lim sup**) of the sequence $\{x_k\}$, and we introduce the following notation:

$$\limsup_{k \rightarrow \infty} x_k = \lim_{n \rightarrow \infty} (\sup\{x_k : k \geq n\}) \quad (3)$$

If $\{x_k\}$ is not bounded above, we write $\limsup_{k \rightarrow \infty} x_k = \infty$.

Similarly, if $\{x_k\}$ is bounded below, its **lower limit** (or **lim inf**), is defined as

$$\liminf_{k \rightarrow \infty} x_k = \lim_{n \rightarrow \infty} (\inf\{x_k : k \geq n\}) \quad (4)$$

If $\{x_k\}$ is not bounded below, we write $\liminf_{k \rightarrow \infty} x_k = -\infty$. The symbols $\limsup$ and $\liminf$ are often written as $\overline{\lim}$ and $\underline{\lim}$.

The following characterization of $\lim$ is often useful:

A CHARACTERIZATION OF THE UPPER LIMIT

Let $\{x_k\}$ be a sequence of real numbers and b^* a (finite) real number. Then $\overline{\lim}_{k \rightarrow \infty} x_k = b^*$ if and only if the following two conditions are satisfied:

- For every $\varepsilon > 0$ there exists an integer N such that $x_k < b^* + \varepsilon$ for all $k > N$.
 - For every $\varepsilon > 0$ and every integer M , there exists an integer $k > M$ such that $x_k > b^* - \varepsilon$.
- (5)

NOTE 3 A similar characterization holds for $\underline{\lim}$. Note that condition (a) means that ultimately all terms of the sequence lie to the left of $b^* + \varepsilon$ on the real line. Condition (b) means that for any $\varepsilon > 0$, however small, infinitely many terms lie to the right of $b^* - \varepsilon$.

EXAMPLE 4 Determine the $\overline{\lim}$ and $\underline{\lim}$ of the following sequences:

$$(a) \{x_k\} = \{(-1)^k\} \quad (b) \{x_k\} = \left\{(-1)^k \left(2 + \frac{1}{k}\right) + 1\right\}$$

Solution: (a) For every n there exists a number $k \geq n$ with $(-1)^k = 1$. Hence $y_n = \sup\{(-1)^k : k \geq n\} = 1$, and so $\lim_{n \rightarrow \infty} y_n = 1$. Thus $\overline{\lim}_{k \rightarrow \infty} x_k = 1$. In the same way we see that $\underline{\lim}_{k \rightarrow \infty} x_k = -1$.

(b) Arguments similar to those in (a) yield: $\overline{\lim}_{k \rightarrow \infty} x_k = 3$, $\underline{\lim}_{k \rightarrow \infty} x_k = -1$. $\blacksquare$

It is not difficult to see that $\lim_{k \rightarrow \infty} x_k \leq \overline{\lim}_{k \rightarrow \infty} x_k$ for every sequence $\{x_k\}$. The following result is also rather easy and we leave the proof to the reader:

THEOREM A.3.6

If the sequence $\{x_k\}$ is convergent, then

$$\overline{\lim}_{k \rightarrow \infty} x_k = \underline{\lim}_{k \rightarrow \infty} x_k = \lim_{k \rightarrow \infty} x_k$$

Conversely, if $\overline{\lim}_{k \rightarrow \infty} x_k = \underline{\lim}_{k \rightarrow \infty} x_k$ and both are finite, then $\{x_k\}$ is convergent.

PROBLEMS FOR SECTION A.3

1. Prove that the sequence $\{x_k\}$ defined by

$$x_1 = 1, \quad x_{k+1} = 2\sqrt{x_k}, \quad k = 1, 2, \dots$$

converges, and find its limit. (Hint: Prove first by induction that $x_k < 4$ for all k .)

2. Prove that for the sequence $\{x_k\}$ in Example 2, $|x_{k+1} - 2| < \frac{1}{2}|x_k - 2|$, and use this to prove that $x_k \rightarrow 2$ as $k \rightarrow \infty$. (Hint: $x_{k+1} - 2 = (x_{k+1}^2 - 4)/(x_{k+1} + 2)$.)

3. Let S be a nonempty set of real numbers bounded above, and let $b^* = \sup S$. Show that there exists a sequence $\{x_n\}$, $x_n \in S$, such that $x_n \rightarrow b^*$. (Hint: See Thm. A.2.1.)

SM 4. Find all possible limits of subsequences of the two sequences defined by

$$x_k = 1 - 1/k + (-1)^k, \quad y_k = (1 + 1/k) \sin(k\pi/3)$$

SM 5. (a) Consider the two sequences with general terms $x_k = \frac{1}{2}(1 + (-1)^k)$, $y_k = \frac{1}{2}(1 - (-1)^k)$. Compute $\lim_{k \rightarrow \infty} x_k$ and $\lim_{k \rightarrow \infty} y_k$ of $\{x_k\}$, $\{y_k\}$, $\{x_k + y_k\}$, and $\{x_k y_k\}$.

(b) Prove that if two sequences $\{x_k\}$ and $\{y_k\}$ are bounded, then

$$(i) \quad \lim_{k \rightarrow \infty} (x_k + y_k) \leq \lim_{k \rightarrow \infty} x_k + \lim_{k \rightarrow \infty} y_k$$

$$(ii) \quad \lim_{k \rightarrow \infty} (x_k y_k) \leq \lim_{k \rightarrow \infty} x_k \cdot \lim_{k \rightarrow \infty} y_k \text{ if } x_k \geq 0 \text{ and } y_k \geq 0 \text{ for all } k$$

Note that the examples in (a) show that the inequality signs $\leq$ in (i) and (ii) cannot be replaced by equals signs.

SM 6. Let $\{x_k\}$ be a sequence such that $|x_{k+1} - x_k| < 1/2^k$ for all $k = 1, 2, \dots$. Prove that $\{x_k\}$ is a Cauchy sequence.

7. Prove that if $\{x_k\}$ converges to both x and y , then $x = y$.

HARDER PROBLEMS

8. Consider the two sequences $\{a_n\}$ and $\{b_n\}$ defined by

$$a_n = (1 + 1/n)^n, \quad b_n = (1 + 1/n)^{n+1}, \quad n = 1, 2, \dots$$

(a) Show that $a_1 < a_2 < a_3 < a_4$, and that $b_1 > b_2 > b_3 > b_4$.

(b) Use induction to prove Bernoulli's inequality,

$$(1 + x)^n \geq 1 + nx \quad \text{for } x \geq -1 \text{ and } n = 1, 2, 3, \dots$$

Show also that for $n > 1$ equality holds only for $x = 0$.

(c) Let $x = -1/n^2$ in the inequality in (b) and multiply by $(n/(n-1))^n$. Deduce that $a_n > a_{n-1}$, so $\{a_n\}$ is strictly increasing.

(d) Let $x = 1/(n^2 - 1)$ in the inequality in (b), and show that $(1 + 1/(n^2 - 1))^n > 1 + n/(n^2 - 1) > 1 + 1/n$. Then multiply by $(1 + 1/n)^n$ and show that $b_n < b_{n-1}$, so $\{b_n\}$ is strictly decreasing.

(e) Of course, $a_n < b_n$ for all n . Explain why the results in (c) and (d) show that $\{a_n\}$ and $\{b_n\}$ both converge. Because $b_n = a_n(1 + 1/n)$, the two sequences have the same limit. The common limit is e , and so $(1 + 1/n)^n < e < (1 + 1/n)^{n+1}$ for all n . For $n = 100$, we get $2.7048 < e < 2.7319$. As you surely know, the irrational number $e \approx 2.718281828$ is one of the most important numbers in mathematics.

9. Prove that every sequence of real numbers has a monotone subsequence.

A.4 Infimum and Supremum of Functions

Suppose that $f(x)$ is defined for all x in B , where $B \subseteq \mathbb{R}^k$. Using (A.2.2) and (A.2.3), we define the **infimum** and **supremum** of the function f over B as

$$\inf_{x \in B} f(x) = \inf\{f(x) : x \in B\}, \quad \sup_{x \in B} f(x) = \sup\{f(x) : x \in B\} \quad (1)$$

EXAMPLE 1 Let $f(x) = e^{-x}$ be defined over $B = (-\infty, \infty)$. Find $\inf_{x \in B} f(x)$ and $\sup_{x \in B} f(x)$.

Solution: The range of $f(x)$ is the interval $(0, \infty)$. Therefore $\inf_{x \in B} f(x) = 0$, while $\sup_{x \in B} f(x) = \infty$.

Example 1 illustrates an important point: $\inf_{x \in B} f(x) = 0$, but for no number x is $f(x) = e^{-x} = 0$.

If a function f is defined over a set B , if $\inf_{x \in B} f(x) = y$, and if there exists a c in B such that $f(c) = y$, then we say that the infimum is **attained** (at the point c) in B . In this case the infimum y is called the **minimum** of f over B , and we often write $\min$ instead of $\inf$. In the same way we write $\max$ instead of $\sup$ when the supremum of f over B is attained in B , and so becomes the maximum.

The following properties are sometimes useful. If the infimum and/or supremum value is attained, $\inf$ and $\sup$ can be replaced by $\min$ and $\max$, respectively.

$$(a) \quad \sup_{x \in B} (-f(x)) = -\inf_{x \in B} f(x) \quad (b) \quad \inf_{x \in B} (-f(x)) = -\sup_{x \in B} f(x) \quad (2)$$

$$\inf_{x \in B} (f(x) + g(x)) \geq \inf_{x \in B} f(x) + \inf_{x \in B} g(x) \quad (3)$$

$$\sup_{x \in B} (f(x) + g(x)) \leq \sup_{x \in B} f(x) + \sup_{x \in B} g(x) \quad (4)$$

$$\inf_{x \in B} (\lambda f(x)) = \lambda \inf_{x \in B} f(x) \quad (\lambda \text{ is a positive real number}) \quad (5)$$

$$\sup_{x \in B} (\lambda f(x)) = \lambda \sup_{x \in B} f(x) \quad (\lambda \text{ is a positive real number}) \quad (6)$$

Property (2)(a) is illustrated in Fig. 1.

Consider the inequality signs in (3) and (4). If $f(x) = x$ and $g(x) = -x$ in $(0, 1]$, then $\inf f(x) = 0$ and $\inf g(x) = -1$, whereas $\inf(f(x) + g(x)) = \inf 0 = 0$. In this case $\inf(f(x) + g(x)) = 0 > \inf f(x) + \inf g(x) = -1$. This is illustrated in Fig. 2.

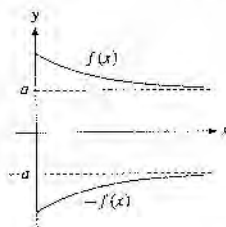


Figure 1 $\sup(-f(x)) = -\inf f(x) = -a$.

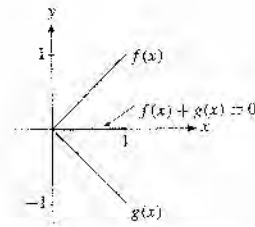


Figure 2 $\inf f(x) = 0$, $\inf g(x) = -1$, $\inf(f(x) + g(x)) = 0$.

We prove only (4): If $\sup f(x)$ or $\sup g(x)$ is ∞ , then the inequality is surely satisfied. Suppose then that $\sup f(x) = p$ and $\sup g(x) = q$, where p and q are both finite numbers. In particular, $f(x) \leq p$ and $g(x) \leq q$ for all x in B , so $f(x) + g(x) \leq p + q$ for all x in B . But then $\sup(f(x) + g(x)) \leq p + q = \sup f(x) + \sup g(x)$, which proves (4).

If $f(x, y)$ is a function defined on a Cartesian product $A \times B$, then

$$\sup_{(x,y) \in A \times B} f(x, y) = \sup_{x \in A} (\sup_{y \in B} f(x, y)) \quad (7)$$

This equality expresses a very important fact: One way to find the supremum of $f(x, y)$ over the set $A \times B$ is as follows: First find the supremum $\sup_y f(x, y)$ of $f(x, y)$ for each given x as y varies over B . This supremum is a function of x , and we take the supremum of this function as x runs through A .

The equality in (7) is also valid if $\sup$ is replaced by $\max$, provided that the relevant suprema are attained, so that the maximum values exist.

Proof of (7): Let $p = \sup_{(x,y) \in A \times B} f(x, y)$ and $q = \sup_{x \in A} (\sup_{y \in B} f(x, y))$. Then $f(x, y) \leq p$ for all x in A and y in B , so $\sup_{y \in B} f(x, y) \leq p$ for all x in A . It follows that $q = \sup_{x \in A} (\sup_{y \in B} f(x, y)) \leq p$. Similarly, note that

$$f(x, y) \leq \sup_{y \in B} f(x, y) \leq \sup_{x \in A} (\sup_{y \in B} f(x, y)) = q$$

for all x in A and all y in B , so $p = \sup_{(x,y) \in A \times B} f(x, y) \leq q$. ■

For infima, the obvious analogue of (7) is

$$\inf_{(x,y) \in A \times B} f(x, y) = \inf_{x \in A} \left\{ \inf_{y \in B} f(x, y) \right\} \quad (8)$$

The result in (7) can be generalized. Let $f(x, y)$ be a real-valued function defined for all x in A , y in B (not necessarily subsets of Euclidean spaces), where A and B are given sets, and let C be a subset of $A \times B$. Then

$$\sup_{(x,y) \in C} f(x, y) = \sup_{x \in C_0} \left(\sup_{y \in C_x} f(x, y) \right) \quad (9)$$

where $C_0 = \{x \in A : (x, y) \in C \text{ for at least one } y \text{ in } B\}$ and $C_x = \{y \in B : (x, y) \in C\}$. The proof is an easy modification of the proof of (7). The result is important for the theory of dynamic programming, discussed in Chapter 12.

On $\liminf$ and $\limsup$ of Functions

This section concludes by defining some limit concepts that are needed in connection with infinite horizon dynamic economic models. They also help to understand the definitions of upper and lower hemicontinuous correspondences in Section 14.1.

Recall first the standard definition of a limit of a function of several variables: Let f be a function defined on a set M in $\mathbb{R}^n$, and suppose that $x^0 \in \text{cl}(M)$, the closure of M . We say that $f(x)$

converges to the number A as x tends to x^0 , if for each $\varepsilon > 0$, there exists a number $\delta > 0$ such that $\|f(x) - A\| < \varepsilon$ for all x in M with $x \neq x^0$ and $\|x - x^0\| < \delta$.

Next we define the **upper** and **lower limits** of f at x^0 as

$$\liminf_{x \rightarrow x^0} f(x) = \lim_{r \rightarrow 0} \left(\inf \{ f(x) : x \in B(x^0; r) \cap M, x \neq x^0 \} \right) \quad (10)$$

$$\limsup_{x \rightarrow x^0} f(x) = \lim_{r \rightarrow 0} \left(\sup \{ f(x) : x \in B(x^0; r) \cap M, x \neq x^0 \} \right) \quad (11)$$

Just as for sequences, we often write $\underline{\lim}$ and $\overline{\lim}$ for $\liminf$ and $\limsup$. With these definitions we obtain the following rules, which are based on the corresponding rules for $\inf$ and $\sup$:

$$\underline{\lim}_{x \rightarrow x^0} (f(x) + g(x)) \geq \underline{\lim}_{x \rightarrow x^0} f(x) + \underline{\lim}_{x \rightarrow x^0} g(x) \quad (\text{if the right-hand side is defined}) \quad (12)$$

$$\overline{\lim}_{x \rightarrow x^0} (f(x) + g(x)) \leq \overline{\lim}_{x \rightarrow x^0} f(x) + \overline{\lim}_{x \rightarrow x^0} g(x) \quad (\text{if the right-hand side is defined}) \quad (13)$$

$$\underline{\lim}_{x \rightarrow x^0} f(x) \leq \overline{\lim}_{x \rightarrow x^0} f(x), \quad \underline{\lim}_{x \rightarrow x^0} f(x) = -\overline{\lim}_{x \rightarrow x^0} (-f(x)), \quad \overline{\lim}_{x \rightarrow x^0} f(x) = -\underline{\lim}_{x \rightarrow x^0} (-f(x)) \quad (14)$$

Note that if $\lim_{x \rightarrow x^0} f(x)$ exists at a point x^0 , then $\underline{\lim}_{x \rightarrow x^0} f(x) = \overline{\lim}_{x \rightarrow x^0} f(x) = \lim_{x \rightarrow x^0} f(x)$. Conversely, if $\underline{\lim}_{x \rightarrow x^0} f(x) = \overline{\lim}_{x \rightarrow x^0} f(x)$, then $\lim_{x \rightarrow x^0} f(x)$ exists and is equal to both.

A function f is called **upper semicontinuous** at a point x^0 in M if $\underline{\lim}_{x \rightarrow x^0} f(x) \leq f(x^0)$. The function f is called **lower semicontinuous** at x^0 if $\overline{\lim}_{x \rightarrow x^0} f(x) \geq f(x^0)$. This definition allows a generalization of the extreme value theorem which says that, if $K \subseteq \mathbb{R}^n$ is a nonempty compact set, and if the function f is upper semicontinuous, the f has a maximum point in the set K . If f is lower semicontinuous, the f has a minimum point in K .

B

APPENDIX

TRIGONOMETRIC FUNCTIONS

God created the integers, all else is the work of man.
—L. Kronecker

Many phenomena appear to repeat themselves with predictable regularity. Examples are alternating electric currents in physics, heartbeat and respiration in physiology, and seasonal variations in economics such as increased demand for heating fuel and warm clothing in winter, as opposed to air-conditioning and cool clothing in summer. Many economists have also looked for regular periodic patterns in macroeconomic variables like national output or interest rates. To describe such phenomena mathematically, one possibility is to use *trigonometric functions*, which are briefly reviewed in Sections B.1 and B.2. The final Section B.3 gives a brief introduction to complex numbers.

B.1 Basic Definitions and Results

Consider the circle in Fig. 1 with radius 1 and centre at the origin in the uv -plane.

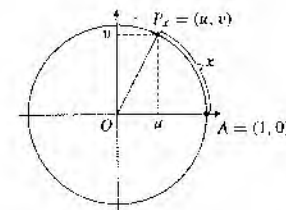


Figure 1 $\sin x = v$ and $\cos x = u$

Let A be the point on the circle with coordinates $(1, 0)$, and let P_x be the point on the circle for which the arc length between A and P_x is x . The point A is P_0 , of course.

The arc x is measured with the same unit of length as the radius. Because the radius of the circle is $r = 1$, the circumference equals $2\pi r = 2\pi$. If $x = \pi/2$, we go one-quarter

of the way round the circle in an anticlockwise direction to arrive at the point $P_{\pi/2}$, which has coordinates $(u, v) = (0, 1)$. For P_x , we go halfway round to the point with coordinates $(u, v) = (-1, 0)$; for $P_{3\pi/2}$, we get $(u, v) = (0, -1)$; for $P_0 = P_{2\pi}$, we have $(u, v) = (1, 0)$; and so on. For the point P_x shown in Fig. 1, where $x \approx 1.1$, we have $u \approx 0.45$ and $v \approx 0.9$. If x is negative, we go a distance $-x$ clockwise around the circle.

In general, as x increases, P_x moves round the unit circle, and the values of u and v oscillate up and down. They repeat themselves as P_x passes through points where it has been before. In particular, $x, x \pm 2\pi, x \pm 4\pi$, and so on, all define the same point on the circle. Thus, $P_x = P_{x+2n\pi}$ for $n = \pm 1, \pm 2, \dots$. This procedure maps each real number x to a point P_x with coordinates (u, v) .

The **sine** function is the rule that maps x to the number v .

The **cosine** function is the rule that maps x to the number u .

It is standard to abbreviate *sine* to *sin* and *cosine* to *cos*. So, referring to Fig. 1, we have

$$\sin x = v \quad \text{and} \quad \cos x = u \quad (1)$$

The circle in Fig. 1 has the equation $u^2 + v^2 = 1$. This implies the important relationship:

$$(\sin x)^2 + (\cos x)^2 = 1 \quad (2)$$

The domains of the functions $\sin$ and $\cos$ are the set of all real numbers. The range of each is the closed interval $[-1, 1]$. Note also that a small change in x will change the point P_x only slightly, so the coordinates u and v will also change only slightly, and $v = \sin x$ and $u = \cos x$ are both continuous functions of x . (In fact, from Fig. 1, we see that a given change in x causes changes in u and v that are smaller in absolute value.)

If x is any number such that $\cos x \neq 0$, we define the **tangent** function by simply dividing $\sin x$ by $\cos x$. It is standard to abbreviate *tangent* to *tan*, so that

$$\tan x = \frac{\sin x}{\cos x} \quad (\text{provided that } \cos x \neq 0) \quad (3)$$

The **cotangent** function, abbreviated *cot*, is defined by $\cot x = \cos x / \sin x$, for all x with $\sin x \neq 0$. It is clear that $\cot x = 1 / \tan x$ when $\tan x$ and $\cot x$ are both defined.

Note that it is common practice to write $\sin^2 x$ for $(\sin x)^2$, $\cos^2 x$ for $(\cos x)^2$, and $\tan^2 x$ for $(\tan x)^2$. Similar notation is also used for higher powers of the trigonometric functions. For example, $\cos^3 x = (\cos x)^3$.

Measuring Angles in Radians

In trigonometry, it is common to define the sine, cosine, and tangent as functions of the *angle*, which is often measured in degrees. Figure 1 shows how the arc length x can be used instead to measure the angle $\angle AOP_x$. Then it is said that the angle is measured in **radians**. In elementary geometry it is common practice to operate with degrees, so one must know

how to convert degrees into radians and vice versa. In fact, $360^\circ = 2\pi$ radians because when $x = 2\pi$, the line OP_x has rotated through 360° . So we have the following:

$$1^\circ = \left(\frac{\pi}{180}\right) \text{ radians} \approx 0.017 \text{ radians}, \quad 1 \text{ radian} = \left(\frac{180}{\pi}\right)^\circ \approx 57.3^\circ \quad (4)$$

Figure 2 illustrates some particularly important angles measured in both degrees and radians.

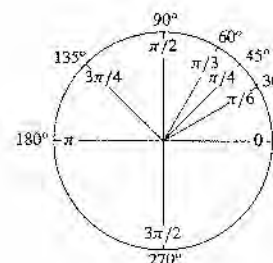


Figure 2

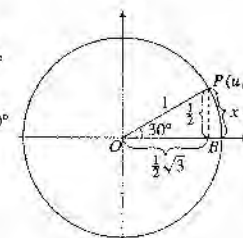


Figure 3

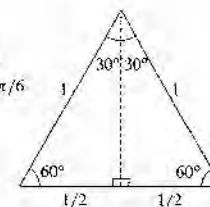


Figure 4

The degree scale for measuring angles is built on an arbitrary choice of unit in that the complete circle is divided into 360° . This corresponds to the ancient Babylonian calendar that divided the year into 360 days. From a mathematical point of view, the radian scale is the most natural one for measuring angles. The reason is that calculus formulas are simpler when angles are measured in radians rather than in degrees.

There is no method for finding exact numerical values of the trigonometric functions in the general case. Programs to calculate numerical approximations are available on most calculators.

For certain special values of x , however, we can compute $\sin x$ and $\cos x$ exactly by using elementary geometry. Consider Fig. 3. Here we have chosen $x = \pi/6$. Then angle $\angle BOP$ is 30° , and triangle BOP is half an equilateral triangle, as Fig. 4 shows more clearly. So in Fig. 3, the length of BP is $\frac{1}{2}$. By Pythagoras' theorem, $(OB)^2 = (OP)^2 - (BP)^2 = 1 - \frac{1}{4} = \frac{3}{4}$, and so $OB = \frac{1}{2}\sqrt{3}$. The coordinates of P are therefore $u = \frac{1}{2}\sqrt{3}$ and $v = \frac{1}{2}$. Hence,

$$\sin \frac{\pi}{6} = \frac{1}{2}, \quad \cos \frac{\pi}{6} = \frac{1}{2}\sqrt{3}, \quad \tan \frac{\pi}{6} = \frac{1}{3}\sqrt{3}$$

Similar geometric considerations establish the other entries in Table 1.

x	0	$\frac{\pi}{6} = 30^\circ$	$\frac{\pi}{4} = 45^\circ$	$\frac{\pi}{3} = 60^\circ$	$\frac{\pi}{2} = 90^\circ$	$\frac{3\pi}{4} = 135^\circ$	$\pi = 180^\circ$	$\frac{3\pi}{2} = 270^\circ$	$2\pi = 360^\circ$
$\sin x$	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{3}}{2}$	1	$\frac{1}{\sqrt{2}}$	0	-1	0
$\cos x$	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	0	$-\frac{1}{\sqrt{2}}$	-1	0	1
$\tan x$	0	$\frac{1}{\sqrt{3}}$	1	$\sqrt{3}$	*	-1	0	*	0

* Not defined.

Table 1 Special values of the trigonometric functions

Graphs of the Trigonometric Functions

By definition of the point P_x in Fig. 1, one has $P_{x+2\pi} = P_x$ for all x , and therefore

$$\sin(x + 2\pi) = \sin x, \quad \cos(x + 2\pi) = \cos x \quad (5)$$

We say that the functions $\sin$ and $\cos$ are **periodic with period 2π** . Also (see Problem 5),

$$\tan(x + \pi) = \tan x \quad (6)$$

so the **tangent function is periodic with period π** .

We noted before that the ranges of $\sin$ and $\cos$ are the interval $[-1, 1]$, so

$$-1 \leq \sin x \leq 1, \quad -1 \leq \cos x \leq 1$$

The graphs of $\sin$ and $\cos$ are shown in Fig. 5. The cosine curve can be obtained by translating the sine curve $\pi/2$ units to the left. This follows from the first of the following formulas:

$$\sin(x + \pi/2) = \cos x, \quad \cos(x + \pi/2) = -\sin x \quad (7)$$

(To prove these formulas, use Problem 4 and formula (8).)

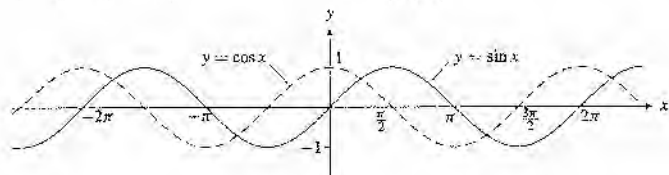


Figure 5

The graph of the tangent function is shown in Fig. 6. Note that its value is positive iff the sine and cosine functions have the same sign. Also, by (3), $\tan x$ is undefined when $x = \frac{1}{2}\pi + n\pi$ for an integer n , because then $\cos x = 0$.

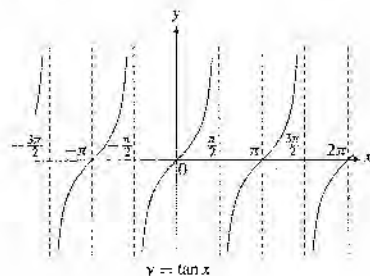


Figure 6

Trigonometric Formulas

There is a plethora of trigonometric formulas that have pestered high school students (and their parents) for generations. Nevertheless, the following formula is particularly useful:

$$\cos(x + y) = \cos x \cos y - \sin x \sin y \quad (8)$$

(For a proof, see Problem 13.) By using this basic equation, similar formulas for $\cos(x - y)$, $\sin(x + y)$, and $\sin(x - y)$ are quite easy to prove (see Problems 3 and 4).

Transformed Trigonometric Functions

We have discussed some important properties of the three basic trigonometric functions: $\sin$, $\cos$, and $\tan$. In economics, they are mainly used in connection with periodic phenomena. Usually transformations of the functions must be used.

So far we have seen that $y = \sin x$ is periodic with period 2π . The graph of the function shows a wavelike curve that is said to have **period** (or **wavelength**) 2π . If, instead, we represent graphically the function given by $y = \sin(x/2)$, we still get a wavelike curve, but the period is now twice as long, namely 4π . The reason is that when x increases from x_0 to $x_0 + 4\pi$, then $x/2$ increases from $x_0/2$ to $x_0/2 + 2\pi$, so $\sin(x/2)$ is periodic with period 4π . More generally, $y = \sin(ax)$ is periodic with period $2\pi/a$, because as x increases by $2\pi/a$, so ax increases by 2π . The value of $y = \sin(ax)$ will still oscillate between -1 and 1 , and we say that the **amplitude** is equal to 1 . To get a periodic function with amplitude A , just put $y = A \sin ax$, which varies between $-A$ and A . Hence,

$$y = A \sin(ax) \quad \text{has period } 2\pi/a \text{ and amplitude } A$$

The reciprocal, $a/2\pi$, of the period is called the **frequency**. It is the number of oscillations per radian.

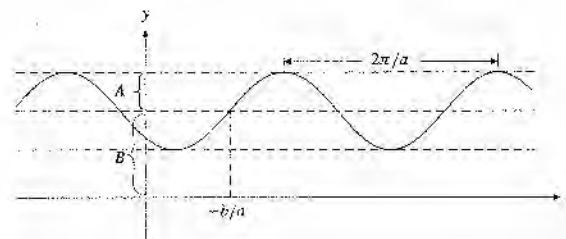


Figure 7

The graph of $y = A \sin(ax)$ intersects the x -axis at $x = 0$. To get a curve translated in the x -direction, let $y = A \sin(ax + b)$. To get a curve that is also translated in the y -direction, let

$$y = A \sin(ax + b) + B \quad (9)$$

The graph of this function is a sine curve with amplitude A and period $2\pi/a$. It is obtained by translating the graph of $y = A \sin(ax)$ a distance $-b/a$ in the x -direction and a distance B in the y -direction. See Fig. 7 (in which $a > 0$ and $b < 0$).

PROBLEMS FOR SECTION B.1

1. Use a diagram like Fig. 3 to verify the values in Table 1 for $x = \pi/4$.
2. Verify that for all x : (a) $\sin(-x) = -\sin x$ (b) $\cos(-x) = \cos x$ (c) $\tan(-x) = -\tan x$
3. Write $\cos(x - y) = \cos[x + (-y)]$, then use the results in (8) and Problem 2 to verify that

$$\cos(x - y) = \cos x \cos y + \sin x \sin y \quad (10)$$

- SM 4. Show that $\cos(y - \pi/2) = \sin y$. From this, it follows that $\sin(y - \pi/2) = \cos(y - \pi) = -\cos y$. Then let $\sin(x + y) = \cos[x + (y - \pi/2)]$ and so prove that

$$\sin(x + y) = \sin x \cos y + \cos x \sin y, \quad \sin(x - y) = \sin x \cos y - \cos x \sin y$$

5. Use the results in Problems 3 and 4 to prove (6) and (7).

- SM 6. Find the following values:

- (a) $\sin(\pi - \pi/6)$ (b) $\cos(\pi + \pi/6)$ (c) $\sin(-3\pi/4)$
 (d) $\cos(5\pi/4)$ (e) $\tan(7\pi/6)$ (f) $\sin(\pi/12)$

- SM 7. Simplify the following expressions:

- (a) $\sqrt{2} \sin\left(x + \frac{1}{4}\pi\right) - \cos x$ (b) $\frac{\sin[\pi - (\alpha + \beta)]}{\cos[2\pi - (\alpha + \beta)]}$ (c) $\frac{\sin(\alpha + x) - \sin(\alpha - x)}{\cos(\alpha + x) - \cos(\alpha - x)}$

- SM 8. Prove that $\sin A - \sin B = 2 \cos \frac{A+B}{2} \sin \frac{A-B}{2}$. (Hint: Put $x + y = A$ and $x - y = B$ in the two formulas in Problem 4, then subtract.)

9. Prove that for all real numbers x and y , $\sin(x + y) \sin(x - y) = \sin^2 x - \sin^2 y$.

10. Draw the graphs of the following functions. Then give their periods and amplitudes.

- (a) $f(x) = \sin(2x)$ (b) $g(x) = 3 \sin(x/2)$ (c) $h(x) = 2 \sin(3x + 4) + 2$

11. Explain why the following functions represent an oscillation that dies out and an oscillation that explodes, respectively. (a) $f(x) = (1/2)^x \sin x$ (b) $g(x) = 2^x \cos 2x$

- SM 12. Find the functions whose graphs are shown in Figs. a to c. In Fig. c the dashed curves have the equations $y = \pm 2e^{-x/\pi}$.

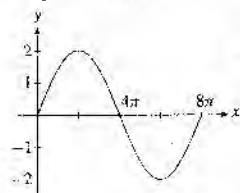


Figure a

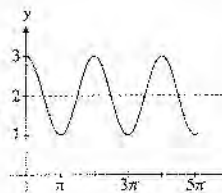


Figure b

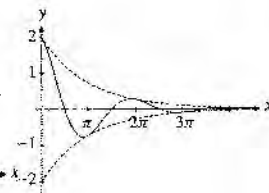
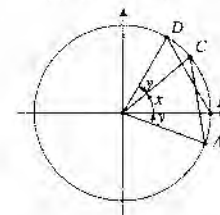


Figure c

- SM 13. In the figure below, the coordinates of the indicated points lying on the unit circle are $A = (\cos y, -\sin y)$, $B = (1, 0)$, $C = (\cos x, \sin x)$, and $D = (\cos(x + y), \sin(x + y))$. Since the line segments AC and BD both subtend the angle $x + y$ at the origin, they must have the same length. Use this fact together with (2) to prove the formula for $\cos(x + y)$ in (8).



B.2 Differentiating Trigonometric Functions

Consider the graph of the sine function in Fig B.1.5. The slope of the graph of $f(x) = \sin x$ at $x = 0$ seems to be 1, as is the value of $\cos x$ at $x = 0$. Also the slope at $x = \pi/2$ is 0, as is $\cos \pi/2$. It is periodic, so its derivative must also be periodic. This helps explain the following, which can be demonstrated along the lines suggested in Problem 12:

$$y = \sin x \Rightarrow y' = \cos x \quad (1)$$

If u is a function of x , the chain rule for differentiation gives

$$y = \sin u, \quad u = u(x) \Rightarrow y' = u' \cos u \quad (2)$$

Let $g(x) = \cos x$. According to (B.1.7), we have $g(x) = \sin(x + \pi/2)$, so (2) yields $g'(x) = \cos(x + \pi/2)$. But $\cos(x + \pi/2) = -\sin x$. Hence,

$$y = \cos x \Rightarrow y' = -\sin x \quad (3)$$

The quotient rule for differentiating $y = \tan x = \sin x / \cos x$ gives (see Problem 2)

$$y = \tan x \Rightarrow y' = \frac{1}{\cos^2 x} = 1 + \tan^2 x \quad (\text{provided that } \cos x \neq 0) \quad (4)$$

Combining these rules of differentiation with those developed earlier allows us to differentiate many expressions involving trigonometric functions.

EXAMPLE 1 Differentiate the following functions:

- (a) $y = \sin 2x$ (b) $y = \sin^2 x + \cos^2 x$ (c) $y = \frac{\sin x}{\cos x + x}$ (d) $y = e^{ax} \sin bx$

Solution:

- (a) Use (2) with $u = 2x$ to obtain $y' = 2 \cos u = 2 \cos 2x$.

(b) $y = (\sin x)^2 + (\cos x)^2 \Rightarrow y' = 2(\sin x) \cos x + 2(\cos x)(-\sin x) = 0$. (Note that $y' = 0$, so that y must be constant. Since $y = 1$ when $x = 0$, this constant must be 1. Hence, we rediscover the relation $\sin^2 x + \cos^2 x = 1$.)

(c) Use the quotient rule for differentiation to obtain

$$y' = \frac{(\cos x + x) \cos x - \sin x(-\sin x + 1)}{(\cos x + x)^2} \\ = \frac{\cos^2 x + x \cos x + \sin^2 x - \sin x}{(\cos x + x)^2} = \frac{1 + x \cos x - \sin x}{(\cos x + x)^2}$$

(d) The product rule yields $y' = ae^{ax} \sin bx + e^{ax} b \cos bx = e^{ax}(a \sin bx + b \cos bx)$. ■

Inverse Trigonometric Functions

Figure 1 illustrates the problem of solving the equation

$$\sin x = y \quad (5)$$

for x . If $y > 1$ or $y < -1$, the equation $\sin x = y$ has no solution, whereas it has infinitely many solutions if $y \in [-1, 1]$.

However, suppose we require that $x \in [-\pi/2, \pi/2]$. In this interval, $\sin x$ is strictly increasing (because $(\sin x)' = \cos x > 0$ in $(-\pi/2, \pi/2)$).

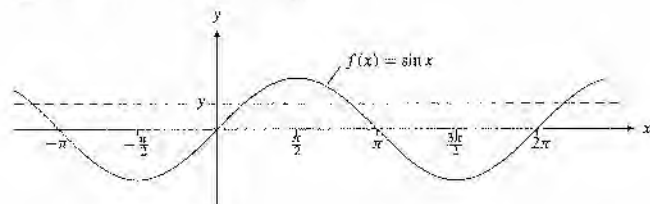


Figure 1

So equation (5) has a unique solution x in this interval for each y in $[-1, 1]$. We denote this solution by $x = \arcsin y$. According to standard terminology we have shown that the function $f(x) = \sin x$, with domain $[-\pi/2, \pi/2]$ and range $[-1, 1]$, has an inverse function g . We call this inverse the **arcsine** function. If we use x as the free variable,

$$g(x) = \arcsin x, \quad x \in [-1, 1] \quad (6)$$

By definition, $\arcsin x$ is that number in $[-\pi/2, \pi/2]$ whose sine is equal to x ($\arcsin x$ is "the angle (arc) whose sine is x "). For instance, we have $\arcsin 1/2 = \pi/6$. The graph of $y = \arcsin x$ is shown in Fig. 2. Since the functions $\sin$ and $\arcsin$ are inverses of each other, the graphs of $y = \sin x$ and $y = \arcsin x$ are symmetric about the line $y = x$.

The derivative of $g(x) = \arcsin x$ is most easily found by implicit differentiation. From the definition of $g(x)$, it follows that $\sin g(x) = x$ for all $x \in (-1, 1)$. If we assume that $g(x)$ is differentiable, differentiating using the chain rule gives $[\cos g(x)] \cdot g'(x) = 1$. So $g'(x) = 1/\cos g(x) = 1/\sqrt{1 - \sin^2 g(x)} = 1/\sqrt{1 - x^2}$. Thus,

$$y = \arcsin x \Rightarrow y' = \frac{1}{\sqrt{1 - x^2}} \quad (-1 < x < 1) \quad (7)$$

It can be shown in the same way that $y = \cos x$ defined on $[0, \pi]$ has an inverse function $y = \arccos x$ defined on $[-1, 1]$, and that

$$y = \arccos x \Rightarrow y' = -\frac{1}{\sqrt{1 - x^2}} \quad (-1 < x < 1) \quad (8)$$

Consider, finally, $y = \tan x$ defined in the interval $(-\pi/2, \pi/2)$. Because $y' = 1/\cos^2 x > 0$, the function is strictly increasing, and the range is $(-\infty, \infty)$. The function, therefore, has an inverse function $y = \arctan x$ that is defined in $(-\infty, \infty)$ and has range $(-\pi/2, \pi/2)$. Using implicit differentiation again, this time in the equation $\tan y = x$ (so that $y = \arctan x$), one obtains

$$y = \arctan x \Rightarrow y' = \frac{1}{1 + x^2} \quad (-\infty < x < \infty) \quad (9)$$

The graph of $y = \arctan x$ is shown in Fig. 3.

