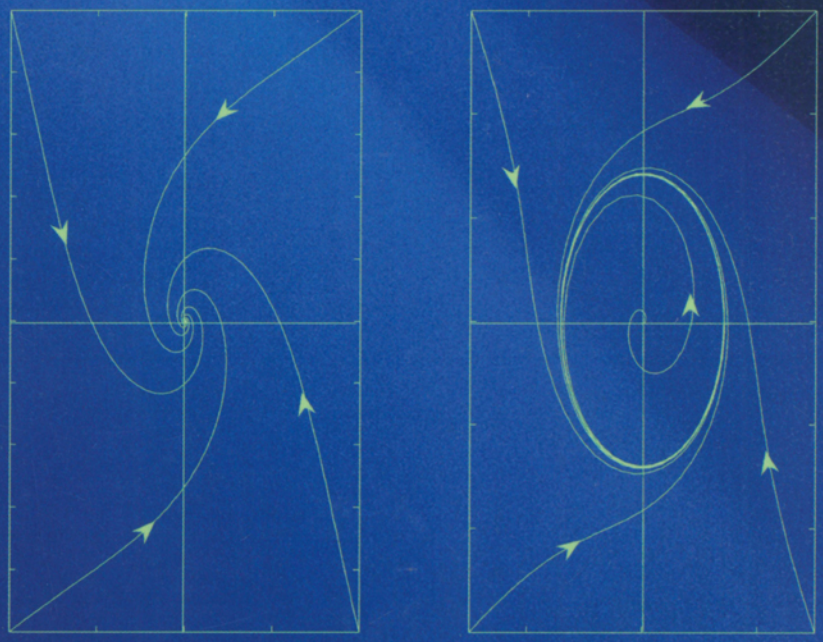
A First Course in Applied Mathematics



Jorge Rebaza

WILEY

WWW. LINK AVAILABLE

A First Course in Applied Mathematics

A First Course in Applied Mathematics

Jorge Rebaza
Department of Mathematics
Missouri State University
Springfield, MO

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

Copyright © 2012 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

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, scanning or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at http://www.wiley.com/go/permission.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representation or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Rebaza, Jorge.
A first course in applied mathematics / Jorge Rebaza.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-118-22962-0
1. Mathematical models. 2. Computer simulation. I. Title.
TA342.R43 2012
510—dc23 2011043340

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To my parents:

Leoncio Rebaza Miñano
and
Santos Vásquez Paredes

CONTENTS

Preface xiii

1 Basics of Linear Algebra 1

1.1 Notation and Terminology 1
1.2 Vector and Matrix Norms 4
1.3 Dot Product and Orthogonality 8
1.4 Special Matrices 9
1.4.1 Diagonal and triangular matrices 9
1.4.2 Hessenberg matrices 10
1.4.3 Nonsingular and inverse matrices 11
1.4.4 Symmetric and positive definite matrices 12
1.4.5 Matrix exponential 14
1.4.6 Permutation matrices 15
1.4.7 Orthogonal matrices 17
1.5 Vector Spaces 21
1.6 Linear Independence and Basis 24
1.7 Orthogonalization and Direct Sums 31
1.8 Column Space, Row Space, and Null Space 34
1.8.1 Linear transformations 40
1.9 Orthogonal Projections 43
1.10 Eigenvalues and Eigenvectors 47

1.11 Similarity 56
1.12 Bezier Curves and Postscript Fonts 59
1.12.1 Properties of Bezier curves 61
1.12.2 Composite Bezier curves 66
1.13 Final Remarks and Further Reading 68
Exercises 69

2 Ranking Web Pages 79

2.1 The Power Method 80
2.2 Stochastic, Irreducible, and Primitive Matrices 84
2.3 Google's PageRank Algorithm 92
2.3.1 The personalization vector 99
2.3.2 Speed of convergence and sparsity 100
2.3.3 Power method and reordering 105
2.4 Alternatives to the Power Method 106
2.4.1 Linear system formulation 107
2.4.2 Iterative aggregation/disaggregation (IAD) 111
2.4.3 IAD and linear systems 117
2.5 Final Remarks and Further Reading 120
Exercises 121

3 Matrix Factorizations 131

3.1 LU Factorization 132
3.1.1 The complex case 137
3.1.2 Solving several systems 137
3.1.3 The $PA = LU$ factorization 139
3.2 QR Factorization 142
3.2.1 QR and Gram–Schmidt 143
3.2.2 The complex case 147
3.2.3 QR and similarity 148
3.2.4 The QR algorithm 149
3.2.5 QR and LU 151
3.3 Singular Value Decomposition (SVD) 155
3.3.1 The complex case 160
3.3.2 Low–rank approximations 161
3.3.3 SVD and spectral norm 164
3.4 Schur Factorization 166
3.4.1 The complex case 171
3.4.2 Schur factorization and invariant subspaces 172
3.4.3 Exchanging eigenblocks 177
3.4.4 Block diagonalization 180

3.5 Information Retrieval 186
3.5.1 Query matching 187
3.5.2 Low-rank query matching 190
3.5.3 Term–term comparison 192
3.6 Partition of Simple Substitution Cryptograms 194
3.6.1 Rank-1 approximation 197
3.6.2 Rank-2 approximation 199
3.7 Final Remarks and Further Reading 203
Exercises 205

4 Least Squares 215

4.1 Projections and Normal Equations 215
4.2 Least Squares and QR Factorization 224
4.3 Lagrange Multipliers 228
4.4 Final Remarks and Further Reading 231
Exercises 231

5 Image Compression 235

5.1 Compressing with Discrete Cosine Transform 236
5.1.1 1-D discrete cosine transform 236
5.1.2 2-D discrete cosine transform 242
5.1.3 Image compression and the human visual system 245
5.1.4 Basis functions and images 247
5.1.5 Low-pass filtering 250
5.1.6 Quantization 254
5.1.7 Compression of color images 257
5.2 Huffman Coding 260
5.2.1 Huffman coding and JPEG 262
5.3 Compression with SVD 267
5.3.1 Compressing grayscale images 268
5.3.2 Compressing color images 268
5.4 Final Remarks and Further Reading 269
Exercises 271

6 Ordinary Differential Equations 277

6.1 One-Dimensional Differential Equations 278
6.1.1 Existence and uniqueness 278
6.1.2 A simple population model 284
6.1.3 Emigration 285
6.1.4 Time-varying emigration 285
6.1.5 Competition 286

6.1.6 Spring systems 287
6.1.7 Undamped equations 293
6.1.8 Damped equations 299
6.1.9 RLC circuits 303
6.2 Linear Systems of Differential Equations 307
6.3 Solutions via Eigenvalues and Eigenvectors 307
6.3.1 Chains of generalized eigenvectors 311
6.4 Fundamental Matrix Solution 312
6.4.1 Nonhomogeneous systems 314
6.5 Final Remarks and Further Reading 316
Exercises 316

7 Dynamical Systems 325

7.1 Linear Dynamical Systems 326
7.1.1 Dynamics in two dimensions 327
7.1.2 Trace-determinant analysis 335
7.1.3 Stable, unstable, and center subspaces 337
7.2 Nonlinear Dynamical Systems 340
7.2.1 Linearization around an equilibrium point 342
7.2.2 Linearization around a periodic orbit 348
7.2.3 Connecting orbits 352
7.2.4 Chaos 360
7.2.5 Bifurcations 365
7.3 Predator–prey Models with Harvesting 374
7.3.1 Boundedness of solutions 376
7.3.2 Equilibrium point analysis 376
7.3.3 Bifurcations 379
7.3.4 Connecting orbits 380
7.3.5 Other models 382
7.4 Final Remarks and Further Reading 385
Exercises 385

8 Mathematical Models 395

8.1 Optimization of a Waste Management System 396
8.1.1 Background 396
8.1.2 Description of the system 397
8.1.3 Development of the mathematical model 398
8.1.4 Building the objective function 399
8.1.5 Building the constraints 400
8.1.6 Numerical experiments 400
8.2 Grouping Problem in Networks 404

8.2.1 Background 405
8.2.2 The N-median approach 406
8.2.3 The probabilistic approach 407
8.2.4 Numerical experiments 409
8.3 American Cutaneous Leishmaniasis 410
8.3.1 Background 410
8.3.2 Development of the mathematical model 413
8.3.3 Equilibria and periodic orbits 414
8.3.4 Stability properties 416
8.3.5 Numerical computations 418
8.4 Variable Population Interactions 420
8.4.1 Model formulation 420
8.4.2 Local stability of equilibria 421
8.4.3 Bifurcations 425

References 431

Index 435

PREFACE

Going back in history, mathematics originated as a practical science, as a tool to facilitate administration of harvest, computation of the calendar, collection of taxes, and so on. But even i early Greek society, the study of mathematics had one main goal: the understanding of humankind's purpose in the universe according to a rational scheme. Thus developed a mathematics investigated more in the spirit of understanding rather than only of utility, and this has been a central and successful focus of mathematics since then.

The constant development of new and more sophisticated technologies, in particular, the very fast progress of software and hardware technology, has contributed to a clear change in how mathematics should be studied and taught nowadays, and in how old and new mathematical theories can now be effectively and efficiently applied to the solution of current real-world problems.

Not everybody agrees on what *Applied Mathematics* means and which subjects or topics it includes. Differential equations may be applied mathematics for some (as it applies notions e.g. from analysis and linear algebra), while for others it is just another subject of pure mathematics, of course with several potential applications. However, we can make a first attempt to list topics:

(Pure) Mathematics: Topology, Abstract Algebra, Analysis, Linear Algebra.

Applied Mathematics: Dynamical Systems, Matrix Computations, Optimization, Financial Mathematics, Numerical Methods.

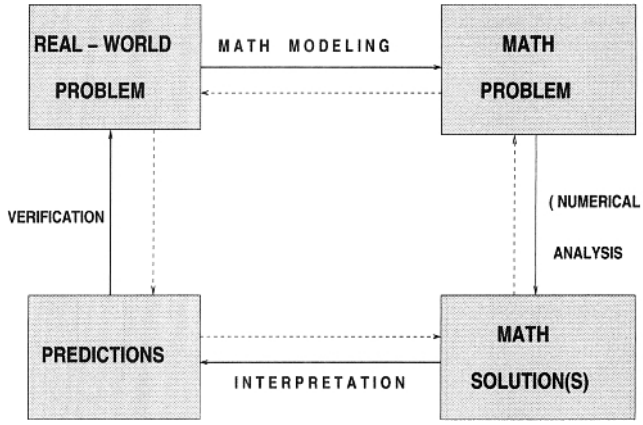


Figure 0.1 Applied Mathematics.

On the other hand, several people would argue that **all** of the subjects above are just mathematics, and not exactly applied mathematics. They would ask for the applications of dynamical systems or of numerical analysis; in other words, they would ask for *applications of the applications*, or "real-world applications".

Here is when the terms Industrial Mathematics and Mathematical Modeling would probably help. Given a real-world ("industrial") problem, e.g., describe the motion of a particle in certain fluid or matter, or rank all web pages in terms of importance, we first try to describe this problem in mathematical terms (a process called mathematical modeling), arriving at a mathematical model (e.g., a set of differential equations or an eigenvalue/eigenvector problem). This mathematical problem is then solved using mathematical tools (e.g. numerical analysis and linear algebra). Finally, this mathematical solution is analyzed together with the original problem and if necessary, the modeling step is modified and the process is repeated to obtain a final solution. Figure 0.1 illustrates this idea.

All the structure and ramifications of every application would not be possible without a solid theory supporting it. Theory is the indispensable and fundamental basis for any application. But theory could be better understood when at the same time corresponding methods or algorithms are used to solve a real-world problem. We could limit ourselves to studying the interesting and challenging questions of existence and uniqueness of solutions, but this is sterile from the point of view of applied mathematics. The beauty and importance of a mathematical concept may be better understood and appreciated when it is applied to a determined problem. For instance, the concept of an eigenvector and the power method to compute a dominant eigenvector are clearly illustrated when we study the mathematics behind the ranking of web pages by search engines like Google.

Through sound theory, a long collection of examples, and numerical computations, in this book we make an attempt to cover all the stages shown in Figure 0.1, and to illustrate how applied mathematics involves each component of the process toward the goal of solving real-world problems. However, only some selected topics in applied mathematics will be

considered, mainly because of limitations of space and time. Thus, for the applications we have in mind, we need to review some mathematical concepts and techniques, especially those in linear algebra, matrix analysis, and differential equations. Some classical definitions and results from analysis will also be discussed and used. Some applications (postscript fonts, information retrieval, etc.) are presented at the end of a chapter as an immediate application of the theory just covered, while those applications that are discussed in more detail (ranking web pages, compression, etc.) will be presented in chapters of their own.