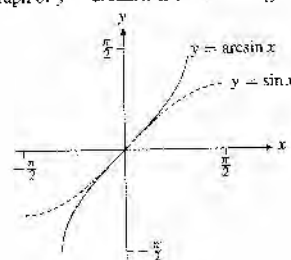


Figure 2

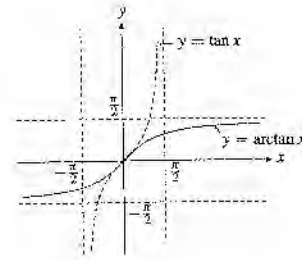


Figure 3



Calculators that have trigonometric functions usually also have their inverses. They are denoted by $\sin^{-1}$, $\cos^{-1}$, and $\tan^{-1}$. If one enters 0.5 and presses the $\sin^{-1}$ key, the answer is 30, because the calculator often uses degrees. If radians are used, the calculator will give the answer $\pi/6$, or rather 0.523598776.

An Important Limit

The derivative of $f(x) = \sin x$ is the limit of the quotient $[\sin(x+h) - \sin x]/h$ as $h \rightarrow 0$. According to (1), at $x = 0$ we have $f'(0) = \cos 0 = 1$, so that $\lim_{h \rightarrow 0} (\sin h/h) = 1$. Changing the variable, we have the following useful limit result:

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1 \quad (10)$$

PROBLEMS FOR SECTION 8.2

1. Find the derivatives of the following functions:

(a) $y = \sin \frac{1}{2}x$ (b) $y = x \cos x$ (c) $y = \tan(x^2)$ (d) $y = e^{2x} \cos x$

2. Prove the differentiation rule in (4). (Hint: Remember that $\sin^2 x + \cos^2 x = 1$.)

3. Find the derivatives of the following functions:

(a) $y = \sin x + \cos x$ (b) $y = x^5 \sin x + \sqrt{x} \cos x + 3$ (c) $y = \frac{\sqrt{x} \cos x}{x^2 + 1}$

SM 4. Compute the following:

(a) $\frac{d}{dx}(1 - \cos 2x)$ (b) $\frac{d}{dt}(at \sin bt)$ (c) $\frac{d}{dt}(\sin[\cos(\sin(at + b))])$

SM 5. Use l'Hôpital's rule (see e.g. EMEA), if necessary, to compute

(a) $\lim_{x \rightarrow 0} \frac{\sin 2x}{x}$ (b) $\lim_{t \rightarrow 0} \frac{\sin mt}{\sin nt}$ ($n \neq 0$) (c) $\lim_{t \rightarrow 0} \frac{1 - \cos t}{t^2}$

SM 6. Find the extreme points of $f(x) = (\sin x - x - 1)^3$ in the interval $I = [0, 3\pi/2]$.

7. Studies of economic cycles often use functions of the form $p(t) = C_0 + C_1 \cos \lambda t + C_2 \sin \lambda t$. Show that $p''(t) + \lambda^2 p(t)$ is a constant K , and find K .

SM 8. Determine the following values:

(a) $\arcsin \frac{1}{2}\sqrt{2}$ (b) $\arccos 0$ (c) $\arccos \frac{1}{2}\sqrt{3}$ (d) $\arctan \sqrt{3}$

SM 9. Find the derivatives of: (a) $\arcsin 2x$ (b) $\arctan(x^2 + 1)$ (c) $\arccos \sqrt{x}$

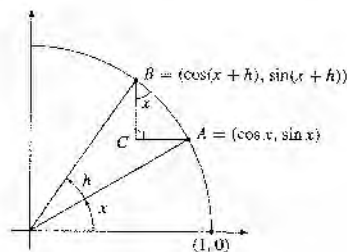
SM 10. Evaluate the following integrals (for the last two integrals, use integration by parts (4.1.1)):

(a) $\int \sin x \, dx$ (b) $\int_0^{\pi/2} \cos x \, dx$ (c) $\int \sin^2 x \, dx$ (d) $\int_0^{\pi} x \cos x \, dx$

SM 11. Evaluate the following integrals by introducing a suitable new variable (see (4.1.2)):

(a) $\int \tan x \, dx = \int \frac{\sin x}{\cos x} \, dx$ (b) $\int \cos x e^{\sin x} \, dx$ (c) $\int \cos^5 x \sin x \, dx$

12. The derivative of $f(x) = \sin x$ is the limit of the quotient $[\sin(x+h) - \sin x]/h$ as $h \rightarrow 0$. From the figure below we see that this quotient is equal to $BC/\text{arc } BA$. If h is small, ACB is almost a right-angled triangle, because the arc BA is almost a straight line. Take the cosine of the angle CBA , which is approximately x . What do you see?



B.3 Complex Numbers

The concept of number can be extended from natural numbers $1, 2, \dots$ via integers $(0, \pm 1, \pm 2, \dots)$, and then rationals (such as 1.414 or $22/7$) to real numbers (including $\sqrt{2}$, e , and π). Each of these extensions expands the set of equations that have solutions. Now, simple quadratic equations like $x^2 + 1 = 0$ and $x^2 + 4x + 8 = 0$ have no solution within the real number system. By introducing complex numbers, however, all quadratic equations become solvable. In fact, within the complex number system, any polynomial equation $x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0 = 0$ has solutions.

The standard formula for solving the equation $x^2 + 4x + 8 = 0$ yields the expressions $-2 + \sqrt{-4}$ and $-2 - \sqrt{-4}$. So far, we have not given any meaning to these expressions. But if we take the liberty of writing $\sqrt{-4} = \sqrt{4}\sqrt{-1} = 2\sqrt{-1}$, we obtain the "solutions"

$$-2 + 2\sqrt{-1} \quad \text{and} \quad -2 - 2\sqrt{-1} \quad (*)$$

Here -2 and 2 are well-known numbers, but $\sqrt{-1}$ is not. By pretending that $\sqrt{-1}$ is a number i whose square is -1 , however, we make i a solution of the equation $i^2 = -1$.

By treating these expressions as if they satisfy the usual algebraic rules, with the additional provision that $\sqrt{-1}\sqrt{-1}$ means -1 , expressions of the type $a + b\sqrt{-1}$ can be used to solve all quadratic equations, even those without real roots.

The symbol $\sqrt{-1}$ can only be given a meaning in an extended system of "numbers", which we call **complex numbers**. Mathematical formalism regards them as 2-vectors (a, b) equipped with the standard addition rule, but with a new multiplication rule.

Informally, instead of writing (a, b) , we usually write this complex number as $a + bi$, where a and b are real numbers and i represents the (so far undefined) symbol $\sqrt{-1}$. Think of i as a symbol that simply identifies which is the second component in the complex number. The real number a is called the **real part**, and the real number b is called the **imaginary part** of the complex number (a, b) . The operations of addition, subtraction and multiplication are defined by

$$(a + bi) + (c + di) = (a + c) + (b + d)i \quad (1)$$

$$(a + bi) - (c + di) = (a - c) + (b - d)i \quad (2)$$

$$(a + bi)(c + di) = (ac - bd) + (ad + bc)i \quad (3)$$

respectively. Formally, rule (1) should be written in the form $(a, b) + (c, d) = (a + c, b + d)$, and rule (3) should be written as $(a, b)(c, d) = (ac - bd, ad + bc)$. What makes the informal expressions $a + bi$ attractive is the fact that (3) is what results if we perform the multiplication $(a + bi)(c + di)$ according to the usual algebraic rules, thus obtaining the expression $ac + (ad + bc)i + bdi^2$, and then finally replace i^2 by -1 . When multiplying complex numbers in practice, we usually perform the computation this way, rather than using rule (3) directly. The complex number $(1, 0)$ is a "unit" in the sense that $(1, 0)(a, b) = (a, b)$.

The way in which we divide two complex numbers can be motivated by the following calculations:

$$\frac{a + bi}{c + di} = \frac{(a + bi)(c - di)}{(c + di)(c - di)} = \frac{(ac + bd) + (bc - ad)i}{c^2 + d^2}$$

The division is defined when $c^2 + d^2 \neq 0$. The formal rule is this:

$$\frac{(a, b)}{(c, d)} = \left(\frac{ac + bd}{c^2 + d^2}, \frac{bc - ad}{c^2 + d^2} \right)$$

In particular, the inverse of $c + di$ is $1/(c + di) = (c - di)/(c^2 + d^2)$. Of course, we have to check that $(c - di)/(c^2 + d^2)$ deserves the name inverse—that is, we have to check that $(c + di)((c - di)/(c^2 + d^2)) = 1 + 0i$, which indeed is the case.

Consider now the problem of giving a meaning to the symbol $i = \sqrt{-1}$. We formally treat this as the problem of finding a complex number (a, b) with the property that $(a, b)(a, b) = (-1, 0)$. It is easy to see that there are two such complex numbers, $(0, 1)$ and $(0, -1)$, and we choose $(0, 1)$ as our i . Then we can formally interpret the symbol $a + bi$ as being $(a, 0)(1, 0) + (b, 0)(0, 1)$, where we simply omit $(1, 0)$ and write a and b instead of $(a, 0)$ and $(b, 0)$.

It is common practice to denote complex numbers by single letters near the end of the alphabet, such as $z = x + yi$ or $w = u + vi$. Two complex numbers, written in this manner, are equal iff both their real and their imaginary parts are equal—that is, $z = w$ iff $x = u$ and $y = v$. If the imaginary part of a complex number is 0, we let $x + 0i = x$. In fact, complex numbers of the form $x + 0i$ behave just like the corresponding real numbers with respect to addition and multiplication. In particular, the numbers $0 (= 0 + 0i)$ and $1 (= 1 + 0i)$ obey the same algebraic rules whether we regard them as complex or as real numbers. Furthermore, $(x, 0)(u, v) = (x + 0i)(u + vi) = (xu, xv) = x(u, v)$, where, for once, we revert to ordinary vector algebra in the last expression, and write the product of a scalar and a vector.

EXAMPLE 1 If $z = 3 + 4i$ and $w = 2 - 5i$, calculate (a) $z + w$ (b) zw (c) z/w

Solution:

$$(a) \ z + w = (3 + 4i) + (2 - 5i) = 5 - i$$

$$(b) \ zw = (3 + 4i)(2 - 5i) = 6 - 15i + 8i - 20i^2 = 26 - 7i$$

$$(c) \ \frac{z}{w} = \frac{3 + 4i}{2 - 5i} = \frac{(3 + 4i)(2 + 5i)}{(2 - 5i)(2 + 5i)} = \frac{6 + 15i + 8i - 20}{4 + 25} = \frac{-14 + 23i}{29}$$

EXAMPLE 2 Prove that if z is a complex number where z^2 is real and nonnegative, then z is real.

Solution: If $z = x + iy$, then $z^2 = (x + iy)^2 = x^2 + 2xyi + (iy)^2 = x^2 - y^2 + 2xyi$. Thus z^2 is real only if $xy = 0$. Then either $x = 0$ or $y = 0$ (or both). Requiring $z^2 \geq 0$ implies that $x^2 \geq y^2$, so that y , at least, must be 0. But then $z = x$, so z is real. ■

Trigonometric Form of Complex Numbers

Each complex number $z = x + yi = (x, y)$ can be represented by a point in the plane. Figure 1 shows how to represent the three particular complex numbers i , $-i$, and $3 + 2i$.

Not surprisingly, the plane representing complex numbers is called the **complex plane**. The horizontal axis, representing numbers of the form $x + 0i$, is called the **real axis**, and the vertical axis, representing numbers of the form $0 + yi$, is called the **imaginary axis**.

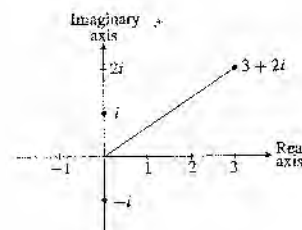


Figure 1 The complex plane.

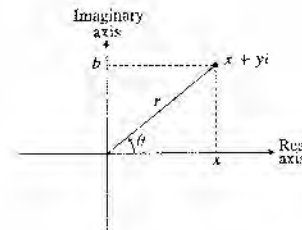


Figure 2 Polar coordinates.

Instead of representing a complex number $z = x + yi$ by the pair (x, y) , we could use **polar coordinates**. As illustrated in Fig. 2, let θ be the angle (measured in radians) between the positive real axis and the vector from the origin to the point (x, y) , and let r be the distance from the origin to the same point. Then $x = r \cos \theta$ and $y = r \sin \theta$, so

$$z = x + yi = r(\cos \theta + i \sin \theta) \quad (4)$$

The last expression is the **trigonometric** (or **polar**) form of the complex number z . The angle θ is called the **argument** of the complex number z . Note that the distance from the origin to the point (x, y) is $r = \sqrt{x^2 + y^2}$. This is called the **modulus** of the complex number, denoted by $|z|$. Hence,

$$|z| = \sqrt{x^2 + y^2} \text{ is the modulus of } z = x + yi \quad (5)$$

If $z = x + iy$, then the **complex conjugate** of z is defined as $\bar{z} = x - iy$. We see that $\bar{z}z = x^2 + y^2 = |z|^2$, where $|z|$ is the modulus of z .

Multiplication and division of complex numbers have neat geometric interpretations if we represent the numbers in trigonometric form. Indeed, applying (3) gives

$$r_1(\cos \theta_1 + i \sin \theta_1)r_2(\cos \theta_2 + i \sin \theta_2) = r_1r_2[\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)] \quad (6)$$

because (B.1.8) and Problem B.1.4 imply that $\cos(\theta_1 + \theta_2) = \cos \theta_1 \cos \theta_2 - \sin \theta_1 \sin \theta_2$ and $\sin(\theta_1 + \theta_2) = \sin \theta_1 \cos \theta_2 + \cos \theta_1 \sin \theta_2$. Thus, the product of two complex numbers is that complex number whose modulus is the product of the moduli of the two factors, and whose argument is the sum of the arguments.

Similarly, we can show that

$$\frac{r_1(\cos \theta_1 + i \sin \theta_1)}{r_2(\cos \theta_2 + i \sin \theta_2)} = \frac{r_1}{r_2}[\cos(\theta_1 - \theta_2) + i \sin(\theta_1 - \theta_2)] \quad (7)$$

If we let $r_1 = r_2 = 1$ and $\theta_1 = \theta_2 = \theta$ in (6), then we obtain $(\cos \theta + i \sin \theta)^2 = \cos 2\theta + i \sin 2\theta$. Similarly, one has

$$\begin{aligned} (\cos \theta + i \sin \theta)^3 &= (\cos \theta + i \sin \theta)^2 (\cos \theta + i \sin \theta) \\ &= (\cos 2\theta + i \sin 2\theta)(\cos \theta + i \sin \theta) = \cos 3\theta + i \sin 3\theta \end{aligned}$$

By induction, we find a famous result, valid for all $n = 1, 2, 3, \dots$:

$$(\cos \theta + i \sin \theta)^n = \cos n\theta + i \sin n\theta \quad (\text{de Moivre's formula}) \quad (8)$$

Vectors with Complex Components

Sometimes we need to consider vectors (and matrices) with complex elements. We add and multiply with scalars (which can be complex numbers) in the obvious way, and the usual algebraic rules still hold. If $\mathbf{z}$ is a vector with complex components $z_1, z_2, \dots, z_n$, we let $\bar{\mathbf{z}}$ denote the vector with components $\bar{z}_1, \bar{z}_2, \dots, \bar{z}_n$, where $\bar{z}_j$ is the complex conjugate of z_j . Then the inner product of $\bar{\mathbf{z}}$ and $\mathbf{z}$ is

$$\bar{\mathbf{z}} \cdot \mathbf{z} = \bar{z}_1 z_1 + \dots + \bar{z}_n z_n = |z_1|^2 + \dots + |z_n|^2 \geq 0 \quad (9)$$

NOTE 1 This has been a very brief introduction to complex numbers. The need to extend the real number system arose in the sixteenth century when various Italian mathematicians derived analytic formulas for the solution to algebraic equations of degree 2, 3, and 4. For a long time, the complex numbers were regarded as “imaginary”, mystical objects. Not any more. Actually, the extension of the number concept from the real numbers to the complex numbers is motivated by the same concern as the extension from the rationals to the reals. In both cases, we want certain equations to have solutions.

Nowadays complex numbers are indispensable in mathematics. Modern science just could not do without them. However, they do not play a very large role in economics. In this book, they allow a convenient description of the solutions to some higher-order difference and differential equations. That also makes them useful in stating results on the stability of solutions to differential equations.

PROBLEMS FOR SECTION 8.3

- If $z = 2 - 5i$ and $w = 3 + 3i$, compute: (a) $z + w$ (b) zw (c) z/w (d) $|z|$
- Represent $z = 2 - 2i$, $w = 1 + 3i$, and $z + w$ as points in the complex plane.
- Write the following numbers in the form $x + yi$:

$$(a) \frac{3+2i}{1-i} \quad (b) \frac{4-3i}{i} \quad (c) \frac{(3-2i)(2-i)}{(-1-i)(3+2i)} \quad (d) \left(\frac{1-i}{1+i} \right)^3$$

- Write the following numbers in trigonometric form:

$$(a) \sqrt{3} + 3i \quad (b) -1 \quad (c) -2 - 2\sqrt{3}i \quad (d) 1 - i$$

ANSWERS

Chapter 1

1.2

- $\begin{pmatrix} 8 \\ 9 \end{pmatrix} = x \begin{pmatrix} 2 \\ 5 \end{pmatrix} + y \begin{pmatrix} -1 \\ 3 \end{pmatrix}$ requires $8 = 2x - y$ and $9 = 5x + 3y$, with solution $x = 3$ and $y = -2$.
- Only the vectors in (b) are linearly independent.
- The determinant of the matrix with the three vectors as columns is equal to 3, so the vectors are linearly independent.
- $x(1, 1, 1) + y(2, 1, 0) + z(3, 1, 4) + w(1, 2, -2) = (0, 0, 0)$ if x, y, z , and w satisfy the equations $x + 2y + 3z + w = 0$, $x + y + z + 2w = 0$, and $x + 4z - 2w = 0$. One solution is $x = -2$, $y = -1$, $z = 1$, $w = 1$, so the vectors are linearly dependent.
- Suppose $\alpha(\mathbf{a} + \mathbf{b}) + \beta(\mathbf{b} + \mathbf{c}) + \gamma(\mathbf{a} + \mathbf{c}) = \mathbf{0}$. Then $(\alpha + \gamma)\mathbf{a} + (\alpha + \beta)\mathbf{b} + (\beta + \gamma)\mathbf{c} = \mathbf{0}$. Because $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are linearly independent, $\alpha + \gamma = 0$, $\alpha + \beta = 0$, and $\beta + \gamma = 0$. It follows that $\alpha = \beta = \gamma = 0$, which means that $\mathbf{a} + \mathbf{b}$, $\mathbf{b} + \mathbf{c}$, and $\mathbf{a} + \mathbf{c}$ are linearly independent. The vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, and $\mathbf{a} + \mathbf{c}$ are linearly dependent because $\mathbf{a} + \mathbf{c} = (\mathbf{a} + \mathbf{b}) + (\mathbf{b} + \mathbf{c})$.
- (a) Suppose $\alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c} = \mathbf{0}$. Taking the inner product of each side with $\mathbf{a}$ yields $\mathbf{a} \cdot (\alpha\mathbf{a} + \beta\mathbf{b} + \gamma\mathbf{c}) = \mathbf{a} \cdot \mathbf{0} = 0$. But $\mathbf{a} \cdot \mathbf{a} = \mathbf{a} \cdot \mathbf{c} = 0$, so $\alpha(\mathbf{a} \cdot \mathbf{a}) = 0$. Because $\mathbf{a} \neq \mathbf{0}$, we conclude that $\alpha = 0$. In a similar way we prove that $\beta = \gamma = 0$, so $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are linearly independent. (b) See SM.
- Because $\mathbf{v}_3 = 0 \cdot \mathbf{v}_1 + 2\mathbf{v}_2$, at least one of the vectors can be written as a linear combination of the others, so the vectors are linearly dependent. But there is no way that $(1, 0)$ can be written as a linear combination of the other two vectors.
- Both statements follow immediately from the definitions.

1.3

- (a) 1. (The determinant of the matrix is 0, so the rank is less than 2. Because not all entries are 0, the rank is 1.) (b) 2 (c) 2 (d) 3 (e) 2 (f) 3
- (a) The determinant is $(x+1)(x-2)$. The rank is 3 if $x \neq -1$ and $x \neq 2$. The rank is 2 if $x = -1$ or $x = 2$. (b) The rank is 3 if $t \neq -4$, $t \neq -2$, and $t \neq 2$. The rank is 2 if $t = -4$, $t = -2$, or $t = 2$. (c) The rank is 2 for all values of x, y, z , and w .
- $\mathbf{A} = \begin{pmatrix} 3 & 1 \\ 6 & 2 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1 & 2 \\ -3 & -6 \end{pmatrix}$. Here $r(\mathbf{AB}) = 0$ and $r(\mathbf{BA}) = 1$.

1.4

- (a) $r(\mathbf{A}) = 1$, $r(\mathbf{A}_h) = 2$. No solutions. (b) $x_1 = 1 + \frac{2}{3}t$, $x_2 = 1 + s - \frac{5}{3}t$, $x_3 = s$, and $x_4 = t$, with s, t arbitrary. Two degrees of freedom. (c) $x_1 = -\frac{4}{3}s$, $x_2 = \frac{5}{3}s$, $x_3 = s$, and $x_4 = 1$, with s arbitrary. One degree of freedom. (d) $r(\mathbf{A}) = 2$, $r(\mathbf{A}_h) = 3$. No solutions.

2. (a) $x_1 = x_2 = x_3 = 0$ is the only solution. There are zero degrees of freedom.
 (b) $x_1 = t$, $x_2 = -t$, $x_3 = -t$, and $x_4 = t$, with t arbitrary. There is one degree of freedom.
3. If $a \neq 0$ and $a \neq 7$, the system has a unique solution. If $a = 0$ and $b = 9/2$, or if $a = 7$ and $b = 10/3$, the system has infinitely many solutions, with one degree of freedom. For other values of the parameters, there are no solutions.
4. $A(\lambda x_1 + (1 - \lambda)x_2) = \lambda Ax_1 + (1 - \lambda)Ax_2 = \lambda b + (1 - \lambda)b = b$. This shows that if $x_1 \neq x_2$ are solutions, then so are all points on the straight line through x_1 and x_2 .
5. Let the n vectors be $a_i = (a_{i1}, \dots, a_{in})$, $i = 1, \dots, n$. If $b = (b_1, \dots, b_n)$ is orthogonal to each of these n vectors, then $a_i \cdot b = a_{i1}b_1 + \dots + a_{in}b_n = 0$, $i = 1, \dots, n$. Because $a_1, \dots, a_n$ are linearly independent, this homogeneous system of equations has only the trivial solution $b_1 = \dots = b_n = 0$, so $b = 0$.
6. (a) $|A_t| = (t - 2)(t + 3)$, so $r(A_t) = 3$ if $t \neq 2$ and $t \neq -3$. Because $\begin{vmatrix} 1 & 3 \\ 2 & 5 \end{vmatrix} \neq 0$, $r(A_2) = 2$, $r(A_{-3}) = 2$.
 (b) $x_1 = -46 + 19s$, $x_2 = 19 - 7s$, $x_3 = s$, $s \in \mathbb{R}$.
7. If b is changed to $b + \Delta b$, the new solution is $x + \Delta x = A^{-1}(b + \Delta b) = A^{-1}b + A^{-1}\Delta b$, and so $\Delta x = A^{-1}\Delta b$. The conclusion follows.

5

1. (a) $-1, -5; \begin{pmatrix} 7 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ (b) $4 \pm 2i$ (c) $5, -5; \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ 3 \end{pmatrix}$ (d) $2, 3, 4; \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$
 (e) $-1, 0, 2; \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$ (f) $0, 1, 3; \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$
2. (a) $X'AX = (ax^2 + ay^2 + bz^2 + 2axy)$, $A^2 = \begin{pmatrix} 2a^2 & 2a^2 & 0 \\ 2a^2 & 2a^2 & 0 \\ 0 & 0 & b^2 \end{pmatrix}$, $A^3 = \begin{pmatrix} 4a^3 & 4a^3 & 0 \\ 4a^3 & 4a^3 & 0 \\ 0 & 0 & b^3 \end{pmatrix}$
 (b) $\lambda = 0$, $\lambda = 2a$, $\lambda = b$ (c) $p(\lambda) = -\lambda^3 + (2a + b)\lambda^2 - 2ab\lambda$ and $-A^3 + (2a + b)A^2 - 2abA = 0$.
3. $Av_1 = 3v_1$ yields $a - c = 3$, $b - e = 0$, and $c - f = -3$. $Av_2 = v_2$ yields $a + 2b + c = 1$, $b + 2d + e = 2$, and $c + 2e + f = 1$. Finally, $Av_3 = 4v_3$ yields $a + b + c = 4$, $b - d + e = -4$, and $c - e + f = 4$. Solving these equations yields $A = \begin{pmatrix} 3 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 3 \end{pmatrix}$.
4. (a) $\lambda = 3$ and $\lambda = 7$. $(A - \lambda I) = (3 - \lambda)^2(7 - \lambda)$. (b) $\begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ -1 \\ 0 \end{pmatrix}$, and $\begin{pmatrix} 1 \\ 0 \\ 0 \\ -1 \end{pmatrix}$ for $\lambda = 3$.
5. (a) $Ax_1 = \begin{pmatrix} 2 \\ 0 \\ 2 \end{pmatrix} = 2x_1$, $Ax_2 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} = -x_2$, $Ax_3 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = x_3$, so all three vectors are eigenvectors for A , with eigenvalues $\lambda_1 = 2$, $\lambda_2 = -1$, and $\lambda_3 = 1$, respectively. (b) If x is an eigenvector for A with eigenvalue λ , then $Bx = A(Ax) = A(\lambda x) = \lambda Ax = \lambda^2 x$, so x is an eigenvector for B with eigenvalue λ^2 . Because $\lambda_1^2 = \lambda_2^2 = \lambda_3^2 = 1$, $Bx_1 = x_1$ and $Bx_3 = x_3$, but $Bx_2 = \lambda_2^2 x_2 = 4x_2 \neq x_2$. (c) See SM.
6. By (1.1.23) and (1.1.14), $|A - \lambda I| = 0 \iff |(\tilde{A} - \lambda \tilde{I})'| = 0 \iff |A' - \lambda I| = 0$. The conclusion follows.
7. Clearly, $\lambda = 0$ is an eigenvalue iff $|A| = 0$. If $\lambda \neq 0$ is an eigenvalue of A , then A has an inverse and $Ax = \lambda x$ implies $x = \lambda A^{-1}x$, or $A^{-1}x = (1/\lambda)x$, which shows that $1/\lambda$ is an eigenvalue of A^{-1} .

8. $|A - I| = \begin{vmatrix} a_{11} - 1 & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - 1 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - 1 \end{vmatrix}$. Add all the last $n - 1$ rows to the first row. Because all the column sums in A are 1, all entries in the first row will be 0. Hence, $|A - I| = 0$, so 1 is an eigenvalue for A .

1.6

1. (a) Eigenvalues are 1 and 3, with corresponding eigenvectors $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Normalizing the eigenvectors, we choose $P = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix}$, and then $P^{-1}AP = \text{diag}(1, 3)$.
 (b) $P = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}$ (c) $P = \begin{pmatrix} 0 & \sqrt{2}/2 & -\sqrt{2}/2 \\ -4/5 & 3\sqrt{2}/10 & 3\sqrt{2}/10 \\ 3/5 & \sqrt{2}/5 & 2\sqrt{2}/5 \end{pmatrix}$
2. (a) The characteristic equation: $(1 - \lambda)(\lambda^2 + \lambda - 3(1 + k)) = 0$. All roots are real $\iff k \geq -13/12$.
 If $k = 3$, the eigenvalues are $-4, 1$, and 3 . (b) $P'A_3P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 3 \end{pmatrix}$, as promised by Theorem 1.6.2.
3. (a) $A^2 = (PDP^{-1})(PDP^{-1}) = PD(P^{-1}P)D^{-1} = PDIDP^{-1} = PD^2P^{-1}$.
 (b) The formula is valid for $m = 1$. Suppose it is valid for $m = k$. Then $A^{k+1} = AA^k = A(PD^kP^{-1}) = PD(P^{-1}P)D^kP^{-1} = PDID^kP^{-1} = PDD^kP^{-1} = PD^{k+1}P^{-1}$, so it holds for $m = k + 1$ as well. By induction, the formula holds for all positive integers m .
4. According to (1), AB and $A^{-1}(AB)A = BA$ have the same eigenvalues.
5. $A^2 = 5A - 5I$, $A^4 = 75A - 100I = \begin{pmatrix} 50 & 75 \\ 75 & 125 \end{pmatrix}$

1.7

1. (a) $a_{11} = -1$, $a_{12} = 1$ (not 2!), and $a_{22} = -6$. Thus $a_{11} < 0$ and $a_{11}a_{22} - a_{12}^2 = 6 - 1 = 5 > 0$, so according to (5)(b), $Q(x_1, x_2)$ is negative definite. (b) $a_{11} = 4$, $a_{12} = 1$, and $a_{22} = 25$. Thus $a_{11} > 0$ and $a_{11}a_{22} - a_{12}^2 = 100 - 1 = 99 > 0$, so according to (5)(a), $Q(x_1, x_2)$ is positive definite.
2. $a_{11}x_1^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + a_{22}x_2^2 + 2a_{23}x_2x_3 + a_{33}x_3^2$
3. (a) $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ (b) $\begin{pmatrix} a & \frac{1}{2}b \\ \frac{1}{2}b & c \end{pmatrix}$ (c) $\begin{pmatrix} 3 & -1 & 3/2 \\ -1 & 1 & 0 \\ 3/2 & 0 & 3 \end{pmatrix}$ 4. $A = \begin{pmatrix} 3 & -1 & 2 & 4 \\ -1 & 1 & 3/2 & 0 \\ 2 & 3/2 & 1 & -1 \\ 4 & 0 & -1 & 1 \end{pmatrix}$
5. (a) Positive definite (b) Positive definite (c) Negative semidefinite (d) Negative definite.
6. Since A is symmetric, by Theorem 1.6.2 all the eigenvalues are real. By Theorem 1.7.2(b), they are all nonnegative iff A is positive semidefinite. Since $|A|$ is 0 iff 0 is an eigenvalue, the conclusion follows from Theorem 1.7.2(a).
7. (a) The determinant of the associated symmetric matrix A is $|A| = 6c - \frac{1}{2}(5 + c)^2 = -\frac{1}{2}(c^2 - 14c + 25) = -\frac{1}{2}(c - c_1)(c - c_2)$, where $c_1 = 7 - 2\sqrt{6} \approx 2.1$ and $c_2 = 7 + 2\sqrt{6} \approx 11.9$. Q is positive definite if $c_1 < c < c_2$, positive semidefinite if $c_1 \leq c \leq c_2$, and indefinite if $c < c_1$ or $c > c_2$.
 (b) Let X be an $n \times n$ matrix. Then $X'AX = X'(B'B)X = (BX)'(BX) = \|BX\|^2 \geq 0$, so A is positive semidefinite. A is positive definite iff $BX \neq 0$ for $X \neq 0$, and this is the case iff $|B| \neq 0$. (See (1.1.34).)

8. (a) $Q(0, \dots, x_i, \dots, 0) = a_{ii}x_i^2$ is positive for $x_i \neq 0$, so $a_{ii} > 0$. (b) Let $R(x_i, x_j) = Q(0, \dots, x_i, \dots, x_j, \dots, 0)$. Then $R(x_i, x_j)$ is a quadratic form in the two variables x_i and x_j with associated symmetric matrix $\mathbf{B} = \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix}$. Since Q is positive definite, so is R , and then $|\mathbf{B}| > 0$, by Theorem 1.7.1(a).
9. By Theorem 1.6.2 all eigenvalues are real. If $\mathbf{A}$ is negative definite, then by Theorem 1.7.2(c) all the eigenvalues $\lambda_1, \dots, \lambda_n$ are negative. But then $\psi(\lambda) = (-1)^n \varphi(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n) = (\lambda + r_1)(\lambda + r_2) \cdots (\lambda + r_n)$, where r_i are all positive. Expanding this product obviously produces a polynomial with positive coefficients only. If, on the other hand, all the coefficients a_i in $\psi(\lambda)$ are positive, then $\psi(\lambda) \geq a_0 > 0$ for all $\lambda \geq 0$. So no positive number can be an eigenvalue.
10. (a) With $\mathcal{L} = 2x_1^2 + 14x_1x_2 + 2x_2^2 - \lambda(x_1^2 + x_2^2 - 1)$, the first-order conditions are $\mathcal{L}'_1 = 4x_1 + 14x_2 - 2\lambda x_1 = 0$, $\mathcal{L}'_2 = 14x_1 + 4x_2 - 2\lambda x_2 = 0$. Dividing each equation by 2, we see that these two equations can be written as $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. So a vector satisfying the first-order conditions is an eigenvector for $\mathbf{A}$, and the Lagrange multiplier is an eigenvalue. (b) The two eigenvalues are 9 and -5 . See SM.

1.8

1. Positive definite subject to the constraint. When $x_2 = -x_1$, $x_1^2 - 2x_1x_2 + x_2^2 = 4x_1^2$, which is positive except when $x_1 = 0$. Using Theorem 1.8.1, $\begin{vmatrix} 0 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \end{vmatrix} = -4 < 0$.
2. (a) Positive definite since $\begin{vmatrix} 0 & 3 & 4 \\ 3 & 2 & -2 \\ 4 & -2 & 1 \end{vmatrix} = -89 < 0$. (b) Negative definite since $\begin{vmatrix} 0 & 5 & -2 \\ 5 & -1 & \frac{1}{2} \\ -2 & \frac{1}{2} & -1 \end{vmatrix} = 19 > 0$.
3. Negative definite subject to the constraints. 4. Positive definite subject to the constraints.
5. The condition is: $\begin{vmatrix} 0 & g'_1(x_1^*, x_2^*) & g'_2(x_1^*, x_2^*) \\ g'_1(x_1^*, x_2^*) & \mathcal{L}''_{11}(x_1^*, x_2^*) & \mathcal{L}''_{12}(x_1^*, x_2^*) \\ g'_2(x_1^*, x_2^*) & \mathcal{L}''_{12}(x_1^*, x_2^*) & \mathcal{L}''_{22}(x_1^*, x_2^*) \end{vmatrix} < 0$.

9

1. $\begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} & a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} \\ \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} \\ \mathbf{A}_{31}\mathbf{B}_{11} + \mathbf{A}_{32}\mathbf{B}_{21} \end{pmatrix}$
2. $\begin{pmatrix} 1 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ -2 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ -3 & 0 \end{pmatrix}$
3. (a) $\begin{pmatrix} -2/11 & 2/11 & 0 & 0 \\ 5/11 & -2/11 & 0 & 0 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 3 & -4 \end{pmatrix}$ (b) $\begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$ (c) $\begin{pmatrix} 1/2 & -1/2 & -1/2 & 0 & 1/2 \\ -1/2 & 1/2 & -1/2 & 0 & 1/2 \\ -1/2 & -1/2 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \\ 1/2 & 1/2 & 1/2 & 0 & -1/2 \end{pmatrix}$
4. Apply first (6) and then (7) to the matrix whose determinant is shown in the middle of (*).
5. Show that $\begin{pmatrix} \mathbf{P} & \mathbf{R} \\ \mathbf{0} & \mathbf{Q} \end{pmatrix} \begin{pmatrix} \mathbf{P}^{-1} & -\mathbf{P}^{-1}\mathbf{R}\mathbf{Q}^{-1} \\ \mathbf{0} & \mathbf{Q}^{-1} \end{pmatrix} = \mathbf{I}$.
6. (a) Use the hint. How to get a term different from 0? (b) Show the equality by direct multiplication. Then use (a).
7. (a) Use the hint, the formula for the product of partitioned matrices, and (6) and (7).
(b) The determinant is $(a_1 - 1)(a_2 - 1) \cdots (a_n - 1)|\mathbf{F}|$. By using the results in (a), we get $|\mathbf{F}| = |\mathbf{I} + \mathbf{BA}| = 1 + \sum_{i=1}^n 1/(a_i - 1)$, and the conclusion follows.

Chapter 2

2.1

1. (a) $\nabla f = (y, 2y + x) = (1, 4)$ at $(2, 1)$ (b) $\nabla g = (e^{xy} + xy e^{xy}, x^2 e^{xy}, -2z) = (1, 0, -2)$ at $(0, 0, 1)$
(c) $\nabla h = (e^t, 2e^{2t}, 3e^{3t}) = (1, 2, 3)$ at $(0, 0, 0)$ (d) $\nabla k = (e^{x+2y+3z}, 2e^{x+2y+3z}, 3e^{x+2y+3z}) = (1, 2, 3)$ at $(0, 0, 0)$
2. (a) $\nabla F(x, y) = (2xf'(x^2 + y^2), 2yf'(x^2 + y^2)) = 2f'(x^2 + y^2)(x, y)$, so $\nabla F(x, y)$ is parallel to (x, y) .
(b) $\nabla G(x, y) = x^{-2}f'(y/x)(-y, x)$ and $\nabla G(x, y) \cdot (x, y) = 0$, so it follows that $\nabla G(x, y) \perp (x, y)$.
3. (a) $3\sqrt{2}/2$ (b) $-2\sqrt{3}/3$
4. Using (6), $f'_a(x) = \lim_{h \rightarrow 0} [f(x + ha) - f(x)]/h = \lim_{h \rightarrow 0} [(x_1 + ha_1)^2 + \cdots + (x_n + ha_n)^2 - x_1^2 - \cdots - x_n^2]/h = \lim_{h \rightarrow 0} [2x_1a_1 + \cdots + 2x_na_n + h(a_1^2 + \cdots + a_n^2)] = 2a_1x_1 + \cdots + 2a_nx_n$. On the other hand, $\nabla f(x) = (2x_1, \dots, 2x_n) = 2\mathbf{x}$, so according to (8), $f'_a(x) = 2\mathbf{x} \cdot \mathbf{a} = 2a_1x_1 + \cdots + 2a_nx_n$.
5. (a) $-(5 \ln 3 + 8/3)/\sqrt{18}$ (b) $(\ln 3 + 2/3, \ln 3 + 2/3, 2/3)$ 6. $\nabla f(0, 0) = (2\sqrt{10}/5, 6\sqrt{10}/5)$
7. $f(x) = b_1x_1 + \cdots + b_nx_n$, so $\nabla f(x) = (b_1, \dots, b_n) = \mathbf{b}$, and $f'_a(x) = \nabla f(x) \cdot \mathbf{a} = \mathbf{b} \cdot \mathbf{a}$.
8. $\text{El}_a f(\mathbf{v}) = \frac{\|\mathbf{v}\|}{f(\mathbf{v})} \nabla f(\mathbf{v}) \cdot \mathbf{a} = \frac{1}{f(\mathbf{v})} \nabla f(\mathbf{v}) \cdot \mathbf{v} = \sum_{i=1}^n \frac{v_i}{f(\mathbf{v})} \frac{\partial f(\mathbf{v})}{\partial v_i} = \sum_{i=1}^n \text{El}_i f(\mathbf{v})$
9. (a) See SM. (b) From $y = 8x^{-2}$ we get $y' = -16x^{-3}$ and $y'' = 48x^{-4} = 3$ when $x = 2$. Using the formula in (a), at $(2, 2)$: $F'_1 = 2xy = 8$, $F'_2 = x^2 = 4$, $F''_{11} = 2y = 4$, $F''_{12} = 2x = 4$, and $F''_{22} = 0$, so $y'' = 4^{-3} \begin{vmatrix} 0 & 8 & 4 \\ 8 & 4 & 4 \\ 4 & 4 & 0 \end{vmatrix} = 3$.

2.2

1. Only (a) and (d) are convex.
2. See Figures A2.2.2(a)–(f).

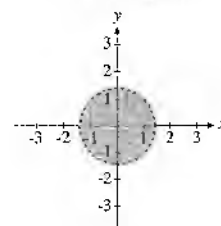


Figure A2.2.2(a) Convex.

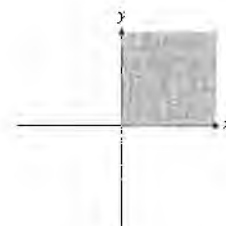


Figure A2.2.2(b) Convex.

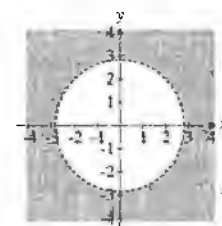


Figure A2.2.2(c) Not convex.

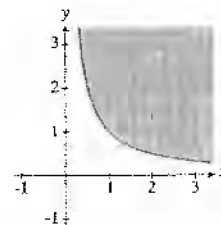


Figure A2.2.2(d) Convex.

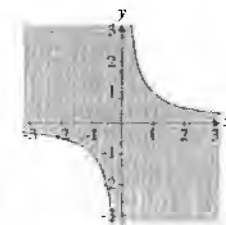


Figure A2.2.2(e) Not convex.

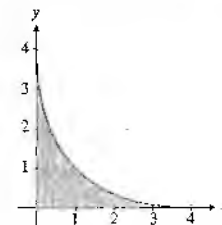


Figure A2.2.2(f) Not convex.

1. S is the intersection of $m + n$ convex sets (half spaces determined by the inequalities), and is therefore convex according to (3).

2. To prove that $aS + bT$ is convex, take $\mathbf{x} \in aS + bT$, $\mathbf{y} \in aS + bT$, and a number $\lambda \in [0, 1]$. Because $\mathbf{x} \in aS + bT$, there are points $\mathbf{x}_1 \in S$ and $\mathbf{x}_2 \in T$ such that $\mathbf{x} = a\mathbf{x}_1 + b\mathbf{x}_2$. Similarly, there are points $\mathbf{y}_1 \in S$ and $\mathbf{y}_2 \in T$ such that $\mathbf{y} = a\mathbf{y}_1 + b\mathbf{y}_2$. Then $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} = \lambda(a\mathbf{x}_1 + b\mathbf{x}_2) + (1-\lambda)(a\mathbf{y}_1 + b\mathbf{y}_2) = a[\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{y}_1] + b[\lambda\mathbf{x}_2 + (1-\lambda)\mathbf{y}_2]$. This belongs to $aS + bT$ because S and T are convex, so $\lambda\mathbf{x}_1 + (1-\lambda)\mathbf{y}_1 \in S$ and $\lambda\mathbf{x}_2 + (1-\lambda)\mathbf{y}_2 \in T$. Therefore $aS + bT$ is convex.

3. Suppose (s_1, t_1) and (s_2, t_2) both belong to $S \times T$, with $s_1, s_2 \in S$ and $t_1, t_2 \in T$. Now, $\lambda(s_1, t_1) + (1-\lambda)(s_2, t_2) = (\lambda s_1 + (1-\lambda)s_2, \lambda t_1 + (1-\lambda)t_2)$. For $\lambda \in [0, 1]$, this belongs to $S \times T$ because $\lambda s_1 + (1-\lambda)s_2 \in S$ and $\lambda t_1 + (1-\lambda)t_2 \in T$, by the convexity of S and T . Hence, $S \times T$ is convex.

4. (a) If $\|\mathbf{x}\| \leq r$, $\|\mathbf{y}\| \leq r$, and $\lambda \in [0, 1]$, then $\|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\| \leq \lambda\|\mathbf{x}\| + (1-\lambda)\|\mathbf{y}\| \leq \lambda r + (1-\lambda)r = r$.
(b) S_1 is convex. Neither S_2 nor S_3 is convex.

5. (a) The set $S = \mathbb{Q}$ of rational numbers has the property, but is not convex. (b) Yes. See SM.

6. See SM.

3

1. (a) Strictly convex. (b) Concave, but not strictly concave. (c) Strictly concave.

2. (a) (i) $f''_{11} = -2 \leq 0$, $f''_{22} = 0 \leq 0$, and $f''_{11}f''_{22} - (f''_{12})^2 = 0 \geq 0$, so f is concave.
(ii) $f(x) = (x-y) + (-x^2)$ is a sum of concave functions, hence concave.
(b) $F(u) = -e^{-u}$ is (strictly) increasing and concave (because $F'(u) = e^{-u} > 0$ and $F''(u) = -e^{-u} < 0$). By Theorem 2.3.5(a), $z = -e^{-f(x,y)}$ is concave.

3. (a) $f''_{11} = 2a$, $f''_{12} = 2b$, $f''_{22} = 2c$, and $f''_{11}f''_{22} - (f''_{12})^2 = 2a2c - (2b)^2 = 4(ac - b^2)$. The result follows from Theorem 2.3.1. (b) Using Theorem 2.3.1 again, f is concave iff $a \leq 0$, $c \leq 0$, and $ac - b^2 \geq 0$; f is convex iff $a \geq 0$, $c \geq 0$, and $ac - b^2 \geq 0$.

4. $f''_{11} = -12$, $f''_{22} = -2$, and $f''_{11}f''_{22} - (f''_{12})^2 = 24 - (2a + 4)^2 = -4a^2 - 16a + 8$. Because $f''_{11} < 0$, the function is never convex. It is concave iff $-4a^2 - 16a + 8 \geq 0$, that is, iff $-2 - \sqrt{6} \leq a \leq -2 + \sqrt{6}$.

5. (a) z is strictly concave. (b) z is strictly convex. (c) Use (8).

6. (a) By Theorem 2.3.3 (b), all the principal minors of order 1 must be ≤ 0 , i.e. $f''_{ii}(x) \leq 0$ for all i .
(b) See SM. (c) Take any $\mathbf{x}$ in the domain of f . If $f(\mathbf{x}) = s$, then $f(2\mathbf{x}) = 2s$ and $f(\frac{1}{2}\mathbf{x} + \frac{1}{2}2\mathbf{x}) = f(\frac{3}{2}\mathbf{x}) = \frac{3}{2}s$. But this contradicts f being strictly concave.

7. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0, 1]$, then $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = \|\lambda\mathbf{x} + (1-\lambda)\mathbf{y}\| \leq \lambda\|\mathbf{x}\| + \|(1-\lambda)\mathbf{y}\| = \lambda\|\mathbf{x}\| + (1-\lambda)\|\mathbf{y}\| = \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$. Hence f is convex. But when $\mathbf{y} = a\mathbf{x}$, then $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = \|\lambda\mathbf{x} + (1-\lambda)a\mathbf{x}\| = \lambda\|\mathbf{x}\| + (1-\lambda)\|\mathbf{y}\|$, so f is not strictly convex.

8. See SM.

9. (a) First, note that $x''_{ij} = a_i a_j z / x_i x_j$ for $i \neq j$, and $x''_{ii} = a_i(a_i - 1)z / x_i^2$. By using rule (1.1.20) repeatedly, we obtain the formula for D_k . (b) Use the hint. (c) If $\sum_{i=1}^n a_i < 1$, then (because each a_i is positive) $\sum_{i=1}^k a_i < 1$ for all k , so the sign of D_k is that of $(-1)^k$. Then use Theorem 2.3.2(b).

4

Inequality (1) reduces to $1 - x^2 - y^2 - (1 - x_0^2 - y_0^2) \leq -2x_0(x - x_0) - 2y_0(y - y_0)$. Rearranging the terms yields the equivalent inequality $0 \leq x^2 - 2xx_0 + x_0^2 + y^2 - 2yy_0 + y_0^2$, or $0 \leq (x - x_0)^2 + (y - y_0)^2$, which is obviously true.

2. Jensen's inequality yields: $\ln(\frac{1}{n}(x_1 + x_2 + \dots + x_n)) \geq \frac{1}{n}(\ln x_1 + \ln x_2 + \dots + \ln x_n) = \frac{1}{n} \ln(x_1 x_2 \dots x_n) = \ln(x_1 x_2 \dots x_n)^{1/n}$. Since $\ln$ is strictly increasing, the required inequality follows.

3. On the right-hand side of (*), the coefficients $\lambda_1 + \lambda_2$ and λ_3 sum to 1, so definition (2.3.1) for $n = 2$ applies. Next note that $\lambda_1/(\lambda_1 + \lambda_2)$ and $\lambda_2/(\lambda_1 + \lambda_2)$ also sum to 1.

4. Because $\lambda(t) \geq 0$, we have $\lambda(t)f(x(t)) - \lambda(t)f(z) \leq f'(z)\lambda(t)x(t) - zf'(z)\lambda(t)$. Integrate each side w.r.t. t to get $\int_a^b \lambda(t)f(x(t)) dt - f(z) \int_a^b \lambda(t) dt \leq f'(z) \int_a^b \lambda(t)x(t) dt - zf'(z) \int_a^b \lambda(t) dt$. But $\int_a^b \lambda(t) dt = 1$ and also $z = \int_a^b \lambda(t)x(t) dt$, so $\int_a^b \lambda(t)f(x(t)) dt - f(\int_a^b \lambda(t)x(t) dt) \leq 0$.

5. See SM. 6. See SM.

2.5

1. From (6): (a) F is strictly concave; (b) F is quasiconcave; (c) F is concave; From (7): (d) F is quasiconvex ($\rho = -2$, $\mu = 1$); (e) F is concave ($\rho = -1/3$, $\mu = 1$); (f) F is concave ($\rho = 1/4$, $\mu = 3/16$).

2. (a) f is linear, so quasiconcave. (b) $x + \ln y$ is a sum of concave functions, so concave and thus quasiconcave. Because $u \mapsto e^u$ is increasing, $f(x, y) = e^{x+\ln y} = ye^x$ must be quasiconcave (Theorem 2.5.2(a)). (c) f is not quasiconcave. (It is quasiconvex in the first quadrant.) (d) f is quasiconcave.

3. (a) $a \geq 0$. (b) g is concave according to Theorem 2.3.5 (a). h is quasiconcave according to Theorem 2.5.2 (a), because f is, in particular, quasiconcave.

4. $f'(x) = -2x/(1+x^2)^2$ and $f(x) \rightarrow -1$ as $x \rightarrow \pm\infty$. See Fig. A2.5.4. $P_a = \{x : -x^2/(1+x^2) \geq a\}$. If $a > 0$, then P_a is empty. If $a = 0$, $P_a = \{0\}$. If $a \in (-1, 0)$, P_a is the closed interval with endpoints $\pm\sqrt{-a/(1+a)}$. Finally, if $a \leq -1$, then $P_a = (-\infty, \infty)$. In all cases, P_a is convex, so f is quasiconcave. Since $f'(0) = 0$, Theorem 2.5.6 does not apply.

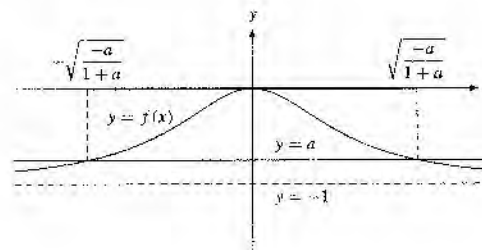


Figure A2.5.4

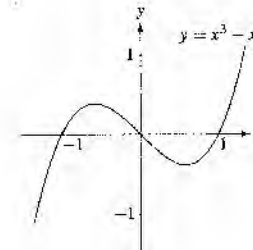


Figure A2.5.6

5. $f'(x) \neq 0$ for all x implies that f is strictly quasiconcave. (In fact, even if f is only a C^1 function, $f'(x) \neq 0$ implies that either f is (strictly) increasing or (strictly) decreasing, and so is quasiconcave according to Example 2. It is then also strictly quasiconcave.)

6. $f(x)$ and $g(x)$ are quasiconcave, but $h(x) = f(x) + g(x) = x^3 - x$ is not. See Fig. A2.5.6. See also SM.

7. (a) Follows from the definition. (b) No, $f(x) = |x - x^*|$ is single-peaked and concave, but not strictly concave.

8. If $f(x) \neq f(x^0)$, then x or x^0 is 0, and the right-hand side of (9) is 0. For $\lambda \in \{0, 1\}$, the left-hand side is 1. The set $\{x : f(x) \geq 1/2\} = (-\infty, 0] \cup (0, \infty)$ is not convex, so f is not quasiconcave.

9. We see that $\varphi''(x) = B_2(x, y)/(F_2^2)^3$, so the conclusion follows.

3. Use Theorem 2.5.1(5). Assume $F(f_1(\mathbf{x}), \dots, f_m(\mathbf{x})) \geq F(f_1(\mathbf{x}^0), \dots, f_m(\mathbf{x}^0))$. Now, by concavity of each f_i , $f_i(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}^0) \geq \lambda f_i(\mathbf{x}) + (1-\lambda)f_i(\mathbf{x}^0)$. Since F is increasing, $F(f_1(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}^0), \dots, f_m(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}^0)) \geq F(\lambda f_1(\mathbf{x}) + (1-\lambda)f_1(\mathbf{x}^0), \dots, \lambda f_m(\mathbf{x}) + (1-\lambda)f_m(\mathbf{x}^0)) \geq F(f_1(\mathbf{x}^0), \dots, f_m(\mathbf{x}^0))$, by quasiconcavity of F .

1. See SM.

6

1. (a) $f(x, y) \approx 1 + xy$ (b) $f(x, y) \approx 1 + x^2 - y^2$ (c) $f(x, y) \approx x + 2y - \frac{1}{2}x^2 - 2xy - 2y^2$

2. (a) $f(x, y) \approx -1 - x - y - \frac{1}{2}x^2 - \frac{1}{2}y^2$ (b) $f(x, y) \approx 1 + x + \frac{1}{2}x^2 + xy$ (c) $f(x, y) \approx x^2 + y^2$

3. $U(x_1, \dots, x_n) = \pi - x_1 - \dots - x_n + \frac{1}{2}(x_1^2 + \dots + x_n^2) + R_3$

4. When $x = y = 0$, $z = 1$. We get $z \approx 1 - x + y + \frac{1}{2}x^2 - 2xy + \frac{1}{2}y^2$.

7

1. (a) f is C^1 everywhere and $f'_y = 3y^2 + 1 \neq 0$, so Theorem 2.7.1 implies that z is defined as a C^1 -function of x and y in a neighbourhood of $(0, 0)$, and $y' = -f'_x/f'_y = 3x^2/(3y^2 + 1) = 0$ at $(0, 0)$. (b) $f'_y = 1 + x \cos(xy) = 1 \neq 0$ at $(0, 0)$, and $y' = -f'_x/f'_y = -(2x + y \cos(xy))/(1 + x \cos(xy)) = 0$ at $(0, 0)$.

2. (a) $F(x, y, z) = x^3 + y^3 + z^3 - xyz - 1$ is obviously C^1 everywhere, and $F'_3(0, 0, 1) = 3 \neq 0$, so by the implicit function theorem the equation defines z as a C^1 function of x and y in a neighbourhood of $(0, 0, 1)$. $g'_1(0, 0) = -F'_1(0, 0, 1)/F'_3(0, 0, 1) = 0$. Likewise, $g'_2(0, 0) = 0$. (b) $g'_1(1, 0) = 2$, $g'_2(1, 0) = 0$.

3. u'_x, v'_x , and w'_x must satisfy $u'_x - v'_x - 3w'_x = 0$, $-2 + u'_x - w'_x = 0$, $2 - u'_x - v'_x + 3w'_x = 0$. The unique solution is $u'_x = 5/2$, $v'_x = 1$, $w'_x = 1/2$.

4. $\partial(f, g)/\partial(u, v) = e^{2u}(\sin^2 v + \cos^2 v) = e^{2u} \neq 0$. (a) No solutions. (b) An infinite set of solutions: $u = \frac{1}{2} \ln 2$, $v = \pi/4 + 2k\pi$ for all integers k .

5. $\partial(F, G)/\partial(u, v) = -2u^2 + 4uv + 2v^2$, so around points where this expression is different from 0, one can express u and v as C^1 functions of x and y : $u = f(x, y)$ and $v = g(x, y)$. At $(x_0, y_0, u_0, v_0) = (2, 1, -1, 2)$ (which does satisfy the equations), $\partial(F, G)/\partial(u, v) = -2$ and $f'_x(2, 1) = 1/2$, $f'_y(2, 1) = 6$, $g'_x(2, 1) = 1$, and $g'_y(2, 1) = 2$.

5. The Jacobian is x_1 . We find that $x_1 = y_1 + y_2$, $x_2 = y_2/(y_1 + y_2)$ (provided $y_1 + y_2 \neq 0$). The transformation maps the given rectangle onto a quadrilateral in the $y_1 y_2$ -plane determined by the inequalities $1 \leq y_1 + y_2 \leq 2$ and $y_1 \leq y_2 \leq 2y_1$.

7. The Jacobian determinant is $ad - bc = 0$. Suppose $a \neq 0$. Then $v = (c/a)u$. The other cases are similar.

3. (a) $J = r$ (b) $T(r, 0) = T(r, 2\pi)$

9. The main condition is that the Jacobian determinant $J = \begin{vmatrix} f'_1 & f'_2 \\ g'_1 & g'_2 \end{vmatrix} \neq 0$. Differentiating the two identities $f(F(u, v), G(u, v)) = u$ and $g(F(u, v), G(u, v)) = v$ w.r.t. u , keeping v constant, yields $f'_1 F'_u + f'_2 G'_u = 1$ and $g'_1 F'_u + g'_2 G'_u = 0$. By Cramer's rule or otherwise we get the desired formulas.

1. (a) $u'_x = 3/2$, $u'_y = 5/(6 \ln 3)$ (b) $f(u) = u - ae^{a(b-1)}$ is strictly increasing because $b \leq 1$. Also $f(0) \leq 0$ and $f(1) \geq 0$. (c) Let $u = (1 + xy)/2$ and $b = x$, and use the result from (b). This gives a unique value of u . Because $u \in [0, 1]$, the first equation then gives a unique value of v .

8

1. (a) Three degrees of freedom. (b) $f'(Y - T) \neq 1$ is sufficient.

2. (a) Direct verification. (b) (i) $v = 2(\ln u)^2 - \ln u$, (ii) $v = (1 - u)/(1 + u)$

3. $\partial v/\partial y = g'_1(\partial \varphi/\partial y) + g'_2 = g'_1(-f'_2/f'_1) + g'_2 = 0$, because $f'_1 g'_2 = f'_2 g'_1$.

4. (a) Easy verification. (b) $w = (u^2 + v^2)/2$

2.9

1. (a) The partial derivatives obviously exist and are continuous for $(x, y) \neq (0, 0)$. Moreover, $f'_1(0, 0) = \lim_{h \rightarrow 0} [f(h, 0) - f(0, 0)]/h = \lim_{h \rightarrow 0} 0 = 0$, and similarly $f'_2(0, 0) = 0$. (b) Every directional derivative clearly exists for $(x, y) \neq (0, 0)$, because f is continuously differentiable at such points. If $\mathbf{a} = (a_1, a_2) \neq \mathbf{0}$, $f'_{\mathbf{a}}(0, 0) = \lim_{h \rightarrow 0} [f(ha_1, ha_2) - f(0, 0)]/h = \lim_{h \rightarrow 0} (a_1 a_2^2/(a_1^2 + h^2 a_2^2)) = a_2^2/a_1$ if $a_1 \neq 0$. If $a_1 = 0$, then $f'_{\mathbf{a}}(0, 0) = 0$. (c) $f(y^2, y) = 1/2$ for all $y \neq 0$ and $f(0, 0) = 0$, so f cannot be continuous at $(0, 0)$. Differentiability implies continuity, so f is not differentiable.

2.10

1. The leading principal minors are 1 and 1, but $\mathbf{x}' \mathbf{A} \mathbf{x} = (x_1 + x_2)^2$, which is 0 when $x_2 = -x_1$.

2. $\frac{c_1}{c_2} = \frac{p_1}{p_2} = \frac{y_1 w_1^\alpha w_2^{1-\alpha}}{y_2 w_1^\beta w_2^{1-\beta}} = \frac{y_1}{y_2} \frac{w_1^{\alpha-\beta} w_2^{\beta-\alpha}}{w_1^{\alpha-\beta} w_2^{\beta-\alpha}} = \left(\frac{w_1}{w_2}\right)^{\alpha-\beta} \frac{y_1}{y_2}$.

Hence $w_1/w_2 = (y_2/y_1)^{1/(\alpha-\beta)}(p_1/p_2)^{1/(\alpha-\beta)}$ if $\alpha \neq \beta$. If p_1/p_2 rises and $\alpha > \beta$, then w_1 rises relative to w_2 in the industry that uses factor 1 more intensively (because $\alpha > \beta$).

Chapter 3

3.1

1. The Hessian matrix is $g''(x, y) = \begin{pmatrix} g''_{11} & g''_{12} \\ g''_{21} & g''_{22} \end{pmatrix} = \begin{pmatrix} 6x & 0 \\ 0 & 6y \end{pmatrix}$. We see that g is strictly convex in its domain since the leading principal minors are positive. Stationary point where $3x^2 - 3 = 0$ and $3y^2 - 2 = 0$, so $(1, \sqrt{6}/3)$ is the only stationary point with $x > 0$, $y > 0$. It is a (global) minimum point for g , and $g_{\min} = g(1, \sqrt{6}/3) = -2 - 4\sqrt{6}/9$.

2. The profit is $\pi(x, y) = 13x + 8y - C(x, y) = 9x + 6y - 0.04x^2 + 0.01xy - 0.01y^2 - 500$. From the first-order conditions, $\pi'_x(x, y) = 9 - 0.08x + 0.01y = 0$ and $\pi'_y(x, y) = 6 + 0.01x - 0.02y = 0$, we get $x = 160$, $y = 380$. Since π is easily seen to be concave, $(160, 380)$ is the maximum point.

3. (a) $v_1^* = \frac{1}{216} p^6 q_1^{-3} q_2^{-3}$, $v_2^* = \frac{1}{144} p^6 q_1^{-2} q_2^{-4}$ (b) $\pi^*(p, q_1, q_2) = \frac{1}{432} p^6 q_1^{-2} q_2^{-3}$. The equalities in (*) follow easily.

4. The first-order conditions for (x^*, y^*) to solve the problem are that $f'_1(x^*, y^*, r) = -2x^* - y^* + 2r = 0$ and $f'_2(x^*, y^*, r) = -x^* - 4y^* + 2r = 0$. It follows that $x^* = 6r/7$ and $y^* = 2r/7$. f is (strictly) concave in (x, y) , so this is the solution. The (optimal) value function is $f^*(r) = f(x^*, y^*, r) = -(x^*)^2 - x^* y^* - 2(y^*)^2 + 2rx^* + 2ry^* = 8r^2/7$, so $df^*(r)/dr = 16r/7$. On the other hand, $f'_3(x^*, y^*, r) = 2x^* + 2y^* = 16r/7$, so $df^*(r)/dr = f'_3(x^*, y^*, r)$.

5. $x^*(r, s) = \frac{1}{2}r^2$ and $y^*(r, s) = \frac{3}{16}s^2$. Moreover, $f^*(r, s) = \frac{1}{2}r^4 + \frac{9}{32}s^4$, so $\partial f^*/\partial r = r^3$ and $\partial f^*/\partial s = 2rx$, so $\partial f^*(x^*, y^*, r, s)/\partial r = 2rx^* = r^3$. Also, $\partial f^*/\partial s = \partial f(x^*, y^*, r, s)/\partial s = 6sy^* = \frac{9}{8}s^3$.

6. (a) $v_i^* = p a_i / q_i - 1$, $i = 1, \dots, n$ (b) Differentiate $\pi(p, q, \mathbf{a}) = p[a_1 \ln(v_1 + 1) + \dots + a_n \ln(v_n + 1)] - q_1 v_1 - q_2 v_2 - \dots - q_n v_n$ w.r.t. all the parameters, and evaluate the partials at v_i^* , $i = 1, \dots, n$. Then compute the value function $\pi^*(p, q, \mathbf{a}) = \pi(v^*, p, q, \mathbf{a})$, and then differentiate w.r.t. all parameters.

3.2

1. The only stationary point is $(0, 0, 0)$. The leading principal minors of the Hessian are $D_1 = 2$, $D_2 = 3$, and $D_3 = 4$, so $(0, 0, 0)$ is a local minimum point by Theorem 3.2.1(a).

2. (a) Stationary points where $x^2 = y$ and $y^2 = x$. Then $y^4 = y$, with the solutions $y = 0$ and $y = 1$. Hence the stationary points are $(x, y) = (0, 0)$ and $(1, 1)$. The quadratic form is $-6h_1h_2$ at $(0, 0)$ and $6(h_1^2 - h_1h_2 + h_2^2)$ at $(1, 1)$. (b) $-6h_1h_2$ is indefinite. Completing the square we see that $6(h_1^2 - h_1h_2 + h_2^2) = 6((h_1 - \frac{1}{2}h_2)^2 + \frac{3}{4}h_2^2) > 0$ for $(h_1, h_2) \neq (0, 0)$, so this form is positive definite. (c) $(0, 0)$ is a saddle point and $(1, 1)$ is a local minimum point according to (1)–(3).
3. (a) $(0, -2, 0)$, $(0, 2, 0)$, and $(\pm\sqrt{3}, -1, 3/2)$ are saddle points. $(0, 0, 2)$ is a local minimum point. (b) $(-1, -1, 2, 3)$ is a saddle point, whereas $(5/3, 5/3, 2, 3)$ is a local maximum point.
4. It is easy to see that $(0, 0)$ is the only stationary point. The second derivatives are $f''_{11}(0, 0) = 2$, $f''_{12}(0, 0) = 0$, and $f''_{22}(0, 0) = 2$, so $(0, 0)$ is a local minimum point. It is not a global minimum point because $f(x, -2) = -x^2 + 4$ tends to $-\infty$ as $x \rightarrow \infty$.

3.3

1. (a) $x = a/6$, $y = a/3$, $z = a/6$, $\lambda = -a/3$. (The Lagrangian $\mathcal{L} = 100 - x^2 - y^2 - z^2 - \lambda(x + 2y + z - a)$ has stationary point where $\mathcal{L}'_x = -2x - \lambda = 0$, $\mathcal{L}'_y = -2y - 2\lambda = 0$, and $\mathcal{L}'_z = -2z - \lambda = 0$. Inserting $y = 2x$ and $z = x$ into the constraint yields $x = a/6$, and then $y = a/3$, $z = a/6$, with $\lambda = -a/3$. $\mathcal{L}$ is concave in (x, y, z) as a sum of concave functions.) (b) $f^*(a) = 100 - a^2/6$. We see that $df^*(a)/da = \lambda$.
2. Maximum $108\sqrt{7}/7$ at $(x, y, z) = (\frac{2}{3}\sqrt{7}, \frac{32}{7}\sqrt{7}, -\frac{27}{7}\sqrt{7})$, with $\lambda_1 = \frac{1}{28}\sqrt{7}$, $\lambda_2 = \frac{6}{7}$.
(b) $\Delta f^* \approx \frac{1}{28}\sqrt{7} \cdot (-1) + 0.1 \cdot \frac{6}{7} \approx -0.009$
3. (a) With $\mathcal{L} = e^x + y + z - \lambda_1(x + y + z - 1) - \lambda_2(x^2 + y^2 + z^2 - 1)$, the first-order conditions are (i) $\partial\mathcal{L}/\partial x = e^x - \lambda_1 - 2\lambda_2x = 0$, (ii) $\partial\mathcal{L}/\partial y = 1 - \lambda_1 - 2\lambda_2y = 0$, (iii) $\partial\mathcal{L}/\partial z = 1 - \lambda_1 - 2\lambda_2z = 0$. From (ii) and (iii), $\lambda_2y = \lambda_2z$, so (A) $\lambda_2 = 0$ or (B) $y = z$, etc. Four candidates: $(0, 0, 1)$ and $(0, 1, 0)$ with $\lambda_1 = 1$, $\lambda_2 = 0$; $(1, 0, 0)$ with $\lambda_1 = 1$, $\lambda_2 = \frac{1}{2}(e - 1)$; $(-\frac{1}{2}, \frac{2}{3}, \frac{2}{3})$ with $\lambda_1 = \frac{1}{4} + \frac{2}{3}e^{-1/3}$ and $\lambda_2 = \frac{1}{2} - \frac{1}{3}e^{-1/3}$. The value of the objective function is highest, and equal to e , at $(1, 0, 0)$. (The maximum exists by the extreme value theorem.) (b) $\Delta f^* \approx \lambda_1 \cdot (0.02) + \lambda_2 \cdot (-0.02) = 0.02 - 0.02 \cdot \frac{1}{2}(e - 1) \approx 0.01(3 - e)$.
4. (a) $\lambda_1^*(m) = \frac{1}{3}(m + 2)$, $\lambda_2^*(m) = \frac{1}{3}(m - 4)$, $\lambda = \frac{2}{3}(m + 5)^{-1}$
(b) $U^*(m) = \frac{2}{3} \ln(m + 5) - \ln 3$, $dU^*/dm = \frac{2}{3}(m + 5)^{-1} = \lambda$
5. We find that $x^* = m/(1 + r)$ and $y^* = m/(1 + r)$, so $f^*(r, m) = 1 - rm^2/(1 + r)$. We find that $\partial f^*(r, m)/\partial r = -m^2/(1 + r)^2 = -(x^*)^2$ and $\partial f^*(r, m)/\partial m = -2mr/(1 + r) = \lambda$.
6. (a) Maximum 1 at $(-3\sqrt{10}/10, \sqrt{10}/10, 0)$ and at $(3\sqrt{10}/10, -\sqrt{10}/10, 0)$ with $\lambda_1 = 1$, $\lambda_2 = 0$.
(b) $\Delta f^* \approx 1 \cdot 0.05 + 0 \cdot 0.05 = 0.05$.
7. (a) With $\mathcal{L} = \sum_{i=1}^n \alpha_i \ln(x_i - a_i) - \lambda(\sum_{i=1}^n p_i x_i - m)$, the first-order conditions are (a) $\mathcal{L}'_j = \alpha_j/(x_j - a_j) - \lambda p_j = 0$, or $p_j x_j^* = p_j a_j + \alpha_j/\lambda$, $j = 1, \dots, n$. Summing these equalities from $j = 1$ to n , yields $m = \sum_{j=1}^n a_j p_j + 1/\lambda$. The expression in the problem for $p_j x_j^*$ follows.
(b) $U^*(p, m) = U(x^*) = \sum_{j=1}^n \alpha_j \ln(\alpha_j(m - \sum_{i=1}^n p_i a_i)/p_j) = \sum_{j=1}^n \alpha_j \ln \alpha_j + \sum_{j=1}^n \alpha_j \ln(m - \sum_{i=1}^n p_i a_i) - \sum_{j=1}^n \alpha_j \ln p_j$. Roy's identity follows by differentiating w.r.t. p_i .
8. (a) The constraints have the unique solution $(x, y) = (\frac{1}{2}\sqrt{2}, \frac{1}{2}\sqrt{2})$, and this pair with $z = 0$ must solve the problem.
(b) With $\mathcal{L} = x^2 + (y - 1)^2 + z^2 - \lambda(x + y - \sqrt{2}) - \mu(x^2 + y^2 - 1)$, $\mathcal{L}'_z = 0$ at $z = 0$, and the constraints give the same solution as before. But the equations $\mathcal{L}'_x = 0$ and $\mathcal{L}'_y = 0$ give a contradiction. (The matrix in (7) is here $\begin{pmatrix} 1 & 1 & 0 \\ 2x & 2y & 0 \end{pmatrix}$, which has rank 1 when $x = y$.)
9. The Lagrangian is $\mathcal{L} = \mathbf{x}'\mathbf{A}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{x} - 1)$. In vector notation, $\mathcal{L}'_x = 2\mathbf{A}\mathbf{x} - 2\lambda\mathbf{x} = 0$, or (*) $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ (see Note 2.3.4), so λ is an eigenvalue of $\mathbf{A}$, with $\mathbf{x}$ as an eigenvector. Multiplying (*) from the left by $\mathbf{x}'$ gives $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'(\lambda\mathbf{x}) = \lambda\mathbf{x}'\mathbf{x} = \lambda$, because $\mathbf{x}'\mathbf{x} = 1$. We conclude that the maximum value of $\mathbf{x}'\mathbf{A}\mathbf{x}$ subject to $\mathbf{x}'\mathbf{x} = 1$ must be the largest eigenvalue of $\mathbf{A}$, whereas the minimum value must be the smallest eigenvalue. Both a maximum and a minimum value exist by the extreme value theorem.

10. See SM.

3.4

1. (a) $(x, y) = (\pm 1, 0)$ with $\lambda = 1/4$ and $(0, \pm\sqrt{2})$ with $\lambda = 1/2$. (b) $B_2(\pm 1, 0) = -64$, so $(\pm 1, 0)$ are local minimum points. $(-1)^2 B_2(0, \pm\sqrt{2}) = 64$, so $(0, \pm\sqrt{2})$ are local maximum points.
(c) The constraint curve is an ellipse, and the problem is to find those points on the curve that have the smallest and the largest (square) distance from $(0, 0)$.
2. $B_1 = -4$, $B_3 = \begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 2 \end{vmatrix} = -12$, so the second-order condition for local minimum is satisfied. (The minimum point is $(1/3, 1/3, 1/3)$.)

3. At $(1, 0, 0)$ with $\lambda_1 = 1$, $\lambda_2 = -1$, we have $(-1)^3 B_3 = 16 > 0$, which gives a local maximum. At $(-1/3, -2/3, -2/3)$ with $\lambda_1 = -1$, $\lambda_2 = 1/3$, we have $(-1)^2 B_3 = 16 > 0$, which gives a local minimum.

3.5

1. The solution is obvious: to maximize $1 - x^2 - y^2$ you must have x and y as small as they are allowed to be, i.e. $x = -2$ and $y = 3$. With $\mathcal{L} = 1 - x^2 - y^2 - \lambda(-x + 2) - \mu(-y + 3)$, the first-order conditions for (x^*, y^*) to solve the problem are: (i) $-2x^* + \lambda = 0$; (ii) $-2y^* + \mu = 0$; (iii) $\lambda \geq 0$ with $\lambda = 0$ if $x^* > 2$; (iv) $\mu \geq 0$ with $\mu = 0$ if $y^* > 3$. Since $x^* \geq 2$, (i) implies $\lambda = 2x^* > 0$, and since $y^* \geq 3$, (ii) implies $\mu = 2y^* > 0$, so from (iii) and (iv) we conclude that $x^* = 2$, $y^* = 3$. This is an optimal solution since $\mathcal{L}$ is concave in (x, y) .
2. (a) See SM. (b) $(x, y) = (3/2, 1/2)$ solves the problem. (c) $V'(5/2) = 4/15$
3. The solution is $(x, y, \lambda_1, \lambda_2) = (\sqrt{2}/2, 2 - \sqrt{2}, 4 - 2\sqrt{2}, 0)$. (Rewrite the problem as: $\max -4 \ln(x^2 + 2) - y^2$ subject to $-x^2 - y \leq -2$, $-x \leq -1$. The necessary Kuhn-Tucker conditions are: (i) $-\frac{8x}{x^2 + 2} + 2\lambda_1 x + \lambda_2 = 0$; (ii) $-2y + \lambda_1 = 0$; (iii) $\lambda_1 \geq 0$ with $\lambda_1 = 0$ if $x^2 + y > 2$; (iv) $\lambda_2 \geq 0$ with $\lambda_2 = 0$ if $x > 1$.)
4. If $a > 1$ and $b > 2$, $(x^*, y^*) = (1, 2)$; if $a > 1$ and $b \leq 2$, $(x^*, y^*) = (1, b)$; if $a \leq 1$ and $b \leq 2$, $(x^*, y^*) = (a, b)$; if $a \leq 1$ and $b > 2$, $(x^*, y^*) = (a, 2)$.
5. Because $(x + y - 2)^2 \leq 0$, the constraint is equivalent to $x + y - 2 = 0$ and the solution is $(x, y) = (1, 1)$. From the Kuhn-Tucker conditions, $y - 2\lambda(x + y - 2) = 0$ and $x - 2\lambda(x + y - 2) = 0$. Letting $(x, y) = (1, 1)$ yields the contradiction $1 = 0$. Note that $g'_1(x, y) = g'_2(x, y) = 2(x + y - 2) = 0$ for $x = y = 1$, so the gradient of g at $(1, 1)$ is $(0, 0)$, which is not a linearly independent. Thus, the constraint qualification does not hold at $(1, 1)$.
6. (a) $(x, y) = (-\frac{1}{5}\sqrt{15}, -\frac{1}{5}\sqrt{15})$ (b) $f(x) = x^5 - x^3$, $x \leq 1$. Max. at $x = -\frac{1}{5}\sqrt{15}$.