This book is intended for advanced undergraduate and beginning graduate students in mathematics, computer science, engineering, and sciences in general, with an interest in mathematical modeling, computational methods, and the applications of mathematics. It contains a little more material that could be covered in one semester. The instructor decides how much time should be devoted to those chapters or sections dealing mainly with a review of some material. It will mostly depend on the background level of the students. Several chapters and sections are independent from the rest, and although different instructors can choose different chapters or sections to cover, here are some guidelines.

Chapters 1, 3, and 4 could be covered sequentially. Within Chapter 1, one could cover in detail, briefly, or even skip some of the first sections before covering the section on Bezier curves or before going to Chapter 2 or 3. The first two sections of Chapter 2 are the additional background needed for the web page ranking problem studied in that chapter. The first two sections of Chapter 3 may be discussed briefly to allow some extra time to study sections three and four; in some cases, section 4 may be skipped; although it will be needed in the second section of Chapter 7. Some material covered in Chapter 1, and especially in Chapter 3, is used in the study of Chapter 4.

Chapter 5 is for the most part independent from the previous chapters, except when viewed as an application of the concepts of linear combination and basis of vector spaces presented in Chapter 1.

The first two sections of Chapter 6 are a review of basic material on differential equations, and more time should be spent instead on sections three and four. This provides a good starting point for the next chapter. In Chapter 7, the first section should be covered relatively quickly as a background for section two on nonlinear systems, which contains a little more advanced material. Section three in this chapter applies the concepts studied in the first two sections, to a concrete problem.

The final chapter is a collection of some mathematical models of a slightly different nature; in particular, the first two sections deal with applications of some basic discrete mathematics and optimization, and for the most part is therefore independent from the previous chapters.

Depending on the background of the students, among other factors, here are a few sequences of sections that could be followed:

- Chapter 1, Chapter 2, 3.1– 3.3, 3.5, 5.1, 5.3, Chapter 6, 7.1, 7.2, 8.1, 8.2
- 1.4, 1.7– 1.9, 1.12, Chapter 2, 3.2– 3.5, Chapter 4, Chapter 5, 6.3, 6.4, Chapter 7
- 1.8, 1.9, 1.12, 2.3, 2.4, 3.3– 3.5, Chapter 4, 7.2, 7.3, Chapter 8

Besides two or more semesters of calculus, we are assuming the student has taken some basic linear algebra and an introductory differential equations course. Most of this material is reviewed in this book, and those sections can be used as a reference when studying new or more advanced topics. We also take care to introduce results that are typically part of a graduate course. An effort has been made to organize the book so that the transition from undergraduate to graduate material is as smooth as possible, and to make this text self-contained. Again, the main goal is to introduce real-world applications where each concept or theory just learned is used.

This book developed from lectures notes for the Applied Mathematics course at Missouri State University, with a typical audience consisting of some juniors and mostly seniors from mathematics and computer science, as well as first–year graduate students. It has been the textbook for this course since fall 2007. Although the content and order of sections covered may vary from semester to semester, typical sections covered are:

- 1.2, 1.4.7, 1.8, 1.12, 2.1– 2.3, 3.1– 3.3, 3.5, 5.1, 5.3, 6.3, 6.4, 7.1, 7.2.1, 7.2.2, 7.2.5.

Selected topics from other sections, including 2.4, 3.4, 3.6, 5.2, 7.3, and 8.1– 8.4 are usually assigned as group projects, and students are required to turn in a paper and give a seminar presentation on the one topic assigned. Different groups get different topics, and group discussions in class are part of the course.

We remark that in general, theory and examples in this book will be illustrated with the help of MATLAB software package. Previous knowledge of MATLAB or programming is not required.

CHAPTER 1

BASICS OF LINEAR ALGEBRA

Undoubtedly, one of the subjects in mathematics that has become more indispensable than ever is linear algebra. Several application problems involve at some stage solving linear systems, the computation of eigenvalues and eigenvectors, linear transformations, bases of vector subspaces, and matrix factorizations, to mention a few. One very important characteristic of linear algebra is that as a first course, it requires only very basic prerequisites so that it can be taught very early at undergraduate level; at the same time, mastering vector spaces, linear transformations and their natural extensions to function spaces is essential for researchers in any area of applied mathematics. Linear algebra has innumerable applications, including differential equations, least-square solutions and optimization, demography, electrical engineering, fractal geometry, communication networks, compression, search engines, social sciences, etc. In the next sections we briefly review the concepts of linear algebra that we will need later on.

1.1 NOTATION AND TERMINOLOGY

We start this section by defining an $m \times n$ **matrix** as a rectangular array of elements arranged in m rows and n columns, and we say the matrix is of order $m \times n$. We usually denote the elements of a matrix A of order $m \times n$ as a_{ij}, where $i = 1, \ldots m$, $j = 1, \ldots, n$, and we

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

write the matrix A as

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

Although the elements of a matrix can be real or complex numbers, here we will mostly consider the entries of a given matrix to be real unless otherwise stated. In most cases, we will take care of stating the complex version of some definitions and results.

Matrix addition and multiplication. Given two arbitrary matrices $A_{m\times n}$ and $B_{m\times n}$, we define the matrix

$$C = A + B$$

by adding the entries of A and B componentwise. That is, we have

$$c_{ij} = a_{ij} + b_{ij},$$

for $i = 1, \ldots, m,\ j = 1, \ldots, n$.

This means that the addition of matrices is well defined only for matrices of the same order.

Now consider two arbitrary matrices $A_{m\times p}$ and $B_{p\times n}$. Then we define the product matrix $C_{m\times n} = A \cdot B$ as

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj},$$

where $i = 1, \ldots, m$, $j = 1, \ldots, n$. This means that to obtain the entry (i, j) of the product matrix, we multiply the i-th row of A with the j-th column of B entry-wise and add their products.

Observe that for the product to be well-defined, the number of columns of A has to agree with the number of rows of B.

■ **EXAMPLE 1.1**

Let $A = \begin{bmatrix} 4 & -5 \\ 3 & 2 \\ 1 & 6 \end{bmatrix}$, $B = \begin{bmatrix} -1 & 2 \\ 0 & 3 \end{bmatrix}$. Then, for instance, to obtain the entry c_{32} of the product matrix $C = AB$, we multiply entrywise the third row of A with the second column of B: $(1)(2) + (6)(3) = 20$. Thus, we get

$$C = A \cdot B = \begin{bmatrix} 4 & -5 \\ 3 & 2 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} -1 & 2 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} -4 & -7 \\ -3 & 12 \\ -1 & 20 \end{bmatrix}.$$

- MATLAB command: $A + B$, $A * B$.

Given a matrix $B_{p\times n}$, we can denote its columns with $b_1, \ldots, b_n$, where each b_i is a p-dimensional vector. We will write accordingly,

$$B = [b_1 \ \cdots \ b_n].$$

In such a case, we can write a product of matrices as

$$AB = A\,[b_1 \ \cdots \ b_n] = [Ab_1 \ \cdots \ Ab_n], \tag{1.1}$$

so that $Ab_1, \ldots, Ab_n$ are the columns of the matrix AB.

Similarly, if we denote with $a_1, \ldots, a_m$ the rows of a matrix $A_{m\times p}$, then

$$AB = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} B = \begin{bmatrix} a_1 B \\ \vdots \\ a_m B \end{bmatrix}.$$

For an arbitrary matrix $A_{m\times n}$, we denote with A^T its **transpose** matrix, that is, the matrix of order $n \times m$, where the rows of A have been exchanged for columns and vice versa. For instance,

$$\text{If} \quad A = \begin{bmatrix} 6 & -4 & 7 \\ -2 & 3 & 8 \end{bmatrix}, \quad \text{then} \quad A^T = \begin{bmatrix} 6 & -2 \\ -4 & 3 \\ 7 & 8 \end{bmatrix}.$$

Remark 1.1 *If $A_{m\times n}$ is a complex matrix, its* **adjoint** *matrix is denoted as A^*, where $A^* = \bar{A}^T$, that is, the conjugate transpose. For instance,*

$$\text{If} \quad A = \begin{bmatrix} 3 & -4+i \\ -2i & 5+2i \end{bmatrix}, \quad \text{then} \quad A^* = \begin{bmatrix} 3 & 2i \\ -4-i & 5-2i \end{bmatrix}.$$

- MATLAB command: A'

The sum and product of matrices satisfy the following properties (see Exercise 1.1):

$$(A+B)^T = A^T + B^T, \qquad (AB)^T = B^T A^T. \tag{1.2}$$

Definition 1.2 *The* **trace** *of a square matrix A of order n is defined as the sum of its diagonal elements, that is,*

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}. \tag{1.3}$$

- MATLAB command: trace(A)

A particular and very special case is that of matrices of order $n \times 1$. Such a matrix is usually called an n-dimensional **vector**. That is, here we consider vectors as column-vectors, and we will use the notation

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = [x_1 \cdots x_n]^T$$

for a typical vector. This substitutes the usual notation $x = (x_1, \ldots, x_n)$, which we reserve to denote a point, and at the same time this notation will allow us to perform matrix-vector operations in agreement with their dimensions. This also closely follows the notation used in MATLAB.

The first two vectors below are column vectors, whereas the third is a row vector.

$$\begin{bmatrix} 9 \\ 4 \\ 3 \end{bmatrix}, \qquad [4 \ -3 \ \ 5]^T, \qquad [1 \ \ 8 \ \ 5].$$

1.2 VECTOR AND MATRIX NORMS

It is always important and useful to have a notion for the "size" or magnitude of a vector or a matrix, just as we understand the magnitude of a real number by using its absolute value. In fact, a norm can be understood as the generalization of the absolute value function to a higher dimensional case. This is especially useful in numerical analysis for estimating the magnitude of the error when approximating the solution to a given problem.

Definition 1.3 *A vector norm, denoted by $\|\cdot\|$, is a real function that satisfies the following properties for arbitrary n-dimensional vectors x and y and for arbitrary real or complex α:*

(i) $\|x\| \geq 0$,

(ii) $\|x\| = 0$ *if and only if* $x = 0$,

(iii) $\|\alpha x\| = |\alpha| \, \|x\|$,

(iv) $\|x + y\| \leq \|x\| + \|y\|$.

There are several norms that can be defined for vectors. Here we mention the most commonly used.

Let $x = [x_1 \cdots x_n]^T, \quad x_i \in \mathbb{R}$. Then, we define

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{n} |x_i| && \text{Sum norm} \\ \|x\|_2 &= \left(\sum_{i=1}^{n} x_i^2\right)^{1/2} && \text{Euclidean norm} \\ \|x\|_\infty &= \max_{i=1,\ldots,n} |x_i| && \text{Maximum norm} \end{aligned} \tag{1.4}$$

- MATLAB commands: norm(x, 1), norm(x) norm(x,inf)

■ **EXAMPLE 1.2**

Let $x = [3 \;\; -2 \;\; 4 \;\; \sqrt{7}]^T$. Then,

$\|x\|_1 = |3| + |-2| + |4| + |\sqrt{7}| \approx 11.6458.$
$\|x\|_2 = \sqrt{9+4+16+7} = 6.$
$\|x\|_\infty = \max\{\,|3|,\, |-2|,\, |4|,\, |\sqrt{7}|\,\} = 4.$

Remark 1.4 *In general, for an arbitrary vector $x \in \mathbb{R}^n$, the following inequalities (see Exercise 1.4) relate the three norms above:*

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1. \tag{1.5}$$

■ **EXAMPLE 1.3**

The unit ball in $\mathbb{R}^n$ is the set $\{x \in \mathbb{R}^n : \|x\| \le 1\}$. The geometrical shape of the unit ball varies according to what norm is used. For the case $n = 2$, the unit balls for the three norms in (1.4) are shown in Figure 1.1.

Note: For simplicity of notation, $\|\cdot\|$ will always denote the Euclidean norm $\|\cdot\|_2$ for vectors, unless otherwise stated.

We now introduce the notion of norms for a matrix, in some sense as a generalization of vector norms.

Definition 1.5 *A matrix norm is a real function that for arbitrary matrices A and B and arbitrary real or complex α, satisfies*

(i) $\|A\| \ge 0$,
(ii) $\|A\| = 0$ *if and only if* $A = 0$,
(iii) $\|\alpha A\| = |\alpha|\|A\|$,

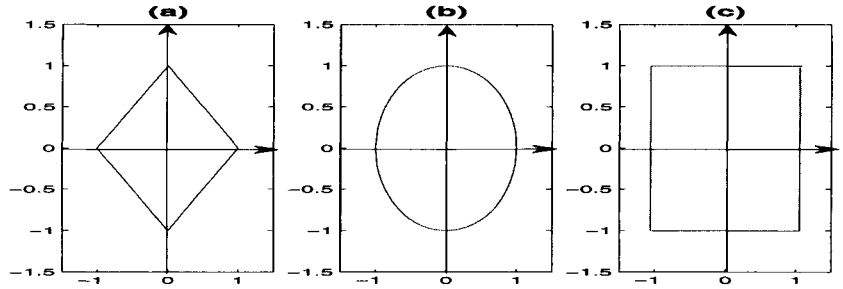


Figure 1.1 Unit ball in (a) $\|\cdot\|_1$, (b) $\|\cdot\|_2$, (c) $\|\cdot\|_\infty$.