3.6

1. Kuhn-Tucker conditions: (i) $-2(x-1)-2\lambda x = 0$; (ii) $-2ye^{x^2} - 2\lambda y = 0$; (iii) $\lambda \geq 0$, with $\lambda = 0$ if $x^2 + y^2 < 1$. From (ii) and $\lambda \geq 0$, $y = 0$. From (i), $x = 1/(1 + \lambda) > 0$. If $\lambda > 0$, then $x < 1$ and so $x^2 < 1$. Then (iii) gives $\lambda = 0$. So $\lambda = 0$, and the solution is $x = 1$, $y = 0$. Because the Lagrangian is concave, this is the optimal solution. Note that for the optimal solution $\lambda = 0$ and $x^2 + y^2 = 1$.
2. $(x, y, \lambda_1, \lambda_2) = (1/2, 1/2, 0, 3/2)$. (With $\mathcal{L} = xy + x + y - \lambda_1(x^2 + y^2 - 2) - \lambda_2(x + y - 1)$, the first-order conditions are (i) $\partial\mathcal{L}/\partial x = y + 1 - 2\lambda_1 x - \lambda_2 = 0$; (ii) $\partial\mathcal{L}/\partial y = x + 1 - 2\lambda_1 y - \lambda_2 = 0$; (iii) $\lambda_1 \geq 0$, with $\lambda_1 = 0$ if $x^2 + y^2 < 2$; (iv) $\lambda_2 \geq 0$, with $\lambda_2 = 0$ if $x + y < 1$. Look at the cases (A) $\lambda_1 = 0$, $\lambda_2 = 0$; (B) $\lambda_1 = 0$, $\lambda_2 > 0$; etc. Case (B) gives the solution. An optimal solution exists by the extreme value theorem.)

3.7

1. (a) For $a \geq 3b$, $(x, y, z) = (b, \frac{1}{2}(a-b), \frac{1}{2}(a-b))$ with $\lambda_1 = e^{-(a-b)/2}$ and $\lambda_2 = e^{-b} - e^{-(a-b)/2}$. For $a < 3b$, $x = y = z = \frac{1}{3}a$ with $\lambda_1 = e^{-a/3}$ and $\lambda_2 = 0$. (b) For $a \geq 3b$, $f^*(a, b) = 1(0) - e^{-b} - 2e^{-(a-b)/2}$ and

- $\partial f^*/\partial a = \lambda_1$, $\partial f^*/\partial b = \lambda_2$. For $a < 3b$, $f^*(a, b) = 100 - 3e^{-a/3}$ and $\partial f^*/\partial a = \lambda_1$, $\partial f^*/\partial b = \lambda_2$.
 (c) Let $g(a) = 100 - 3e^{-a/3}$, $h(a) = 99 - 2e^{-a/2}$. Then $F^*(a) = g(a)$ if $a < 0$, $F^*(a) = h(a)$ if $a \geq 0$. The functions g and h are both concave. Moreover, $g(0) = h(0)$ and $g'(0) = h'(0)$, so their graphs have a common tangent at the point corresponding to $a = 0$. It follows that F^* is concave.
1. $f^*(r) = (1+r)^2$ for $r \geq 0$, $f^*(r) = (1-r)^2$ for $r < 0$. (Graph the parabola $y = (x-r)^2$ over the interval $[-1, 1]$ for different values of r . Note that $df^*(r)/dr \rightarrow 2$ as $r \rightarrow 0^+$, while $df^*(r)/dr \rightarrow -2$ as $r \rightarrow 0^-$.)
1. (a) $x = \pm \frac{1}{2} s \sqrt{2}$, $y = 0$, with $\lambda = 0$, $\mu = \frac{1}{2}$, $f^*(r, s) = \frac{1}{2} s^2$. (b) $x = 0$, $y = \pm \frac{1}{2} r$, with $\lambda = \frac{1}{4}$, $\mu = 0$, $f^*(r, s) = \frac{1}{4} r^2$. (c) The admissible set is the area between two ellipses, and the problem is to find the squares of the largest and the smallest distances from the origin to a point in this admissible set.
1. See SM.

8

1. (a) $1 - x^2 - y^2 \leq 1$ for all $x \geq 0$, $y \geq 0$, so the optimal solution must be $x = y = 0$.
 (b) With $\mathcal{L} = 1 - x^2 - y^2$, the Kuhn-Tucker conditions are: (i) $\partial \mathcal{L}/\partial x = -2x \leq 0$ ($= 0$ if $x > 0$); (ii) $\partial \mathcal{L}/\partial y = -2y \leq 0$ ($= 0$ if $y > 0$). The only solution is obviously $x = y = 0$. The Lagrangian is concave.
1. (a) $(x, y) = (1, 1/2)$ (b) $(x, y) = (2\alpha/(\alpha + \beta), \beta/(\alpha + \beta))$
1. (a) The solution is: For $c \leq 0$, $x^* = 0$, $y^* = \sqrt{6}/3$, with $\lambda = \sqrt{6}/12$; for $c > 0$, $x^* = c\sqrt{6}/\sqrt{3c^2 + 1}$, $y^* = \sqrt{6}/3\sqrt{3c^2 + 1}$, with $\lambda = \sqrt{6}/3\sqrt{3c^2 + 1}/12$. (b) For $c \leq 0$, $f^*(c) = cx^* + y^* = \sqrt{6}/3$. For $c > 0$, $f^*(c) = cx^* + y^* = (\sqrt{6}/3)\sqrt{3c^2 + 1} \rightarrow \sqrt{6}/3$ as $c \rightarrow 0$, so f^* is continuous also at $c = 0$. Note that $df^*(c)/dc = x^*$ in both cases, so (3.7.5) holds.
1. (a) With $\mathcal{L} = \ln(1+x) + y - \lambda(px + y - m)$, the Kuhn-Tucker conditions are: (i) $1/(1+x^*) - \lambda p \leq 0$ ($= 0$ if $x^* > 0$); (ii) $1 - \lambda \leq 0$ ($= 0$ if $y^* > 0$); (iii) $\lambda \geq 0$ with $\lambda = 0$ if $px^* + y^* < m$. (b) $x = -1 + 1/p$, $y = m + p - 1$, with $\lambda = 1$ is the solution for all $p \in (0, 1]$ and $m > 1$. $\mathcal{L}(x, y)$ is concave in (x, y) . (Hint: $\lambda \geq 1 > 0$, so $px^* + y^* = m$. Look first at the case $x^* > 0$, $y^* > 0$, which gives the solution.)

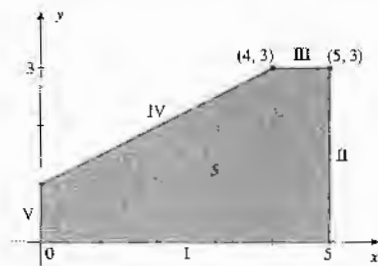


Figure A3.8.5

See Fig. A3.8.5. There are no stationary points in the interior of S . The Kuhn-Tucker conditions are: (i) $1 - (x + y) - 1/4 - \lambda_1 + \lambda_3 \leq 0$ ($= 0$ if $x > 0$); (ii) $1 - (x + y) - 1/3 - \lambda_2 - 2\lambda_3 \leq 0$ ($= 0$ if $y > 0$); (iii) $\lambda_1 \geq 0$ with $\lambda_1 = 0$ if $x < 5$; (iv) $\lambda_2 \geq 0$ with $\lambda_2 = 0$ if $y < 3$; (v) $\lambda_3 \geq 0$ with $\lambda_3 = 0$ if $-x + 2y < 2$. The solution is $x = 3/4$, $y = 0$.

See SM.

9

Use (A) to prove that $\sum_{j=1}^m \lambda_j g_j(\bar{x}) \geq \sum_{j=1}^m \lambda_j g_j(x^*)$. Then argue why the last inequality is an equality. See SM for details.

3.10

1. (a) With $\mathcal{L}(x, y, z, a) = x^2 + y^2 + z^2 - \lambda_1(2x^2 + y^2 + z^2 - a^2) - \lambda_2(x + y + z)$, the necessary conditions are: (i) $\partial \mathcal{L}/\partial x = 2x - 4\lambda_1 x - \lambda_2 = 0$; (ii) $\partial \mathcal{L}/\partial y = 2y - 2\lambda_1 y - \lambda_2 = 0$; (iii) $\partial \mathcal{L}/\partial z = 2z - 2\lambda_1 z - \lambda_2 = 0$; (iv) $\lambda_1 \geq 0$ with $\lambda_1 = 0$ if $2x^2 + y^2 + z^2 < a^2$. Solution: $(x^*, y^*, z^*) = (0, \pm \frac{1}{2}\sqrt{2}a, \mp \frac{1}{2}\sqrt{2}a)$, with $\lambda_1 = 1$, $\lambda_2 = 0$ both solve the problem. (b) $f^*(a) = a^2$ so $df^*(a)/da = 2a$, and $\partial \mathcal{L}(x^*, y^*, z^*, a)/\partial a = 2\lambda_1 a = 2a$.

3.11

1. See SM.

Chapter 4

4.1

1. (a) $x - x^3 + C$ (b) $-\frac{1}{3}x^3 + C$ (c) $\int (1 - x^2)^2 dx = \int (1 - 2x^2 + x^4) dx = x - \frac{2}{3}x^3 + \frac{1}{5}x^5 + C$
2. (a) $2500/3$ (b) $\int_0^{10} (-2te^{-2t} - e^{-2t}) dt = 1 - 21e^{-20}$ (c) $320/3 + 11 \ln 11 \approx 133$. (Introduce $u = t + 1$ as a new variable, or use polynomial division: $(10t^2 - t^3)/(t + 1) = -t^3 + 11t - 11 + 11/(t + 1)$.)
3. (a) $64/3 - 12\sqrt{3}$. (Substitute $u = \sqrt{4 - x^2}$.) (b) $2 - 6 \ln \frac{7}{6}$. (Substitute $u = 3 + \sqrt{t + 8}$. Then $u - 3 = \sqrt{t + 8}$ and $(u - 3)^2 = t + 8$. Differentiation yields $2(u - 3) du = dt$, etc.) (c) $\frac{8}{3}e^3 + \frac{4}{3}$. (Integration by parts.)
4. (a) $\int \frac{x^{2n} - 2x^n x^m + x^{2m}}{\sqrt{x}} dx = \frac{2x^{2n+1/2}}{4n+1} - \frac{4x^{n+m+1/2}}{2n+2m+1} + \frac{2x^{2m+1/2}}{4m+1} + C$ (b) $\frac{1}{3} - \ln(e^{1/3} + 1) + \ln 2$. (Substitute $u = e^x + 1$, $dx = du/e^x = du/(u - 1)$.) (c) $272/15$. (Substitute $u = \sqrt{x - 1}$.)
5. (a) $1 + \ln \frac{9}{4}$ (b) $\frac{1}{2}$. (Substitute $u = 1 + \sqrt{x}$, then integration by parts.) (c) $45/2 - 3 \ln 4$. (Substitute $u = 1 + x^{1/2}$.)

4.2

1. (a) $F'(x) = \int_1^2 e^{xt} dt = \frac{1}{x}(e^{2x} - e^x)$ (b) $F'(x) = \int_1^x \frac{1}{t} dt = \frac{1}{x}(e - 1)$ (c) $F'(x) = \int_0^1 \frac{-te^{-t}}{(1 + xt)^2} dt$
 (d) $F'(x) = \int_3^8 \frac{2t^3}{(1 - xt)^3} dt$
2. $F'(a) = (\alpha e^a - e^a + 1)/2a^2$ for $\alpha \neq 0$, $F'(0) = 1/4$.
3. (a) $F'(x) = 16x^3$ (b) $F'(x) = x^6 + 3x^5 + 5x^4$ (c) $F'(x) = 2x - \frac{1}{2}x^{-1/2} \cos(x - x^2) + 4x^3 \int_{\sqrt{x}}^1 \sin(t^2 - x^2) dt$
4. $I = e^{-\alpha g(p)} f(g(p)) g'(p) - \int_0^{g(p)} t e^{-\alpha t} f(t) dt$
5. $M'(t) = \int_{-\infty}^{\infty} x e^{tx} f(x) dx$, and so $M'(0) = \int_{-\infty}^{\infty} x f(x) dx$. By induction, $M^{(n)}(t) = \int_{-\infty}^{\infty} x^n e^{tx} f(x) dx$ and $M^{(n)}(0) = \int_{-\infty}^{\infty} x^n f(x) dx$.
6. $\dot{z}(t) = y(t) - \delta x(t)$ 7. $dF(\sigma_k)/d\sigma_k = \int_{-\infty}^{\infty} U'(y_k + \sigma_k z)(d\mu_k/d\sigma_k + z) f(z, 0, 1) dz$
8. $\dot{V}(t) = (1/T(0))\{G(t, t) + \int_{-\infty}^t (\partial G/\partial t) d\tau\}$. Here $\partial G/\partial t = -k(t)f(t - \tau)$ and $G(t, t) = k(t) \int_0^{\infty} f(\xi) d\xi = k(t)T(0)$. The conclusion follows.
9. We have $z(t) = \int_1^{2t} F(\tau, t) d\tau$, where $F(\tau, t) = x(\tau)e^{-\int_{\tau}^t r(s) ds}$. Leibniz's formula gives $\dot{z}(t) = 2F(2t, t) - F(t, t) + \int_1^{2t} (\partial F/\partial t)(\tau, t) d\tau = 2x(2t)e^{-\int_t^{2t} r(s) ds} - x(t) + \int_1^{2t} x(\tau)e^{-\int_{\tau}^t r(s) ds} r(t) d\tau = 2x(2t)p(t) - x(t) + \int_1^{2t} F(\tau, t)r(t) d\tau = 2p(t)x(2t) - x(t) + r(t)z(t)$, and therefore $\dot{z}(t) - r(t)z(t) = 2p(t)x(2t) - x(t)$.
10. (a) $g'(Q) = c + h \int_0^2 f(D) dD - p \int_0^Q f(D) dD$, $g''(Q) = (h + p)f(Q) \geq 0$ for all Q , and therefore g is convex. (b) $F(Q^*) = (p - c)/(h + p)$

1.3

- (a) $\frac{1}{2}\sqrt{\pi/a}$. (Substitute $t = \sqrt{ax}$ and use (3).) (b) 1. (Substitute $t = x/\sqrt{2}$ and use (3).)
- From (2), $\Gamma(\frac{1}{2}) = \Gamma(\frac{1}{2} + 1) = \frac{1}{2}\Gamma(\frac{1}{2}) \cdot \frac{1}{2}\sqrt{\pi}$. For the induction proof, see SM.
- $n! = \Gamma(n+1) = \sqrt{2\pi n} (n+1)^{n+1/2} e^{-(n+1)} e^{\theta/12(n+1)} = s_n \sqrt{2\pi n} (n/e)^n$, where $s_n = (n+1)^{n+1/2} n^{-n-1/2} e^{-1} e^{\theta/12(n+1)} = \sqrt{1+1/n} (1+1/n)^n e^{-1} e^{\theta/12(n+1)} \rightarrow 1$ as $n \rightarrow \infty$.
- $dt = -dz/z$, $t = 0$ gives $z = 1$, and $t = \infty$ gives $z = 0$. Then $\Gamma'(x) = \int_0^\infty e^{-t} t^{x-1} dt = \int_0^\infty t^{x-1} (-e^{-t}) dt = \int_0^1 (\ln(1/z))^{x-1} dz$.
- (a) Introduce $u = \lambda x$ as a new variable. (b) $M(t) = \lambda^\alpha (\lambda - t)^{-\alpha}$. Then $M'(0) = \alpha/\lambda$, and in general, $M^n(0) = \alpha(\alpha+1) \cdots (\alpha+n-1)/\lambda^n$.

1.4

- (a) $\int_0^2 (\int_0^{2x} (2x+3y+4) dx) dy = \int_0^2 (\int_{1/2}^{3/2} (x^2+3xy+4x)) dy = \int_0^2 (5+3y) dy = \int_0^2 (5y + \frac{3}{2}y^2) dy = 16$
(b) $\frac{1}{2}ab^2(3a-b)$ (c) $16\ln 2 - 3\ln 3 - 5\ln 5$ (d) $1/8 - 1/4\pi$
- $\frac{1}{b}(e^b - e^{b/a}) + \frac{1}{a} \rightarrow 1$ 3. $k_a = 2 + 4/(a^2 + 3a) > 2$ for all $a > 0$.
- $t = -16$. (The inner integral is $\int_{-2}^1 (x^2 y^3 - (y+1)^2) dy = -\frac{15}{4}x^2 - 3$) 5. See SM.

1.5

- (a) $11/120$. See Fig. A4.5.1(a). (Here $\int_{x^2}^{y^2} (x^2 + xy) dy = \int_{y=x^2}^{y=y^2} (x^2 y + \frac{1}{2}xy^2) = \frac{2}{3}x^3 - x^4 - \frac{1}{2}x^5$. This gives $\int_0^1 (\int_{x^2}^{y^2} (x^2 + xy) dy) dx = \int_0^1 (\frac{2}{3}x^3 - x^4 - \frac{1}{2}x^5) dx = \int_0^1 (\frac{2}{3}x^4 - \frac{1}{2}x^5 - \frac{1}{12}x^6) = \frac{11}{120}$.)
(b) $\int_0^1 (\int_{y^2}^{y^2} (x^2 + xy) dx) dy = \int_0^1 (\frac{1}{3}y\sqrt{y} + \frac{1}{2}y^2 - \frac{5}{6}y^3) dy = \frac{11}{120}$. See Fig. A4.5.1(b).

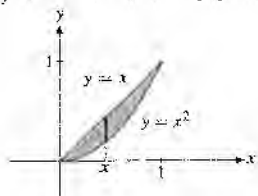


Figure A4.5.1(a)

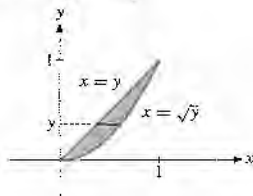


Figure A4.5.1(b)

- $V = \int_0^1 (\int_0^y xy^2 dx) dy + \int_1^2 (\int_{y^2-1}^1 xy^2 dx) dy = \int_0^1 (y^4/2) dy + \int_1^2 (y^2 - y^3/2) dy = 67/120$
- The double integral gives the area of A . 4. $V = \int_0^1 (\int_0^x e^{x^2} dy) dx = \int_0^1 x e^{x^2} dx = \frac{1}{2}e^{x^2} = \frac{1}{2}(e-1)$

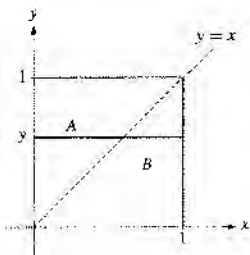


Figure A4.5.6(a)

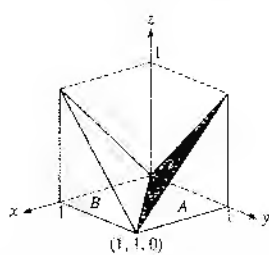


Figure A4.5.6(b)

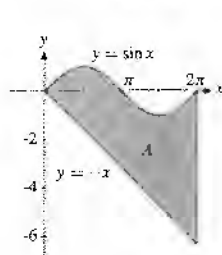


Figure A4.5.7

$$5. \int_0^3 (\int_0^{\sqrt{25-y^2}} 2x dx) dy + \int_0^4 (\int_0^{3y/4} 2x dx) dy = 14/3 + 12 = 50/3$$

- See Fig. A4.5.6(a). $\int_0^1 \int_0^1 |x-y| dx dy = \frac{1}{3}$. Geometrically we compute the volume of two pyramids (see Fig. A4.5.6(b)). Each pyramid has a triangular base (B and A) with area $1 \cdot \frac{1}{2} = \frac{1}{2}$ and height 1. So the volume of each is $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$. The total volume is $2 \cdot \frac{1}{6} = \frac{1}{3}$.
- Figure A4.5.7 shows the set A . We get $\iint_A 2y \cos x dx dy = \int_0^{2\pi} (\int_{-x}^{x} 2y \cos x dy) dx = -4\pi$. (You need integration by parts to get $\int x^2 \cos x dx = x^2 \sin x + 2x \cos x - 2 \sin x + C$.)
- (a) $\int_0^{F-\theta} e^{a\theta} e^{bT} dT = e^{a\theta} [\frac{1}{b} e^{bT}]_0^{F-\theta} = \frac{1}{b} (e^{bF} e^{(a-b)\theta} - e^{a\theta})$. Then $f = \frac{1}{b} \int_0^F (e^{bF} e^{(a-b)\theta} - e^{a\theta}) d\theta = \frac{e^{bF} - e^{aF}}{a-b}$.
(b) See Figs. A4.5.8(a) and A4.5.8(b). $I = \int_0^T (\int_0^{F-T} e^{a\theta} e^{bT} d\theta) dT$.

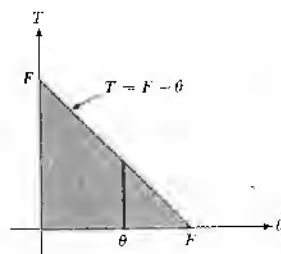


Figure A4.5.8(a)

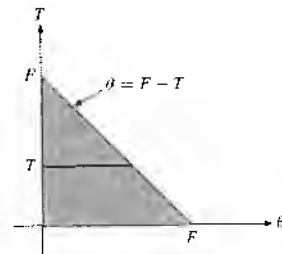


Figure A4.5.8(b)

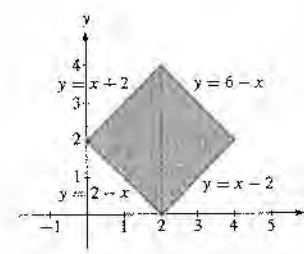


Figure A4.7.1

- (a) $\int_0^a (\int_0^b f(\xi_1, \xi_2) d\xi_1) d\xi_2$, where $a = 1/q_1$, $b = 1/q_2 - q_1 \xi_1/q_2$. (b) $\int_0^c (\int_0^d f(\xi_1, \xi_2) d\xi_1) d\xi_2$, where $c = 1/q_2$, $d = 1/q_1 - q_2 \xi_2/q_1$. (c) $\partial g/\partial q_1 = -(1/q_1) \int_0^{1/q_1} \xi_1 f(\xi_1, 1/q_2 - q_1 \xi_1/q_2) d\xi_1$

4.6

- (a) 2. See SM. (b) $\int_0^2 (\int_0^1 (2x - y + 1) dx) dy = \int_0^2 (\int_0^1 (x^2 - xy + x)) dy = \int_0^2 (2 - y) dy = 2$.

4.7

- (a) See Fig. A4.7.1. $I = \int_0^2 (\int_{x^2}^{x^2+x} (x+xy) dy) dx + \int_2^4 (\int_{x^2-2}^{x^2-x} (x+xy) dy) dx = \int_0^2 6x^2 dx + \int_2^4 (-6x^2 + 24x) dx = 16 + 32 = 48$. (b) $\int_2^4 [\int_{x^2-2}^{x^2-x} (\frac{1}{2}u + \frac{1}{2}v + \frac{1}{4}u^2 - \frac{1}{4}u^2) \frac{1}{2} du] dv = \int_2^4 (\frac{1}{2}v^2 + v - \frac{3}{2}) dv = 48$
- See SM.
- (a) $\pi/64$. ($\iint_A x^2 dx dy = \int_0^{2\pi} (\int_0^{1/2} (r^2 \cos^2 \theta) r dr) d\theta = \int_0^{2\pi} \cos^2 \theta d\theta \int_0^{1/2} r^3 dr = \pi/64$.) (b) $\pi/64$
- (a) $ad - bc$ (b) $\text{area}(T(A')) = \text{area}(A) \cdot \sqrt{2}\sqrt{8} = 4$, $|J| = \frac{1}{2}$ and $\text{area}(A') = 2 \cdot 4 = 8$, so the formula is confirmed.
- (a) $\pi/2$. (Using polar coordinates: $\iint_{A_1} (1 - x^2 - y^2) dx dy = \int_0^{2\pi} d\theta \int_0^1 (1 - r^2) r dr = \pi/2$.) (b) $144/25$

4.8

- (a) $\frac{1}{2}\pi$. (Use polar coordinates, let $A_n = \{(x, y) : 1 \leq x^2 + y^2 \leq n^2\}$, and compute $\iint_{A_n} (x^2 + y^2)^{-3} dx dy$.)
(b) Convergence to $\pi/(p-1)$ for $p > 1$. Divergence for $p \leq 1$.
- Let $I(z) = \int_{-\infty}^{\infty} F(x, z) dx$, where $F(x, z) = \int_{-\infty}^{\infty} f(x, y) dy$. Then $I'(z) = \int_{-\infty}^{\infty} (\partial F(x, z)/\partial z) dx = \int_{-\infty}^{\infty} f(x, z - x) dx$.

(a) Introducing polar coordinates, $\iint_{x^2+y^2 \leq 1} k\sqrt{1-x^2-y^2} dx dy = k \int_0^{2\pi} d\theta \int_0^1 \sqrt{1-r^2} \cdot r dr = 2\pi k/3$, so $k = 3/2\pi$. (b) $f_X(x) = (3/2\pi) \int_{x^2+y^2 \leq 1} \sqrt{1-x^2-y^2} dy = (3/2\pi) \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \sqrt{1-x^2-y^2} dy = (3/2\pi) \frac{1}{2} \pi a^2 = \frac{3}{4}(1-x^2)$, where $a = \sqrt{1-x^2}$.

One iterated integral is $\int_1^\infty \left(\int_1^\infty \frac{y-x}{(y+x)^3} dy \right) dx = \lim_{a \rightarrow \infty} \left[\lim_{b \rightarrow \infty} I(b, a) \right] = \frac{1}{2}$, whereas the other is $\int_1^\infty \left(\int_1^\infty \frac{y-x}{(y+x)^3} dx \right) dy = \lim_{a \rightarrow \infty} \left[\lim_{b \rightarrow \infty} I(b, a) \right] = -\frac{1}{2}$, so the double integral is not well defined.

Put $F(x, y) = \int_{-\infty}^x G(u, y) du$, with $G(u, y) = \int_{-\infty}^y f(u, v) dv$. Then $F_1'(x, y) = G(x, y)$ and $F_{12}''(x, y) = G_2'(x, y) = f(x, y)$. If $F(x, y) = \frac{1}{4}(2 - e^{-x})(2 - e^{-y})$ we easily find that $F_{12}''(x, y) = \frac{1}{4}e^{-x}e^{-y} = \frac{1}{4}e^{-x-y}$.

(a) Using the sets A_n in Example 2(i), $I_n = \int_0^{2\pi} \left(\int_0^n r/(1+r^2)^{3/2} dr \right) d\theta = \int_0^{2\pi} d\theta \int_0^n r/(1+r^2)^{3/2} dr = 2\pi(1 - 1/\sqrt{1+n^2}) \rightarrow 2\pi$ as $n \rightarrow \infty$. (b) $\frac{1}{2}\pi^{3/2}$.

(a) π . (Introducing polar coordinates and defining A_n as in Example 3, we get $\iint_{A_n} x^2(x^2+y^2)^{-3/2} dx dy = \int_0^{2\pi} \left(\int_{1/n}^1 (r^2 \cos^2 \theta) r^{-3} r dr \right) d\theta = \int_0^{2\pi} \cos^2 \theta d\theta \int_{1/n}^1 dr = \pi(1 - 1/n) \rightarrow \pi$ as $n \rightarrow \infty$.) (b) π .

Chapter 5

1

If $x(t) = Ce^{-t} + \frac{1}{2}e^t$, then $\dot{x}(t) + x(t) = -Ce^{-t} + \frac{1}{2}e^t + Ce^{-t} + \frac{1}{2}e^t = e^t$.

If $x = Cr^2$, then $\dot{x} = 2Cr$, and so $t\dot{x} = 2Cr^2 = 2x$. The curve $x = Cr^2$ passes through $(1, 2)$ if $C = 2$. Hence, $x = 2r^2$ is the desired solution.

Differentiate $xe^{tx} = C$ implicitly to obtain $\dot{x}e^{tx} + x[e^{tx}(x + t\dot{x})] = 0$. Cancelling e^{tx} and rearranging gives $(1 + tx)\dot{x} = -x^2$.

(a) Differentiation of $x^2 = 2at$ w.r.t. t gives $2x\dot{x} = 2a$, and further $2t\dot{x}^2 + a = 2ta^2/x^2 = 2a$.

(b) Differentiation yields $\frac{1}{2}e^{2t}2t - e^{2t}\dot{x}(x+1) + e^{-x}\dot{x} = 0$, and the result follows.

(c) Differentiation yields $-x^2 + 2(1-t)x\dot{x} = 3t^2$. Simplifying and using $(1-t)x^2 = t^3$ yields the given equation.

If $x = Ct - C^2$, then $\dot{x} = C$, so $\dot{x}^2 = C^2$ and $t\dot{x} - x = tC - Ct + C^2 = C^2$. If $x = \frac{1}{4}t^2$, then $\dot{x} = \frac{1}{2}t$, so $\dot{x}^2 = \frac{1}{4}t^2$, and $t\dot{x} - x = \frac{1}{4}t^2$. We conclude that $x = Ct - C^2$ is not the general solution.

Since $\dot{x} = (1+x^2)t$, $\dot{x} < 0$ for $t < 0$ and $\dot{x} > 0$ for $t > 0$. Thus $t = 0$ is a global minimum point, and because $x(0) = 0$, one has $x(t) \geq 0$ for all t . Differentiating the equation w.r.t. t yields $\ddot{x} = 2xt + (1+x^2) = 2xt(1+x^2)t + (1+x^2) = (1+x^2)(2xt^2 + 1)$. Clearly $\ddot{x} > 0$ for all t , so $x = x(t)$ is convex.

2

See Fig. A5.2.1. (The solutions are $x = Ct$, for $t \neq 0$, with C an arbitrary constant.)

The desired integral curve is the upper semicircle in Fig. A5.2.2. See SM.

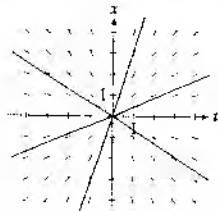


Figure A5.2.1

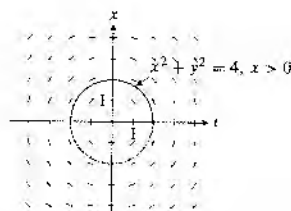


Figure A5.2.2

5.3

1. $x^2 dx = (t+1) dt$, so $\int x^2 dx = \int (t+1) dt$. Integration yields $\frac{1}{3}x^3 = \frac{1}{2}t^2 + t + C$, or $x = \sqrt[3]{\frac{1}{2}t^2 + t + 3C}$.

If $x = 1$ when $t = 1$, then $1 = \sqrt[3]{\frac{1}{2} + 3 + 3C}$, so $\frac{3}{2} + 3 + 3C = 1$, and hence $C = -7/6$.

2. (a) Direct integration yields $x = \frac{1}{4}t^4 - \frac{1}{2}t^2 + C$. (b) Direct integration yields $x = te^t - e^t - \frac{1}{2}t^2 + C$.

(c) $e^t dx = (t+1) dt$, so that $\int e^t dx = \int (t+1) dt$ and $e^t = \frac{1}{2}t^2 + t + C$. The solution is $x = \ln(\frac{1}{2}t^2 + t + C)$.

3. (a) $x = Cte^{-t}$; $C = 1$. (Separate: $dx/x = \{(1/t) - 1\} dt$. Integrate: $\ln|x| = \ln|t| - t + C_1$. Hence, $|x| = e^{\ln|t| - t + C_1} = e^{\ln|t|} e^{-t} e^{C_1} = C_2 |t| e^{-t} = Cte^{-t}$, where $C = \pm C_2 = \pm e^{C_1}$.) (b) $x = C\sqrt[3]{1+t^3}$; $C = 2$.

(c) $x = \sqrt{t^2 - 1}$. (General solution: $x^2 - t^2 = C$.) (d) $x = \frac{1 - e^{-2t}}{1 + e^{-2t}}$. (Hint: $e^{2t}\dot{x} = (x+1)^2$.)

4. $x = Ce^{-\int a(t) dt}$. If $a(t) = a + bc^t$, then $\int a(t) dt = at + (b/\ln c)c^t$. This implies $x = Ce^{-at} e^{-b/\ln c c^t} = C(e^{-a})^t (e^{-b/\ln c})^{c^t} = Cp^t q^{c^t}$, with $p = e^{-a}$ and $q = e^{-b/\ln c}$.

5. In both cases $\dot{N}$ depends on both N and t . (For instance, in Fig. A in the text, $N(t_1) = N(t_2)$, but $\dot{N}(t_1) \neq \dot{N}(t_2)$.)

6. (a) $x = Cr^a$ (b) $x = Crt^b e^{at}$ (c) $x = Cht^b/(1 - aCr^a)$

7. (a) Separable. Let $f(t) = 1$ and $g(x) = x^2 - 1$. (b) Separable, because $xt + t = t(x+1)$.

(c) Not separable. $xt + t^2$ cannot be written in the form $f(t)g(x)$. (Note that $t(x+t)$ does not count as a separation because both factors would then depend on t .) (d) Separable, because $e^{x-t} = e^x e^{-t}$.

(e) Not separable. It is impossible to write $\sqrt{t^2 + x}$ in the form $f(t)g(x)$. (f) Not separable in general. (Looks simple, but no method is known for solving this equation, except numerically or in special cases.)

8. (a) $K = \left[\frac{A\alpha_0^a a^b}{\alpha v + \varepsilon} (1 - b + c)e^{(a+b+c)t} + C \right]^{-1/(1-b+c)}$ (b) $[\alpha x - \beta]^{(\beta/\alpha)} |x - a|^{-\alpha} = K e^{(\alpha a - \beta)x}$

9. (a) $K/L = \left[K_0^a/L_0^a e^{a\lambda t} + (sA/\lambda)(1 - e^{-a\lambda t}) \right]^{1/a} \rightarrow (sA/\lambda)^{1/a}$ and $X/L = A(K/L)^{1-a} \rightarrow A(sA/\lambda)^{(1-a)/a}$ as $t \rightarrow \infty$. (b) $\dot{K} = sAb^a(t+a)^{pa} K^{1-a}$, so $K(t) = \left[K_0^a + sAb^a \left((t+a)^{pa+1} - a^{pa+1} \right) / (pa+1) \right]^{1/a}$. Hence, $K/L = \left[K_0^a/b^a(t+a)^{pa} + (sA/\lambda)(pa+1) \left((t+a) - a^{pa+1}/(t+a)^{pa} \right) \right]^{1/a} \rightarrow \infty$ as $t \rightarrow \infty$.

10. Using the given identity, (*) implies $\int (1/y + \alpha y^{a-1}/(1 - \alpha y^a)) dy = \int dx/x$. Integration (with $x > 0$, $y > 0$) yields $\ln y - (1/\alpha) \ln |1 - \alpha y^a| = \ln x + C_1$. Multiplying both sides by α leads to $\ln y^\alpha - \ln |1 - \alpha y^a| = \ln x^\alpha + C_1 \alpha$, or $\ln |y^\alpha/(1 - \alpha y^a)| = \ln e^{C_1 \alpha} x^\alpha$. Hence, $y^\alpha/(1 - \alpha y^a) = Cx^\alpha$, with $C = \pm e^{C_1 \alpha}$. Putting $\beta = 1/C$ and solving for y yields (**).

5.4

1. $x = Ce^{-t/2} + \frac{1}{2}$. The equilibrium state $x^* = \frac{1}{2}$ is stable. See Fig. A5.4.1.

2. Formula (3) yields immediately: (a) $x = Ce^{-t} + 10$ (b) $x = Ce^{3t} - 9$ (c) $x = Ce^{-5t/4} + 20$

3. Applying (4) with $a = -1$ and $b(t) = t$ yields $x = Ce^t + e^t \int t e^{-t} dt$. Integrating by parts, $\int t e^{-t} dt = -te^{-t} + \int e^{-t} dt = -te^{-t} - e^{-t}$, and so the solution is $x = Ce^t - t - 1$.

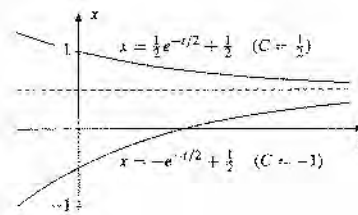


Figure A5.4.1

4. (a) $x = Ce^{3t} - 5/3$. For $C = 8/3$, the integral curve passes through $(0, 1)$.
 (b) $x = Ce^{-2t/3} - 8$. For $C = 9$, the integral curve passes through $(0, 1)$.
 (c) $x = Ce^{-2t} + e^{-2t} \int e^{2t} t^2 dt = Ce^{-2t} + \frac{1}{2}t^2 - \frac{1}{2}t + \frac{1}{2}$. For $C = 3/4$, the integral curve passes through $(0, 1)$.
5. $dx/dt = b - ax$, so $\int dx/(b - ax) = \int dt + A$, or $(-1/a) \ln |b - ax| = t + A$, etc.
6. (a) $\dot{x} + (2/t)x = -1$. Apply (6) with $a(t) = 2/t$ and $b = -1$. Then $\int a(t) dt = \int (2/t) dt = 2 \ln |t| = \ln |t|^2 = \ln t^2$, and so $\exp(\int a(t) dt) = \exp(\ln t^2) = t^2$. Then $x = (1/t^2)[C + \int t^2(-1) dt] = Ct^{-2} - \frac{1}{2}t$.
 (b) $x = Ct + t^2$ (c) $x = C\sqrt{t^2 - 1} + t^2 - 1$ (d) $x = Ct^2 + 2a^2/3t$
7. Clearly, $\dot{x}(0) = 0$. Moreover, $\ddot{x}(t) = 2x(t) + 2t\dot{x}(t) + 1 + t^2 + 2t^2$. But then $\ddot{x}(0) = 1$, and $x = 0$ is a local minimum point. (It is not necessary to solve the equation, but the solution is $x(t) = e^{t^2} - 1 - \frac{1}{2}t^2$.)
8. Substituting T for t_0 and x_T for x_0 , equation (7) yields the answer.
9. $\dot{x} = (a - 1)\alpha x - (a - 1)\beta$ with solution $x(t) = [x(0) - \beta/\alpha]e^{a(a-1)t} + \beta/\alpha$. $N(t) = [x(t)/A]^{1/(a-1)}$, $X(t) = A[N(t)]^a$. If $0 < a < 1$, then $x(t) \rightarrow \beta/\alpha$, $N(t) \rightarrow (\beta/\alpha A)^{1/(a-1)}$, and $X(t) \rightarrow A(\beta/\alpha A)^{a/(a-1)}$ as $t \rightarrow \infty$.
10. (a) $x(t) = X(t)/N(t)$ increases with t if $\alpha\sigma \geq \rho$. When $\sigma = 0.3$ and $\rho = 0.03$, this implies that $\alpha \geq 0.1$ ($= 10\%$).
 (b) See SM. (c) Foreign aid must grow faster than the population.

5

1. Separating the variables, $3x^2 dx = -2t dt$, so $\int 3x^2 dx = -2 \int t dt$, which gives $x^3 + t^2 = C$, or $x = \sqrt[3]{C - t^2}$. Alternatively, with $f(t, x) = 2t$ and $g(t, x) = 3x^2$, we have $f'_x = g'_t = 0$ so the equation is exact. From (8), $h(t, x) = \int_0^x 2t d\tau + \int_0^x 3\tau^2 d\tau = t^2 + x^3 - t_0^2 - x_0^3$, so the solution is given implicitly by $t^2 + x^3 - t_0^2 - x_0^3 = C_1$, i.e. $t^2 + x^3 = C$.
2. $x = -\frac{1}{2}t + \sqrt{\frac{1}{4}t^2 + C}$. (We are in case II, with $\beta(x) = x$.)