(iv) $\|A+B\| \leq \|A\| + \|B\|$,
(v) $\|AB\| \leq \|A\|\,\|B\|$.

Let A be an $m \times n$ matrix. Here are some of the most common matrix norms

$$\begin{aligned}
\|A\|_1 &= \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| && \text{Maximum column sum norm} \\
\|A\|_F &= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2} && \text{Frobenius norm} \\
\|A\|_\infty &= \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}| && \text{Maximum row sum norm}
\end{aligned} \tag{1.6}$$

- MATLAB command: norm(A,1), norm(A,'fro'), norm(A,inf)

Remark 1.6 *The Frobenius norm is also known as the Euclidean norm for matrices, and in some textbooks it is denoted as* $\|\cdot\|_2$. *We will reserve this notation for a different matrix norm.*

■ **EXAMPLE 1.4**

Let $A = \begin{bmatrix} -1 & 3 & 2 \\ 2 & 0 & -2 \\ 4 & -3 & 1 \end{bmatrix}$. Then,

$$\begin{aligned}
\|A\|_1 &= \max\{7,6,5\} = 7. \\
\|A\|_F &= (1+9+4+4+0+4+16+9+1)^{1/2} \approx 6.9282. \\
\|A\|_\infty &= \max\{6,4,8\} = 8.
\end{aligned}$$

Remark 1.7 *There are two other ways of defining the Frobenius norm, which are useful for certain computations, as we illustrate later on. Let $A_{m\times n}$ be an arbitrary matrix, and denote its columns with $a_1, \ldots, a_n$. Then,*

$$\text{(a)} \quad \|A\|_F^2 = \|a_1\|_2^2 + \cdots + \|a_n\|_2^2. \tag{1.7}$$

$$\text{(b)} \quad \|A\|_F^2 = \operatorname{tr}(A^T A). \tag{1.8}$$

It seems natural to think that vector norms can be directly generalized to obtain matrix norms (after all, $m \times n$ matrices can be thought of as vectors of dimension $m \cdot n$). However, not all vector norms directly become matrix norms (see Exercise 1.11).

A matrix norm can be **induced** or constructed directly from a vector norm by defining for $1 \le p \le \infty$

$$\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p, \tag{1.9}$$

where the norms on the right–hand side represent the vector norm. This is the way to correctly extend or use a vector norm into obtaining a matrix norm.

Given a vector x of norm $\|x\|$, when it is multiplied by A, we get the new vector Ax of norm $\|Ax\|$. Thus, we can interpret the matrix norm (1.9) as a natural way to measure how much the vector x can be stretched or shrunk by A.

The definition of the p-norm of a matrix in (1.9) is not easy to implement or compute in general. Luckily enough, there are alternative ways to compute such a norm for some particular values of p.

■ **EXAMPLE 1.5**

It can be proved (see Exercise 1.12) that for $p = 1$ and $p = \infty$, the p-norms in (1.9) can be directly computed through the corresponding definitions in (1.6).

■ **EXAMPLE 1.6**

For $p = 2$, the p-norm in (1.9) can be computed as the square root of the largest eigenvalue of the symmetric matrix $A^T A$. That is,

$$\|A\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^T A\}. \tag{1.10}$$

For instance, for the matrix of Example 1.4, we get $\|A\|_2 \approx 5.9198$.

Remark 1.8 *It is possible to show that an equivalent way to define the matrix norm (1.9) is*

$$\|A\|_p = \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}. \tag{1.11}$$

Remark 1.9 *In general, inequalities for matrix norms similar to the ones in (1.5) are not true. However, it is still true that*

$$\|A\|_2 \leq \|A\|_F, \tag{1.12}$$

where $\|\cdot\|_2$ *is given by (1.9).*

Finally, we mention another important property that relates matrix and vector norms. We say that the norm of an $m \times n$ matrix A is **consistent** with the norm of an n-dimensional vector x if

$$\|Ax\| \leq \|A\| \, \|x\|, \tag{1.13}$$

for any $x \in \mathbb{R}^n$. Then, it is straightforward to see that the p-norms defined by (1.9) or (1.11) are all consistent.

1.3 DOT PRODUCT AND ORTHOGONALITY

Quite often we will need to use the so-called **dot product** or **inner product** between two vectors $x = [x_1 \cdots x_n]^T$ and $y = [y_1 \cdots y_n]^T$, defined as

$$x \cdot y = < x, y > = x_1 y_1 + \cdots + x_n y_n. \tag{1.14}$$

As vectors can also be considered to be $n \times 1$ matrices, it is common to represent the inner product as matrix multiplication. Thus, we will consider all these notations to be equivalent:

$$x \cdot y = < x, y > = x^T y. \tag{1.15}$$

Observe that with the definition above,

$$\|x\|_2^2 = x^T x.$$

For arbitrary vectors $x, y, z \in \mathbb{R}^n$ and arbitrary real scalar c, the dot product satisfies the following properties:

(1) $< x, x > \geq 0$,
(2) $< x, x > = 0$ if and only if $x = 0$,
(3) $< cx, y > = c < x, y >$,
(4) $< x + y, z > = < x, z > + < y, z >$,
(5) $< x, y > = < y, x >$.

Remark 1.10 *In the complex case, we have*

$$\begin{aligned} (3a) < x, cy > &= \bar{c} < x, y > \\ (5a) < x, y > &= \overline{< y, x >}. \end{aligned} \tag{1.16}$$

Following (1.15), we also see that for arbitrary $A_{m\times n}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$,

$$< Ax, y >= (Ax)^T y = x^T A^T y =< x, A^T y > . \tag{1.17}$$

Remark 1.11 *The dot product introduced here is a particular case of the general inner product function studied in the context of inner product vector spaces, where the elements are not restricted to real n-dimensional vectors.*

There is a special kind of vectors that is very useful in several instances in linear algebra and matrix computations. These are the so-called **orthonormal vectors**, which are orthogonal (perpendicular) to each other and they are unit, that is, x and y are orthonormal if

$$x^T y = 0, \quad \text{and} \quad \|x\| = \|y\| = 1.$$

For example, the following set of vectors is orthonormal:

$$v_1 = \begin{bmatrix} \frac{2}{\sqrt{5}} & 0 & -\frac{1}{\sqrt{5}} \end{bmatrix}^T, \quad v_2 = \begin{bmatrix} \frac{1}{\sqrt{5}} & 0 & \frac{2}{\sqrt{5}} \end{bmatrix}^T, \quad v_3 = [0 \ -1 \ \ 0]^T.$$

In fact, we can readily verify that

$$v_1^T v_2 = v_1^T v_3 = v_2^T v_3 = 0, \quad \text{and} \quad \|v_1\| = \|v_2\| = \|v_3\| = 1.$$

1.4 SPECIAL MATRICES

We will be using matrices at almost every place in this book, matrices of different kinds and properties. Here we list some of the most common types of matrices we will encounter.

1.4.1 Diagonal and triangular matrices

A square matrix A of order n is called **diagonal** if $a_{ij} = 0$ for all $i \neq j$.

■ **EXAMPLE 1.7**

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -9 \end{bmatrix}.$$

- MATLAB command: `diag`

A square matrix A of order n is called **upper** (resp. **lower**) **triangular** if $a_{ij} = 0$ for all $i > j$, (resp. for all $i < j$).

■ **EXAMPLE 1.8**

$$A = \begin{bmatrix} 4 & 2 & 5 \\ 0 & 1 & 3 \\ 0 & 0 & 9 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & 0 & 0 \\ 5 & 1 & 0 \\ 7 & 1 & 9 \end{bmatrix}.$$

The matrix A is upper triangular, and B is lower triangular.

- MATLAB commands: triu, tril

Remark 1.12 *If the matrix A is rectangular, say order $m \times n$, we say it is* upper (resp. lower) trapezoidal *if $a_{ij} = 0$ for all $i > j$ (resp. for all $i < j$).*

A more general case of triangular matrices is that of **block triangular matrices**. For example, the following 5×5 matrix is block upper triangular

$$A = \left[\begin{array}{cc|cc|c} -8 & 1 & 0 & 9 & 3 \\ 4 & -2 & 8 & 5 & 5 \\ \hline 0 & 0 & 5 & 7 & 1 \\ 0 & 0 & -9 & -8 & 4 \\ \hline 0 & 0 & 0 & 0 & -6 \end{array}\right] = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix},$$

where for instance, the zero in the (2,1) entry of the matrix on the right represents the corresponding 2×2 zero block of the matrix on the left. In a similar way, a block diagonal matrix can be defined. We will encounter this type of matrices in Section 1.10 when we compute eigenvalues of a matrix and when compute special vector subspaces later in Chapter 7.

1.4.2 Hessenberg matrices

A matrix $A_{n\times n}$ is called an upper Hessenberg matrix if $a_{ij} = 0$ for all $i > j + 1$. They take the form:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ 0 & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ 0 & 0 & a_{43} & \cdots & a_{4n-1} & a_{4n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn-1} & a_{nn} \end{bmatrix}. \tag{1.18}$$

In other words, it is an upper triangular matrix with an additional subdiagonal below the main diagonal.

A matrix A is called lower Hessenberg if A^T is upper Hessenberg.

- MATLAB command: hess(A)

1.4.3 Nonsingular and inverse matrices

In the set of real numbers $\mathbb{R}$, the number 1 is the multiplicative identity, meaning that $a \cdot 1 = 1 \cdot a = a$, for any real number a. This is generalized to what is called the **identity matrix** of order n, denoted by I. It consists of a diagonal of ones, and every other entry is zero (see (1.21) below), with the property that

$$A\,I = I\,A = A,$$

for any matrix A.

- MATLAB command: eye(n).

A square matrix A of order n is called **nonsingular** if its determinant is not zero: $\det(A) \neq 0$. Otherwise it is called *singular.*

■ **EXAMPLE 1.9**

$$\det \begin{bmatrix} 4 & 2 & 9 \\ 1 & 1 & 0 \\ 7 & 1 & 3 \end{bmatrix} = -48. \text{ Hence, the matrix is nonsingular.}$$

- MATLAB command: det(A).

Remark 1.13 *Two very useful properties of determinants are the following:*

$$\begin{aligned} \det(AB) &= \det A \det B, \\ \det A^{-1} &= 1/\det A. \end{aligned} \tag{1.19}$$

In one dimension, every real number $a \neq 0$ has a unique multiplicative inverse $1/a$, and it is obvious that $a\left(\frac{1}{a}\right) = \left(\frac{1}{a}\right)a = 1$. This has a natural generalization to matrices. A nonsingular matrix $A_{n\times n}$ has a unique $n \times n$ **inverse matrix**, denoted by A^{-1} with the property that

$$A\,A^{-1} = A^{-1}A = I. \tag{1.20}$$

For this reason, nonsingular matrices are also called invertible.

- MATLAB command: inv(A).

■ **EXAMPLE 1.10**

Let $A = \begin{bmatrix} 4 & 0 & 9 \\ 1 & -1 & 0 \\ -2 & 1 & 0 \end{bmatrix}$. Since $\det(A) = -9$, the matrix is nonsingular and has an inverse: $A^{-1} = \begin{bmatrix} 0 & -1 & -1 \\ 0 & -2 & -1 \\ 1/9 & 4/9 & 4/9 \end{bmatrix}$, and we can verify that in fact we have

$$A\,A^{-1} = A^{-1}A = I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1.21}$$

Computing the inverse of a matrix is not an easy task. There is more than one analytical or exact way to do this, e.g., using the adjoint method, we have

$$A^{-1} = \text{adjoint}(A)/\det(A).$$

However, this method is rarely used in practice, and we do not discuss it here. A second and more efficient method is through Gauss elimination (see Section 3.1). In general, the computation of the inverse of a matrix has to be done numerically, and great care has to be taken due to potentially large accumulation of errors for some matrices. Thus, in practice it is customary to avoid computing the inverse of a matrix explicitly, and some other options must be used. We will discuss these issues later on.

Remark 1.14 *For the inverse of a product of nonsingular matrices, we have*

$$(A_1 A_2 \cdots A_k)^{-1} = A_k^{-1} A_{k-1}^{-1} \cdots A_1^{-1}. \tag{1.22}$$

1.4.4 Symmetric and positive definite matrices

A square matrix A of order n is called **symmetric** if $A = A^T$, that is, the matrix equals its transpose. If $A = (a_{ij})$, then we can also say A is symmetric if $a_{ij} = a_{ji}$, for $i, j = 1, \ldots, n$.