6

1. (a) Substituting $z = x^{-1}$ leads to $\dot{z} - (2/t)z = -1$ whose solution is $z = Ct^2 + t$. Thus $x = (Ct^2 + t)^{-1}$.
 (b) $x = (Ce^{2t} - e^t)^{-2}$ (c) $x = (1 + \ln t + Ct)^{-1}$.
2. With $x = w/t$, $\dot{x} = (t\dot{w} - w)/t^2$. Inserted into the given equation this gives $t(1 + w)\dot{w} = w$, which is separable: $\int (1/w + 1) dw = \int (1/t) dt$, so $\ln |w| + w = \ln t + C$, or $\ln |tx| + w = \ln t + C$. With $t > 0$, we get $\ln t + \ln |x| + tx = \ln t + C$, so $\ln |x| + tx = C$. In addition, $x(t) = 0$ is a solution.
3. $K = [Ce^{-\alpha\delta t(1-b)} + \alpha A n_0^a(1-b)e^{(\alpha\alpha+\delta)t}]/(\alpha v + \varepsilon + \alpha\delta(1-b))^{1/(1-b)}$ 4. $K = [Ce^{2t/(1-\gamma)} - \gamma b/\gamma_2]^{1/(1-\gamma)}$
5. (a) $x = 0$ is a solution. With $x = tz$, $\dot{x} = z + t\dot{z}$, which inserted into the equation yields $(*) \dot{z} = -f(t)z^2$.
 (b) Equation $(*)$ reduces to $\dot{z} = -t^3 z^2/(t^4 + 2)$, which is separable. The general solution: $x = 4t/(\ln(2 + t^4) + C)$. $C = 4 - \ln 3$ gives the solution through $(1, 1)$.
6. (a) With $x = zt$, $\dot{x} = \dot{z}t + z$, which leads to $t\dot{z} = g(z) - z$. (b) Write the equation as $\dot{z} = \frac{1}{2}(x/z) + \frac{1}{3}(x/z)^{-2}$, which leads to the equation $t\dot{z} = \frac{1}{2}z^{-2} - \frac{2}{3}z$, or $[z^2/(1 - 2z^3)]dz = 3(1/t)dt$. Introducing the new variable $u = 1 - 2z^3$, leads to $z = \sqrt[3]{\frac{1}{2} + C/t^2}$, and finally $x = tz = \sqrt[3]{\frac{1}{2}t^3 + Ct}$.
7. $x = (At^2 - t)/(At^2 + 1)$.
8. If $x = u + 1/z$, then $\dot{x} = \dot{u} - \dot{z}/z^2$. If $u = u(t)$ is a particular solution, the equation is converted into the linear form $\dot{z} + [Q(t) + 2u(t)R(t)]z = -R(t)$. For the equation $t\dot{x} = x - (x - t)^2$ and the particular solution $u = t$, we get $t\dot{z} + z = 1$, whose solution is $tz = C + t$. Thus $x = t + t/(t + C)$ is the general solution.

7

1. (a) $x = t$ is unstable. See Fig. A5.7.1(a). (b) $x = 12$ is stable. See Fig. A5.7.1(b). (c) $x = -3$ is stable; $x = 3$ is unstable. See Fig. A5.7.1(c).

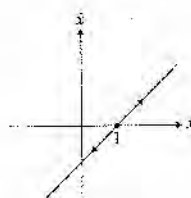


Figure A5.7.1(a)

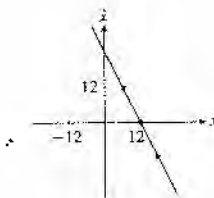


Figure A5.7.1(b)

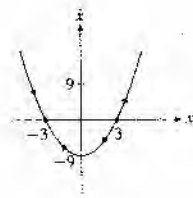


Figure A5.7.1(c)

2. (a) $\dot{x} = (x - 1)(x + 1)^2$. Here $x = 1$ is unstable, whereas $x = -1$ is neither stable nor unstable. (It is stable on the right, but unstable on the left.) (b) No equilibrium states. (c) The only equilibrium state $x = 0$ is unstable.
3. (a) $x(t) = (1 + Ae^t)/(1 - Ae^t)$. See SM. See Fig. A5.7.3(a) for some integral curves. (b) $x = -1$ is stable; $x = 1$ is unstable. See Fig. A5.7.3(b).

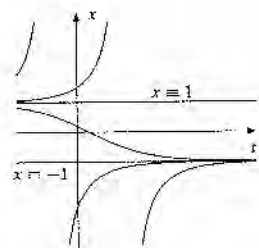


Figure A5.7.3(a)

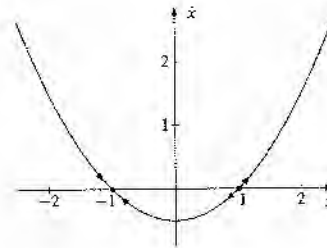


Figure A5.7.3(b)

4. (a) $\partial k^*/\partial s = f(k^*)/[\lambda - sf'(k^*)] > 0$ and $\partial k^*/\partial \lambda = -k^*/[\lambda - sf'(k^*)] < 0$ when $\lambda > sf'(k^*)$.
 (b) $c = (Y - \bar{K})/L = (1 - s)Y/L = (1 - s)f(k)$. But $sf(k^*) = \lambda k^*$, so when $k = k^*$ we have $c = f(k^*) - \lambda k^*$.
 (c) $0 = \dot{k}/k = \dot{K}/K - \dot{L}/L = \dot{K}/K - \lambda$ in the stationary state.

5.8

1. $F(t, x) = 2x/t$ is not continuous at $t = 0$, so the conditions in Theorem 5.8.1 are not satisfied at $(0, 0)$.
2. $F(t, x) = t^2 + e^{-x^2}$ and $F'_x(t, x) = -2xe^{-x^2}$ are continuous everywhere. Let the set Γ be defined by $\Gamma = \{(t, x) : |t| \leq a, |x| \leq a(a^2 + 1)\}$. Then $M = \max_{(t, x) \in \Gamma} (t^2 + e^{-x^2}) = a^2 + 1$, $r = a$ and the conclusion follows.
3. $x(t) = 1 + \frac{1}{11}t + \frac{1}{22}t^2 + \dots = e^t$. $x_0(t) = 1$, $x_1(t) = 1 + \int_0^t ds = 1 + t$, $x_2(t) = 1 + \int_0^t (1 + s) ds = 1 + t + \frac{1}{2}t^2$, etc.
4. $x = 1/(1 + e^{-t})$ 5. See SM.

Chapter 6

6.1

1. (a) $\dot{x} = \int t dt = \frac{1}{2}t^2 + A$, so $x = \int (\frac{1}{2}t^2 + A) dt = \frac{1}{6}t^3 + At + B$ (b) $x = -\sin t + At + B$ (see Appendix B).
 (c) $x = e^t + \frac{1}{12}t^4 + At + B$

$x = \frac{1}{12}t^4 - \frac{1}{6}t^3 + 2t + 1$. ($\dot{x} = \frac{1}{3}t^3 - \frac{1}{2}t^2 + A$, $x = \frac{1}{12}t^4 - \frac{1}{6}t^3 + At + B$. $x(0) = 1$ gives $B = 1$, $\dot{x}(0) = 2$ gives $A = 2$.)

$x \approx Ae^t - \frac{1}{3}t^2 - t + B$ from Example 2. With $x(0) = 1$ and $x(1) = 2$ we get $A + B = 1$ and $Ae + B = 7/2$, so that $A = 5/(2e - 2)$ and $B = 1 - A$.

(a) $x = Ae^{-2t} + B + 4t$ (b) $x = Ae^{2t} + B + te^{2t}$ (c) $x = Ae^t + B - \frac{1}{3}t^3 - t^2 - 2t$

(a) Let $w(y) = u'(y)$. Then $-w'(y)/w(y) = \lambda$, so that $w'(y) = -\lambda w(y)$, which has the solution $w(y) = Ae^{-\lambda y}$. Integration gives $u(y) = \int w(y) dy = -(A/\lambda)e^{-\lambda y} + B$ if $\lambda \neq 0$. For $\lambda = 0$ we get $u(y) = Ay + B$.

(b) Let $w(y) = u'(y)$. Then $-yw'(y)/w(y) = k$, which is separable, with the solution $w(y) = Ay^{-k}$. Then for $k \neq 1$, $u(y) = \int Ay^{-k} dy = Ay^{1-k}/(1-k) + B$. For $k = 1$, $w(y) = A \ln y + B$.

(a) See SM. (b) (i) Solutions are given implicitly by $x^3 + Ax + B = 6t$. In addition, $x = C$ is a solution. (ii) $x = Be^{At}$.

(a) $u'_t = \alpha e^{\alpha t} e^{ax}$, $u'_{xx} = \alpha^2 e^{\alpha t} e^{ax}$, $u'_t = \alpha^2 e^{\alpha t} e^{ax}$, so $u'_{xx} = u'_t$.

(b) $u'_t = g'(y)t^{-1/2}$, $u'_{xx} = g''(y)t^{-1/2}$, and $u'_t = -\frac{1}{2}g'(y)t^{-3/2}$, so $g''(y)/g'(y) = -\frac{1}{2}y$. With $w(y) = g'(y)$ this gives $w'(y)/w(y) = -\frac{1}{2}y$, with the solution $\ln|w(y)| = -\frac{1}{4}y^2 + C$. Hence, $|w(y)| = e^{-\frac{1}{4}y^2 + C} = e^{-\frac{1}{4}y^2} e^C$, so $w(y) = Ae^{-\frac{1}{4}y^2}$, where $A = \pm e^C$. It follows that $u(y) = A \int e^{-\frac{1}{4}y^2} dy + B$.

1

(a) With $u_2 = te^t$, $\dot{u}_2 = e^t + te^t$, $\ddot{u}_2 = e^t + e^t + te^t = 2e^t + te^t$, so $\ddot{u}_2 - 2\dot{u}_2 + u_2 = 0$. In the same way we verify that $u_1 = e^t$ is a solution. If $u_2 = ku_1$ for all t , then $t = k$ for all t , which is absurd. Thus the general solution is $x(t) = Ae^t + Bte^t$. (b) One particular solution is $u^*(t) = 3$, so the general solution is $x(t) = Ae^t + Bte^t + 3$.

Easy verification by differentiating $\sin t$ and $\cos t$ twice. General solution: $x = A \sin t + B \cos t$.

(a) $x = Ae^{2t} + Be^{-3t}$ (b) $x = Ae^{2t} + Be^{-3t} - t - \frac{1}{6}$

By direct differentiation, $u_1 = e^{at}$ and $u_2 = e^{a/(1-a)t}$ are easily seen to be solutions. They are not proportional, so the general solution is $x(t) = Ae^{at} + Be^{a/(1-a)t}$, where A and B are arbitrary constants.

General solution: $x = A(t+a)^{-1} + B(t+b)^{-1}$.

(a) $x = Ae^{\sqrt{3}t} + Be^{-\sqrt{3}t}$ (b) $x = e^{-2t}(A \cos 2t + B \sin 2t)$ (c) $x = A + Be^{-3t/3}$

(d) $x \approx e^{-t/2}(A + Bt)$ (e) $x = Ae^{-3t} + Be^{2t} - 4/3$ (f) $x = Ae^{-t} + Be^{-2t} + (1/42)e^{5t}$

(a) $x = Ae^t + Be^{-t} - \frac{1}{2} \sin t$ (b) $x = Ae^t + Be^{-t} - \frac{1}{2}te^{-t}$ (c) $x = Ae^{5t} + Bte^{5t} + \frac{2}{13}t + \frac{3}{135}$

(a) $x = -(6+t)e^{-t} + t^2 - 4t + 6$ (b) $x = \frac{1}{2} \sin 2t + (\pi/2 + 1/4) \cos 2t + t + 1/4$

$u^* = kt + L_0 + [\beta + \alpha(1-\beta)]k/\delta^*$ is a particular solution. Oscillations if $\gamma^2[\beta + \alpha(1-\beta)]^2 + 4\gamma\delta^* < 0$.

Using formula (B.1.8), $C \cos(\beta t + D) = C \cos \beta t \cos D - C \sin \beta t \sin D = A \cos \beta t + B \sin \beta t$ provided that $A = C \cos D$ and $B = -C \sin D$. This requires $C = \sqrt{A^2 + B^2}$ and $D = \tan^{-1}(-B/A)$.

If $x = ue^{rt}$, then $\dot{x} = e^{rt}(\dot{u} + ru)$ and $\ddot{x} = e^{rt}[\ddot{u} + (2r+a)\dot{u} + (r^2 + ar + b)u] = e^{rt}\ddot{u}$ because $r = -a/2$ and r is a root in the characteristic equation. Thus $x = ue^{rt}$ satisfies the given equation provided $\ddot{u} = 0$. The last equation has the general solution $u = At + B$, so the conclusion follows.

When $(a-1)^2 = 4b$, $u_1 = t^{\frac{1}{2}(1-a)}$ and $u_2 = (\ln t)t^{\frac{1}{2}(1-a)}$ are not proportional and satisfy equation (8). Let us prove that u_2 satisfies the equation: We get $\dot{u}_2 = [1 + \frac{1}{2}(1-a) \ln t]t^{\frac{1}{2}(1-a)-1}$ and $\ddot{u}_2 = -\frac{1}{2}(1-a^2)(\ln t)t^{\frac{1}{2}(1-a)-2}$, and it follows easily that $t^2\ddot{u}_2 + ta\dot{u}_2 + bu_2 = 0$.

(a) $x = At^{-1} + Bt^{-3}$ (b) $x = At + Bt^3 - t^3$

9. Characteristic equation: $r^2 + 2ar - 3a^2 = (r-a)(r+3a) = 0$. General solution of the homogeneous equation: For $a \neq 0$, $x = C_1e^{at} + C_2e^{-3at}$; for $a = 0$, $x = P + Qt$. Particular solution when $a \neq 0$: $u^* = Re^{bt}$ with $R = 100/(b^2 + 2ab - 3a^2)$ for $b \neq a$ and $b \neq -3a$. For $a \neq 0$ and $b = a$ and $b = -3a$, $u^* = Ste^{bt}$ with $S = 25/a$ and $S = -25/a$, respectively. For $a = 0$ and $b = 0$: $u^* = 50t^2$, and for $a = 0$ and $b \neq 0$: $u^* = (100/b^2)e^{bt}$.

10. $\ddot{p} + \lambda^2 p = a(d_0 - s_0)$ where $\lambda = [a(s_1 - d_1)]^{1/2}$. Solution: $p = A \cos \lambda t + B \sin \lambda t + (d_0 - s_0)/(s_1 - d_1)$.

6.4

1. It is clear from the answers to Problem 6.3.1 that (b), (d), and (f) are globally asymptotically stable, the others are not. This is easily seen to agree with (6.4.3).

2. According to (6.4.3), the equation is globally asymptotically stable iff $1 - a^2 > 0$ and $2a > 0$, i.e. iff $0 < a < 1$.

3. For $\lambda = \gamma(a - \alpha) > 0$ the solution is $p(t) = Ae^{rt} + Be^{-rt} - k/r^2$, where $r = \sqrt{\lambda}$; for $\lambda = 0$ the solution is $p(t) = At + B + \frac{1}{2}kt^2$; for $\lambda < 0$ the solution is $p(t) = A \cos \sqrt{-\lambda}t + B \sin \sqrt{-\lambda}t - k/\lambda$. The equation is not stable for any values of the constants.

6.5

1. (a) $x = Ae^t + Be^{-t} - t$, $y = Ae^t - Be^{-t} - 1$. (Differentiating the first equation w.r.t. t gives $\dot{x} = \dot{y}$, and substituting from the second equation we get $\ddot{x} = x + t$. The methods of Section 6.3 give the solution for x . Then from the first equation we get $y = \dot{x}$.) (b) $x = Ae^{\sqrt{2}t} + Be^{-\sqrt{2}t}$, $y = A(\sqrt{2} - 1)e^{\sqrt{2}t} - B(\sqrt{2} + 1)e^{-\sqrt{2}t}$. ($\ddot{x} - 2\dot{x} = 0$.) (c) $x = Ae^{-t} + Be^{2t} + t - \frac{2}{3}$, $y = Ae^{-t} - \frac{1}{3}Be^{3t} + \frac{2}{3}t - \frac{7}{9}$. ($\ddot{x} - 2\dot{x} - 3x = -3t$.)

2. (a) $x = \frac{a}{a+b}e^{(a+b)t} - \frac{a-b}{2(a+b)}$, $y = \frac{b}{a+b}e^{(a+b)t} + \frac{a-b}{2(a+b)}$ (b) $x = t + 1$, $y = -2t^2 - 2t + 1$
(c) $x = -\frac{1}{2} \cos 2t + \cos t - 1/2$, $y = -\frac{1}{2} \sin 2t + \sin t$

3. $x = Ae^{(1+\sqrt{2})t} + Be^{(1-\sqrt{2})t}$, $p = A\sqrt{2}e^{(\sqrt{2}-1)t} - B\sqrt{2}e^{(-\sqrt{2}-1)t}$. ($\ddot{x} - 2\dot{x} - x = 0$.)

4. $\pi = Ae^{r_1 t} + Be^{r_2 t}$, $\sigma = (\alpha - r_1)Ae^{r_1 t} + (\alpha - r_2)Be^{r_2 t}$, where $r_{1,2} = \frac{1}{2}(\alpha - 1/\beta) \pm \frac{1}{2}\sqrt{(\alpha - 1/\beta)^2 - 4}$.

5. $dy/dx = \dot{y}/\dot{x} = x/y$, a separable equation whose solution curve through $x = 1$, $y = \sqrt{2}$ is $y^2 = 1 + x^2$. Then $\dot{x} \approx t$, whose solution through $t = 1$, $x = 1$ is $x = \frac{1}{2}(1 + t^2)$, implying that $y = \sqrt{1 + \frac{1}{4}(1 + t^2)^2}$.

6.6

1. (a) $A = \begin{pmatrix} 1 & -8 \\ 2 & -4 \end{pmatrix}$. Hence, $\text{tr}(A) = -3$ and $|A| = 12$, so the system is globally asymptotically stable.
(b) The trace is equal to 0, so the system is not globally asymptotically stable.
(c) The trace is equal to -3 and the determinant is 8, so the system is globally asymptotically stable.

2. (a) $A = \begin{pmatrix} a & -1 \\ 1 & a \end{pmatrix}$. $\text{tr}(A) = 2a$ and $|A| = a^2 + 1$, so the system is globally asymptotically stable iff $a < 0$.
(b) $A = \begin{pmatrix} a & 4-2a \\ 1 & 2a \end{pmatrix}$. $\text{tr}(A) = 3a$ and $|A| = 2a^2 + 2a - 4 = 2(a-1)(a+2)$, so the system is globally asymptotically stable iff $a < -2$. (Use a sign diagram.)

3. (i) $x = Ae^t + Be^{-t} + 5$, $y = -Be^{-t} + 2$. (ii) See SM.

1. (a) $x = Ae^{(a-2)t} + Be^{(a+2)t} + \frac{2\beta - a\alpha}{a^2 - 4}$, $y = -Ae^{(a-2)t} + Be^{(a+2)t} + \frac{2\alpha - a\beta}{a^2 - 4}$.
(b) $(x^*, y^*) = \left(\frac{2\beta - a\alpha}{a^2 - 4}, \frac{2\alpha - a\beta}{a^2 - 4}\right)$ is globally asymptotically stable iff $a < -2$.
(c) $x = e^{-3t} + 2$, $y = -e^{-3t} + 3$.

7

1. In all three cases $(0, 0)$ is the only equilibrium point. (a) $x = Ae^t + Be^{-t}$, $y = Ae^t - Be^{-t}$. See Fig. A6.7.1(a).
(b) See Problem 6.5.1(b). See Fig. A6.7.1(b). (c) $x = Ae^{-t} + Be^{-3t}$, $y = \frac{1}{2}Ae^{-t} + Be^{-3t}$. See Fig. A6.7.1(c).

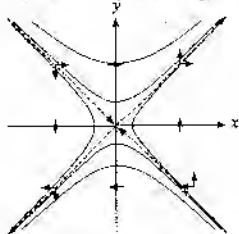


Figure A6.7.1(a)

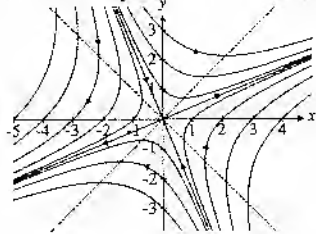


Figure A6.7.1(b)

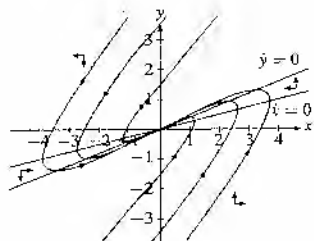


Figure A6.7.1(c)

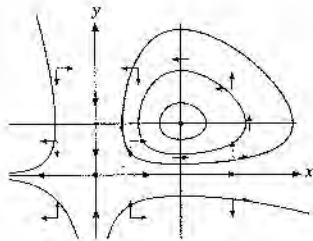


Figure A6.7.2

1. See Fig. A6.7.2, which also shows some solution curves. Note that the solution curves that start in the first quadrant cannot escape from it, and appear to be closed curves. (See Example 7.5.2.)
2. See Fig. A6.7.3(a). (The nullclines are $C = AK^2 = 2\sqrt{K}$ and $K = (\alpha A/r)^{1/(1-\alpha)} = 400$.) A more detailed phase diagram with some solution curves is given in Fig. A6.7.3(b).

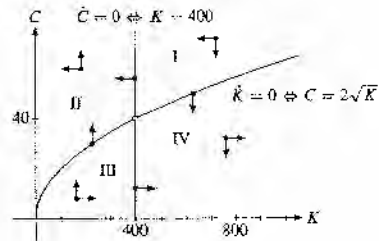


Figure A6.7.3(a)

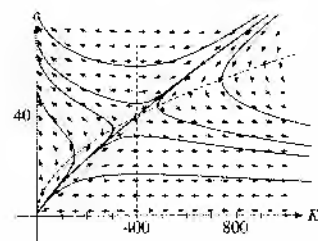


Figure A6.7.3(b)

4. (a) The equilibrium is $(0, 0)$, which is not stable. See Figures A6.7.4(a) and (b).
(b) $x(t) = -e^{-t}$, $y(t) = (z(t))^{-1}$, where $z(t) = e^{e^{-t}}(e^{-1} + \int_0^t e^{-\tau} d\tau)$. Here $(x(t), y(t)) \rightarrow (0, 0)$ as $t \rightarrow \infty$.

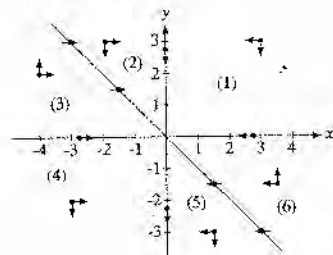


Figure A6.7.4(a)

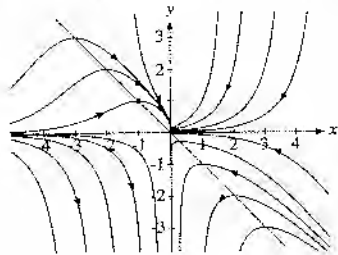


Figure A6.7.4(b)

5. (a) See Fig. A6.7.5(a). (b) See Fig. A6.7.5(b).

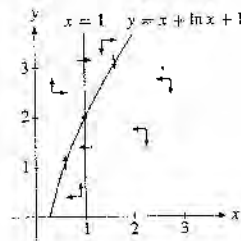


Figure A6.7.5(a)

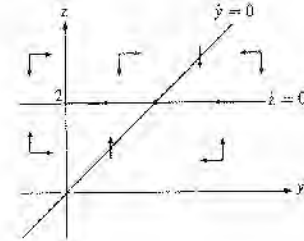


Figure A6.7.5(b)

6. $(x(t), y(t)) = (e^{-t}, e^{\frac{1}{2}(e^{-2t}-1)}) \rightarrow (0, e^{-\frac{1}{2}})$ as $t \rightarrow \infty$.

6.8

1. $A(x, y) = \begin{pmatrix} f_1'(x, y) & f_2'(x, y) \\ g_1'(x, y) & g_2'(x, y) \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ -2x+8 & -2 \end{pmatrix}$, so $A(6, 6) = \begin{pmatrix} -1 & 1 \\ -4 & -2 \end{pmatrix}$.

Since $\text{tr}(A(6, 6)) = -3$, and $|A(6, 6)| = 6$, the equilibrium point $(6, 6)$ is locally asymptotically stable.

2. (a) $A(0, 0) = \begin{pmatrix} -1 & 0 \\ 2 & -2 \end{pmatrix}$, $\text{tr}(A) = -3$, and $|A| = 2$, so $(0, 0)$ is locally asymptotically stable.
(b) $A(1, 1) = \begin{pmatrix} 4 & -2 \\ 1 & 0 \end{pmatrix}$, $\text{tr}(A) > 0$, so $(1, 1)$ is not locally asymptotically stable.
(c) $A(0, 0) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, so $\text{tr}(A) = 0$, and Theorem 6.8.1 does not apply.
(d) $A(0, 0) = \begin{pmatrix} 2 & 8 \\ -1 & -3 \end{pmatrix}$, $\text{tr}(A) = -1$ and $|A| = 2$, so $(0, 0)$ is locally asymptotically stable.

3. $\text{tr}(A) = -k < 0$, $|A| = m^2 > 0$, and $f_2'g_1' = -m^2 \neq 0$, so by Olech's theorem, $(0, 0)$ is globally asymptotically stable.

4. With $f(q, p) = a(p - c(q))$ and $g(q, p) = b(D(p) - q)$, the matrix A in Theorem 6.8.1, evaluated at (q^*, p^*) , is $A = \begin{pmatrix} -ac'(q^*) & a \\ -b & bD'(p^*) \end{pmatrix}$. If $D'(p^*) < 0$ and $c'(q^*) > 0$, then $\text{tr}(A) < 0$ and $|A| > 0$, so that (q^*, p^*) is locally asymptotically stable.

5. Direct application of Olech's theorem.

6. $(K^*, P^*) = \left(\left(\frac{1}{\delta} \right)^{1/(1-\alpha)}, \frac{1}{\gamma} \left(\frac{1}{\delta} \right)^{\delta/(1-\alpha)} \right)$. The matrix $A(K^*, P^*)$ is $\begin{pmatrix} -(1-\alpha)\delta & 0 \\ \beta(K^*)^{\beta-1} & -\gamma \end{pmatrix}$.

Thus $\text{tr}(A) = -(1-\alpha)\delta - \gamma < 0$ and $|A| = (1-\alpha)\delta\gamma > 0$, so (K^*, P^*) is locally asymptotically stable.

$K(t) = \left[(K_0^{1-\alpha} - s/\delta)e^{-s(1-\alpha)t} + s/\delta \right]^{1/(1-\alpha)} \rightarrow (s/\delta)^{1/(1-\alpha)} = K^*$ as $t \rightarrow \infty$.

9.

1. (a) $A = \begin{pmatrix} -1/2 & 1 \\ 0 & 1 \end{pmatrix}$. Since $|A| = -1/2 < 0$, the equilibrium point $(4, 2)$ is a local saddle point.

The eigenvalues of A are $\lambda_1 = -1/2$ and $\lambda_2 = 1$. An eigenvector associated with $\lambda_1 = -1/2$ is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

(b) See Fig. A6.9.1. The solution curves that converge to the equilibrium point are given by $x(t) = (x(0) - 4)e^{-t/2} + 4$, $y = 2$. (One for $x(0) < 4$, one for $x(0) > 4$.)

2. The equilibrium (k^*, c^*) is defined by the equations $f'(k^*) = r + \delta$, $c^* = f(k^*) - \delta k^*$. It is a saddle point.

3. (a) $(x_0, y_0) = (4/3, 8/3)$. It is a (local) saddle point. (b) See Fig. A6.9.3.

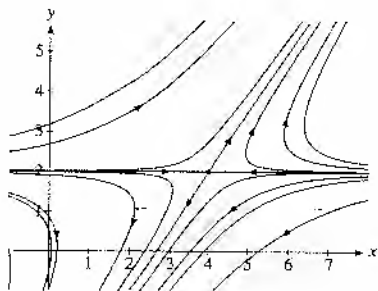


Figure A6.9.1

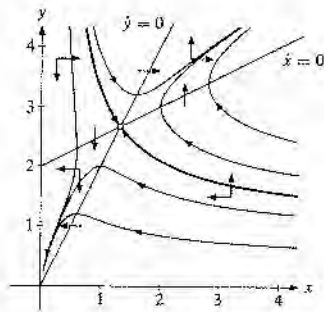


Figure A6.9.3

4. (a) $(9/4, 3/2)$ is locally asymptotically stable, $(9/4, -3/2)$ is a saddle point.
(b) See Figure A6.9.4.

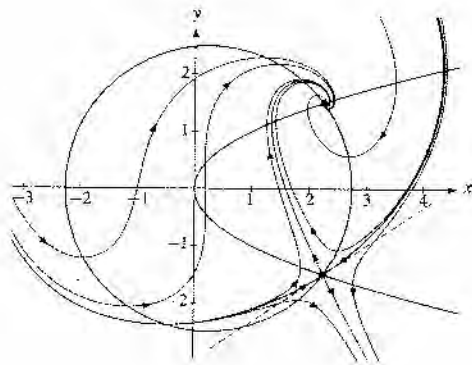


Figure A6.9.4

Chapter 7

7.1

1. $x = C_1 e^t + C_2 e^{-t} + C_3 e^{2t} + 5$

2. $x = Ae^t + Be^{-t} - \frac{1}{2}te^{-t}$. (The two required equations are here: (i) $\dot{C}_1(t)e^t + \dot{C}_2(t)e^{-t} = 0$, (ii) $\dot{C}_1(t)e^t - \dot{C}_2(t)e^{-t} = e^t$. From (i), $\dot{C}_2(t) = -\dot{C}_1(t)e^{2t}$, which inserted into (ii) yields $\dot{C}_1(t) = \frac{1}{2}e^{-2t}$, and thus $C_1(t) = -\frac{1}{4}e^{-2t} + A$. Then $\dot{C}_2(t) = -\frac{1}{2}$ and thus $C_2(t) = -\frac{1}{2}t + B_1$. The general solution is then $x(t) = C_1(t)e^t + C_2(t)e^{-t} = Ae^t + Be^{-t} - \frac{1}{2}te^{-t}$, where $B = B_1 + \frac{1}{4}$.)

3. $x = A \sin t + B \cos t + \sin t \int \frac{\cos t}{t} dt - \cos t \int \frac{\sin t}{t} dt$. (The integrals cannot be evaluated in terms of elementary functions.)

7.2

1. (a) $x = (C_1 + C_2 t + C_3 t^2)e^{-t} + 3$ (b) $x = C_1 e^{2t} + C_2 t e^{2t} + e^{-t/2}(C_3 \cos \frac{1}{2}\sqrt{3}t + C_4 \sin \frac{1}{2}\sqrt{3}t) + \frac{1}{2}t - \frac{1}{4}$

2. $x \approx (t-1)e^t + e^{-t}(t+1)^2$. (The characteristic polynomial is $r^3 - r^2 - r + 1 = (r-1)^2(r+1)$. Note that every particular solution must contain a term of the form At^2e^{-t} .)

3. $\ddot{K} - p\dot{K} + qK = 0$, where $p = \gamma_1\kappa + \gamma_2 + \mu$ and $q = (\gamma_1\kappa + \gamma_2)\mu - (\gamma_1\sigma + \gamma_3)\mu_0$. The characteristic equation is $r(r^2 - pr + q) = 0$, and $r^2 - pr + q = 0$ has two different real roots not equal to 0 provided $p^2 > 4q$ and $q \neq 0$.

7.3

1. Using (3)(c), $a_1 = 3 > 0$, $a_3 = 1 > 0$, and $a_1a_2 - a_3 = 8 > 0$. The equation is globally asymptotically stable.

2. $a_1 = 4 > 0$, $a_3 = 2 > 0$, and $a_1a_2 - a_3 = 18 > 0$. Hence the equation is globally asymptotically stable. The general solution is $x = C_1 e^{-t} + C_2 t e^{-t} + C_3 e^{-2t}$, which approaches 0 as $t \rightarrow \infty$.

7.4

1. $x_1 = Ae^{-t} + Be^{-2t} + Ce^{2t}$, $x_2 = Ae^{-t} - Be^{-2t} + Ce^{2t}$, $x_3 = -Ae^{-t} + 2Ce^{2t}$. (We find that $\ddot{x}_1 + \dot{x}_1 - 4x_1 - 4x_3 = 0$, with characteristic polynomial $r^3 + r^2 - 4r - 4 = (r+1)(r^2 - 4)$.)

7.5

1. (a) $V(x, y) = x^2 + y^2$ is positive definite and $\dot{V} = 2x\dot{x} + 2y\dot{y} = 2x(-y - x^3) + 2y(x - y^3) = -2x^4 - 2y^4 < 0$ for all $(x, y) \neq (0, 0)$. According to Theorem 7.5.2 the equilibrium point $(0, 0)$ is locally asymptotically stable. Since $V(x, y) = x^2 + y^2 \rightarrow \infty$ as $\|(x, y) - (0, 0)\| = \|(x, y)\| = \sqrt{x^2 + y^2} \rightarrow \infty$, the equilibrium point is globally asymptotically stable according to Theorem 7.5.3.

(b) $V(x, y) = 12x^2 + 12xy + 20y^2$ is easily seen to be positive definite and $\dot{V} = 24x\dot{x} + 12\dot{x}y + 12x\dot{y} + 40y\dot{y} = 24x(-\frac{7}{4}x + \frac{1}{4}y) + 12(-\frac{7}{4}x + \frac{1}{4}y)y + 12x(\frac{3}{4}x - \frac{5}{4}y) + 40y(\frac{3}{4}x - \frac{5}{4}y) = -33x^2 - 47y^2 < 0$ for all $(x, y) \neq (0, 0)$. According to Theorem 7.5.2 the equilibrium point $(0, 0)$ is locally asymptotically stable. In fact, Theorem 7.5.3 implies that $(0, 0)$ is globally asymptotically stable. ($V(x, y) = (2x + 3y)^2 + 11y^2$.)

(c) The matrix in Theorem 7.5.1 is here $A = \begin{pmatrix} -\frac{7}{4} & \frac{1}{4} \\ \frac{3}{4} & -\frac{5}{4} \end{pmatrix}$. The eigenvalues satisfy $\begin{vmatrix} -\frac{7}{4} - \lambda & \frac{1}{4} \\ \frac{3}{4} & -\frac{5}{4} - \lambda \end{vmatrix} = 0$, i.e. $\lambda^2 + 3\lambda + 2 = 0$, and we see that both eigenvalues have negative real part. (Using Theorem 6.8.1 we can use the fact that the trace of A is negative ($= -3$) and the determinant is positive ($= 2$).

2. The equilibrium point is $p_0 = b/c$. $V(p) > 0$ for $p \neq p_0$. Moreover, $\dot{V}(p(t)) = -(2a/(pc))(b - pc)^2 < 0$ for all $p \neq p_0$, and we conclude that $p_0 = b/c$ is a locally asymptotically stable equilibrium point.

3. $V(x)$ is positive definite. Moreover, $\dot{V}(x) = \sum_{i=1}^n -u_i'(x)u_i'(x) = -(\nabla u(x))^2 < 0$ for $x \neq 0$, and we conclude that 0 is locally asymptotically stable.

4. See SM. 5. (x_0, y_0) is locally asymptotically stable.

7

(a) By integration, $z = \frac{1}{4}x^4 + \frac{1}{2}x^2y^2 - e^t y + \varphi(y)$, where φ is an arbitrary differentiable function.

(b) $z = 3x + \varphi(y - 2x)$, where φ is an arbitrary differentiable function. (c) $z = \frac{x}{1 + x\varphi(1/y - 1/x)}$

The equations in (3) are $dy/dx = 1/y$ and $dz/dx = z/2y^2$. The first equation yields $\frac{1}{2}y^2 = x + C_1$, and then the second equation reduces to $dz/dx = z/4(x + C_1)$. This is a separable differential equation with solution $\ln z^4 = \ln(x + C_1) + C_2 = \ln \frac{1}{2}y^2 + C_2$, or $z = C_3\sqrt{y}$ for a new constant C_3 . We can write these the solutions as $y^2 - 2x = C_4$, $z/\sqrt{y} = C_3$. The solution of the given partial differential equation is then given implicitly by $\Phi(y^2 - 2x, z/\sqrt{y}) = 0$, and then $z = \sqrt{y}\varphi(y^2 - 2x)$.

(a) $z = x + \varphi(xy)$, where φ is an arbitrary differentiable function. (b) The condition $f(x, 1) = x^2$ implies that $x + \varphi(x) = x^2$. Thus $\varphi(x) = -x + x^2$ for all x , and hence $f(x, y) = x - xy + (xy)^2$.

Using the definition of elasticity, we get $xz'_1 - yz'_2 = xz$. The solution is $\ln z = x + \varphi(xy)$, or $z = e^x \psi(xy)$, where $\psi(xy) = e^{\varphi(xy)}$.

The equations in (3) are $dx_2/dx_1 = -f(x_1)$ and $dU/dx_1 = 0$, with the solutions $x_2 = -F(x_1) + C_1$ and $U = C_2$, where F is an indefinite integral of f and C_1 and C_2 are constants. The solutions of the equation are therefore given by $\Phi(x_2 + F(x_1), U) = 0$, and thus $U = \varphi(x_2 + F(x_1)) = \varphi(x_2 + \int f(x_1) dx_1)$.

The equations in (3) reduce to $dy/dx = y/x$ and $dz/dx = nz/x$, with the solutions $y/x = C_1$, $z/x^n = C_2$. The general solution of the equation is therefore $\Phi(y/x, z/x^n) = 0$, or $z = x^n \varphi(y/x)$. We see that z as a function of x and y is homogeneous of degree n .

See SM.

The equations in (8) reduce to $dx_2/dx_1 = -f(x_1, x_2)$, $dx_3/dx_1 = 0$, and $dz/dx_1 = 0$. The first equation has a solution of the form $g(x_1, x_2) = C_1$, and the others, $x_3 = C_2$, $z = C_3$. Thus the general solution is of the form $H(g(x_1, x_2, x_3), x_3, z) = 0$, or $z = G(g(x_1, x_2, x_3), x_3)$.

Chapter 8

1

(i) $J(x) = \int_0^1 [(e^2 - 1)^2 t^2 + (e^2 - 1)^2] dt = (e^2 - 1)^2 \int_0^1 (\frac{1}{3}t^3 + t) dt = (4/3)(e^2 - 1)^2$

(ii) $J(x) = \int_0^1 [(e^{1+t} - e^{1-t})^2 + (e^{1+t} + e^{1-t})^2] dt = e^4 - 1$. We find that $e^4 - 1 < (4/3)(e^2 - 1)^2$.