■ **EXAMPLE 1.11**

The following matrices are symmetric:

$$A = \begin{bmatrix} 1 & 4 \\ 4 & 9 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & -2 & 7 \\ -2 & 3 & 5 \\ 7 & 5 & 6 \end{bmatrix}.$$

■ **EXAMPLE 1.12**

For an arbitrary matrix A, the matrices $A + A^T$ and $A\,A^T$ are symmetric.

Remark 1.15 *If $A_{n\times n}$ is complex, we say A is* **Hermitian** *if $A = A^*$. For instance, the following matrix is Hermitian:*

$$A = \begin{bmatrix} 1 & 4-2i \\ 4+2i & -9 \end{bmatrix}.$$

In the set of real numbers, we can define a number a to be positive if $a\,x^2 > 0$, for all numbers $x \neq 0$. The generalization of this definition for matrices leads to the so–called **positive definite** matrices. A square matrix A is positive definite if

$$x^T A\,x > 0, \tag{1.23}$$

for all vectors $x \neq 0$. If instead we have $x^T A\,x \geq 0$, then the matrix is said to be positive semidefinite.

In several applications, especially optimization, we will encounter matrices that are both, symmetric and positive definite. That is why several textbooks simply assume positive definite matrices to be symmetric; we call such matrices symmetric positive definite (**spd**). See also Remark 1.16 below.

- MATLAB command: chol(A) (Choleski)

■ **EXAMPLE 1.13**

The following matrices are spd:

$$A = \begin{bmatrix} 4 & 10 \\ 10 & 26 \end{bmatrix}, \quad B = \begin{bmatrix} 100 & 15 & 0.01 \\ 15 & 2.3 & 0.01 \\ 0.01 & 0.01 & 1 \end{bmatrix}.$$

It is obvious that A is symmetric. To see why it is positive definite, we can verify that $x^T A\,x > 0$ for any vector $x \neq 0$:

$$[x_1 \quad x_2] \begin{bmatrix} 4 & 10 \\ 10 & 26 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (2x_1 + 5x_2)^2 + x_2^2 > 0.$$

■ **EXAMPLE 1.14**

The matrix $A^T A$ is symmetric positive semidefinite, for any matrix $A_{m\times n}$.

■ **EXAMPLE 1.15**

The matrix $A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$ with $a > 0$ is positive definite.

Remark 1.16 *It is important to state that a positive definite matrix does not need to be symmetric, as in the case of the matrix* $A = \begin{bmatrix} 1 & 0 \\ 2 & 2 \end{bmatrix}$.

Definition 1.17 *Let A be a square matrix of order n, and let $m < n$. A* **principal submatrix** *of A is a matrix obtained by deleting any $n - m$ rows and corresponding columns. A* leading principal submatrix *is obtained by deleting the last $n - m$ rows and columns.*

Notation. We can use and index set $I \subseteq \{1, \ldots, n\}$ to represent which rows and columns of A are used to form the principal submatrix. For instance, if

$$A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}, \quad \text{then } A(\{1,3\}) = \begin{bmatrix} 1 & 7 \\ 3 & 9 \end{bmatrix}, \qquad A(\{1,2\}) = \begin{bmatrix} 1 & 4 \\ 2 & 5 \end{bmatrix}.$$

The last one is a leading principal submatrix of A.

Remark 1.18 *A positive definite matrix has all its principal submatrices positive definite. Thus, in particular, all the diagonal elements of a positive definite matrix are positive.*

1.4.5 Matrix exponential

A central tool in the theory and applications of differential equations, dynamical systems, and control theory, is the **matrix exponential** of a square matrix A of order n. This special type of matrix is defined as

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!} = I + \frac{A}{1!} + \frac{A^2}{2!} + \cdots + \frac{A^k}{k!} + \cdots. \tag{1.24}$$

This is an extension of the real exponential function

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots + \frac{x^k}{k!} + \cdots \qquad (x \in \mathbb{R})$$

to n dimensions. The series (1.24) always converges for any square matrix A.

■ **EXAMPLE 1.16**

Let $A = \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix}$. Since $A^0 = I$ and

$$A^2 = \begin{bmatrix} 3^2 & 0 \\ 0 & 5^2 \end{bmatrix}, \; A^3 = \begin{bmatrix} 3^3 & 0 \\ 0 & 5^3 \end{bmatrix}, \ldots, A^k = \begin{bmatrix} 3^k & 0 \\ 0 & 5^k \end{bmatrix},$$

we have

$$e^A = \begin{bmatrix} \sum_{k=0}^{\infty} \frac{3^k}{k!} & 0 \\ 0 & \sum_{k=0}^{\infty} \frac{5^k}{k!} \end{bmatrix} = \begin{bmatrix} e^3 & 0 \\ 0 & e^5 \end{bmatrix}.$$

■ **EXAMPLE 1.17**

Let $A = \begin{bmatrix} 0 & 1 & -2 \\ 0 & 0 & 4 \\ 0 & 0 & 0 \end{bmatrix}$. Then, $A^2 = \begin{bmatrix} 0 & 0 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $A^3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$,
and A^k is the zero matrix, for any $k \geq 3$. Then $e^A = I + A + \frac{1}{2}A^2$, and therefore
$e^A = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$.

In Chapter 7, we will consider several other examples of a matrix exponential.

- MATLAB command: expm(A)

1.4.6 Permutation matrices

In matrix computations, we often need to perform permutations of rows or columns of a matrix; that is, we exchange the entries of two given rows or columns, respectively. This can be simply achieved through matrix multiplication by a special type of square matrix.

A **permutation matrix** P can be defined as the identity matrix with some of its rows reordered as needed.

■ **EXAMPLE 1.18**

The following are permutation matrices:

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad P_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \quad P_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

Observe that for the matrices above, we have

$$P_1^2 = I, \qquad P_2^4 = I, \qquad P_3^3 = I.$$

Thus, to recover the identity matrix, we need to multiply the permutation matrix P by itself as many times as the number of rows that have been reordered in the identity (or, as the number of zeros on the diagonal of P).

In general, due to the way they are defined, permutation matrices satisfy the important property

$$P^T P = I = P P^T. \tag{1.25}$$

Therefore, they are invertible, and $P^{-1} = P^T$.

Multiplying an arbitrary matrix A by a permutation matrix P from the left will result in row permutations in A, and multiplying from the right will result in column permutations.

■ **EXAMPLE 1.19**

Let $A = \begin{bmatrix} 1 & 5 & 8 \\ 2 & 0 & 5 \\ 3 & 3 & 6 \end{bmatrix}$, and $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. Then, since the permutation matrix P is the identity after permuting its rows or columns 2 and 3, we can permute the corresponding rows (or columns) of A by multiplying:

$$PA = \begin{bmatrix} 1 & 5 & 8 \\ 3 & 3 & 6 \\ 2 & 0 & 5 \end{bmatrix}, \quad \text{or} \quad AP = \begin{bmatrix} 1 & 8 & 5 \\ 2 & 5 & 0 \\ 3 & 6 & 3 \end{bmatrix}.$$

It is clear that if P is a permutation matrix, then P^T also is (but in general, $P^T \neq P$, as we see from the matrices P_2, P_3 in Example 1.18). Then, we can simultaneously interchange rows and columns of a given matrix A through the operation $P^T A P$. In the example below, AP is the matrix A with the first and third columns interchanged. Then, the product $P^T A P$ interchanges the first and third row of AP.

■ **EXAMPLE 1.20**

Let $A = \begin{bmatrix} 0 & 7 & 9 & 0 \\ 3 & 0 & 0 & 0 \\ 6 & 4 & 0 & 2 \\ 1 & 0 & 5 & 0 \end{bmatrix}$ and $P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. Then,

$$P^T A P = \begin{bmatrix} 0 & 4 & 6 & 2 \\ 0 & 0 & 3 & 0 \\ 9 & 7 & 0 & 0 \\ 5 & 0 & 1 & 0 \end{bmatrix}.$$

- MATLAB command: colperm(A), flipud(A), fliplr(A)

1.4.7 Orthogonal matrices

And last, but definitely not least, we consider the type of matrix that is essential and the favorite one in matrix computations, numerical analysis, and applied mathematics in general, due to its very important and remarkable properties.

An $m \times n$ matrix Q is called **orthogonal** if

$$Q^T Q = I. \tag{1.26}$$

This definition implies (see Exercise 1.38) that the columns of an orthogonal matrix $Q_{m\times n}$ are orthonormal, in the $\| \ \|_2$ norm.

In particular, if the matrix Q is square, then $Q^T Q = Q\,Q^T = I$. In this case, the following statements are equivalent:

(a) Q is orthogonal.
(b) $Q^{-1} = Q^T$.
(c) The columns of Q are orthonormal.
(d) The rows of Q are orthonormal.

■ **EXAMPLE 1.21**

Permutation matrices are obviously orthogonal. See the property (1.25).

■ **EXAMPLE 1.22**

The following matrices (the first for any real α) are also orthogonal.

$$Q = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}, \quad Q = \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

Remark 1.19 *In the complex case, we say a matrix $U_{m\times n}$ is* **unitary** *if*

$$U^* U = I,$$

where as before U^* denotes the adjoint of U.

The definition (1.26) and the statements (a)–(d) above, immediately set orthogonal matrices apart from most other matrices, and it would be enough to consider them very important in theory and applications. But there is more.

A very important and useful fact about orthogonal matrices is that they preserve the norm of a vector or a matrix. More precisely, we have the following theorem.

Theorem 1.20 *Let Q be an orthogonal matrix. For any vector $x \in \mathbb{R}^n$ and an arbitrary matrix A we have*

$$\|Q\,x\|_2 = \|x\|_2, \quad \text{and} \quad \|Q\,A\|_2 = \|A\|_2. \tag{1.27}$$

Proof. First, observe that $\|Qx\|_2^2 = x^T Q^T Q x = x^T x = \|x\|_2^2$. Using this fact, we now have

$$\|QA\|_2 = \max_{\|x\|_2=1} \|QAx\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \|A\|_2.$$

□

The next important property of an orthogonal matrix is especially related to what is known as sensitivity in the solution of several problems. First we need the following definition.

Definition 1.21 *The* **condition number** *of a square matrix A is defined as*

$$\text{cond}\,(A) = \|A\|\,\|A^{-1}\|. \tag{1.28}$$

This number represents how well or ill-posed a matrix can be, in the sense that how much we can rely on the computations performed with such a matrix, such as solving systems of equations, due to potential accumulation of round–off errors. Computations involving matrices with large condition numbers are potentially very inaccurate; roughly speaking, with $\text{cond}(A) \approx 10^k$, one can expect to lose about k digits of precision in numerical computations.

For an arbitrary matrix A, the condition number has the property

$$\text{cond}(A) \geq 1,$$

and for any orthogonal matrix Q, we have

$$\text{cond}(Q) = 1,$$

the minimum possible. See Exercise 1.45.

It is important to remark that due to these properties, errors are not magnified when performing matrix computations with orthogonal matrices. For instance, when reducing a matrix A to a triangular form, operations with orthogonal matrices such as successive

products of the type QA are performed safely because the condition number of QA is the same as that of A. In general, with products of the type MA, there is the actual risk that the condition number of MA is much larger than that of A, making the computations numerically unstable and unreliable. This does not happen with orthogonal matrices. See also Exercises 1.42 and 1.43.

- MATLAB command: cond(A)

Remark 1.22 *It is important to stress again that orthogonal matrices are the ideal tools to use in numerical computations. They are stable, errors are not magnified, and there is no need to compute inverses. In addition, their condition number is 1, the best possible.*

Now we introduce one of the most important orthogonal matrices that are extensively used in matrix computations and that we will use explicitly in Chapter 3. This matrix has the additional and convenient property of symmetry.

Definition 1.23 (Householder matrix). *Let u be a unit vector in $\mathbb{R}^n$. Then the matrix*

$$H = I - 2\,u\,u^T \tag{1.29}$$

is called Householder matrix.

Theorem 1.24 *The matrix H in (1.29) is orthogonal and symmetric.*

Proof. Since $(uu^T)^T = uu^T$, we have $H^T = H$ and hence H is symmetric. In addition, since $u^T u = 1$, we get

$$\begin{aligned} H^T H &= (I - 2uu^T)^T (I - 2uu^T) = (I - 2uu^T)\,(I - 2uu^T) \\ &= I - 4uu^T + 4u\,u^T u\,u^T = I, \end{aligned}$$

and therefore H is orthogonal.

□

A Householder matrix (1.29) is also known as *Householder reflection* because given an arbitrary vector $x \in \mathbb{R}^n$, the vector Hx is a reflection of x with respect to the hyperplane $u^\perp$, which is the set of all vectors perpendicular to u. In other words, the vectors x and Hx have exactly the same orthogonal projection onto that hyperplane. See e.g. Figure 1.2.

■ **EXAMPLE 1.23**

Let $u = [\frac{1}{\sqrt{2}} \quad \frac{1}{\sqrt{2}}]^T$. Then, the associated Householder matrix is

$$H = I - 2\,u\,u^T = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}.$$

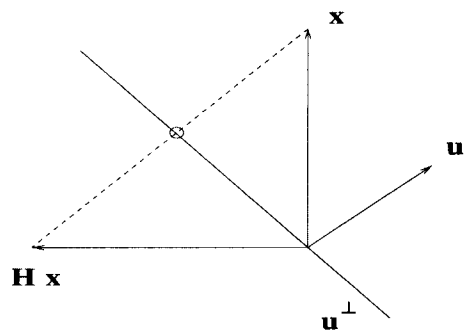


Figure 1.2 Householder reflection of x.

Now let $x = [0 \quad 2]^T$; then $Hx = [-2 \quad 0]^T$, which as we can see from Figure 1.2 is a reflection of x with respect to $u^\perp$.

Remark 1.25 *For an arbitrary vector $u \neq 0$, a Householder matrix is defined as*

$$H = I - \frac{2\,u\,u^T}{u^T u}. \tag{1.30}$$

Orthonormal extensions. One important application of Householder matrices is extending a given vector $x \in \mathbb{R}^n$ to a set of n orthonormal vectors. The idea is to get an orthogonal matrix H, with its first column being the vector x, normalized, if necessary. This tool will be crucial, e.g., when proving several theorems in Chapter 3.

Let $x \in \mathbb{R}^n$ be a vector with $\|x\|^2 = x^T x = 1$. Define the vector

$$u = x - e_1.$$

First observe that $u^T x = x^T x - e_1^T x = 1 - e_1^T x$ and that $u^T u = x^T x - x^T e_1 - e_1^T x + e_1^T e_1 = 2(1 - e_1^T x)$, so that $u^T u = 2u^T x$. Then,

$$\begin{aligned} Hx &= \left(I - \tfrac{2uu^T}{u^T u}\right) x = x - \tfrac{2u(u^T x)}{u^T u} = x - \tfrac{(2u^T x)u}{u^T u} \\ &= x - u = e_1. \end{aligned}$$

The columns of this Householder matrix H are of course orthonormal, but according to the equality $Hx = e_1$, we have that $x = H^T e_1 = He_1$. That is, x is the first column of H.

■ **EXAMPLE 1.24**

$$\text{Let} \quad x = \frac{1}{6}\begin{bmatrix} 4 \\ 0 \\ -4 \\ 2 \end{bmatrix}. \quad \text{Take} \quad u = x - e_1 = \frac{1}{6}\begin{bmatrix} -2 \\ 0 \\ -4 \\ 2 \end{bmatrix}.$$

Then, the matrix

$$H = I - \frac{2uu^T}{u^Tu} = \frac{1}{6}\begin{bmatrix} 4 & 0 & -4 & 2 \\ 0 & 6 & 0 & 0 \\ -4 & 0 & -2 & 4 \\ 2 & 0 & 4 & 4 \end{bmatrix}$$

is orthogonal, with x as its first column.

We will see some more applications of Householder matrices in Chapter 3.

1.5 VECTOR SPACES

Although vector spaces can be very general, we restrict our attention to vector spaces in finite dimensions, and unless otherwise stated, we will always consider the scalars to be real numbers.

We know that if we add, subtract, multiply, or divide (except division by zero) two real numbers, we get a real number again. We know that there is the number zero and the number one, with the properties that $x + 0 = 0 + x = x$ and $x \cdot 1 = 1 \cdot x = x$. We also know that the real numbers have the commutative and distributive properties; namely: $x + y = y + x$, $x \cdot y = y \cdot x$ and $x + (y + z) = (x + y) + z$, and so on. Because of all these properties, working with real numbers is simple and flexible, and we usually take all these properties for granted.

As linear algebra deals not only with real numbers but also, and especially, with vectors and matrices, we would like to provide vectors and matrices with properties similar to those of real numbers, so that we can handle them in a simple manner. It happens that this is possible not only for vectors and matrices but also for several other objects. Thus, in general, we have the following definition.

Definition 1.26 *A set V of elements (called vectors) having two operations, addition and scalar multiplication, is a* **vector space** *if the following properties are satisfied:*

1. $u + v \in V$, *for all* $u, v \in V$.
2. $cu \in V$, *for any scalar* c *and* $u \in V$.
3. $u + v = v + u$, *for all* $u, v \in V$.
4. $u + (v + w) = (u + v) + w$, *for all* $u, v, w \in V$.
5. *There exists a* zero *vector such that* $u + 0 = u$, *for all* $u \in V$.
6. $1u = u$, *for all* $u \in V$.

7. *For every u in V, there exists a* negative *vector $-u$, such that $u + (-u) = 0$.*

8. $c(u + v) = cu + cv$, *for every scalar c and all $u, v \in V$.*

9. $(c_1 + c_2)u = c_1u + c_2u$, *for every scalars c_1, c_2 and all $u \in V$.*

10. $c_1(c_2u) = (c_1c_2)u$, *for every scalars c_1, c_2 and all $u \in V$.*

■ **EXAMPLE 1.25**

The real numbers with the usual addition and multiplication obviously form a vector space.

■ **EXAMPLE 1.26**

The set $\mathbb{R}^n$ with the usual operations of addition and scalar multiplication forms a vector space. This is somehow the "canonical" example of a vector space. In fact, when we think about vectors, we tend to think about objects of the form $[x_1 \cdots x_n]^T$, with the usual geometric representation as arrows. But vectors are elements of more general sets.

■ **EXAMPLE 1.27**

The set of matrices with the usual addition and scalar multiplication (a real number times a matrix), forms a vector space.

■ **EXAMPLE 1.28**

The set of functions with the real numbers as their domain, and with the usual addition and scalar multiplication, forms a vector space.

Given a vector space V, some of its subsets may be vector spaces on their own, with the same operations of addition and scalar multiplication. More formally, we can characterize them through the following definition.

Definition 1.27 *Let V be a vector space and U a nonempty subset of V. Then, we say U is a vector* **subspace** *of V if*

(i) $u + v \in U$, *for all $u, v \in U$.*

(ii) $cu \in U$, *for all scalars c and all $u \in U$.*

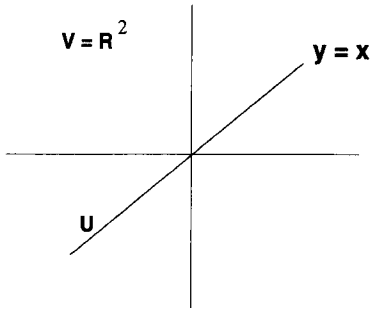


Figure 1.3 The line $y = x$ is a subspace U of $V = R^2$.

This definition simply says that operations performed with elements of U will generate vectors that will stay within U.

■ **EXAMPLE 1.29**

Let $V = \mathbb{R}^2$. Then V is a vector space (see Example 1.26), and let $U = \{(x, y) \in R^2 : y = x\}$. In other words, U is the line $y = x$. Clearly, U is a subset of V. We need to verify the conditions in Definition 1.27.

(i) Let u, v be arbitrary elements in U. Then, $u = (u_1, u_1)$ and $v = (v_1, v_1)$, for some real numbers, u_1, v_1. Thus, $u + v = (u_1 + v_1, u_1 + v_1)$, and hence, $u + v \in U$.

(ii) Let c be any real scalar, and u an arbitrary vector in U. Then, $c\,u = c\,(u_1, u_1) = (cu_1, cu_1)$, and therefore, $c\,u \in U$.

Hence, U is a subspace of V. See Figure 1.3.

■ **EXAMPLE 1.30**

Let again $V = \mathbb{R}^2$, and let $U = \{(x, y) \in R^2 : y = x + 1\}$. Then, U is not a vector subspace of V (see Exercise 1.49).

Observe that because condition (ii) in Definition 1.27 must be true for $c = 0$, a vector subspace (and in general any vector space) must contain the zero vector. The subset U in Example 1.30 fails to be a vector subspace because it does not contain the origin.

■ **EXAMPLE 1.31**

Let V be the vector space of all continuous functions on an interval $[a,\ b]$. Then, the set U of all polynomials of degree at most n, defined on the interval $[a,\ b]$, is a vector subspace of U.

■ **EXAMPLE 1.32**

Let V be the vector space of square matrices of order n. Then, the set U of symmetric matrices of order n is a vector subspace of V.

■ **EXAMPLE 1.33**

(Sum of subspaces). If U and W are subspaces of V, then the set

$$U + W = \{u + w : \ u \in U, \ w \in W\} \tag{1.31}$$

(called the sum of U and W) is a subspace of V.

Here is a specific example of a sum of subspaces.

■ **EXAMPLE 1.34**

Let V be the vector space of square real matrices of order 2, and let

$$U = \left\{ \begin{bmatrix} 0 & a \\ 0 & b \end{bmatrix} : a, b \in \mathbb{R} \right\}, \qquad W = \left\{ \begin{bmatrix} 0 & 0 \\ c & d \end{bmatrix} : c, d \in \mathbb{R} \right\}$$

Clearly, U and W are subspaces of V, and

$$U + W = \left\{ \begin{bmatrix} 0 & e \\ f & g \end{bmatrix} : e, f, g \in \mathbb{R} \right\},$$

which is a vector subspace of V.

■ **EXAMPLE 1.35**

A very important example of subspace is that of the column space of a matrix, which is introduced in Section 1.8.

1.6 LINEAR INDEPENDENCE AND BASIS

The concepts introduced in this section are related to a fundamental notion in linear algebra: a **linear combination** of a set of vectors $v_1, \ldots, v_k$, which is an expression of the form

$$c_1 v_1 + \cdots + c_k v_k,$$

where the $c_1, \ldots, c_k$ are scalars.

Definition 1.28 *Let V be a vector space, and $U \subseteq V$ a vector subspace of V. A set of vectors $S = \{v_1, \ldots, v_k\}$ in V is said to* **span** *U if any vector $x \in U$ can be written as a linear combination of elements in S; that is,*

$$x = c_1 v_1 + \cdots + c_k v_k, \tag{1.32}$$

for some scalars $c_1, \ldots, c_k$. This linear combination does not need to be unique.

■ **EXAMPLE 1.36**

Let $V = \mathbb{R}^2$. Then, the set

$$S_1 = \{[-1 \quad 1]^T, [0 \quad 1]^T, [2 \quad 2]^T\}$$

spans V; that is, any vector in $\mathbb{R}^2$ can be expressed as a linear combination of the vectors in S_1. For instance,

$$\begin{bmatrix} 3 \\ -4 \end{bmatrix} = -5 \begin{bmatrix} -1 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 2 \\ 2 \end{bmatrix}. \tag{1.33}$$

However, the combination in (1.33) is not unique. Indeed, instead of the scalars $-5, 3, -1$ in (1.33), we could also use $-7, 7, -2$ and the new combination would still gives us $[3 \quad -4]^T$. This indicates again that in general the linear combination (1.32) need not be unique.

■ **EXAMPLE 1.37**

Again, let $V = \mathbb{R}^2$, and let $S_2 = \{[-1 \quad 1]^T, [2 \quad 2]^T\}$. Observe that S_2 is formed with only two vectors of S_1 from Example 1.36, but it also spans V. In particular, we can write

$$\begin{bmatrix} 3 \\ -4 \end{bmatrix} = -\frac{7}{2} \begin{bmatrix} -1 \\ 1 \end{bmatrix} - \frac{1}{4} \begin{bmatrix} 2 \\ 2 \end{bmatrix}. \tag{1.34}$$

Thus, it seems that although S_1 does span V, it has one element too many. This is not the case with S_2. Furthermore, the combination (1.34) is unique.

■ **EXAMPLE 1.38**

If the sets of vectors T_1, T_2 span the subspaces V and W respectively, then the set $T_1 \bigcup T_2$ spans the sum subspace $V + W$.

In fact, let $T_1 = \{v_1, \ldots, v_p\}$ and $T_2 = \{w_1, \ldots, w_q\}$, and let $x \in V + W$ be arbitrary. Then $x = v + w$, for some $v \in V$ and $w \in W$. Then, we can write $x = (\alpha_1 v_1 + \cdots + \alpha_p v_p) + (\beta_1 w_1 + \cdots + \beta_q w_q)$, which implies that x lies in the span of $T_1 \bigcup T_2$.