2

With $F(t, x, \dot{x}) = 4xt - \dot{x}^2$, $F'_t = 4t$ and $F'_x = -2\dot{x}$, so the Euler equation is $4t - (d/dt)(-2\dot{x}) = 0$, or $\ddot{x} = -2t$. The general solution is $x = -\frac{1}{3}t^3 + At + B$. The boundary conditions yield $A = -1$ and $B = 2$. The function F is (for t fixed) concave in $(x, \dot{x})$, so we have found the solution.

Here $F(t, x, \dot{x}) = t\dot{x} + \dot{x}^2$, $F'_t = 0$ and $F'_x = t + 2\dot{x}$, so the Euler equation is $-(d/dt)(t + 2\dot{x}) = 0$. Hence $t + 2\dot{x} = A$ for some constant A , i.e. $\dot{x} = \frac{1}{2}A - \frac{1}{2}t$, and then $x = \frac{1}{2}At - \frac{1}{4}t^2 + B$. The boundary conditions yield $A = -3/2$, $B = 1$. The function F is convex in $(x, \dot{x})$, so we have found the solution.

(a) $\ddot{x} - x = e^t$ (b) $\ddot{x} - ax + a = 0$ (c) $\ddot{x} - ax + (a - 2)x = 0$ (d) $\ddot{x} + (1/t)\dot{x} = 1$

Euler equation: $\ddot{x} = 0$. Solution: $x = t + 1$.

Euler equation: $\ddot{x} - \frac{1}{x}x = \frac{1}{2}t$. General solution: $x(t) = Ae^{\frac{1}{2}\sqrt{2}t} + Be^{-\frac{1}{2}\sqrt{2}t} - t$. With $x(0) = 0$ and $x(1) = 1$, $A = -B = 2/(e^{\frac{1}{2}\sqrt{2}} - e^{-\frac{1}{2}\sqrt{2}})$. With $F(t, x, \dot{x}) = x^2 + tx + tx\dot{x} + \dot{x}^2$, we have $F'_{xx} = 2 > 0$ and $F'_{xx}F'_{xx} - (F'_{x\dot{x}})^2 = 4 - t^2 > 0$ for $t \in [0, 1]$. F is (strictly) convex, so we have found the solution.

6. $x(t) = x_0 + \frac{x_1 - x_0}{t_1 - t_0}(t - t_0)$, the straight line through the two points.

7. Euler equation: $t^2\ddot{x} + 2t\dot{x} - \frac{1}{2}x = 0$. General solution: $x(t) = At^{a_1} + Bt^{a_2}$, where $a_{1,2} = \frac{1}{2}(-1 \pm \sqrt{3})$. (See equation (6.3.8).) The boundary conditions yield: $A = -B = (2^{a_1} - 2^{a_2})^{-1}$. $F(t, x, \dot{x}) = x^2 + tx\dot{x} + t^2\dot{x}^2$ is convex, so we have found the solution.

8. Let $V = [N(\dot{x}) + \dot{x}f(x)]e^{-rt}$. Then the Euler equation is $\dot{x}f'(x)e^{-rt} - \frac{d}{dt}[N'(\dot{x}) + f(x)]e^{-rt} = 0$, i.e. $\dot{x}f'(x)e^{-rt} - \frac{d}{dt}[N'(\dot{x})]e^{-rt} + rN'(\dot{x})e^{-rt} - f'(x)\dot{x}e^{-rt} - f(x)(-r)e^{-rt} = 0$, which reduces to the equation given in the problem.

8.3

1. $x(t) = a(t - t^2)$ is admissible and makes the integral equal to $11a^2/30$, which tends to ∞ as $a \rightarrow \infty$. The conclusion follows.

2. (a) The Euler equation takes the form $(d/dt)[e^{-rt}U'(\bar{c} - \dot{x}e^{rt})] = 0$, so $U'(\bar{c} - \dot{x}e^{rt}) = Ke^{-rt}$ for some constant K . (b) $x = A + [B + t/v]e^{-rt}$, where A and B are determined by the equations $A + B = x_0$, $A + [B + T/v]e^{-rT} = 0$. This solves the problem because $U''(c) = -ve^{-rc} < 0$, so U is concave.

3. The Euler equation is $(d/dt)(t\dot{x}) = 0$, so $x = A + B \ln t$ for $t > 0$. When $a \in (0, 1)$ the only solution satisfying the boundary conditions is $x(t) = 1 - \ln t / \ln a$. The integrand is convex in $(x, \dot{x})$, so this is the solution. For $a = 0$ there are no solutions to the necessary conditions, so there is no optimal solution in this case.

4. (a) $\ddot{y} - (2/\sigma^2)\dot{y} + (1/\sigma^2)y = (\bar{z}/\sigma^2)(\dot{t}(t) - \sigma \dot{t}(t))$ (b) $y = Ae^{t/\sigma} + Bte^{t/\sigma} + \bar{z}_0 e^{at}/(1 - \alpha\sigma)$

8.4

1. $4\ddot{K} - 15\dot{K} + 14K = 0$. The solution is $K = Ae^{2t} + Be^{\frac{7}{2}t}$, where $A = P(K_T - K_0e^{\frac{7}{2}T})$, $B = P(K_0e^{2T} - K_T)$ with $P = (e^{2T} - e^{\frac{7}{2}T})^{-1}$.

2. (a) The Euler equation is $\ddot{x} - \frac{1}{10}\dot{x} = -\frac{1}{200}t$. The solution of the problem is $x = A + Be^{t/10} + \frac{1}{40}t^2 + \frac{1}{2}t$, where $A = -B = (\frac{1}{40}T^2 + \frac{1}{2}T - S)/(e^{T/10} - 1)$. (b) As in part (a), but with $A = -B = 12.5/(1 - e)$.

3. The Euler equation is $U'_C(f'_K - \delta) + (d/dt)U'_C = 0$, or $U'_C(f'_K - \delta) + U'_{CC}\dot{C} + U'_{CK} = 0$. We deduce that $\dot{C}/C = [(U'_{CK}/U'_C) - (f'_K - \delta)]/\bar{\omega}$, where $\bar{\omega}$ is the elasticity w.r.t. consumption of the marginal utility of consumption.

4. (a) $D + (p - b'(D))\partial D/\partial p - (d/dt)[(p - b'(D))\partial D/\partial \dot{p}] = 0$

(b) $p = C_1e^{\lambda t} + C_2e^{-\lambda t} + k$, where $\lambda = \frac{\sqrt{A^2 - A/\alpha}}{B}$, $k = \frac{\beta A + 2\alpha AC - C}{2A(1 - \alpha A)}$, and C_1 and C_2 are determined by the equations $p(0) = C_1 + C_2 + k$, $p(T) = C_1e^{\lambda T} + C_2e^{-\lambda T} + k$, where $p(0)$ and $p(T)$ are the given values of $p(t)$ for $t = 0$ and for $t = T$.

8.5

1. $F(t, x, \dot{x}) = t\dot{x} + \dot{x}^2$ is convex in $(x, \dot{x})$. Euler equation: $-\frac{d}{dt}(t + 2\dot{x}) = 0$, and so $t + 2\dot{x} = A$ for some constant A . Integrating with $x(0) = 1$ yields $x(t) = -\frac{1}{4}t^2 + \frac{1}{2}At + 1$. Case (i): Condition (2) is $(F'_x)_{t=1} = 1 + 2\dot{x}(1) = 0$, which reduces to $A = 0$, and the solution is $x = -\frac{1}{4}t^2 + 1$. Case (ii): Condition (3) is here $(F'_x)_{t=1} = 1 + 2\dot{x}(1) \geq 0$ ($= 0$ if $x(1) > 1$), which reduces to $A \geq 0$ ($= 0$ if $x(1) > 1$). With $A = 0$, $x(1) = -\frac{1}{4} + 1 < 1$. Thus $A > 0$ and $x(1) = 1$, so the solution is $x = -\frac{1}{4}t^2 + \frac{1}{2}t + 1$.

2. (a) $x = Ae^{rt} + Be^{-rt}$, with $A = -B = \frac{1}{e^{\beta} - e^{-\beta}}$. (b) (i) $x = 0$ (ii) $x = \frac{2(e^{2t} - e^{-2t})}{e^{\beta} - e^{-\beta}}$

3. Replace G in the integrand by $G = (r_1/r_2)\bar{Y} - (1/r_2)\dot{\bar{Y}}$. Euler equation: $\ddot{\bar{Y}} = m^2\bar{Y}$ with $m^2 = (\alpha_2r_1^2 + \alpha_1r_2^2)/\alpha_2$. Solution: $\bar{Y} = Ae^{mt} + Be^{-mt}$, where $A = (r_1 + m)\bar{Y}_0/(e^{2mT}(m - r_1) + (m + r_1))$, $B = \bar{Y}_0 - A$.

4. $A = Ke^{rt} + (r - \rho)t/r + L$. The constants K and L are determined by $A(0) = A_0$, $A(T) = A_T$.

5. (a) The conditions are $-(d/dt)[C'_t(t, \dot{x})e^{-rt}] = 0$ and $C'_t(5, \dot{x}(5)) \geq 0$ ($= 0$ if $x(5) > 1500$).

(b) $x(t) = 300t$, implying that planting takes place at the constant rate of 300 hectares per year.

Chapter 9

The Hamiltonian is $H = e^t x - u^2 + p(-u)$, so $H'_x = e^t$ and $H'_u = -2u - p$. Because $U = (-\infty, \infty)$, (7) implies that $u^*(t) = -\frac{1}{2}p(t)$. (5) reduces to $\dot{p} = -e^t$, $p(2) = 0$, from which it follows that $p(t) = -e^t + e^2$. Hence, $u^*(t) = \frac{1}{2}e^t - \frac{1}{2}e^2$. From $\dot{x}^* = -u^*(t) = -\frac{1}{2}e^t + \frac{1}{2}e^2$ and $x^*(0) = 0$, we find $x^*(t) = \frac{1}{2}(e^2t - e^t + 1)$. The Hamiltonian is a sum of concave functions, so $u^*(t) = \frac{1}{2}e^t - \frac{1}{2}e^2$ is the optimal control.

$u^*(t) = 0$, $x^*(t) = e^t$, $p(t) = 0$. (With $H = 1 - u^2 + p(x+u)$, H is concave in (x, u) and $H'_x = p$, $H'_u = -2u + p$. If (x^*, u^*) solves the problem, then $(H'_u)^* = 0$ gives $u^*(t) = \frac{1}{2}p(t)$. Moreover, $\dot{p}(t) = -(H'_x)^* = -p(t)$, with $p(1) = 0$ implies $p(t) = 0$, and so $u^*(t) = 0$. Then from $\dot{x}^*(t) = x^*(t)$ and $x^*(0) = 1$, we get $x^*(t) = e^t$. This solution is also obvious without using the maximum principle, because $1 - u(t)^2$ is largest when $u(t) = 0$.)

$u^*(t) = -\frac{1}{2}t + \frac{1}{2}$, $x^*(t) = \frac{1}{4}t^2 - \frac{1}{2}t$, with $p(t) = t - 1$. (The Hamiltonian is $H = -(x+u^2) - pu$. Note that the first minus sign is inserted because we minimize the criterion.)

$u^*(t) = t - 10$, $x^*(t) = \frac{1}{2}t^2 - 10t$, with $p(t) = 4(t - 10)$.

$u^*(t) = \frac{1}{2}(e^{T-t} - 1)$, $x^*(t) = \frac{1}{4}e^{T+t} - \frac{1}{4}e^{T-t} - \frac{1}{2}e^t + \frac{1}{2}$, $p(t) = e^{T-t} - 1$.

(a) $H = qf(K) - c(I) + p(I - \delta K)$. (7): $c'(I^*(t)) = p(t)$, (5): $\dot{p}(t) - \delta p(t) = -qf'(K^*(t))$, $p(T) = 0$.
(b) $\dot{K} = 0.04K = -0.5$. General solution: $K = Ae^{0.2t} + Be^{-0.2t} + 12.5$. $K(0) = 10$ and $p(10) = 0$ determine the constants A and B . The Hamiltonian is concave in (K, I) .

$u^*(t) = 1$ with $x^*(t) = t$ is the obvious optimal solution. We find that $V(T) = \int_0^T x^*(t) dt = \int_0^T t dt = \frac{1}{2}T^2$. The Hamiltonian is $H = x + pu$, so $\dot{p} = -H'_x = -1$, with $p(T) = 0$. Thus, $p = T - t$. The optimal control must maximize $p(t)u = (T - t)u$ for $u \in [0, 1]$, so $u^*(t) = 1$. As before, $x(t) = t$. The Hamiltonian is linear and hence concave in (x, u) .

$u^*(t) = A(e^t + e^{-t})$, $x^*(t) = A(e^t - e^{-t})$, $p(t) = 2u^*(t)$, where $A = e/(e^2 - 1)$.

(a) $u^*(t) = 0$, $x^*(t) = x_0$. (b) $u^*(t) = -1$ and $x^*(t) = x_0 - t$ in $[0, \sqrt{T^2 - 4}]$;

$u^*(t) = -\frac{1}{4}(T^2 - t^2)$ and $x^*(t) = \frac{1}{12}t^3 - \frac{1}{4}tT^2 + x_0 + (\frac{1}{6}T^2 - \frac{2}{3})\sqrt{T^2 - 4} \ln(\sqrt{T^2 - 4}, T]$

(a) $(u^*(t), x^*(t)) = \begin{cases} (1, t) & \text{if } t \in [0, 2] \\ (0, 2) & \text{if } t \in (2, 10] \end{cases}$, with $p(t) = 2 - t$. $V = \int_0^{10} t dt + \int_2^{10} 2 dt = 2 + 16 = 18$.

(b) $(u^*(t), x^*(t)) = \begin{cases} (1, t + x_0) & \text{if } t \in [0, x_1 - x_0] \\ (0, x_1) & \text{if } t \in (x_1 - x_0, T] \end{cases}$, with $p(t) = x_1 - x_0 - t$.

$V = \int_0^{x_1-x_0} (t+x_0) dt + \int_{x_1-x_0}^T x_1 dt = Tx_1 + x_0x_1 - \frac{1}{2}x_0^2 - \frac{1}{2}x_1^2$.

(a) With $H = -(u^2 + x^2) + pau$, (6) gives $\dot{p} = 2x^*$, and (5) implies $u^* = 0$ if $-2u^* + ap < 0$, $u^* = 1$ if $-2u^* + ap > 0$. Condition (7)(c) yields $p(T) = 0$. If $a \geq 0$, $u^*(t) = 0$, $x^*(t) = 1$, and $p(t) = -2(T - t)$.

(b) If $a < 0$, $u^*(t) = \frac{1}{1 + e^{2at}}(e^{at} - e^{2at}e^{-at})$, $p(t) = \frac{2u^*(t)}{a}$, $x^*(t) = \frac{e^{at} + e^{2at}e^{-at}}{1 + e^{2at}}$.

$x^*(t) = 5 - \frac{3e^{0.1t}}{2(e^{0.5} - 1)}$, $x^*(t) = \frac{15(e^{0.1t} - 1)}{e^{0.5} - 1} - 5t + 10$, with $p(t) = \frac{3}{e^{0.5} - 1}$.

(a) With $H = -(ax + bu^2) + pu$ we have $\dot{p} = -H'_x = a$, and $u^*(t)$ maximizes $-bu^2 + p(t)u$ for $u \geq 0$.

(b) If $B \geq aT^2/4b$: $u^*(t) = a(2t - T)/4b + B/T$, $x^*(t) = at(t - T)/4b + Bt/T$, $p(t) = at + 2bB/T - aT/2$.

If $B < aT^2/4b$: $u^*(t) = \begin{cases} 0 & \text{if } t \in [0, t^*] \\ a(t - t^*)/2b & \text{if } t \in (t^*, T] \end{cases}$, $x^*(t) = \begin{cases} 0 & \text{if } t \in [0, t^*] \\ a(t - t^*)^2/4b & \text{if } t \in (t^*, T] \end{cases}$, and

$v(t) = a(t - t^*)$, where $t^* = T - 2\sqrt{bB/a}$.

8. With $r^* = \frac{1}{2}(1 + \sqrt{5})$, the solution is:

$$u^*(t) = \begin{cases} 1 & \text{if } t \in [0, t^*] \\ 0 & \text{if } t \in (t^*, 2] \end{cases}, \quad x^*(t) = \begin{cases} t + 1 & \text{if } t \in [0, t^*] \\ \frac{1}{2}(3 + \sqrt{5}) & \text{if } t \in (t^*, 2] \end{cases}, \text{ with}$$

$$p(t) = \begin{cases} -t^2 - 2t + \frac{3}{2}(3 + \sqrt{5}) & \text{if } t \in [0, t^*] \\ -(3 + \sqrt{5})t + 6 + 2\sqrt{5} & \text{if } t \in (t^*, 2] \end{cases}$$

9. (a) Show that $p_0 + p(t)$ must be 0. Hence, $p(t) = -1$, and $H = 0$.

(b) $\max \int_0^2 u(t) dt = \max \int_0^2 \dot{x}(t) dt = \max x(2) = 1$.

10. (a) Obviously, $u^*(t) = 0$ is the only admissible control, so it is optimal.

(b) With $H = -p_0u + pu^2$, the equation $H'_u = -p_0 + 2pu = 0$ has the solution $u = 0$ only if $p_0 = 0$.

9.5

1. The Euler equation is $\ddot{x} = -e^{-t}$. The solution is $x = -e^{-t} + e^{-1}t + 1$. To solve it as a control problem, put $\dot{x} = u$.

2. $x^*(t) = Ae^{\frac{1}{2}t} + (1 - A)e^{-\frac{1}{2}t}$, with $A = (4e^{\sqrt{2}} - 1)/(e^{2\sqrt{2}} - 1)$.

3. The Euler equation is $-\frac{d}{dt}[(-2 - 2\dot{x})e^{-t/10}] = 0$, which implies $\dot{x} = Ae^{t/10} - 1$, for some constant A . Integrate and use the boundary conditions to get the solution $x = 1 - t$. (Alternative form of the Euler equation: $\ddot{x} - \frac{1}{10}\dot{x} - \frac{1}{10} = 0$.) The control problem is: $\max \int_0^1 (-2u - u^2)e^{-t/10} dt$, $\dot{x} = u$, $x(0) = 1$, $x(1) = 0$. We find that $u^*(t) = -1$, $p(t) = 0$, and $x(t) = 1 - t$.

4. The Euler equation reduces to $\frac{d}{dt}[ae^{(a-r)t} + 2\dot{x}e^{(b-r)t}] = 0$, so $ae^{(a-r)t} + 2\dot{x}e^{(b-r)t} = C$ for some constant C . Then $(\partial\pi/\partial\dot{x})e^{-rt} = (-ae^{at} - 2\dot{x}e^{bt})e^{-rt} = -C$, and we can let $c = -C$.

With $u = \dot{x}$, the Hamiltonian is $H = \pi(u)e^{-rt} + pu$, and provided there is an interior solution, $\partial H^*/\partial u = \pi'(u)e^{-rt} + p(t) = 0$, and $\dot{p}(t) = -\partial H^*/\partial x = 0$, so $p(t)$ is a constant. The conclusion follows.

5. (a) With $H = U(x) - b(x) - gz + pax$, the conditions are: (i) x^* maximizes $U(x) - b(x) - gz^* + p(t)ax$ for $x \geq 0$; (ii) $\dot{p} = g$, $p(T) = 0$. From (ii) we immediately get $p(t) = g(T - t)$. Moreover, $\partial H^*/\partial x = 0$ yields (*). (b) H is concave in (z, x) , so Mangasarian's theorem applies. We find that $dx^*/dt = -ag/(U'' - b'') > 0$.

9.6

1. (a) $u^*(t) = T - t$, $x^*(t) = x_0 + Tt - \frac{1}{2}t^2$, with $p(t) = T - t$.

(b) $V(x_0, T) = x_0T + \frac{1}{6}T^3$ and the relevant equalities in (5) are easily verified.

2. $H^*(T) = x^*(T) + p(T)u^*(T) = x^*(T) = T$, so using the expression for $V(T)$ found in Problem 9.4.1, it follows that $V'(T) = H^*(T)$.

3. Using the results in the answers to Problem 9.4.4 (b), we find: $\partial V/\partial x_0 = x_1 - x_0 = p(0)$, $\partial V/\partial x_1 = T + x_0 - x_1 = -p(T)$, $\partial V/\partial T = x_1$, and $H^*(T) = x^*(T) + p(T)u^*(T) = x_1$.

4. (a) For $T \leq \frac{1}{2} \ln 2$ we have $u^*(t) = 0$ and $x^*(t) = 1$, with $p(t) = 2e^{-2t}(1 - e^{-2(T-t)})$.

For $T > \frac{1}{2} \ln 2$, let $t^* = \ln(\sqrt{\frac{1}{16}}e^{4T} + e^{2T}) - \frac{1}{2}e^{2T}$. Then for $t \in [0, t^*]$, $u^*(t) = 1$ and $x^*(t) = e^t$, with $p(t) = 4(e^{-t} - e^{-2t}) + 1$. For $t \in (t^*, T]$, $u^*(t) = 0$ and $x^*(t) = e^{t^*}$, with $p(t) = 2e^{t^*}(e^{-2t} - e^{-2T})$.

(b) For $T \leq \frac{1}{2} \ln 2$, $V(T) = 1 - e^{-2T}$ and $H^*(T) = 2e^{-2T}$, so $V'(T) = H^*(T)$. For $T > \frac{1}{2} \ln 2$, $V'(T) = 2t^* - e^{t^*} - e^{2t^* - 2T} + 2$, so $V'(T) = (2 - e^{t^*} - 2e^{2t^* - 2T})\frac{dt^*}{dT} + 2e^{2t^* - 2T} = 2e^{2t^* - 2T} = H^*(T)$.

5. (a) We get $x^*(t) = x_0$. For $x_0 < 0$, $u^* = 0$ maximizes $H = x_0u$. For $x_0 > 0$, $u^* = 1$ maximizes $H = x_0u$.

(b) $V(x_0) = 0$ when $x_0 < 0$ and $V(x_0) = x_0$ when $x_0 \geq 0$, so V is not differentiable at 0.

9.7

1. (a) $u^*(t) = p(t) = t + x_1 - x_0 - \frac{1}{2}$, $x^*(t) = \frac{1}{2}t^2 + (x_1 - x_0 - \frac{1}{2})t + x_0$.

(b) Verified by differentiating under the integral sign.

(a) $(s^*(t), k^*(t)) = \begin{cases} (1, (\frac{1}{2}t + 1)^2) & \text{if } t \in [0, 4] \\ (0, 9) & \text{if } t \in (4, 10] \end{cases}$, $p(t) = \begin{cases} 6/(t+2) & \text{if } t \in [0, 4] \\ (10-t)/6 & \text{if } t \in (4, 10] \end{cases}$

(b) $\hat{H}(t, k, p) = \max_{x \in [0, 1]} [\sqrt{k} + \sqrt{k(p-1)}x] = \begin{cases} \sqrt{k}p & \text{if } p > 1 \\ \sqrt{k} & \text{if } p \leq 1 \end{cases}$

Hence $\hat{H}$ is concave in k for $k > 0$. The solution candidate in (a) is therefore optimal.

(a) $u^*(t) = \frac{\alpha - 2\beta}{e^{(\alpha-2\beta)T} - 1} e^{2(\alpha-\beta)t}$, $x^*(t) = \frac{e^{(\alpha-2\beta)T} - e^{(\alpha-2\beta)t}}{e^{(\alpha-2\beta)T} - 1} e^{\alpha t}$, $p(t) = \sqrt{\frac{e^{(\alpha-2\beta)T} - 1}{4(\alpha-2\beta)}} e^{-\alpha t}$.

(Hint: Argue why $u^*(t) > 0$.) (b) The solution is still as in (a).

We find that $g'(0) = \nabla f(x^0) \cdot (x - x^0)$. Since g has a maximum at $t = 0$, $g'(0)$ cannot be > 0 , so $g'(0) \leq 0$. The conclusion follows. (If f is concave, the implication can be reversed.)

3

(a) $u^*(t) = \frac{1}{2} - t$, $x^*(t) = \frac{1}{2}t(1-t)$, $p(t) = u^*(t)$, $T^* = \frac{1}{2}$. (With $H = x - t^3 - \frac{1}{2}u^2 + pu$, $H'_t = 1$ and $H'_u = -u + p$. Then $\bar{p} = -1$, with $p(T) = 0$, so $p(t) = T - t$. Also, $H'_t = 0$ yields $u^*(t) = p(t) = T - t$ and then $\dot{x}^*(t) = T - t$, with $x^*(0) = 0$, so $x(t) = Tt - \frac{1}{2}t^2$. With $p(T) = u^*(T) = 0$, $H^*(T) = x^*(T) - T^3 = T^2 - \frac{1}{2}T^2 - T^3 = 0$, so $T^* = T = \frac{1}{2}$.) (b) $u^*(t) \equiv 6$, $T^* = 8/3$

$T = 2(Bb/a)^{1/2}$

(a) $u^*(t) = \frac{1}{2}ae^{(\alpha-\beta)t} - \frac{1}{2}\bar{p}e^{(\alpha-\beta)t}$. For $\alpha = \beta = 0$, $x^*(t) = K - \frac{1}{2}at + \frac{\bar{p}}{2a}(e^{\alpha t} - 1)$ with $\bar{p} = \frac{2r}{e^{\alpha T} - 1}(\frac{1}{2}aT - K)$.
(b) $u^*(T^*) = \frac{1}{2}(a - z)$, where $z = \bar{p}e^{\alpha T^*}$. The condition $H^*(T^*) = 0$ reduces to $z^2 - 2az + a^2 - 4c = 0$, with (the only admissible) solution $z = a - 2\sqrt{c}$, which is positive because $a^2 > 4c$. The equation for determining T^* is $\varphi(T^*) = arT^* - (a - 2\sqrt{c}) + (a - 2\sqrt{c})e^{-\alpha T^*} - 2rK = 0$. (Look at $\varphi(0)$, $\varphi(\infty)$, and the sign of $\varphi'(T^*)$.)

9

With $H^c = -x^2 - \frac{1}{2}u^2 + \lambda(x + u)$, $\partial H^c/\partial u = -u + \lambda$ and $\partial H^c/\partial x = -2x + \lambda$. $\partial(H^c)^*/\partial u = 0$ yields $u^*(t) = \lambda(t)$. Moreover, $\bar{\lambda} - 2\lambda = -\partial(H^c)^*/\partial x = 2x^* - \lambda$ with $\lambda(T) = 0$. Thus x^* and λ must be solutions to $\dot{x} = x + \lambda$, $\dot{\lambda} = 2x + \lambda$. Derive $\ddot{x} - 2\dot{x} - x = 0$, with solution: $x = Ae^{(1+\sqrt{2})t} + Be^{(1-\sqrt{2})t}$. Then $u^* = \lambda = \dot{x}^* - x^* = A\sqrt{2}e^{(\sqrt{2}+1)t} - B\sqrt{2}e^{(1-\sqrt{2})t}$. $x^*(0) = 1$ and $\lambda(T) = 0$ yields the same values of A and B as before.

See the answer to Problem 9.4.6.

$H^c = -2u - u^2 + \lambda u$. $\partial(H^c)^*/\partial u = 0$ when $u^* = \frac{1}{2}\lambda - 1$, and $\dot{\lambda} - \frac{1}{10}\lambda = -\partial(H^c)^*/\partial x = 0$. We find $u^* = -1$, $x^* = 1 - t$, and $\lambda = 0$.

10

$u^*(t) = 1 - \frac{1}{2}t$, $x^*(t) = -\frac{1}{4}t^2 + t + 1$, $p(t) = 2$. ($H = 1 - tu - u^2 + pu$, $S(x) = 2x + 3$. $H'_x = 0$ implies $p(t) = \bar{p}$, a constant. But $p(1) = \bar{p} = S'(x^*(1)) = 2$, so $\bar{p} = 2$. $H'_u = 0$ yields $u^* = 1 - \frac{1}{2}t$, etc.)

$\ddot{A}^* - rA^* + (\rho - r)U'/U'' = 0$, $A^*(0) = A_0$, $\varphi'(A^*(T)) = U'(rA^*(T) + w - u(T))$

If $t \in [0, 9]$, then $s^*(t) = 1$, $k^*(t) = (\frac{1}{2}t + 1)^2$, and $p(t) = 11/(t + 2)$.

If $t \in (9, 10]$, then $s^*(t) = 0$, $k^*(t) = 30.25$, and $p(t) = \frac{20}{11} - \frac{1}{11}t$.

(a) $(u^*(t), x^*(t)) = \begin{cases} (1, t + \frac{1}{2}) & \text{if } t \in [0, \frac{1}{2}] \\ (0, 1) & \text{if } t \in (\frac{1}{2}, 1] \end{cases}$, with $p(t) = -t + 3/2$. (b) Same solution as in (a).

(a) $u^*(t) = -2ae^t$, $x^*(t) = a(3e^{2T}t - e^t)$, $p(t) = 2u^*(t)$, where $a = \frac{\lambda_0}{3e^{2T} - 1}$. (b) $V(x_0, T) = \frac{-2(\lambda_0)^2}{3e^{2T} - 1}$

6. $u^*(t) = at - \frac{aT(e^{\alpha t}(1+r) - 1)}{e^{\alpha T}(1+r) - 1}$, $x^*(t) = at - \frac{aT(e^{\alpha t} - 1)}{e^{\alpha T}(1+r) - 1}$, $\lambda(t) = \frac{-2arTe^{\alpha t}}{e^{\alpha T}(1+r) - 1}$

7. The only problem is to prove that $p(t_1) = S'(x^*(t_1))$: from $p = q + S'(x^*)$ we get $\dot{q} = \dot{p} - S''(x^*)\dot{x}^* = \dot{p} - S''(x^*)g^*$. Moreover, $\partial H^*_t/\partial x = \partial f^*/\partial x + S''(x^*)g^* - S'(x^*)\partial g^*/\partial x + q\partial g^*/\partial x$. Hence $\dot{q} = -\partial H^*_t/\partial x$ implies that $\dot{p} = -\partial H^*/\partial x$.

9.11

1. $u^*(t) = 2e^{-0.1t}$, $x^*(t) = 10e^{-0.1t}$

2. $(u^*(t), x^*(t)) = \begin{cases} (1, e^{1-e^{-t}}) & \text{if } t \in [0, \ln 2] \\ (0, e^{1/2}) & \text{if } t \in (\ln 2, \infty) \end{cases}$, $p(t) = \begin{cases} 2e^{-t/2} - 1 & \text{if } t \in [0, \ln 2] \\ 2e^{-t} & \text{if } t \in (\ln 2, \infty) \end{cases}$

3. $V = (1 - \delta)^{-1}[(\rho - r)/\delta + r]^{-3}(A_0 + w/r)^{1-\delta}$. As ρ increases, V decreases. As w increases, V increases. We see that $\partial V/\partial A_0 = \bar{\lambda} = \lambda(0)$.

4. $u^*(t) = \begin{cases} 1 & \text{if } t \in [-1, 0] \\ 0 & \text{if } t \in (0, \infty) \end{cases}$, $x^*(t) = \begin{cases} e - e^{-t} & \text{if } t \in [-1, 0] \\ e - 1 & \text{if } t \in (0, \infty) \end{cases}$, $p(t) = e^{-t}$, $\lambda(t) = 1$

9.12

1. (a) $H^c = ax - \frac{1}{2}u^2 + \lambda(-bx + u)$. System (2) is $\dot{x} = -bx + \lambda$, $\dot{\lambda} = -a + (b + r)\lambda$. The equilibrium point is $(\bar{x}, \bar{\lambda}) = (a/b(b + r), a/(b + r))$. (b) The phase diagram is similar to Figure 9.12.1. The solution is $x^*(t) = (x_0 - a/(b(b + r)))e^{-(b+r)t} + a/(b(b + r))$, $u^*(t) = \lambda(t) = \bar{\lambda}$. (c) Easy verification. (Differentiate under the integral sign.)

2. $x^*(t) = e^{(1-\sqrt{2})t}$, $u^*(t) = -\sqrt{2}e^{(1-\sqrt{2})t}$, $p(t) = -\sqrt{2}e^{(1-\sqrt{2}-1)t}$, $\lambda(t) = -\sqrt{2}e^{(1-\sqrt{2})t}$

3. (a) With $H^c = \ln C + \lambda(AK^a - C)$, $\partial(H^c)^*/\partial C = 0$ implies $C^*\lambda = 1$, and so $\dot{C}^*/C^* + \dot{\lambda}/\lambda = 0$. Also, $\dot{\lambda} - r\lambda = -\partial(H^c)^*/\partial K = -\lambda\alpha A(K^*)^{\alpha-1}$, or $\dot{\lambda}/\lambda = r - \alpha A(K^*)^{\alpha-1}$. (b) See SM. (c) The solution curve converges to the equilibrium point. For sufficient conditions, see Note 9.11.3.

4. (a) $\dot{x} = x + \lambda$, $\dot{\lambda} = 2x - 2$, with equilibrium $(x_0, \lambda_0) = (1, -1)$. (b) $u^*(t) = 1 - e^{-t}$, $x^*(t) = 1 - \frac{1}{2}e^{-t}$.

Chapter 10

10.1

1. If $y(t) = \int_{t_0}^t G(\tau, x(\tau), \dot{x}(\tau)) d\tau$, then $\dot{y}(t) = G(t, x(t), \dot{x}(t))$, so the control problem is:

$$\max \int_{t_0}^{t_1} F(t, x, u) dt, \quad \begin{cases} \dot{x} = u & x(t_0) = x^0, x(t_1) = x^1 \\ \dot{y} = G(t, x, u) & y(t_0) = 0, y(t_1) = K \end{cases}$$

2. With vector notation, $\frac{d}{dt}H^* = \frac{d}{dt}H(t, x^*, u^*, p) = \frac{\partial H^*}{\partial t} + \frac{\partial H^*}{\partial x} \cdot \dot{x}^* + \frac{\partial H^*}{\partial u} \cdot \dot{u}^* + \frac{\partial H^*}{\partial p} \cdot \dot{p}^*$. Because Theorem 10.1.1 yields $\partial H^*/\partial x = -\bar{p}$, $\partial H^*/\partial u = 0$, and $\partial H^*/\partial p = g^* = \dot{x}^*$, equation (1) follows immediately.

3. $f(\lambda x_1 + (1-\lambda)x_2) \geq F(\lambda x_1 + (1-\lambda)x_2, \lambda u_1 + (1-\lambda)u_2) \geq \lambda F(x_1, u_1) + (1-\lambda)F(x_2, u_2) = \lambda f(x_1) + (1-\lambda)f(x_2)$. Concavity of g is proved in a similar manner.

4. Define $y(t) = \int_{t_0}^t h(\tau, x(\tau), u(\tau)) d\tau$. Replace the integral constraint by $\dot{y} = h(t, x, u)$, $y(t_0) = 0$, $y(t_1) = K$.

1.2

$u^*(t) = -1$ in $[0, 4 - \sqrt{2}]$, $u^*(t) = 1$ in $(4 - \sqrt{2}, 4]$, $p_1(t) = t - 4$, $p_2(t) = -\frac{1}{2}(t - 4)^2$.
($H = 10 - x_1 + p_1x_2 + (1 + p_2)u$, and $u^*(t) = 1$ if $p_2(t) > -1$, $u^*(t) = 0$ if $p_2(t) < -1$. $p_1 = 1$ and $p_1(4) = 0$, so $p_1(t) = t - 4$. Since $p_2 = -p_1$ and $p_2(4) = 0$, we have $p_2(t) = -\frac{1}{2}(t - 4)^2$. Note that $p_2(t) > -1$ iff $t > 4 - \sqrt{2}$.)

$x_1^*(t) = x_1^0 e^{at}$ and $x_2^*(t) = x_2^0$ for t in $[0, T - 2/a]$,
 $x_1^*(t) = x_1^0 e^{a(T-2)} e^{a(t-T+2/a)}$ and $x_2^*(t) = x_2^0 + ax_1^0 e^{a(T-2)}(t - (T - 2/a))$ for t in $(T - 2/a, T]$.

(a) $u_1^*(t) = \begin{cases} 1 & \text{if } t \in [0, T - 2] \\ 0 & \text{if } t \in (T - 2, T] \end{cases}$, $u_2^*(t) = \begin{cases} 1 & \text{if } t \in [0, T - 5] \\ 0 & \text{if } t \in (T - 5, T] \end{cases}$, $x_1^*(t) = \begin{cases} t & \text{if } t \in [0, T - 2] \\ T - 2 & \text{if } t \in (T - 2, T] \end{cases}$

$x_2^*(t) = \begin{cases} t & \text{if } t \in [0, T - 5] \\ T - 5 & \text{if } t \in (T - 5, T] \end{cases}$, $p_1(t) = \frac{1}{2}(T - t)$, $p_2(t) = \frac{1}{2}(T - t)$

(b) $u_1^*(t) = u_2^*(t) = 1$ and $x_1^*(t) = x_2^*(t) = t$ for all t , $p_1(t) = 3 + \frac{1}{2}(T - t)$, $p_2(t) = 2 + \frac{1}{2}(T - t)$.

$u_1 = 1$ in $[0, T - 2/b]$, $u_2 = 1$ in $(T - 2/b, T - c/a]$, $u_1 = u_2 = 0$ in $(T - c/a, T]$.

$u^*(t) = 0$ in $[0, t_*)$, $u^*(t) = 1$ in (t_*, t_{**}) , and $u^*(t) = 0$ in $(t_{**}, T]$, where $t_* = T - \frac{1}{bc}(1 + \sqrt{1 - 2bc})$ and $t_{**} = T - \frac{1}{bc}(1 - \sqrt{1 - 2bc})$.

(a) Necessary conditions: (i) $\dot{p}_1 = -\partial H/\partial K = -p_1 f_K'$, (ii) $\dot{p}_2 = 0$, (iii) $\partial H/\partial u = p_1 f_u' - p_2 = 0$, (iv) $\partial H/\partial c = U'(c)e^{-rt} - p_1 = 0$. (b) Differentiating (iv) w.r.t. t and using (i) and (iv) again yields $U''(c)\dot{c}e^{-rt} - rU'(c)e^{-rt} = -p_1 f_K' = -U'(c)f_K' e^{-rt}$. Using the definition of $\tilde{c}$ and rearranging, we find the first equation in (b). The second equality in (b) is obtained by differentiating (iii) w.r.t. t and using (i).

$u^*(t) = \begin{cases} 1 & \text{if } t \leq 1 \\ 0 & \text{if } t > 1 \end{cases}$, $x^*(t) = \begin{cases} t+1 & \text{if } t \leq 1 \\ 2 & \text{if } t > 1 \end{cases}$, $y^*(t) = \begin{cases} t & \text{if } t \leq 1 \\ 1 & \text{if } t > 1 \end{cases}$, $p_1(t) = 2 - t$, $p_2(t) = -\frac{1}{2}$.