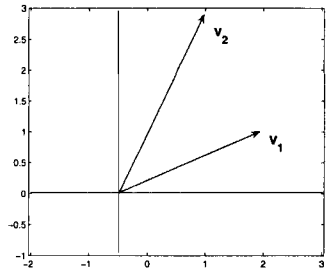


Figure 1.4 v_1 and v_2 in Example 1.39 are linearly independent.

The fundamental difference between the sets S_1 and S_2 of Examples 1.36 and 1.37 is in a central concept in linear algebra given in the following definition.

Definition 1.29 *A set of vectors* $\{v_1, ..., v_n\}$ *is said to be* **linearly independent** *if*

$$c_1 v_1 + \ldots + c_n v_n = 0 \quad \text{implies} \quad c_1 = \cdots = c_n = 0. \tag{1.35}$$

In simple words, this definitions says that if $v_1, ..., v_n$ are linearly independent, none of them can be written as a combination of the others. (Otherwise, say $c_1 \neq 0$; then, we can write v_1 as the linear combination $v_1 = -\frac{c_2}{c_1} v_2 - \cdots - \frac{c_n}{c_1} v_n$.

■ **EXAMPLE 1.39**

Let $V = R^2$ and let $v_1 = [2 \quad 1]^T$ and $v_2 = [1 \quad 3]^T$. To show linear independence, assume that $c_1 v_1 + c_2 v_2 = [0 \quad 0]^T$. Then, this equality gives the system

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

whose solution is $c_1 = c_2 = 0$. Therefore, v_1 and v_2 are linearly independent. Observe that in this case, linear independence means the vectors are not parallel or. multiple of each other. See Figure 1.4.

■ **EXAMPLE 1.40**

The set S_1 in Example 1.36 is not linearly independent. In fact, any of the vectors in S_1 can be written as a combination of the other two. For instance,

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -1 \\ 1 \end{bmatrix} + \frac{1}{4} \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

However, the set S_2 in Example 1.37 is linearly independent.

■ **EXAMPLE 1.41**

Let $V = P_2$ be the vector space of real polynomials of degree at most 2. The set $\{ f_1 = x^2 + 3x - 1,\ f_2 = x + 3,\ f_3 = 2x^2 - x + 1 \}$ is linearly independent. In fact, if $c_1 f_1 + c_2 f_2 + c_3 f_3 = 0$, we get the system

$$\begin{aligned} c_1 \quad + \quad 2c_3 &= 0 \\ 3c_1 + c_2 - c_3 &= 0 \\ -c_1 + 3c_2 + c_3 &= 0, \end{aligned}$$

whose solution is $c_1 = c_2 = c_3 = 0$.

A set S of vectors could span a vector space V but not necessarily be linearly independent, and some other set may be linearly independent but not span V. If a given set of vectors is linearly independent and at the same time spans a given vector space, then it is a very special set.

Definition 1.30 *A set of vectors $\mathcal{B} = \{v_1, ..., v_n\}$ is said to be a* **basis** *of a vector space V if it spans V and is linearly independent.*

The most important application of this definition is that if x is an arbitrary element of V, then it can be written in a *unique* form as a linear combination of the elements of the basis. That is,

$$x = c_1 v_1 + \cdots + c_n v_n, \tag{1.36}$$

for some *unique* scalars $c_1, \ldots, c_n$. In other words, the scalars or coefficients c_i in (1.36) uniquely determine the vector x on a given basis. For a different basis, there will be different coefficients, but again this combination is unique, on that basis. From Definition 1.30 we can say that a basis $\mathcal{B}$ is a genuine (although not unique) representative of a vector space V.

■ **EXAMPLE 1.42**

The set of vectors

$$\mathcal{B} = \{ e_1 = [1 \quad 0 \cdots 0]^T,\ e_2 = [0 \quad 1 \cdots 0]^T, \ldots,\ e_n = [0 \quad 0 \cdots 1]^T \}$$

is a basis for $V = \mathbb{R}^n$. It is known as the canonical or standard basis of $\mathbb{R}^n$.

Observe for instance that

$$\begin{bmatrix} 8 \\ 3 \\ 4 \end{bmatrix} = 8 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 3 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 4 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Thus, in the basis $\mathcal{B} = \{e_1, e_2, e_3\}$, the (unique) coefficients in the linear combination are exactly the entries of the given vector. That is why it is called the canonical or standard basis.

- MATLAB commands: orth(A).

■ **EXAMPLE 1.43**

The set of vectors

$$\mathcal{B} = \{ v_1 = [1 \quad 0 \quad 2]^T, v_2 = [-2 \quad 1 \quad 8]^T, v_3 = [0 \quad 2 \quad 0]^T \}$$

is a basis of $\mathbb{R}^3$, so that any vector $x \in \mathbb{R}^3$ can be written as a unique combination of such basis vectors. By instance,

$$\begin{bmatrix} 8 \\ 3 \\ 4 \end{bmatrix} = 6 \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} - 1 \begin{bmatrix} -2 \\ 1 \\ 8 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}.$$

This means that on the basis $\mathcal{B}$ above, the vector $[8 \quad 3 \quad 4]^T$ is fully represented by its coordinates $\{6, -1, 2\}$. Similarly, on the basis $\mathcal{B}$ of Example 1.42, the same vector is represented by its corresponding coordinates $\{8, 3, 4\}$.

■ **EXAMPLE 1.44**

The set of matrices

$$\mathcal{B} = \left\{ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \cdots \quad, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right\} \tag{1.37}$$

is a basis for the vector spaces of squares matrices of order 4, and just as in Example 1.42, it is known as the canonical or standard basis of such space. This basis is particularly important (among other things) in image compression, which we study in Chapter 5.

■ **EXAMPLE 1.45**

The set $\mathcal{B} = \{ f_1 = 1, f_2 = x, f_3 = x^2 \}$ is a basis for the vector space P_2 of the polynomials of degree at most two.

It is not difficult to see that the set $\mathcal{B}$ in Example 1.45 spans P_2. Let us see that it is also linear independent. Assume that $c_1 f_1 + c_2 f_2 + c_3 f_3 = 0$. This gives $c_1 \cdot 1 + c_2 x + c_3 x^2 = 0$. Comparing coefficients, we immediately get $c_2 = c_3 = 0$, leaving us with $c_1 \cdot 1 = 0$. Thus, c_1 must also be zero. This proves $\mathcal{B}$ is linearly independent.

The next is a particular case of a very interesting and useful class of polynomials called **Bernstein polynomials**, which form a basis for the vector space P_n of polynomial of degree at most n.

■ **EXAMPLE 1.46**

The following set of polynomials:

$$\begin{aligned} B_{0,3}(t) &= (1-t)^3, \\ B_{1,3}(t) &= 3t(1-t)^2, \\ B_{2,3}(t) &= 3t^2(1-t), \\ B_{3,3}(t) &= t^3. \end{aligned}$$

is a basis for the vector space P_3 of the real polynomials of degree at most three. In fact, any polynomial of degree at most three can be expressed as a unique combination of the Bernstein polynomials above. By instance, if $p(t) = 25t^3 - 21t^2 - 3t + 2$, then

$$p(t) = 2B_{0,3}(t) + B_{1,3}(t) - 7\,B_{2,3}(t) + 3\,B_{3,3}(t).$$

We will study Bernstein polynomials and their applications in detail later in Section 1.12.

Intuition tells us that the real line $\mathbb{R}$ is one-dimensional, the plane $\mathbb{R}^2$ is two-dimensional, and the space $\mathbb{R}^3$ is three-dimensional. This coincides with the number of vectors in the bases of each of of these spaces; e.g., $\mathcal{B} = \{e_1, e_2\}$ is the canonical basis of R^2. These ideas generalize into the following definition.

Definition 1.31 *If a vector space V has a basis consisting of m vectors, then we say the* **dimension** *of V is m, and we write* $\dim(V) = m$.

■ **EXAMPLE 1.47**

In Example 1.29, $\dim(V) = 2$ and $\dim(U) = 1$.

Remark 1.32 *If U is a subspace of V, then* $\dim(U) \leq \dim(V)$, *and if* $\dim(U) = \dim(V)$, *then we must have $U = V$.*

■ **EXAMPLE 1.48**

The basis in Example 1.45 has 3 elements; therefore, the dimension of the vector space P_2 of real polynomials of degree at most 2, is 3. Similarly, the basis in

Example 1.46 indicates that the dimension of the vector space P_3, is 4. In general, the dimension of P_n is $n+1$.

The next result expresses the relationship between the dimensions of two vectors subspaces and that of its sum.

Theorem 1.33 *If V and W are subspaces of U, then*

$$\dim(V+W) = \dim(V) + \dim(W) - \dim(V \cap W). \tag{1.38}$$

Proof. Let $\mathcal{R} = \{u_1, \ldots, u_r\}$ be a basis of $V \cap W$. For some positive integers m and n, extend this set $\mathcal{R}$ to form the set $B_V = \{u_1 \ldots, u_r, v_1, \ldots, v_m\}$, a basis of V, and the set $B_W = \{u_1, \ldots, u_r, w_1, \ldots, w_n\}$, a basis of W. In Example 1.38, we saw that the set $\mathcal{B} = B_V \cup B_W$ spans $V+W$. Let us prove that $\mathcal{B}$ is also linearly independent. Thus, assume that

$$\sum_{i=1}^{r} a_i u_i + \sum_{j=1}^{m} b_j v_j + \sum_{k=1}^{n} c_k w_k = 0 \tag{1.39}$$

for some scalars a_i, b_j, c_k. This implies that

$$\sum_{i=1}^{r} a_i u_i + \sum_{j=1}^{m} b_j v_j = -\sum_{k=1}^{n} c_k w_k.$$

Since the left–hand side is a vector in V, then the vector $\sum_{k=1}^{n} c_k w_k$ lies in $V \cap W$ and therefore can be written as $\sum_{k=1}^{n} c_k w_k = \sum_{i=1}^{r} d_i u_i$, for some scalars d_i. Thus,

$$\sum_{k=1}^{n} c_k w_k - \sum_{i=1}^{r} d_i u_i = 0.$$

But since $\mathcal{B}_{\mathcal{W}}$ is linearly independent, we must have $c_k = 0$ and $d_i = 0$, for all $k = 1, \ldots, n$ and $i = 1, \ldots, r$. Then from (1.39), we get

$$\sum_{i=1}^{r} a_i u_i + \sum_{j=1}^{m} b_j v_j = 0.$$

Similarly, since $\mathcal{B}_{\mathcal{V}}$ is linearly independent, we must have $a_i = 0$ and $b_j = 0$, for all $i = 1, \ldots, r$ and $j = 1, \ldots, m$. Thus, $\mathcal{B}$ is linearly independent and therefore a basis of $V+W$. Finally, using the definition of dimension, we get

$$\dim(V+W) = r+m+n = (r+m)+(r+n)-r = \dim V + \dim W - \dim(V \cap W).$$

□

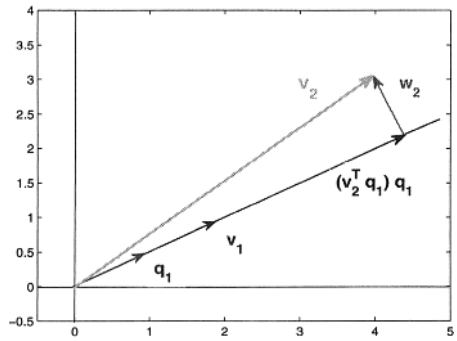


Figure 1.5 Gram-Schmidt: w_2 is orthogonal to q_1.

1.7 ORTHOGONALIZATION AND DIRECT SUMS

Given a set of linearly independent vectors, it is always possible to transform it into an orthonormal set. In particular, this means that given a basis of a vector space, we can always transform such basis into an orthonormal one. In fact, due to some technical advantages, several applications start by considering only orthonormal bases. Here we introduce a well–known procedure to orthonormalize a linearly independent set of vectors.

Gram–Schmidt process: Given m linearly independent vectors

$$\{v_1, ..., v_m\}$$

in $\mathbb{R}^n$, $(m \leq n)$, an orthonormal set of vectors $\{q_1, ..., q_m\}$ can be obtained by defining

$$\begin{aligned} & & & q_1 = \frac{v_1}{\|v_1\|}, \\ w_2 &= v_2 - (v_2^T q_1)\, q_1, & & q_2 = \frac{w_2}{\|w_2\|}, \\ w_3 &= v_3 - (v_3^T q_1)\, q_1 - (v_3^T q_2)\, q_2, & & q_3 = \frac{w_3}{\|w_3\|}, \\ & \vdots & & \vdots \\ w_m &= v_m - \sum_{k=1}^{m-1} (v_m^T q_k)\, q_k, & & q_m = \frac{w_m}{\|w_m\|}. \end{aligned} \tag{1.40}$$

Observe that in general the vector $v_2 - q_1$ need not be orthogonal to q_1. This is why in the definition of w_2, the vector q_1 is first rescaled by a factor $v_2^T q_1$ so that w_2 is orthogonal to q_1. A similar idea is applied to the remaining vectors w_i. See Figure 1.5.