1.3

$u^*(t) = 1$ in $[0, -\ln r]$, $u^*(t) = 0$ in $(-\ln r, \infty)$

(a) $x_1^*(t) = x_1^0$, $x_2^*(t) = ax_1^0 t$, $u^*(t) = 0$. (b) No solution exists. (If $\dot{x}_1 = bx_1$, where $r < b < a$, the objective function becomes infinite.)

1.4

(i) $(x(t), u(t)) = (e^t - 1, 1)$ is admissible; (ii) $N(t, x)$ is the rectangle $\{(r, s) : r \leq x, x - 1 \leq s \leq x + 1\}$, which surely is convex. (Draw a graph.) (iii) When $|u| \leq 1$, $|x + u| \leq |x| + 1$, so condition (2) in Note 2 is also satisfied. We conclude from Theorem 10.4.1 that the problem has an optimal solution.

The existence of a solution is secured by Theorem 10.4.1. The solution is $u^*(t) = 1$ if $t \in [0, 1/2]$, $u^*(t) = 0$ if $t \in (1/2, 1]$. ($x^*(t) > 0$ and $p(t) = 0$, so there must be an interval $(t^*, 1]$ where $u^*(t) = 0$. Note that $p(t)$ is strictly decreasing and $x^*(t)$ is increasing.)

$u^*(t) = 0$, $x^*(t) = t + 4$ and $p(t) = -t^2 - 8t + 105/16$ for t in $[0, \frac{3}{4}]$; $u^*(t) = 2$, $x^*(t) = -3t + 7$ and $p(t) = 3t^2 - 14t + 141/16$ for t in $(\frac{3}{4}, 1]$. (Hint: $u = u^*(t)$ maximizes $-p(t)u^2$ for u in $[-1, 2]$. Then $u^*(t)$ must be 2 if $p(t) < 0$ and 0 if $p(t) > 0$. Since $|u| \leq 2$, one has $\dot{x} \geq -3$, and therefore $x(t) \geq x(0) - 3t = 4 - 3t \geq 1$ for t in $[0, 1]$. Thus any admissible $x(t)$ is positive and $\dot{p}(t) = -2x^*(t) < 0$, so $p(t)$ is strictly decreasing, etc.)

1.6

(a) $H = -\frac{1}{2}u^2 - x - pu$ and $\mathcal{L} = H + q(x - u)$. H is concave in (x, u) and $h(t, x, u) = x - u$ is linear and therefore quasiconcave. Here (i) $\partial \mathcal{L}^*/\partial u = -u^*(t) - p(t) - q(t) = 0$. (ii) $q(t) \geq 0$ ($= 0$ if $x^*(t) > u^*(t)$), (iii) $\dot{p}(t) = 1 - q(t)$, with $p(2) = 0$. (b) $u^*(t) = x^*(t) = e^{-t}$, $p(t) = (t^* - \frac{1}{2}(t^*)^2)e^{-t} - 1 - \frac{1}{2}e^{-t}$, $q(t) = -e^{-t} - p(t)$ for t in $[0, t^*]$; $u^*(t) = 2 - t$, $x^*(t) = \frac{1}{2}t^2 - 2t + 2 + t^* - \frac{1}{2}(t^*)^2$, $p(t) = t - 2$, $q(t) = 0$ for t in $(t^*, 2]$, with t^* determined by $e^{-t^*} = 2 - t^*$. ($t^* \approx 1.84$.) (Note that one has to check that $q(t) \geq 0$ for t in $[0, t^*]$, and that $x^*(t) > u^*(t)$ for t in $(t^*, 2]$.)

2. Let $t^* \approx 0.44285$ be the solution of $2 - t^* = e^{t^*}$. Then,
for t in $[0, t^*]$: $x^*(t) = u^*(t) = e^t$, $p(t) = (e^{t^*} + \frac{1}{2}e^{2t^*})e^{-t} + \frac{1}{2}e^t - 1$, and $q(t) = p(t) - e^t$;
for t in $(t^*, 2]$: $x^*(t) = -\frac{1}{2}(t^2 - (t^*)^2) + 2(t - t^*) + e^{t^*}$, $u^*(t) = p(t) = 2 - t$, and $q(t) = 0$.

3. $u^*(t) = c$, $x^*(t) = (x_0 - c/a)e^{at} + c/a$, $p(t) = e^{a(t'-t)}$, $q_1(t) = e^{a(t'-t)} - 1$, $q_2(t) = 0$ for t in $[0, t']$;
 $u^*(t) = ax^*(t)$, $x^*(t) = (x_0 - c/a)e^{at} + c/a$, $p(t) = a(t' - t) + 1$, $q_1(t) = 0$, $q_2(t) = a(t - t')$ for t in $[t', T]$;
where $t' = \max\{T - 1/a, t^0\}$, $t^0 = (1/a)\ln[(x_0 - c/a)/(x_0 - c/a)]$.

4. For t in $[0, \ln 2]$: $u^*(t) = 1$, $x^*(t) = e^t - 1$, $p(t) = q_1(t) = (4 - 2\ln 2)e^{-t} - 1$, $q_2(t) = 0$, $q_3(t) = 0$.
For t in $(\ln 2, 1]$: $u^*(t) = 1 + 2\ln 2 - 2t$, $x^*(t) = 2t + 1 - 2\ln 2$, $p(t) = 1 - t$, $q_1(t) = 0$, $q_2(t) = 0$, $q_3(t) = 1 - t$.

10.7

1. In $[0, \sqrt{2}]$: $u^*(t) = 0$, $x^*(t) = 1 - \frac{1}{2}t^2$, $p(t) = t + 1 - \sqrt{2}$, and $q(t) = 0$. In $(\sqrt{2}, 5]$: $u^*(t) = t$, $x^*(t) = 0$, $q(t) = 1$ with $p(t) = 1$ in $(\sqrt{2}, 5)$ and $p(5) = 0$, with $\beta = 1$.

2. $(u^*(t), x^*(t)) = \begin{cases} (-1, 1 - t) & \text{if } t \in [0, 1] \\ (0, 0) & \text{if } t \in (1, 2] \end{cases}$.

3. $u^*(t) = \frac{1}{2}(t - 2)$, $x^*(t) = \frac{1}{4}(t - 2)^2$, $p(t) = t - 2$, and $q(t) = 0$ in $[0, 2]$;
 $u^*(t) = 0$, $x^*(t) = 0$, $p(t) = 0$, $q(t) = 1$ in $(2, 10]$, with $\beta = 0$.

4. (a) $(u^*(t), x^*(t)) = \begin{cases} (2, 1 + 2t) & \text{if } t \in [0, 1] \\ (0, 3) & \text{if } t \in (1, 3] \end{cases}$, $p(t) = \bar{p} = -3$.

(b) It obviously pays to keep $u(t)$ as large as possible. Therefore we suggest

for t in $[0, 2]$: $u^*(t) = 1$, $x^*(t) = 1 + t$, $p(t) = t - 4$, $q(t) = 1$;
for t in $(2, 3]$: $u^*(t) = 0$, $x^*(t) = 3$, $p(t) = -2$, $q(t) = 0$; and $\beta = 0$.

It is easy to verify that all the conditions in Theorem 10.7.1 are satisfied. Thus we have found an optimal solution.

Chapter 11

11.1

1. (a) According to (4), $x_t = 5 \cdot 2^t - 4$. (b) $x_t = (1/3)^t + 1$. (c) $x_t = (-3/5)(-3/2)^t - 2/5$. (d) $x_t = -3t + 3$.

2. (a) Monotone convergence to x^* from below. (b) Damped oscillations around x^* . (c) Monotonically increasing towards ∞ . (d) Explosive oscillations around x^* . (e) $x_t = x^*$ for all t . (f) Oscillations around x^* with constant amplitude. (g) Monotonically (linearly) increasing towards ∞ . (h) Monotonically (linearly) decreasing towards $-\infty$. (i) $x_t = x_0$ for all t .

3. $y_{t+1} = ay_t$, such that $y_t = y_0 a^t$, etc.

4. (a) Because the parameters are positive, y_{t+1} is positive provided $y_t > 0$. Since y_0 is positive, so is y_t a.s.o.
(b) Substituting $y_t = 1/x_t$ gives $x_{t+1} = (a/c)x_t + b/c$. When $a = 2$, $b = 3$, and $c = 4$, $x_{t+1} = (1/2)x_t + 3/4$. When $x_0 = 1/y_0 = 2$, $x_t = (1/2)^{t+1} + 3/2$, and so $y_t = [(1/2)^{t+1} + 3/2]^{-1}$. Then $y_t \rightarrow 2/3$ as $t \rightarrow \infty$.

5. $x_1 = \sqrt{x_0 - 1} = \sqrt{5 - 1} = 2$, $x_2 = \sqrt{x_1 - 1} = \sqrt{2 - 1} = 1$, and $x_3 = \sqrt{x_2 - 1} = \sqrt{0} = 0$. Then $x_t = \sqrt{0 - 1} = \sqrt{-1}$, which is not a real number.

11.2

1. $a_t \approx (1.2)^t \cdot 1000 + 50 \sum_{k=1}^t (1.2)^{t-k} \approx (1.2)^t \cdot 1000 + 50 \frac{(1.2)^t - 1}{(1.2) - 1} \approx 1250(1.2)^t - 250$

2. According to $(**)$ in Example 1, the yearly repayment is $a = \frac{0.07 \cdot 100,000}{1 - (1.07)^{-30}} \approx 8058.64$. In the first year the interest payment is $0.07H = 7000$, and so the principal repayment is $\approx 8058.64 - 7000 = 1058.64$. In the last year, the

interest payment is $0.07b_{29} \approx 8058.64[1 - (1.07)^{-1}] \approx 527.20$, and so the principal repayment is $\approx 8058.64 - 527.20 \approx 7531.44$.

(a) Let the remaining debt on 1 January in year n be L_n . Argue why $L_{n-1} = L_n + \frac{1}{2}rL_{n-1}$, $n = 1, 2, \dots$
 (b) $(1-r/2)^{10}L = (1/2)L$ implies that $r = 2 - 2 \cdot 2^{-1/10} \approx 0.133934$. (c) See SM.

See SM.

3

(a) $x_{t+1} = A + B2^{t+1} = A + 2B2^t$ and $x_{t+2} = A + B2^{t+2} = A + 4B2^t$, so $x_{t+2} - 3x_{t+1} + 2x_t = A + 4B2^t - 3A - 6B2^t + 2A + 2B2^t = 0$ for all t . (b) Direct verification as in (a).

$x_t = A + Bt$ is a solution: $x_{t+2} - 2x_{t+1} + x_t = A + B(t+2) - 2[A + B(t+1)] + A + Bt = A + Bt + 2B - 2A - 2Bt - 2B + A + Bt = 0$. Substituting $t = 0$ and $t = 1$ in $x_t = A + Bt$ yields $A = x_0$ and $A + B = x_1$, with solution $A = x_0$ and $B = x_1 - x_0$. So $x_t = A + Bt$ is the general solution of the given equation.

$x_t = A3^t + B4^t$ is a solution. Substituting $t = 0$ and $t = 1$ yields $A + B = x_0$ and $3A + 4B = x_1$, with solution $A = 4x_0 - x_1$ and $B = -3x_0 + x_1$. So $x_t = A3^t + B4^t$ is the general solution of the given equation.

With $x_t = (A + Bt)2^t + 1$, we have $x_{t+2} - 4x_{t+1} + 4x_t = [A + B(t+2)]2^{t+2} + 1 - 4[A + B(t+1)]2^{t+1} + 1 + 4[(A + Bt)2^t + 1] = 4A2^t + 4Bt2^t + 8B2^t + 1 - 8A2^t - 8Bt2^t - 8B2^t - 4 + 4A2^t + 4Bt2^t + 4 = 1$. Substituting $t = 0$ and $t = 1$ in $x_t = A2^t + Bt2^t + 1$ yields $A + 1 = x_0$ and $2A + 2B + 1 = x_1$, with solution $A = x_0 - 1$ and $B = \frac{1}{2}x_1 - x_0 + \frac{1}{2}$. So $x_t = A2^t + Bt2^t + 1$ is the general solution of the given equation.

See SM.

(a) $x_t = A + Bt + u_t^*$, where $u_t^* = -\sum_{k=1}^t kc_{k-1} + t \sum_{k=1}^{\infty} c_{k-1}$. (b) $u_t^* = \frac{1}{6}(t-2)t(t-1)$. You should verify that u_t^* is a particular solution to the equation.

4

(a) $x_t = A2^t + B4^t$ (b) $x_t = A4^t + Bt4^t$ (c) $x_t = A\sqrt{3}^t \cos \theta t + B\sqrt{3}^t \sin \theta t$, where $\cos \theta = -\sqrt{3}/3$
 (d) $x_t = \left(\frac{\sqrt{6}}{3}\right)^t (A \cos \frac{\pi}{2}t + B \sin \frac{\pi}{2}t) + \frac{4}{3}$

(a) $x_t = (A + Bt)(-1)^t + 2^t$ (b) $x_t = A + B2^t + \frac{1}{4}5^t + \frac{3}{10} \cos \frac{\pi}{2}t + \frac{1}{10} \sin \frac{\pi}{2}t$

(a) $Y_t^* = b/(1-a)$ (b) $m^2 - a(1+c)m + ac = 0$. Two different real roots, a double real root, or two complex roots, according as $a(1+c)^2 - 4c > 0, = 0$, or < 0 .

If $a \neq -2$, $D = c/(a+2)$. If $a = -2$, $D = \frac{1}{2}c$. S. $D_n = Am_1^n + Bm_2^n$, where $m_{1,2} = 2(ab+1 \pm \sqrt{1+2ab})$

$x_t = u_t(-a/2)^t = (A + Bt)(-a/2)^t$, which is the result claimed for case II in Theorem 11.4.1.

(a) Stable since $|a| = 0 < 1 - \frac{1}{3}$ and $b = -\frac{1}{3} < 1$. (b) Not stable. (c) Stable.

(a) The first two equations state that consumption and capital are proportional to the net national product in the previous period. The third equation states that the net national product, Y_t , is divided between consumption, C_t , and net investment, $K_t - K_{t-1}$. (b) First replace t by $t+2$ in the last displayed equation in the problem to obtain $Y_{t+2} = C_{t+2} + K_{t+2} - K_{t+1}$. But $C_{t+2} = cY_{t+1}$, $K_{t+2} = \sigma Y_{t+1}$, and $K_{t+1} = \sigma Y_t$, so we obtain $Y_{t+2} = (c + \sigma)Y_{t+1} + \sigma Y_t = 0$. Explosive oscillations occur when $(c + \sigma)^2 < 4\sigma$ and $\sigma > 1$.

(a) $Y_t^* = \frac{a(1+g)^t}{(1+g)^2 - b(1+g) - k\beta}$ (b) $(b+k)^2 < 4k$ (c) $r = \sqrt{k}$. Damped oscillations if $k < 1$.

10. (a) The first equation states that the proportional increase in wages is equal to the proportional increase in the price index one period earlier, whereas the second equation relates prices to current wages. (b) The expression for P_t from (ii) is inserted in (i). Then the equation gives (iii). (c) $W_t = (W_0 - c\gamma/(1 - c\beta))(c\beta)^t + c\gamma/(1 - c\beta)$, ($1 \neq c\beta$). The equation is globally asymptotically stable if $|c\beta| < 1$. Thus $W_t \rightarrow \frac{c\gamma}{1 - c\beta}$ as $t \rightarrow \infty$.

11. See SM.

11.5

- (a) $x_t = A + Bt + C(-2)^t$. (The characteristic polynomial is $(r-1)^2(r+2)$.)
 (b) $x_t = (A + Bt) \cos \frac{\pi}{2}t + (C + Dt) \sin \frac{\pi}{2}t + 2$. (The characteristic polynomial is $(r^2 + 1)^2$.)
- (a) Stable. (b) Not stable. (c) Stable. (d) Not stable. 3. A triangle with corners at $(-2, 1)$, $(2, 1)$, and $(0, -1)$.
- The equation is stable on $0 < k < 1$ and $0 < b < 1$.
- Both characteristic roots complex: $\beta\sigma < 4(1 - \alpha)\alpha$. Stability: $(1 + \alpha)\beta\sigma < 4\alpha$ and $\alpha < 1$.

11.6

- (a) $x_t = \frac{2}{3} - \frac{1}{3}(-1)^t$, $y_t = \frac{3}{4} + \frac{1}{4}(-1)^t$. (From the equations we deduce that $x_{t+2} = x_t$, etc.)
 (b) $x_t = -\frac{1}{3} + \frac{1}{3}(-2)^t + \frac{1}{2}t(3-t)$, $y_t = -\frac{1}{3} + \frac{1}{3}(-2)^t$, $z_t = \frac{2}{3} + \frac{1}{3}(-2)^t + \frac{1}{2}t(t-1)$
- (a) For $\lambda = \sqrt{ab}$, $x_t = \lambda^t(A + B(-1)^t)$, $y_t = \frac{1}{a}\lambda^{t+1}(A - B(-1)^t)$.
 (b) $x_t = \lambda^t(A + B(-1)^t) + \frac{ad+ck}{k^2-ab}k^t$, $y_t = \frac{1}{a}\lambda^{t+1}(A - B(-1)^t) + \frac{bc+dk}{k^2-ab}k^t$, where $\lambda = \sqrt{ab}$.
- (a) $y_{t+1} - 0.92y_t + 0.18894y_{t-1} = 0$ (b) Solutions of the characteristic equation: $m_1 \approx 0.61$, $m_2 \approx 0.31$. Thus $y_t \approx A(0.61)^t + B(0.31)^t$, $i_t \approx 1.47y_{t+1} - 0.72y_t$, etc.

11.7

- Let $g(x) = f(x) - x$. Then $g(\xi_1) = \xi_2 - \xi_1$ and $g(\xi_2) = \xi_1 - \xi_2$ have opposite signs, and the intermediate value theorem implies that g has a zero x^* somewhere between ξ_1 and ξ_2 . Then $f(x^*) = x^*$, so x^* is an equilibrium state.
- (a) $x^* \approx -2.94753$. See SM. (b) Since $x = e^t - 3 \Leftrightarrow x = \ln(x+3)$, the positive root of $f(x) = x$ is a (stable) equilibrium state for $x_{t+1} = g(x_t)$, where $g(x) = \ln(x+3)$. Starting this time with $x_0 = 1$, we get $x_1 = 1.38629$, $x_2 = 1.47848$, $x_3 = 1.49928$, $x_4 = 1.50392$, $\dots$, converging to $x^* \approx 1.50524$.
- The cycle points are $\xi_1 = (25 - 3\sqrt{5})/10 \approx 1.82918$, $\xi_2 = (25 + 3\sqrt{5})/10 \approx 3.17082$. Since $f'(\xi_1)f'(\xi_2) = -4/5$, the cycle is locally asymptotically stable. The equilibrium states are $x_1 = (15 + \sqrt{145})/10 \approx 0.29584$, $x_2 = (15 + \sqrt{145})/10 \approx 2.70416$, with $f'(x_1) = 1 + \sqrt{29/5}$, $f'(x_2) = 1 - \sqrt{29/5} \approx -1.40832$. It follows that both equilibria are locally unstable. It is also clear that x_2 lies between ξ_1 and ξ_2 .

Chapter 12

12.1

- (a) $J_2(x) = 1 - x^2$ for $u_2^*(x) = 0$, $J_1(x) = 2 - 5x^2/3$ for $u_1^*(x) = x/5$, $J_0(x) = 3 - 21x^2/11$ for $u_0^*(x) = 5x/11$. It follows that $u_0^*(5) = 25/11$. Hence $x_1^* = 5 - 25/11 = 30/11$, so $u_1^*(30/11) = 10/11$, and finally, $u_2^*(x) = 0$ and $x_2^* = 20/11$. (b) $x_1 = 5 - u_0$, $x_2 = 5 - u_0 - u_1$. The sum in (*) is $S(u_0, u_1, u_2) = -22 - 2u_0^2 - (5 - u_0)^2 - 2u_1^2 - (5 - u_0 - u_1)^2 - 2u_2^2$. This concave function has a maximum for $u_0 = 25/11$, $u_1 = 10/11$, $u_2 = 0$.
- With $\beta = (1+r)^{-1}$, $J_T(x) = \beta^T \sqrt{x}$ with $u_T^*(x) = 1$; $J_{T-1}(x) = \beta^{T-1} \sqrt{x} \sqrt{1 + \rho\beta^2}$ with $u_{T-1}^*(x) = \frac{1}{1 + \rho\beta^2}$; $J_{T-2}(x) = \beta^{T-2} \sqrt{x} \sqrt{1 + \rho\beta^2(1 + \rho\beta^2)}$ with $u_{T-2}^*(x) = 1/(1 + \rho\beta^2(1 + \rho\beta^2))$.

(a) With $\beta = (1+r)^{-1}$, $J_T(x) = \beta^T x$ for $u_T^*(x) = 1$. For $\beta\rho < 1$, $J_{T-1}(x) = \beta^{T-1}x$ with $u_{T-1}^*(x) = 1$; for $\beta\rho \geq 1$, $J_{T-1}(x) = \rho\beta^T x$ with $u_{T-1}^*(x) = 0$. (b) For $\beta\rho < 1$, $P_t = \beta^t$; for $\beta\rho \geq 1$, $P_t = \beta^t \rho^{T-t}$. For $\beta\rho < 1$, $J_0(x) = x$ and $u_1 = \dots = u_T = 1$; for $\beta\rho \geq 1$, $J_0(x) = \beta^T \rho^T x$ and $u_1 = \dots = u_{T-1} = 0$, $u_T = 1$.

(a) and (b): $J_{T-n}(x) = (2n+3)x^2$ with $u_T^*(x) = 0$ and $u_{T-n}^*(x) = 1$ for $n = 1, \dots, T$.

$J_T(x) = \ln x$, $u_T^*(x)$ is arbitrary, $J_{T-1}(x) = \ln x + C$ with $C = \ln(3/2) \approx 1/3$. The optimal controls are $u_0^*(x) = u_1^*(x) = \dots = u_{T-1}^*(x) = 1/2$. The difference equation for x_t^* is $x_{t+1}^* = \frac{1}{2}x_t^*$, with $x_0^* = x_0$. The solution of this first-order difference equation is $x_t^* = (\frac{1}{2})^t x_0$.

(a) $J_T(x) = \max_{u \in \mathbf{R}} [x - u^2] = x$ for $u_T^*(x) = 0$. $J_s(x) = \max_{u \in \mathbf{R}} [x - u^2 + J_{s+1}(2(x \pm u))]$ for $s = 0, 1, \dots, T-1$.

(b) Use induction. $u_t^*(x) = 2^{T-t} - 1$ for $t = 0, 1, \dots, T$, and $V = J_0(x_0) = J_0(0) = \sum_{j=0}^T (2^j - 1)^2$.

(a) $J_T(x) = -\alpha e^{-\gamma x}$, $J_{T-1}(x) = -2\sqrt{\alpha}e^{-\gamma x}$, $J_{T-2}(x) = -2^{3/2}\alpha^{1/4}e^{-\gamma x}$.

(b) The difference equation for α_t is $\alpha_{t-1} = 2\sqrt{\alpha_t}$, $t = T, \dots, 1$.

(a) $J_T(x) = \sqrt{x}$ for $u = u_T^*(x) = 1$; $J_{T-1}(x) = \sqrt{x} \max_{u \in [0,1]} [\sqrt{u} + \sqrt{\rho} \sqrt{1-u}] = \sqrt{x} \sqrt{1+\rho}$ with $u_{T-1}^*(x) = 1/(1+\rho)$; $J_{T-2}(x) = \sqrt{x} \sqrt{1+\rho+\rho^2}$ with $u_{T-2}^*(x) = 1/(1+\rho+\rho^2)$. (b) Use induction.

2

(a) $F(t, x_t, x_{t+1}) = 1 - x_t^2 - 2(x_t - x_{t+1})^2$ for $t = 0, 1$; $F(2, x_2, x_3) = 1 - x_2^2$. (b) The Euler equation is $x_2 - (5/2)x_1 + x_0 = 0$ (for $t = 0$), $x_2 - (2/3)x_1 = 0$ (for $t = 1$). With $x_0 = 5$ we find $x_1 = 30/11$ and $x_2 = 20/11$, as in Problem 12.1.1.

The problem is $\max \sum_{t=0}^{T-1} (-2/3)(x_{t+1}/x_t - 1) + \ln x_T$. The Euler equation gives $(-2/3)x_{T-1}^{-1} + x_T^{-1} = 0$ and $(2/3)x_{t+1}/x_t^2 - (2/3)1/x_t = 0$, which yields (a) $x_T = (3/2)x_{T-1}$ and (b) $x_{t+2} = x_{t+1}^2/x_t$ for $t < T-2$. For $t = T-2$, using (a) and (b), yields $x_{T-1} = (3/2)x_{T-2}$. Next, inserting the latter result in (b) when $t = T-3$, yields $x_{T-2} = (3/2)x_{T-3}$. And so on.

3

Inserting $J(x) = -\alpha e^{-x}$ into the Bellman equation yields $-\alpha e^{-x} = \max_{u \in \mathbf{R}} \{-e^{-u} - \frac{1}{2}e^{-x} - \alpha\beta e^{-2x+u}\}$. The maximizing u is $u^* = x - \frac{1}{2} \ln(\alpha\beta)$, and the equation reduces to $\alpha = 2\sqrt{\alpha\beta} + \frac{1}{2}$. Then $\sqrt{\alpha} = \sqrt{\beta} + \sqrt{\beta+1/2}$, and so $\alpha = (\sqrt{\beta} + \sqrt{\beta+1/2})^2$. For optimality, see SM.

(a) $(3\alpha - 2)(1 + \alpha\beta) = 3\alpha\beta$, and the only positive solution is $\alpha = \frac{5\beta - 3 + \sqrt{(5\beta - 3)^2 + 24\beta}}{6\beta}$.

$u^*(x) = -\alpha\beta x/(1 + \alpha\beta)$. (b) Note 2 applies for $x \in \mathcal{X}(x_0) \subseteq [-x_0, x_0]$, $u \in [-x_0, x_0]$.

4

$H = 1 - (x^2 + 2u^2) + p(x - u)$ for $t = 0, 1$ and $H = 1 - (x^2 + 2u^2)$ for $t = 2$. Condition (3) yields $p_0 = -4u_0^*$ and $p_1 = -4u_1^*$, and condition (4) gives $p_0 = -2x_1^* + p_1$ and $p_1 = -2x_2^*$. Now, $x_1^* = x_0^* - u_0^* = 5 - u_0^*$ and $x_2^* = x_1^* - u_1^* = 5 - u_0^* - u_1^*$. From all these equations we get $u_0^* = 25/11$, $u_1^* = 10/11$, $u_2^* = 0$. (Start by eliminating p_0 and p_1 .) The Hamiltonian is concave in (x, u) , so we have found the solution.

(a) $I = \sum_{t=0}^T (u_t^2 - 2x_t^2) = u_0^2 - 2x_0^2 + \sum_{t=1}^T (u_t^2 - 2x_t^2) = u_0^2 + \sum_{t=1}^T (u_t^2 - 2u_{t-1}^2) = -u_0^2 - u_1^2 - u_2^2 - \dots - u_{T-1}^2 + u_T^2$. I is maximized when $u_0^* = u_1^* = \dots = u_{T-1}^* = 0$ and $u_T^* = \pm 1$. (b) $p_t = 0$. (c) The Hamiltonian is minimized.

5

(a) $I = 12 - 2u_0 - 2u_0^2 - u_1 - 2u_1^2 - 2u_2^2 - 2u_3 - u_3^2 - v_1 - v_1^2 - v_2^2$. Since I is concave as a sum of concave functions, the unique stationary point $(u_0, v_0, u_1, v_1, u_2, v_2) = (-\frac{1}{2}, -1, -\frac{1}{4}, -\frac{1}{2}, 0, 0)$ solves the problem.

(b) $J_2(x, y) = 1 + x - y$ for $u_2^* = v_2^* = 0$; $J_1(x, y) = \frac{19}{8} + 2x - 2y$ for $u_1^* = -\frac{1}{4}$, $v_1^* = -\frac{1}{2}$, and $J_0(x, y) = \frac{39}{8} + 3x - 3y$ for $u_0^* = -\frac{1}{2}$, $v_0^*(x, y) = -1$. (c) See SM.

2. The Hamiltonian is $H = -x^2 - u^2 + py + q(y + u)$ for $t < T$ and $H = -x^2 - u^2$ for $t = T$, so it is concave in (x, y, u) . The conditions in Theorem 12.5.1, using Note 3, are therefore sufficient:

(i) $-2u_t^* + q_t = 0$ for $t < T$, $-2u_T^* = 0$ (ii) $p_{T-1} = -2x_T^* + p_T$, with $p_T = 0$

(iii) $q_{t-1} = p_t + q_t$ for $t < T$, $q_{T-1} = q_T$, with $q_T = 0$ (iv) $x_{t+1}^* = y_t^*$, $y_{t+1}^* = y_t^* + u_t^*$.

Derive the equation $x_{t+1}^* - 3x_{t+1}^* + x_t^* = 0$. For $t < T$, the solution is $x_t^* = Am_1^t + Bm_2^t$, where $m_{1,2} = \frac{1}{2}(3 \pm \sqrt{5})$. Moreover, $x_0^* = x_0^0$, $x_1^* = y_0^* = y_0^0$, and $y_1^* = x_{t+1}^*$, $p_t = -2x_{t+1}^*$, $q_t = 2(x_{t+2}^* - x_{t+1}^*)$, and $u_t^* = \frac{1}{2}q_t$, with $u_T^* = 0$.

3. $x_t^* = \beta^t x_0$, $u_t^* = \beta^{t+1} x_0$. (With $H = \beta^t \ln(x - u) + pu$, the conditions are: (i) $\partial H^*/\partial u = p_t - \beta^t/(x_t^* - u_t^*) = 0$, (ii) $p_{t-1} = \partial H^*/\partial x = \beta^t/(x_t^* - u_t^*)$, (iii) $\lim_{t \rightarrow \infty} p_t(x_t - x_t^*) \geq 0$, (iv) $x_{t+1}^* = u_t^*$, $x_0^* = x_0$. From (i) and (ii) we get $p_{t-1} = p_t$, so $p_t = \bar{p}$, a constant. From (i) and (iv), $x_{t+1}^* = x_t^* - \beta^t/\bar{p}$, with solution $x_t^* = x_0 + (\beta^t - 1)/\bar{p}(1 - \beta)$. The first term in (iii), $\bar{p}x_t$, is positive. Moreover, $-\bar{p}x_t^* = -\bar{p}[x_0 + (\beta^t - 1)/\bar{p}(1 - \beta)] \rightarrow 0$ iff $\bar{p} = 1/x_0(1 - \beta)$, and then $x_t^* = \beta^t x_0$. Note that $x_t^* - u_t^* = \beta^t x_0(1 - \beta) > 0$.)

12.6

1. $\alpha_{t-1} = 2(\alpha_t K)^{1/2}$, $u_{t-1} = x - (1/2\gamma) \ln(\alpha_t K)$; $\alpha_T = \delta$. (If $J_t(x) = -\alpha_t e^{-\gamma x}$, writing $x_{t-1} = x$, $u_{t-1} = u$, and $V_t = V$, the optimality equation is $J_{t-1}(x) = -\alpha_{t-1} e^{-\gamma x} = \max_u \{-e^{-\gamma u} - \alpha_t K e^{-\gamma(2x-u+V)}\} = \max_u \{-e^{-\gamma u} - \alpha_t K e^{-\gamma(2x-u)}\}$. The first-order condition is $\gamma e^{-\gamma u} - \gamma \alpha_t K e^{-2\gamma x + \gamma u} = 0$, with solution $u = u_{t-1} = x - (1/2\gamma) \ln(\alpha_t K)$, so $-\alpha_{t-1} e^{-\gamma x} = -e^{\frac{1}{2} \ln(\alpha_t K)} e^{-\gamma x} - \alpha_t K e^{-\gamma x} e^{-\frac{1}{2} \ln(\alpha_t K)}$. Hence, $\alpha_{t-1} = 2(\alpha_t K)^{1/2}$, $\alpha_T = \delta$.)

2. Optimal control: $C_t = \frac{A_t}{k_t}$ and w_t , where w_t is a solution of $E\left[\frac{r_t - V_t}{1 + V_t + (r_t - V_t)w_t}\right] = 0$ (no explicit formula is available). Here k_t is governed by $k_t = 1 + k_{t+1}/(1 + \theta)$, $k_T = k$. Value function: $J_t(A_t) = (1 + \theta)^{-1} k_t \ln A_t + b_t$, where for $t < T$, $b_t = -\ln k_t + (1 + \theta)^{-1} k_{t+1} \left[\ln \frac{k_{t+1}}{1 + \theta} - \ln \left(1 + \frac{k_{t+1}}{1 + \theta}\right) + d_t \right] + (1 + \theta)^{-1} b_{t+1}$, with $b_T = 0$ and $d_t = E \ln[(1 + r_t)w_t + (1 + V_t)(1 - w_t)]$.

3. $J_t(x_t) = k_t + a_t x_t$, $a_t = a 2^{t-T}$, $k_{t-1} = k_t + 2/a_t$, $k_T = 0$, $u_t = 4/a_t^2$.

4. The optimality equations are $-b_{t-1}x^2 = J_{t-1}(x, 0) = \max_u \{-u^2 - (1/4)a_t(x + u)^2 - (3/4)b_t u^2\}$, and $-a_{t-1}x^2 = J_{t-1}(x, 1) = \max_u \{-u^2 - (3/4)a_t(x + u)^2 - (1/4)b_t u^2\}$. The optimal controls are $u_{t-1}(x, 0) = -a_t a_t x$, where $d_t = 1/(4 + a_t + 3b_t)$, with $b_{t-1} = a_t^2 d_t^2 + (1/4)a_t(4 + 3b_t)^2 d_t^2 + (3/4)b_t a_t^2 d_t^2$, and $u_{t-1}(x, 1) = -3a_t c_t x$, where $c_t = 1/(4 + 3a_t + b_t)$, and $a_{t-1} = 9a_t^2 c_t^2 + (3/4)a_t(4 + b_t)^2 c_t^2 + (9/4)b_t a_t^2 c_t^2$.

5. (a) and (b): $J_t(x) = \max\{x^2, a_t + 2x^2\}$ where $a_{t-1} = -1 + \frac{1}{4}a_t$, $a_{T-1} = -1$, $u_t = 0$ if $a_t + x^2 \leq 0$, $u_t = 1$ if $a_t + x^2 > 0$.

6. $u_t = x_t/(1 + a_{t+1}^2)$, $a_t = (1 + a_{t+1}^2)^{1/2}$, $a_T = a/2$.

7. $u_t = \bar{u} = (\hat{q} - \bar{p})/(\hat{q} + \bar{p})$, where $\hat{q} = q^{-1/\alpha}$, $\bar{p} = p^{-1/\alpha}$, $J_t(x) = A_t x^{1-\alpha}$, $A_{t-1} = A_t[p(1 + \bar{u})^{1-\alpha} + q(1 - \bar{u})^{1-\alpha}]$.

8. $x_1 = \frac{1}{2} - v_0$, $x_2 = 1 - v_0 - v_1$, $x_3 = \frac{3}{2} - v_0 - v_1 - v_2$.

12.7

1. (a) $a = [1 - 2\beta - \sqrt{1 + 4\beta^2}]/2\beta$, $b = \hat{a}\beta d/(1 - \beta)$. (b) See SM.

2. The optimal control is $u_t(x) = (1 - \alpha)(1 + \alpha)^{-1}x$, and the value function is $J(x) = a \ln x + b$, where $a = (2 - \alpha)^{-1}$, $b = [\alpha d + \alpha a \ln(\alpha a) - (1 + \alpha a) \ln(1 + \alpha a)](1 - \alpha)^{-1}$, and $d = E \ln V$.

Chapter 13

13.1

1. $d(x, y) \geq |x_j - y_j|$ for all j , so if $d(x, y) < r$, then $|x_j - y_j| < r$, which means that $-r < x_j - y_j < r$.

2. By the triangle inequality, $d(x, x) \leq d(x, y) + d(y, x)$, so $d(x, x) - d(x, y) \leq d(x, y)$. Moreover, $d(x, y) \leq d(x, x) + d(x, y)$, so $-d(x, x) \leq d(x, x) - d(x, y)$. This proves the desired inequality. (Recall that $|u| \leq b \iff -b \leq u \leq b$.)

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^n (x_j - y_j)^2} \leq \sum_{j=1}^n \sqrt{(x_j - y_j)^2} = \sum_{j=1}^n |x_j - y_j|$$

S_1 is open and bounded. S_2 is closed and bounded. S_3 is closed and unbounded. S_4 is neither open nor closed, but bounded. S_5 is open and unbounded.

See Figure A13.1.5. S is closed because it contains all its boundary points, which are the points on the curve.

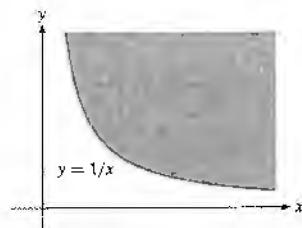


Figure A13.1.5

(a) The points $(1/n, 1)$, $n = 1, 2, \dots$, all belong to E , but $\lim_n (1/n, 1) = (0, 1)$ does not. Hence E is not closed.
(b) F is closed. (The point $(0, 0)$ is the only accumulation point of F —i.e. the only point with the property that every open ball around the point contains infinitely many points from F . Fortunately, $(0, 0) \in F$.)

Closedness: The points $(2n, 1)$ are boundary points of A and B , but do not belong to either of those sets, so A and B are not closed. C is closed because it contains all its boundary points. Openness: A and C are not open since no ball (two-dimensional disk) at all can be contained in A or C . B is obviously open, because it is the union of a family of open rectangles.

$\partial S = \text{cl}(S) \cap \text{cl}(\mathbb{R}^n \setminus S)$ is the intersection of two closed sets, and is therefore closed. 9. See SM.

(a) The interior of an open set is the set itself (see the remark just before Theorem 13.1.1 in the main text). Hence, for every U in $\mathcal{U}$, we have $U = \text{int}(U) \subseteq \text{int}(S)$ (by Problem 9(a)). Also, $\text{int}(S)$ is open and contained in S , so $\text{int}(S) \in \mathcal{U}$. It follows that $\text{int}(S) \subseteq \bigcup_{U \in \mathcal{U}} U \subseteq \text{int}(S)$, so $\text{int}(S) = \bigcup_{U \in \mathcal{U}} U$.

(b) Similarly, every closed set is its own closure (see Problem 9(b)), so for every F in $\mathcal{F}$, we have $F = \text{cl}(F) \supseteq \text{cl}(S)$. Moreover, $\text{cl}(S)$ is closed and contains S , so $\text{cl}(S) \in \mathcal{F}$. It follows that $\text{cl}(S) \supseteq \bigcap_{F \in \mathcal{F}} F \supseteq \text{cl}(S)$, and so $\text{cl}(S) = \bigcap_{F \in \mathcal{F}} F$.

Let $A_k = \mathbb{C}B_{1/k}(0) = \{\mathbf{x} : \|\mathbf{x}\| \geq 1/k\}$. Since $B_{1/k}(0)$ is open, A_k is closed for each $k = 1, 2, \dots$. But the union $A = \bigcup_k A_k = \mathbb{C}(\bigcap_k B_{1/k}(0)) = \{\mathbf{x} : \mathbf{x} \neq \mathbf{0}\} = \{\mathbf{x} : \mathbf{x} \neq \mathbf{0}\}$ is not closed, since $\mathbf{0}$ belongs to $\text{cl}(A)$ but not to A itself.

Let $\{F_i\}_{i \in I}$ be an arbitrary family of closed sets. To prove (b), note that $\bigcap_i F_i = \mathbb{C}(\bigcup_i (\mathbb{C}F_i))$. Since each $\mathbb{C}F_i$ is open, the union $\bigcup_i (\mathbb{C}F_i)$ is open, and therefore its complement is closed. Part (c) follows in the same way from the identity $\bigcup_i F_i = \mathbb{C}(\bigcap_i (\mathbb{C}F_i))$. (In (c) we must assume that the index set I is finite.)

$\partial \mathbb{Q} = \mathbb{R}$, because arbitrarily close to any number in $\mathbb{R}$, we can find a rational number. Hence, $\bar{\mathbb{Q}} = \mathbb{Q} \cup \partial \mathbb{Q} = \mathbb{R}$. Since every open interval contains irrational numbers, $\text{int}(\mathbb{Q}) = \emptyset$.

If $\emptyset \neq S \neq \mathbb{R}^n$, let $\mathbf{a} \in S$, $\mathbf{b} \in \mathbb{C}S$. For μ in $[0, 1]$, let $\mathbf{c}_\mu = \mathbf{a} + \mu(\mathbf{b} - \mathbf{a})$. Then $\mathbf{c}_0 = \mathbf{a} \in S$, but $\mathbf{c}_1 = \mathbf{b} \notin S$. Let $\lambda = \sup\{\mu \in [0, 1] : \mathbf{c}_\mu \in S\}$. Then $\mathbf{c}_\lambda \in \partial S \subseteq \text{cl} S$, but $\mathbf{c}_\lambda \notin \text{int} S$. Hence, $\text{int} S \neq \text{cl} S$, so S cannot be both open and closed.

(a) and (c) are false, (b) and (d) are true. See SM.

2

(a) $\mathbf{x}_k \rightarrow (0, 1)$ (b) $\mathbf{x}_k$ does not converge. (c) $\mathbf{x}_k \rightarrow (\frac{1}{3}, 0)$ (d) $\mathbf{x}_k \rightarrow (1, e)$. (Recall that $(1 + 1/k)^k \rightarrow e = 2.71828 \dots$ as $k \rightarrow \infty$.)

2. Suppose that the sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}$, and that $\mathbf{y} \neq \mathbf{x}$. Let $r = d(\mathbf{x}, \mathbf{y}) > 0$. There exists a natural number K such that for all $k > K$, $\mathbf{x}_k$ lies in $B(\mathbf{x}; r/2)$, which is disjoint from $B(\mathbf{y}; r/2)$. It follows that the sequence cannot converge to $\mathbf{y}$.

3. Suppose $\{\mathbf{x}_k\}$ is a sequence in $\mathbb{R}^n$ converging to $\mathbf{x}$. Given $\varepsilon > 0$, choose a natural number N such that $\|\mathbf{x}_k - \mathbf{x}\| < \varepsilon/2$ for all $k > N$. Then for $k > N$ and $m > N$, $\|\mathbf{x}_k - \mathbf{x}_m\| = \|(\mathbf{x}_k - \mathbf{x}) + (\mathbf{x} - \mathbf{x}_m)\| \leq \|\mathbf{x}_k - \mathbf{x}\| + \|\mathbf{x} - \mathbf{x}_m\| < \varepsilon/2 + \varepsilon/2 = \varepsilon$. Therefore $\{\mathbf{x}_k\}$ is a Cauchy sequence.

4. If S is unbounded, then for every k there exists a point $\mathbf{x}_k$ in S with $\|\mathbf{x}_k\| > k$. No subsequence of $\{\mathbf{x}_k\}$ can be convergent.

5. *Hint:* Show that if the sequence $\{\mathbf{x}_k\}$ does not converge to $\mathbf{x}^0$, then it has a subsequence converging to a point different from $\mathbf{x}^0$. Remember that the set X is compact. See SM.

6. You need to show that $A \times B$ is closed and bounded. See SM for details.

13.3

1. Define the continuous functions f and g from $\mathbb{R}^2$ into $\mathbb{R}$ by $f(x, y) = 2x - y$ and $g(x, y) = x - 3y$. Both these functions are continuous, and so the sets $f^{-1}((-\infty, 2))$ and $g^{-1}((-\infty, 5))$ are both open. S is open as the intersection of these open sets.

2. $S = \bigcap_i g_i^{-1}((-\infty, 0])$. Use Theorem 13.3.4(b) and Theorem 13.1.2(b).

3. (a) $S = \mathbb{R}$, $f(x) = e^x$, $f(S) = (0, \infty)$ (b) $S = \mathbb{R}$, $f(x) = e^{x^2}$, $f(S) = [1, \infty)$ (c) $S = (0, 1)$, $f(x) = 1/x$

4. For any $\varepsilon > 0$, let $\delta = \varepsilon$. Then if $d(\mathbf{x}, \mathbf{y}) < \delta$, it follows from Problem 13.1.2 that $|f(\mathbf{x}) - f(\mathbf{y})| = |d(\mathbf{x}, \mathbf{a}) - d(\mathbf{y}, \mathbf{a})| \leq d(\mathbf{x}, \mathbf{y}) < \varepsilon$.

5. The intersection of S and $\bar{B}(\mathbf{y}; r)$ is closed and bounded, hence compact. Then $h(\mathbf{x})$ attains a minimum at some point $\mathbf{x}^0$ in $S \cap \bar{B}(\mathbf{y}; r)$, and $\mathbf{x}^0$ is the required point.

6. If f is not continuous at $\mathbf{x}^0$, then there exists an $\varepsilon > 0$ such that for every $k = 1, 2, \dots$ there exists a point $\mathbf{x}_k$ in $B(\mathbf{x}^0, 1/k)$ with $\|f(\mathbf{x}_k) - f(\mathbf{x}^0)\| > \varepsilon$. The sequence $\{\mathbf{x}_k\}$ converges to $\mathbf{x}^0$, but $\{f(\mathbf{x}_k)\}$ does not converge to $f(\mathbf{x}^0)$.

7. See SM. 8. You need to consider the ε - δ definition of continuity. See SM.

13.4

1. (a) S is compact and f is continuous, so by Theorem 13.4.1, V is continuous.

(b) The closure of S is $\text{cl}(S) = \{(y, z) : y \geq 0, z \geq 0, y^2 + z^2 \leq 4\}$. This set is compact, i.e. closed and bounded. Hence, for any x , $f(x, y, z)$ will attain a maximum value at some point (y_x, z_x) in $\text{cl}(S)$. Obviously, $y_x > 0$ and $z_x > 0$, so $(y_x, z_x) \in S$. This implies that the supremum value of $f(x, y, z)$ over S is attained at (y_x, z_x) . Hence, $V(x) = \max_{(y,z) \in S} f(x, y, z)$, and by Theorem 13.4.1, V is continuous.

2. (a) Theorem 13.4.1 implies that V_1 is continuous. (b) Let $\varphi(x, u) = e^{-xu^2} - (u - x)^2$, so that $V_2(x) = \max_{u \in \mathbb{R}} \varphi(x, u)$. It is clear that if $x < 0$, then $V_2(x) = \infty$. If $x \geq 0$, then $\varphi'_x(x, u) \geq 0$ for $u \leq 0$ and $\varphi'_x(x, u) < 0$ for $u > x$. Therefore $V_2(x) = \max_{u \in \mathbb{R}} \varphi(x, u) = \max_{u \in [0, x]} \varphi(x, u)$, so for $x > 0$, Theorem 13.4.2 yields continuity.

3. For each fixed x , study where $f(x, y)$ is increasing with respect to y and where it is decreasing. See SM.

4. We have $f'_y(x, y) = \frac{xe^y}{1 + xe^y} - 2y$, so it is clear that $f'_y(x, y) > 0$ for all $y \leq 0$. Hence, for each value of x , the function $f(x, y)$ attains its maximum with respect to y in Y at some point of $[0, 1]$. (Since $f(x, y)$ is strictly concave with respect to y , the maximum point is also unique.) Therefore $V(x) = \max_{y \in [0, 1]} f(x, y)$. Since $[0, 1]$ is compact, Theorem 13.4.1 shows that V is continuous.

.5

(a) $S + T$ is the rectangle (Cartesian product of intervals) $(a, b) \times (c, d)$ with corners (a, c) , (b, c) , (b, d) , and (a, d) . (b) $S + T$ is S translated so that its south-west corner coincides with T —i.e. $S + T$ is the rectangle $[b, b + a] \times [b, b + a]$.

(c) $\text{co}(S)$ = the set of all points in the triangle with vertices at $(0, 0)$, $(a, 0)$, and $(0, a)$.

(d) $\text{co}(S)$ is the closed square with the four points as extreme points.

Let $\lambda_i = n_i/n$ for $i = 1, \dots, n$. Then the λ_i are all nonnegative and their sum is 1. Hence $\mathbf{z} = \sum_{i=1}^n \lambda_i \mathbf{x}_i$ is a convex combination of the points $\mathbf{x}_i$. It is called the *barycentre*, or *centre of gravity*, of the $\mathbf{x}_i$ (with weights $\lambda_i = n_i/n$).

$\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} = \sum_{i=1}^p \lambda \lambda_i \mathbf{u}_i + \sum_{j=1}^q (1 - \lambda) \mu_j \mathbf{v}_j$, and $\sum_{i=1}^p \lambda \lambda_i + \sum_{j=1}^q (1 - \lambda) \mu_j = \lambda + (1 - \lambda) = 1$.

An interior point evidently lies “properly inside” some line segment in the convex set.

Mathematical induction.

See SM. (Hint: Every point in $\text{co}(S)$ can be written as a convex combination of exactly $n + 1$ points (not necessarily distinct) in S .)

6

Use Theorem 13.6.1 to find a hyperplane that strictly separates S from the origin.

For every λ in $[0, 1]$, we have $\lambda \mathbf{x}_k + (1 - \lambda) \mathbf{y}_k \in S$, and therefore $\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} = \lim_k (\lambda \mathbf{x}_k + (1 - \lambda) \mathbf{y}_k) \in \bar{S}$.

Hint: If $\mathbf{x}$ is interior in $\bar{S}$, there exists an n -cube contained in $\bar{S}$, with $\mathbf{x}$ as its centre point. Arbitrarily close to each corner $\mathbf{z}_j$ of the cube there is a point $\mathbf{y}_j$ from S , and when the $\mathbf{y}_j$ are sufficiently close to $\mathbf{z}_j$, then evidently $\mathbf{x}$ lies in the interior of the convex hull of the points $\mathbf{y}_j$. See SM for a different argument.

$\mathbf{y}$ is *not* necessarily a boundary point of $\bar{S}$. If $S \approx \mathbb{Q}$, the set of rational numbers, then $\bar{S} = \bar{\mathbb{Q}} = \mathbb{R}$, the whole real line, and $\sqrt{2}$ is certainly not a boundary point of $\mathbb{R}$. For a different example, consider the closed unit disk $D = \{(x, y) : x^2 + y^2 \leq 1\}$ in $\mathbb{R}^2$, and let D_* be the punctured unit disk $D_* = \{(x, y) : 0 < x^2 + y^2 \leq 1\}$, i.e. D with the origin removed. Then $(0, 0)$ is a boundary point of D_* , but not of its closure $\bar{D}_* = D$.

The statement is false. Consider two intersecting straight lines in $\mathbb{R}^2$.

7

(a) We easily find that $(\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{7} \begin{pmatrix} 12 & 9 \\ 2 & 12 \end{pmatrix}$, and from Theorem 13.7.2 it follows that $\mathbf{A}$ is productive.

(b) Let $\mathbf{1}$ be the $n \times 1$ matrix (column vector) with all elements equal to 1. Then $\mathbf{A}\mathbf{1}$ is the vector of row sums in $\mathbf{A}$, so $\mathbf{A}\mathbf{1} \ll \mathbf{1}$. Hence $\mathbf{A}$ is productive by the definition in (1).

(a) $\lambda = 3$ and $(1, 1)'$ is an associated eigenvector. (b) $\lambda = \frac{1}{6}(2 + \sqrt{2})$. An associated eigenvector is $(3, \sqrt{2})'$.

(c) $\lambda = 4$, and $(2, 3, 1)'$ is an associated eigenvector.

Chapter 14

1

The domain of F is closed, and the graph is closed. F is obviously not upper hemicontinuous at $x = 0$.

(a) The graph is closed, so the correspondence has the closed graph property. It is not lower hemicontinuous.

(b) Closed graph property, but not lower hemicontinuous. (c) Not closed graph property, but lower hemicontinuous.

The domain of F is closed, and the graph is closed. To show l.h.c., for any $y^0 \in F(x^0)$, if $x_n \rightarrow x^0$, choose $y_n = \min\{y^0, 1/x_n\}$ (let $y_n \rightarrow y^0$ if $x_n = 0$). Then $y_n \rightarrow y^0$, with $y^0 \in F(x^0)$.

4. See SM.

5. For a sufficiently large number $\alpha > 0$, both $F(x^0)$ and $G(x^0)$ are contained in the ball $B(0; \alpha)$. Then there exists a $\delta > 0$ such that $F(x) \subseteq B(0; \alpha + 1)$ and $G(x) \subseteq B(0; \alpha + 1)$ whenever $\|x - x^0\| < \delta$. Then $H(x) \subseteq B(0; 2\alpha + 2)$ for such x . Hence, by Theorem 14.1.3 it suffices to prove that $H(x)$ has a closed graph at x^0 . Let $x_n \rightarrow x^0$, $h_n \rightarrow h$, $h_n \in H(x_n)$. Then $h_n = f_n + g_n$ for some $f_n \in F(x_n)$ and $g_n \in G(x_n)$. A subsequence (f_{n_j}, g_{n_j}) converges to some point $(f, g) \in F(x^0) \times G(x^0)$. Then $h = f + g \in H(x^0)$.

6. Use the characterization of l.h.c. in (14.1.8). See SM for details.

7. The sequence property described in the problem implies that F has a closed graph at x^0 . Thus by Theorem 14.1.2, it suffices to show that $F(x)$ is locally bounded near x^0 . If it is not, then for every k there is a pair x^k, y^k such that $\|x^k - x^0\| < 1/k$, $\|y^k\| > k$, and $y^k \in F(x^k)$. But $\{y^k\}$ contains no convergent subsequence. (It is also true that if F is compact-valued and upper hemicontinuous at x^0 , then the sequence property in the problem holds.)

8. For each i we have $g_i(x, y) < b_i$ and $g_i(x, y') \leq b_i$, so for every λ in $(0, 1)$ we get $g_i(x, \lambda y + (1 - \lambda)y') \leq \lambda f_i(x, y) + (1 - \lambda)g_i(x, y') < \lambda b_i + (1 - \lambda)b_i = b_i$. It follows that $\lambda y + (1 - \lambda)y' \in \mathcal{P}^0(x)$ for all λ in $(0, 1)$. Therefore $y' = \lim_{\lambda \rightarrow 0} (\lambda y + (1 - \lambda)y') \in \mathcal{P}^0(x)$, and by Example 14.1.5, $\mathcal{P}$ is l.h.c.

9. For a point (x, y) with $x \in X$ and $y \in \mathcal{P}(x)$, let $\mathbf{A} = \{\partial g_i(x, y)/\partial x_j\}_{i \in I}$, where $I = \{i : g_i(x, y) = b_i\}$. Then $\mathbf{A}z = \mathbf{1}$ has a solution z . Hence, for all $\delta > 0$ that are small enough, $g_i(x - \delta z, y) < 0$ for all i in I , and also for the other i 's, so $y \in \mathcal{P}^0(x)$.

10. See SM. 11. Use Problem 13.5.7. See SM.

14.2

$$1. V(x) = \begin{cases} -x & \text{if } x \leq 0 \\ x^2/4 - x & \text{if } x \in (0, 2] \\ -1 & \text{if } x > 2 \end{cases}, \quad Y^*(x) = \begin{cases} \{0\} & \text{if } x \leq 0 \\ \{\sqrt{x/2}, \sqrt{x/2}\} & \text{if } x \in (0, 2] \\ \{-1, 1\} & \text{if } x > 2 \end{cases}.$$

2. It is easily seen that $\mathcal{B}(p, m)$ is continuous for $m \geq 0$, $p \gg 0$. If some $p_i = 0$, $\mathcal{B}(p, m)$ becomes unbounded.

3. For $p \gg 0$, $m > 0$, the demand correspondence is upper hemicontinuous and the indirect utility function is continuous. Quasiconcavity entails convexity of each set $\xi(p, m)$.

14.4

1. $\frac{1}{2}(x^* + 1) = x^*$ would imply $x^* = 1$. The interval $(0, 1)$ is not closed, so Brouwer's theorem does not apply.

2. Both E and B are compact. T_B has the origin as its only fixed point, whereas T_E has no fixed point. (Compare the sheep example illustrated in Figs. 14.4.3 and 14.4.4.)

3. Suppose $x \in \Delta^{n-1}$. Then the i th component of the vector $\mathbf{A}x$ is equal to $\sum_{j=1}^n a_{ij}x_j$, which is ≥ 0 because all the x_j and all the a_{ij} are nonnegative. Moreover, the sum of all the components of $\mathbf{A}x$ is $\sum_{i=1}^n (\sum_{j=1}^n a_{ij}x_j) = \sum_{j=1}^n x_j \sum_{i=1}^n a_{ij} = \sum_{j=1}^n x_j = 1$. Thus the linear, and therefore continuous, transformation $x \mapsto \mathbf{A}x$ maps Δ^{n-1} into itself. By Brouwer's theorem there exists an x^* in S such that $\mathbf{A}x^* = x^*$. Thus $\lambda = 1$ is an eigenvalue for $\mathbf{A}$, and x^* is an eigenvector.

4. F has the closed graph property, but has no fixed point. Kakutani's theorem does not apply because $F(1)$ is not convex. See Fig. A14.4.4.

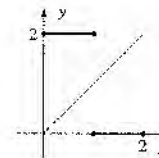


Figure A14.4.4

5. We shall show that x^* is a fixed point for f .

- (1) If $x^* > 0$, then for every natural number n , we can find an x_n in A with $x^* - 1/n < x_n \leq x^*$. Since $f(x_n) \geq x_n$, we get $f(x^*) \geq \lim_{n \rightarrow \infty} f(x_n) \geq \lim_{n \rightarrow \infty} x_n = x^*$.
- (2) If $x^* < 1$, then $f(x) < x$ for every $x > x^*$, and so $f(x^*) \leq \lim_{n \rightarrow \infty} f(x_n) \leq \lim_{n \rightarrow \infty} x_n = x^*$.
- (3) It follows from (1) and (2) that, if $0 < x^* < 1$, then $f(x^*) \leq x^* \leq f(x^*)$, so $f(x^*) = x^*$.
- (4) It also follows that, if $x^* = 0$, then $0 \leq f(0) = f(x^*) \leq x^* = 0$, and if $x^* = 1$, then $1 \geq f(x^*) \geq x^* = 1$. Hence, in every case, we get $f(x^*) = x^*$.

4.5

1. The budget set does not change if p is replaced by λp , with $\lambda > 0$.

appendix A

1.

1. Suppose that $\{\{a\}, \{a, b\}\} = \{\{c\}, \{c, d\}\}$. There are two cases to consider: $a = b$ and $a \neq b$. If $a = b$, then $\{a, b\} = \{a\}$, and so $\{\{c\}, \{c, d\}\} = \{\{a\}, \{a, b\}\} = \{\{a\}\}$. But then $\{c, d\} = \{c\} = \{a\}$, so $c = d = a = b$. If $a \neq b$, then $\{a, b\}$ is a two-element set, and we must have $\{c\} = \{a\}$ and $\{c, d\} = \{a, b\}$. This shows that $c = a$, and therefore $d = b$.

2. $\text{dom}(R^{-1}) = \text{range}(R)$ because $b \in \text{dom}(R^{-1}) \iff bR^{-1}a$ for an a in $A \iff aRb$ for an a in $A \iff b \in \text{range}(R)$. If we now apply the equation $\text{range}(R) = \text{dom}(R^{-1})$ to R^{-1} instead of R , we get $\text{range}(R^{-1}) = \text{dom}((R^{-1})^{-1}) = \text{dom}(R)$ because $(R^{-1})^{-1} = R$.

3. $D_f = \{0, 1\}$, $f(0) = 0$, $f(1) = 0$, $S_1 = \{0\}$, $S_2 = \{1\}$.

4. A relation is a linear ordering if and only if it is (i) reflexive, (ii) transitive, (iii) anti-symmetric, and (iv) complete. For each of these four properties it is easy to see that if a relation R has that property, then so has R^{-1} .

5. The inclusions $S \subseteq f^{-1}(f(S))$ and $f(f^{-1}(T)) \subseteq T$ follow immediately from the definitions of direct and inverse images. If $x \in f^{-1}(T)$, then $f(x) \in T$ and $x \in \text{range}(f)$, and it follows that $f^{-1}(T) \subseteq f^{-1}(T \cap \text{range}(f))$. The opposite inclusion is obvious, since $T \cap \text{range}(f) \subseteq T$. To show that the first two inclusions cannot always be replaced by equality signs, define $f: \mathbb{R} \rightarrow \mathbb{R}$ by $f(x) = x^2$, and let $S = [0, 2]$ and $T = [-1, 1]$. Then $f^{-1}(f(S)) = f^{-1}([0, 4]) = [-2, 2] \neq S$, and $f(f^{-1}(T)) = f([-1, 1]) = [0, 1] \neq T$.

2.

1. $\sup A = 7$, $\inf A = -3$; $\sup B = 1$, $\inf B = 0$; $\sup C = \infty$, $\inf C = \sqrt{3}$.

2. It follows from $r^2 > 2$ that $s = (2 + r^2)/2r < (r^2 + r^2)/2r = r$. Moreover, $s^2 - 2 = (4 + 4r^2 + r^4)/4r^2 - 2 = (4 - 4r^2 + r^4)/4r^2 = (2 - r^2)^2/4r^2 > 0$.

3. $\sup S = \infty \iff S$ is not bounded above $\iff$ for every real number b there is an x in S with $x > b$.

3.

1. $x_1 = 1 < 4$, and if $x_k < 4$, then $x_{k+1} = 2\sqrt{x_k} < 2\sqrt{4} = 4$. By induction $x_k < 4$ for all k . Moreover, $x_k > 0$ for all k and $x_{k+1} = 2\sqrt{x_k} = \sqrt{4x_k} > \sqrt{x_k^2} = x_k$. Hence $\{x_k\}$ is increasing. By Theorem A.3.1 the sequence is convergent. Let its limit be x . By taking the limit as $k \rightarrow \infty$ in $x_{k+1} = 2\sqrt{x_k}$, we obtain $x = 2\sqrt{x}$, or $x^2 = 4x$, and thus $x = 4$.

2. Since all the terms in the sequence are positive, so $x_{k+1} + 2 > 2$ and $|x_{k+1} - 2| = \left| \frac{x_k^2 + 4}{x_{k+1} + 2} - 2 \right| = \left| \frac{x_k - 2}{x_{k+1} + 2} \right| < \frac{|x_k - 2|}{2}$ for $k \geq 1$. A straightforward induction argument now shows that $|x_k - 2| \leq |x_1 - 2|/2^{k-1}$ for all $k \geq 1$, and it follows that $\lim_{k \rightarrow \infty} x_k = 2$.
3. According to Theorem A.2.1(b), for each natural number n , because $1/n > 0$, there exists a number in A , call it x_n , such that $x_n > b^* - 1/n$. Because $x_n \leq b^*$, we have $|x_n - b^*| < 1/n$. It follows that $x_n \rightarrow b^*$ as $n \rightarrow \infty$.
4. See SM.
5. (a) Note that $x_k + y_k = 1$ and $x_k y_k \rightarrow \frac{1}{4}(1 - (-1)^{2k}) \rightarrow 0$ for all k . The required limits are: $\lim_{k \rightarrow \infty} x_k = 1$, $\lim_{k \rightarrow \infty} y_k = 1$, $\lim_{k \rightarrow \infty} (x_k + y_k) = 1$, $\lim_{k \rightarrow \infty} (x_k y_k) = 0$, $\lim_{k \rightarrow \infty} x_k = 0$, $\lim_{k \rightarrow \infty} y_k = 0$, $\lim_{k \rightarrow \infty} (x_k + y_k) = 1$, $\lim_{k \rightarrow \infty} (x_k y_k) = 0$. (b) See SM.
6. See SM.
7. Suppose that $x \neq y$. Let $\varepsilon = |x - y|$. Since $\varepsilon/2 > 0$, there exist numbers N and M such that $|x_n - x| < \varepsilon/2$ for all $n > N$ and $|x_n - y| < \varepsilon/2$ for all $n > M$. Then for $n > \max\{N, M\}$, we get $|x - y| = |x - x_n + x_n - y| \leq |x_n - x| + |x_n - y| < \varepsilon/2 + \varepsilon/2 = \varepsilon$, a contradiction.
8. (a) $a_1 = 2$, $a_2 = 2.25$, $a_3 \approx 2.3704$, $a_4 \approx 2.4414$; $b_1 = 4$, $b_2 = 3.375$, $b_3 \approx 3.1605$, $b_4 \approx 3.0518$.
(b) The inequality is valid for $n = 1$. Suppose $(1+x)^n \geq 1+nx$ for $x \geq -1$. Then $(1+x)^{n+1} = (1+x)^n(1+x) \geq (1+nx)(1+x) = 1 + (n+1)x + nx^2 \geq 1 + (n+1)x$. The last inequality is strict if $x \neq 0$.
(c) With $x = -1/n^2 \neq 0$, the inequality yields $(1 - 1/n^2)^n > 1 - 1/n$. Multiplying by $(n/(n-1))^n$ yields $(1 + 1/n)^n > (1 + 1/(n-1))^{n-1}$. (d) Left to the reader. (e) The sequence $\{a_n\}$ is increasing and bounded above (by any b_n). The sequence $\{b_n\}$ is decreasing and bounded below (by any a_n). Thus both sequences converge.
9. Let $\{x_k\}$ be a sequence of real numbers. For any natural number k , let us call the k th term of the sequence a *tail peak* if $x_k > x_j$ for all $j > k$. (A) If $\{x_k\}$ contains an infinite number of tail peaks, they form a (strictly) decreasing subsequence of $\{x_k\}$. (B) If $\{x_k\}$ contains only finitely many tail peaks (maybe none at all), let k_1 be the index of a term after the last tail peak. There is then a $k_2 > k_1$ with $x_{k_1} \leq x_{k_2}$, a $k_3 > k_2$ with $x_{k_2} \leq x_{k_3}$, etc. This process yields an increasing subsequence of $\{x_k\}$.

Appendix B

B.1

1. See Fig. AB.1.1. $OB = BF = \frac{1}{2}\sqrt{2}$ by Pythagoras's theorem. Hence, $\sin 45^\circ = \sin \pi/4 = BP/OP = \frac{1}{2}\sqrt{2} = \cos \pi/4$, whereas $\tan 45^\circ = \tan \pi/4 = \sin(\pi/4)/\cos(\pi/4) = 1$.
2. Look at Fig. AB.1.2. If the coordinates of P_x are (u, v) , then P_{-x} has coordinates $(u, -v)$, so $\sin(-x) = -v = -\sin x$, and $\cos(-x) = u = \cos x$. Then by definition, $\tan(-x) = \sin(-x)/\cos(-x) = -\sin x/\cos x = -\tan x$.

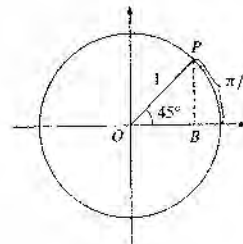


Figure AB.1.1

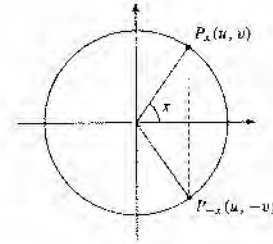


Figure AB.1.2

$$\cos(x - y) = \cos(x + (-y)) = \cos x \cos(-y) - \sin x \sin(-y) = \cos x \cos y + \sin x \sin y$$

$\cos(y - \pi/2) = \sin y$ follows directly from (10). For the rest see SM.

$$\tan(x + \pi) = \frac{\sin(x + \pi)}{\cos(x + \pi)} = \frac{\sin x \cos \pi + \cos x \sin \pi}{\cos x \cos \pi - \sin x \sin \pi} = \frac{-\sin x}{-\cos x} = \tan x$$

$$\sin(x + \frac{1}{2}\pi) = \sin x \cos \frac{1}{2}\pi + \cos x \sin \frac{1}{2}\pi = (\sin x) \cdot 0 + (\cos x) \cdot 1 = \cos x$$

$$\cos(x + \frac{1}{2}\pi) = \cos x \cos \frac{1}{2}\pi - \sin x \sin \frac{1}{2}\pi = -\sin x$$

(a) $1/2$. (Draw a figure similar to Figure B.1.3 in the text, or use the formula for $\sin(x - y)$ in Problem 4.)

(b) $-\sqrt{3}/2$ (c) $-\sqrt{2}/2$ (d) $-\sqrt{2}/2$ (e) $\sqrt{3}/3$

$$(f) \sin(\pi/12) = \sin(\pi/3 - \pi/4) = \sin(\pi/3) \cos(\pi/4) - \cos(\pi/3) \sin(\pi/4) = \frac{1}{4}(\sqrt{6} - \sqrt{2}).$$

$$(a) \sqrt{2} \sin(x + \pi/4) - \cos x = \sqrt{2}(\sin x \cos \pi/4 + \cos x \sin \pi/4) - \cos x = \sqrt{2}(\sin x \cdot 1/\sqrt{2} + \cos x \cdot 1/\sqrt{2}) - \cos x = \sin x \quad (b) \tan(\alpha + \beta) \quad (c) -\cos u / \sin a$$

Note that $x + y = A$ and $x - y = B$ imply $x = \frac{1}{2}(A + B)$ and $y = \frac{1}{2}(A - B)$. The desired formula then follows easily from the hint.

$$\sin(x + y) \sin(x - y) = (\sin x \cos y + \cos x \sin y)(\sin x \cos y - \cos x \sin y) = \sin^2 x \cos^2 y - \cos^2 x \sin^2 y = \sin^2 x(1 - \sin^2 y) - (1 - \sin^2 x) \sin^2 y = \sin^2 x - \sin^2 y$$

(a) See Fig. AB.1.10(a). Period π , amplitude 1. (b) See Fig. AB.1.10(b). Period 4π , amplitude 3.

(c) See Fig. AB.1.10(c). Period $2\pi/3$, amplitude 2.

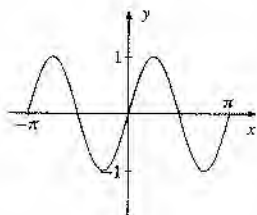


Figure AB.1.10(a)

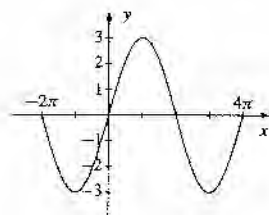


Figure AB.1.10(b)

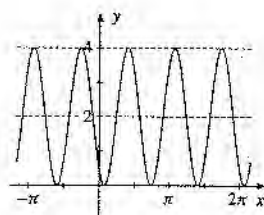


Figure AB.1.10(c)

(a) Because $|f(x)| = |(1/2)^x \sin x| \leq (1/2)^x$ for all x , and $(1/2)^x \rightarrow 0$ as $x \rightarrow \infty$, the oscillations die out.

(b) Because $2^x \rightarrow \infty$ as $x \rightarrow \infty$, the oscillations explode.

$$(a) y = 2 \sin \frac{1}{4}x \quad (b) y = 2 + \cos x \quad (c) y = 2e^{-x/\pi} \cos x$$

$(AC)^2 = (BD)^2$ yields $(\cos x - \cos y)^2 + (\sin x - (-\sin y))^2 = (\cos(x + y) - 1)^2 + \sin^2(x + y)$. Expanding and using equation (2) three times eventually gives formula (8).

$$(a) y' = \frac{1}{2} \cos \frac{1}{2}x \quad (b) y' = \cos x - x \sin x \quad (c) y' = \frac{2x}{\cos^2 x^2} \quad (d) y' = e^{2x}(2 \cos x - \sin x)$$

$$y = \tan x = \frac{\sin x}{\cos x} \text{ gives } y' = \frac{\cos x \cos x - \sin x(-\sin x)}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x} = 1 + \tan^2 x. \text{ Since } \cos^2 x + \sin^2 x = 1, \text{ we get } y' = 1/\cos^2 x \text{ as an alternative answer.}$$

$$(a) \cos x - \sin x \quad (b) 5x^4 \sin x + x^5 \cos x + (1/2\sqrt{x}) \cos x - \sqrt{x} \sin x$$

$$(c) \frac{1}{(x^2 + 1)^2} \left[\left(\frac{1}{2\sqrt{x}} - \frac{3x\sqrt{x}}{2} \right) \cos x - \sqrt{x}(1 + x^2) \sin x \right]$$

$$4. (a) a \sin ax \quad (b) a \sin bt + abt \cos bt \quad (c) -a \cos\{\cos[\sin(at + b)]\} \sin[\sin(at + b)] \cos(at + b)$$

$$5. (a) 2 \quad (b) m/n \quad (c) 1/2$$

$$6. \text{Maximum at } x = 0, \text{ minimum at } x = 3\pi/2. (f'(x) = 3(\sin x - x - 1)^2(\cos x - 1).)$$

$$7. p'(t) = -\lambda C_1 \sin \lambda t + \lambda C_2 \cos \lambda t \text{ and } p''(t) = -\lambda^2 C_1 \cos \lambda t - \lambda^2 C_2 \sin \lambda t, \text{ so } p''(t) + \lambda^2 p(t) = C_0 \lambda^2. \text{ Thus, } K = C_0 \lambda^2.$$

$$8. (a) \pi/4 \quad (b) \pi/2 \quad (c) \pi/6 \quad (d) \pi/3. \quad 9. (a) \frac{2}{\sqrt{1-4x^2}} \quad (b) \frac{2x}{1+(x^2+1)^2} \quad (c) \frac{1}{2\sqrt{x}\sqrt{1-x}}$$

$$10. (a) -\cos x + C \quad (b) \int_0^{\pi/2} \cos x \, dx = \sin x \Big|_0^{\pi/2} = \sin(\pi/2) - 0 = 1 \quad (c) \text{Integrating by parts, } I = \int \sin^2 x \, dx = \sin x(-\cos x) - \int \cos x(-\cos x) \, dx = -\sin x \cos x + \int \cos^2 x \, dx = -\sin x \cos x + \int (1 - \sin^2 x) \, dx. \text{ Hence, } I = -\sin x \cos x + x - I + C. \text{ Solving for } I \text{ gives } I = \frac{1}{2}(x - \sin x \cos x) + C_1. \quad (d) \int_0^\pi x \cos x \, dx = \left[\frac{\pi}{2} \sin x - \frac{1}{2} \cos x \right]_0^\pi = 0 + \frac{1}{2} \cos \pi - \frac{1}{2} \cos 0 = -2.$$

$$11. (a) -\ln |\cos x| + C \quad (b) e^{\sin x} + C \quad (c) -\frac{1}{5} \cos^5 x + C$$

12. For small h , $[\sin(x + h) - \sin x]/h \approx \cos x$, and passing to the limit as $h \rightarrow 0$ we get $(\sin x)' = \cos x$.

B.3

$$1. (a) z + w = 5 - 2i \quad (b) zw = 21 - 9i \quad (c) z/w = (-3 - 7i)/6 \quad (d) |z| = \sqrt{2^2 + (-5)^2} = \sqrt{29}$$

$$2. \text{See Fig. AB.3.2.} \quad 3. (a) \frac{1}{2}(1 + 5i) \quad (b) -3 - 4i \quad (c) (31 + 27i)/26 \quad (d) i$$

$$4. (a) 2\sqrt{3}(\cos(\pi/3) + i \sin(\pi/3)) \quad (b) \cos \pi + i \sin \pi \quad (c) 4(\cos(4\pi/3) + i \sin(4\pi/3)) \quad (d) \sqrt{2}(\cos(7\pi/4) + i \sin(7\pi/4))$$

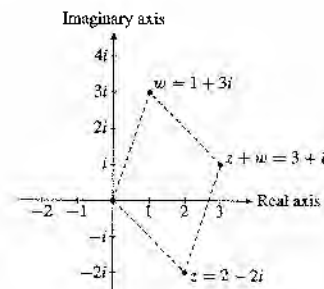


Figure AB.3.2

REFERENCES

- Arkin, V. I. and L. V. Evstigneev (1987) *Stochastic Models of Control and Economic Dynamics*. Academic Press.
- Arrow, K. J. (1951) "An extension of the basic theorems of classical welfare economics", in J. Neyman (ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press.
- Arrow, K. J. and M. Kurz (1970) *Public Investment, the Rate of Return, and Optimal Fiscal Policy*. The Johns Hopkins Press.
- Atkinson, A. B. (1971) "Capital taxes, the redistribution of wealth and individual savings". *Review of Economic Studies*, 38.
- Aubin, J.-P. and H. Frankowska (1990) *Set-Valued Analysis*. Birkhäuser.
- Barro, R. J. and X. Sala-i-Martin (1995) *Economic Growth*. McGraw-Hill.
- Berman, A. and R. J. Plemmons (1994) *Nonnegative Matrices in the Mathematical Sciences*. Society for Industrial and Applied Mathematics.
- Bertsekas, D. P. (1976) *Dynamic Programming and Stochastic Control*. Academic Press.
- Blanchard, O. and S. Fischer (1989) *Lectures on Macroeconomics*. MIT Press.
- Cesari, L. (1983) *Optimization - Theory and Applications*. Springer-Verlag.
- Chipman, J. S. (1950) "The multi-sector multiplier". *Econometrica*, 18.
- Clarke, F. H. (1983) *Optimization and Nonsmooth Analysis*. John Wiley & Sons.
- Coddington, E. A. and N. Levinson (1955) *Theory of Ordinary Differential Equations*. McGraw-Hill.
- Dorfinan, R. (1969) "An economic interpretation of optimal control theory". *American Economic Review*, LIX, 5.
- Faddeeva, V. N. (1959) *Computational Methods of Linear Algebra*. Dover Publications, Inc.
- Fleming, W. H. and R. W. Rishel (1975) *Deterministic and Stochastic Control*. Springer-Verlag.