- MATLAB command: orth(A).

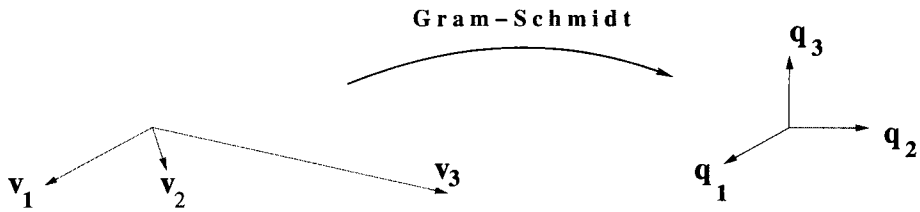


Figure 1.6 Gram–Schmidt orthonormalization.

■ **EXAMPLE 1.49**

Let $v_1 = [2 \ -1 \ \ 0]^T$, $v_2 = [1 \ \ 0 \ -1]^T$, and $v_3 = [3 \ \ 7 \ -1]^T$. Then, following (1.40),

$$\|v_1\| = \sqrt{5}, \qquad q_1 = \tfrac{1}{\sqrt{5}}[2 \ -1 \ \ 0]^T.$$

$$w_2 = v_2 - \tfrac{2}{\sqrt{5}} q_1 = \tfrac{1}{5}[1 \ \ 2 \ -5]^T, \quad \|w_2\| = \tfrac{\sqrt{30}}{5},$$

$$q_2 = \tfrac{w_2}{\|w_2\|} = \tfrac{1}{\sqrt{30}}[1 \ \ 2 \ -5]^T.$$

$$w_3 = v_3 - \tfrac{-1}{\sqrt{5}} q_1 - \tfrac{22}{\sqrt{30}} q_2 = \tfrac{1}{3}[8 \ \ 16 \ \ 8]^T, \quad \|w_3\| = \tfrac{8}{3}\sqrt{6},$$

$$q_3 = \tfrac{w_3}{\|w_3\|} = \tfrac{1}{\sqrt{6}}[1 \ \ 2 \ \ 1]^T.$$

The set $\{q_1, q_2, q_3\}$ is orthonormal. See Figure 1.6.

Direct Sums: In some special cases, every element of a vector space can be expressed as a unique sum of two vectors, each one lying in one of two different vector subspaces. More formally, we have the following definition.

Definition 1.34 *Let V be a vector space, and let U and W be two subspaces of V. Then, we say that U and W form a* **direct sum** *of V, denoted by $V = U \oplus W$, if*

$$U \oplus W = \{v \in V : v = u + w, \quad u \in U, w \in W, \quad U \cap W = 0\}. \tag{1.41}$$

This definition implies that V is spanned by U and W, and that every vector in V can be uniquely expressed as $v = u + w$.

Remark 1.35 *If $V = U \oplus W$, then U and W are called* complementary spaces. *The vectors u and w in (1.41) are called the* projection *of v onto U along W and the* projection *of v onto W along U, respectively.*

■ **EXAMPLE 1.50**

Let $V = \mathbb{R}^3$, and consider two subspaces: U the xy plane, and W the line spanned by $w = [1 \quad 1 \quad 1]^T$. Then, $V = U \oplus W$. In fact, any vector $v = [x \quad y \quad z]^T$ can be expressed as a unique sum of a vector of U and a vector of W:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x - z \\ y - z \\ 0 \end{bmatrix} + \begin{bmatrix} z \\ z \\ z \end{bmatrix}.$$

Observe that $U \cap W = \{0\}$.

■ **EXAMPLE 1.51**

In Example 1.34, we can observe that $U \cap W \neq \{0\}$ and, therefore, this implies that $U + W$ is not a direct sum.

For a very useful example of direct sums, consider now an arbitrary vector subspace U of a vector space V. We can define another vector subspace, the **orthogonal complement of** U, which consists of all vectors that are orthogonal to each vector in U (and it is usually called the "**perp**" of U):

$$U^{\perp} = \{v \in V : v^T u = 0, \quad \forall u \in U\}. \tag{1.42}$$

A direct sum of subspaces can be obtained by using the perp of a vector subspace. In fact, we have the following result.

Theorem 1.36 *Let V be a vector space, and let U be an arbitrary vector subspace of V. Then,*

$$V = U \oplus U^{\perp}.$$

Proof. *Let v be an arbitrary vector in V, and let $\{u_1, \ldots, u_m\}$ be an orthonormal basis of U. Then, similar to the way we defined the vectors in the Gram–Schmidt process (1.40), let*

$$u = (v^T u_1)u_1 + \cdots + (v^T u_m)u_m.$$

Then, the vector $w = v - u$ is orthogonal to each u_i, $i = 1, \ldots, m$, and therefore, it is in $U^{\perp}$. Hence, $v = u + w$, where $u \in U$ and $w \in U^{\perp}$. Also, it is clear that the only intersection of U and $U^{\perp}$ is the zero vector, for if $u \in U$ and $u \in U^{\perp}$, then

$$\|u\|^2 = u^T u = 0.$$

This proves that in fact, $V = U \oplus U^{\perp}$.

□

■ **EXAMPLE 1.52**

If $V = \mathbb{R}^3$, and U is the XY-plane, then $U^\perp$ is the Z-axis. Then, it is clear that $V = U \oplus U^\perp$. In fact, every vector $[x \quad y \quad z]^T$ can be uniquely written as $[x \quad y \quad z]^T = [x \quad y \quad 0]^T + [0 \quad 0 \quad z]^T$.

■ **EXAMPLE 1.53**

A matrix $A_{n\times n}$ is said to be **skew symmetric** if $A = -A^T$. If we let S be the subspace of $n \times n$ symmetric matrices and K the subspace of $n \times n$ skew symmetric matrices, then

$$R^{n\times n} = S \oplus K.$$

Here is an example of a skew symmetric matrix: $A = \begin{bmatrix} 0 & 2 & 5 \\ -2 & 0 & -4 \\ -5 & 4 & 0 \end{bmatrix}$.

Note. In the complex case, we say that a matrix is A skew hermitian if

$$A = -A^*.$$

Later on we will see some particular and useful examples of direct sums.

1.8 COLUMN SPACE, ROW SPACE, AND NULL SPACE

Associated with every $m \times n$ matrix A, there are three fundamental vector spaces: the column space of A, col(A), the row space of A, row(A), and the null space of A, $N(A)$, which are important in several applications, e.g. in the solution or least squares solution of a linear system $Ax = b$, in web page ranking, and in information retrieval, topics that we consider in detail later on. Here we will use the concepts of linear independence, span, and so on, that we have introduced in the previous sections.

Definition 1.37 *The* **column space** *of a matrix $A_{m\times n}$ is defined as the vector subspace of $\mathbb{R}^m$ spanned by the columns of A.*

This definition says that col(A) is the vector subspace formed by the vectors obtained through all possible linear combinations of the columns of A.

Another appropriate and useful definition of the column space is given as

$$\mathrm{col}(A) = \{\, y \in R^m : y = A\,x, \quad \text{for } x \in R^n \,\}. \tag{1.43}$$

In other words, the column space of A is the set of vectors that can be expressed as Ax, for some vector $x \in \mathbb{R}^n$. In this sense, the column space is also known as the **range** of A, or the **image** of A, and it is usually denoted as $R(A)$, or $\mathrm{Im}(A)$.

That is,

$$\mathrm{col}(A) = R(A).$$

Remark 1.38 *Observe that each column of $A_{m\times n}$ is an m-dimensional vector; that is why the col(A) is a subspace of $\mathbb{R}^m$.*

Remark 1.39 *Combining Definition 1.37 and (1.43), we conclude that for any $x \in R^n$, Ax is a combination of the columns of A.*

Definition 1.40 *The dimension of the column space of a matrix $A_{m\times n}$ (the number of linearly independent columns of A) is called the* **rank** *of A.*

- MATLAB command: rank(A).

■ **EXAMPLE 1.54**

Let $A = \begin{bmatrix} 1 & -2 \\ 5 & 3 \\ -4 & 7 \end{bmatrix}$, and let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ be arbitrary. Then,

$$A\,x = \begin{bmatrix} x_1 - 2x_2 \\ 5x_1 + 3x_2 \\ -4x_1 + 7x_2 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 5 \\ -4 \end{bmatrix} + x_2 \begin{bmatrix} -2 \\ 3 \\ 7 \end{bmatrix}.$$

This clearly illustrates that in fact Ax is a combination of the columns of A, for any x, as noted in Remark 1.39. We also observe that in this case, since the two columns of A are linearly independent, the col(A) is 2-dimensional; that is, geometrically it is a plane in $\mathbb{R}^3$ spanned by the two columns of the matrix A, and thus, rank(A) $= 2$.

■ **EXAMPLE 1.55**

Let $A = \begin{bmatrix} 2 & -1 & 0 \\ 0 & 1 & -1 \\ 2 & 0 & -1 \end{bmatrix}$. By definition, the col(A) is the set of vectors of the form Ax, for any x. In this case,

$$A\,x = \begin{bmatrix} 2x_1 - x_2 \\ x_2 - x_3 \\ 2x_1 - x_3 \end{bmatrix} = x_1 \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix} + x_2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix}.$$

It is very important to observe in Example 1.55 that although it is true that the col(A) is spanned by (all) the columns of A, and A has three columns, they are not linearly independent. In fact, it is easy to verify that for instance, the third column is a combination of the first two, and that these first two columns are linearly independent. This means that the rank of A, and the dimension of col(A) is two. Thus, we can say that the col(A) is also spanned by just the first two columns of A, and hence it has dimension two. In other words, for Example 1.55, the column space of A is the subspace of vectors that can be obtained via linear combinations of the form

$$\text{col}(A) = c_1 \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix} + c_2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix},$$

where c_1, c_2 are arbitrary scalars. The col(A) is therefore a plane in $\mathbb{R}^3$.

Remark 1.41 *In Example 1.55, we have expressed col(A) as a combination of the first two columns, but it can also be expressed as a combination of the last two, or the first and the last. In any case, there are only two linearly independent columns.*

- MATLAB command: orth

Naturally, we can also consider the subspace spanned by the columns of A^T, or in other words, by the rows of A, and we denote this space by row(A).

Definition 1.42 *The* **row space** *of a matrix $A_{m\times n}$ is defined as the vector subspace of $\mathbb{R}^n$ spanned by the rows of A.*

■ **EXAMPLE 1.56**

For the matrix A of Example 1.54, we have seen that dim col(A) $= 2$. If we now denote the first, second, and third rows of that matrix with r_1, r_2, and r_3 respectively, then we can prove that $r_2 = -47\, r_1 - 13\, r_3$, that is, r_2 is a linear combination of the other two rows, which means that the three rows are not linearly independent. And in fact, we must have dim row $(A) = 2$. In this case, the row space of A is the set of linear combinations of the form

$$\text{row}(A) = c_1 [1 \quad -2] + c_2 [-4 \quad 7],$$

where c_1, c_2 are arbitrary scalars. Thus, row (A) spans the whole space $\mathbb{R}^2$.

It is a very remarkable fact that for any matrix $A_{m\times n}$, the row space has the same dimension as the column space. More formally,

Theorem 1.43 *Let A be an $m \times n$ matrix. Then,*

$$\operatorname{rank}(A) = \dim \operatorname{col}(A) = \dim \operatorname{row}(A). \tag{1.44}$$

Proof. Denote the rows of A by $r_1, r_2, \ldots, r_m$, and let $\dim \operatorname{row}(A) = k$. Assume that $\mathcal{B} = \{v_1, \ldots, v_k\}$ is a basis of $\operatorname{row}(A)$. Then, each row vector r_i, $i = 1, \ldots, m$ can be written as

$$r_i = c_{i1}v_1 + c_{i2}v_2 + \cdots + c_{ik}v_k,$$

for some scalars $c_{i1}, \ldots, c_{ik}$. But the j–th component of each row vector r_i is nothing else than a_{ij}. Thus, with respect to each j–th component ($j = 1, \ldots, n$), the above equation can be written as

$$a_{ij} = c_{i1}v_{1j} + c_{i2}v_{2j} + \cdots + c_{ik}v_{kj}, \quad i = 1, \ldots, m,$$

which is the same as

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix} = v_{1j} \begin{bmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{m1} \end{bmatrix} + v_{2j} \begin{bmatrix} c_{12} \\ c_{22} \\ \vdots \\ c_{m2} \end{bmatrix} + \cdots + v_{kj} \begin{bmatrix} c_{1k} \\ c_{2k} \\ \vdots \\ c_{mk} \end{bmatrix}, \qquad j = 1, \ldots, n.$$

This tells us that the columns of A are linear combinations of the k vectors c. Therefore,

$$\dim \operatorname{col}(A) \leq k = \dim \operatorname{row}(A).$$

By applying the ideas above to A^T, we can also show that $\dim \operatorname{row}(A) \leq \dim \operatorname{col}(A)$, and therefore conclude that indeed $\dim \operatorname{col}(A) = \dim \operatorname{row}(A)$.

□

We next state an important result on the rank of product of matrices.

Theorem 1.44 *Let A and B be two arbitrary matrices whose product AB is well defined. Then,*

(a) $\operatorname{rank}(AB) \leq \operatorname{rank}(A)$, *and* $\operatorname{rank}(AB) \leq \operatorname{rank}(B)$.

(b) If B is nonsingular, then $\operatorname{rank}(AB) = \operatorname{rank}(A)$.

If A is nonsingular, then $\operatorname{rank}(AB) = \operatorname{rank}(B)$.

Proof. (a) If we denote the columns of B by $b_1, \ldots, b_n$, then we have

$$AB = A[b_1 \cdots b_n] = [Ab_1 \cdots Ab_n].$$

By recalling Remark 1.39, we observe that the columns of $AB : Ab_1, \ldots, Ab_n$ are linear combinations of the columns of A, and therefore, we must have that $\operatorname{rank}(AB) \leq \operatorname{rank}(A)$. The second inequality is proved similarly using rows.

(b) If B is nonsingular, we have

$$\operatorname{rank}(A) = \operatorname{rank}(A\,B\,B^{-1}) \leq \operatorname{rank}(AB).$$

By combining this with the first inequality in (a), we conclude that $\text{rank}(AB) = \text{rank}(A)$. The proof of the last statement is similar.

□

We now define the null space of A, denoted as $N(A)$.

Definition 1.45 *The* **null space** *of a matrix $A_{m\times n}$ is defined as the vector subspace of $\mathbb{R}^n$:*

$$N(A) = \{\, x \in R^n : Ax = 0 \,\}. \tag{1.45}$$

■ **EXAMPLE 1.57**

Let $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$. From the definition, we see that the null space is the set $\{[x_1 \quad x_2 \quad x_3]^T : x_1 + 2x_2 + 3x_3 = 0\}$. Geometrically, this is a plane (hence, $N(A)$ is two-dimensional) passing through $(0, 0, 0)$ and with normal vector $[1 \quad 2 \quad 3]^T$. On the other hand, since the three columns of A are just multiples of each other, it is clear that the col(A) is one-dimensional and is spanned by the same vector $[1 \quad 2 \quad 3]^T$. So, for this particular example, col(A) and $N(A)$ are orthogonal.

Note: It is very interesting to learn that in general, col(A) and $N(A^T)$ are orthogonal. In the example above, the matrix A is symmetric ($A^T = A$); that is why col(A) is also orthogonal to $N(A)$. See Theorem 1.48 below and the discussion after it.

Furthermore, there are some orthogonality relationships between these subspaces as the following theorem states.

Theorem 1.46 *Let A be an arbitrary $m \times n$ matrix. Then,*

$$\text{col}(A)^\perp = N(A^T), \quad \text{and} \quad \text{col}(A^T)^\perp = N(A). \tag{1.46}$$

Proof. Let v be any vector in col$(A)^\perp$. For an arbitrary vector $x \in \mathbb{R}^n$, we know that $Ax \in$ col(A); then we have

$$(Ax)^T v = 0 \iff x^T(A^T v) = 0 \iff A^T v = 0 \iff v \in N(A^T).$$

This implies that, in fact, col$(A)^\perp = N(A^T)$. The second part of the theorem is proved similarly.

□

In general, an arbitrary matrix $A_{m\times n}$ generates the four subspaces: col (A), col (A^T), $N(A)$, and $N(A^T)$. These vector subspaces decompose $\mathbb{R}^m$ and $\mathbb{R}^n$ in direct sums, and their dimensions are related by closed formulas.

We can now state the **Fundamental Theorem of Linear Algebra.**

Theorem 1.47 *Let $A_{m\times n}$ be an arbitrary matrix. Then,*

$$\mathbb{R}^m = \mathrm{col}(A) \oplus N(A^T), \qquad \mathbb{R}^n = \mathrm{col}(A^T) \oplus N(A). \tag{1.47}$$

Proof. The vector spaces col(A) and $N(A^T)$ are subspaces of R^m. Similarly, the vector spaces col(A^T) and $N(A)$ are subspaces of R^n. Now, to obtain the result, simply combine Theorems 1.36 and 1.46.

□

An immediate consequence of Theorems 1.43 and 1.47 is the following corollary.

Corollary 1.8.1 *If $A_{m\times n}$ is a matrix with rank$(A) = r$, then*

$$\dim \mathrm{col}(A) = \dim \mathrm{col}(A^T) = r, \quad \dim N(A) = n - r, \quad \dim N(A^T) = m - r. \tag{1.48}$$

At the same time, Corollary 1.8.1 implies the following important result.

Theorem 1.48 (Dimension Formula). *Consider an $m \times n$ matrix A. Then,*

$$\dim \mathrm{col}(A) + \dim N(A) = n. \tag{1.49}$$

■ **EXAMPLE 1.58**

Let $A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$. Observe that we can write the third column a_3 as a combination of the first two a_1, a_2 as $a_3 = 2a_2 - a_1$. And since these first two columns are linearly independent, this implies that col(A) is 2-dimensional. Geometrically then, col(A) is the plane $-x + 2y - z = 0$ passing through the origin, with normal vector $[-1 \quad 2 \quad -1]^T$. From Theorem 1.48, the null space $N(A)$ is one-dimensional. In fact, by solving the system $Ax = 0$, we observe that all solutions are of the form $[x_1 \quad x_2 \quad x_3]^T$, with $x_2 = -2x_1$, $x_3 = x_1$. That is, $N(A)$ is a line spanned by the vector $[-1 \quad 2 \quad -1]^T$. Using this vector, we observe that $N(A)$ is orthogonal to every row of A and therefore $N(A) \perp \mathrm{col}\,(A^T)$, in agreement with Theorem 1.46, and finally, because $N(A)^\perp = \mathrm{col}\,(A^T)$, we also have $\mathbb{R}^3 = \mathrm{col}(A^T) \oplus N(A)$, as in (1.47).

■ **EXAMPLE 1.59**

Let $A = \begin{bmatrix} 2 & 0 & 1 & 3 & 0 \\ 0 & 3 & 1 & 2 & 0 \\ 2 & 3 & 2 & 5 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$. In this case the columns 1, 2, and 5 are linearly independent and they form a basis of col(A). That is, any vector in col(A) (in particular, the third and fourth columns of A) can be expressed as a unique combination of columns 1, 2, and 5. This means that geometrically, col(A) is a 3-dimensional hyperplane (in $\mathbb{R}^4$) spanned by those three columns, and therefore $N(A)$ is a two-dimensional subspace of $\mathbb{R}^5$.

As stated before, one of the applications of column spaces, row spaces, and null spaces is in the solution of linear systems. We can now put together the links between the solution of $Ax = b$ and the vector spaces col(A) and $N(A)$ in a theorem.

Theorem 1.49 *Given an $m \times n$ matrix A, consider solving the system $Ax = b$. Then,*

1. (Existence) *The system has a solution if and only if $b \in$ col(A).*
2. (Uniqueness) *The system has at most one solution for every b if and only if $N(A) = \{0\}$.*

Proof. *Existence*: We have already seen that for arbitrary x, the vector $Ax \in$ col(A). Then, for $Ax = b$ to have a solution, b must lie in the same subspace.

Uniqueness: If $N(A) \neq \{0\}$, then besides $x = 0$, there is another solution to $Ax = b$, with $b = 0$, which is a contradiction. On the other hand, assuming $N(A) = \{0\}$, if there is a b for which $Ax = b$ has more than one solution, that is $Ax_1 = b$ and $Ax_2 = b$, with $x_1 \neq x_2$, then $A(x_1 - x_2) = Ax_1 - Ax_2 = b - b = 0$, which means that $x_1 - x_2 = 0$, or $x_1 = x_2$.

□

Remark 1.50 *Observe that by combining Theorem 1.48 with the the uniqueness part of Theorem 1.49, we conclude that the system $Ax = b$ has a unique solution if and only if dim col(A) $= n$, which means A has to be a full rank matrix, or which is the same, all the columns of A must be linearly independent.*

1.8.1 Linear transformations

Given two arbitrary vector spaces V and W, we can define a function

$$T : V \longrightarrow W$$

that for any two vectors $u, v \in V$ and any scalar c, it satisfies

(a) $T(u+v) = T(u) + T(v)$,
(b) $T(cu) = cT(u)$.

Such a function T is called a **linear transformation** from V to W. In the case $V = W$, it is called a linear operator.

These transformations play an important role in the theory of linear algebra and we refer the reader to any standard textbook in linear algebra for a full discussion on general linear transformations. For completion, here we just give a few remarks on how some terminology for a particular class of linear transformations is closely related to the theory introduced in this chapter.

More precisely, we consider linear transformations

$$T : R^n \longrightarrow R^m. \tag{1.50}$$

Given a vector $x \in \mathbb{R}^n$, the linear transformation T transform such vector x into a vector $y \in \mathbb{R}^m$ so that $T(x) = y$.

The important fact here is that any linear transformation of the form (1.50) has a matrix representation and vice versa, any matrix $A_{m\times n}$ corresponds to a linear transformation (1.50).

■ **EXAMPLE 1.60**

Define

$$T(x_1,\ x_2,\ x_3) = (2x_1 - x_2,\ \ 3x_1 + 4x_2 - x_3).$$

We can write this transformation as $T(x_1, x_2, x_3) = (y_1,\ y_2)$, where

$$\begin{aligned} y_1 &= 2x_1 - x_2, \\ y_2 &= 3x_1 + 4x_2 - x_3. \end{aligned}$$

Or, in matrix-vector notation as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 0 \\ 3 & 4 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

That is,

$$y = Ax.$$

Just as the transformation T, the matrix A takes a 3-dimensional vector x, and via multiplication, it transforms it into a 2-dimensional vector y.

The example above illustrates the fact that

To any linear transformation $T : \mathbb{R}^n \to \mathbb{R}^m$, there corresponds a matrix $A_{m\times n}$.

Remark 1.51 *The correspondence between a linear transformation and a matrix is true not only for the spaces $\mathbb{R}^n$ and $\mathbb{R}^m$, but for any vector spaces V of dimension n and W of dimension m.*

Since there is a correspondence between linear transformations and matrices, then the same is true for some definitions in the two cases.

Given a linear transformation

$$T : V \longrightarrow W,$$

we say that the vector space V is the **domain** and the vector space W is the **codomain** of T.

The set

$$R(T) = \{ y \in W : \ y = T(x) \}$$

is called the **range** or **image** of T.

The set

$$\mathrm{Ker}(T) = \{ x \in V : \ T(x) = 0 \}$$

is called the **kernel** of T.

From these definitions, we immediately see that the range of T is exactly the column space of the corresponding matrix A. Similarly, the kernel of T is exactly the null space of the matrix A.

Linear Transformations	Matrices
range	column space
kernel	null space

Accordingly, the **rank** of the linear transformation T is defined as the dimension of its range. The dimension of the kernel is called the **nullity**.

Some of the definitions and results on matrices we have introduced so far, and some others presented later, have their equivalent counterpart in the context of linear transformations. We speak of zero, identity linear transformations, similarity, eigenvalues, and so on. For a linear transformation $T : V \to W$, where V is a vector space of dimension n and W a vector space of dimension m, the dimension formula (1.49) now reads as

$$\dim R(T) + \dim \mathrm{Ker}(T) = n. \tag{1.51}$$

Remark 1.52 *The set of all linear operators on $\mathbb{R}^n$, denoted by $L(\mathbb{R}^n)$, forms a normed vector space, with* $\|T\| = \max_{\|x\|=1} \|T(x)\|$.

1.9 ORTHOGONAL PROJECTIONS

Definition 1.53 *A projection matrix is a square matrix P such that $P^2 = P$. A projection matrix onto a subspace S is a projection matrix for which range$(P) = S$.*

If P is a projection matrix onto a subspace S, then the condition $P^2 = P$ says that for an arbitrary vector x, since Px is already in S, a new application of P, that is $P(Px)$, should not move the vector Px at all: $P(Px) = Px$.

■ **EXAMPLE 1.61**

Consider the vector subspace S of $\mathbb{R}^2$ to be the line $y = x$. Then, a projection matrix onto S is given by

$$P = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

In fact, for an arbitrary vector $v = [x \quad y]^T$, we get $Pv = [y \quad y]^T$, which lies on S. Furthermore, it is easy to verify that $P^2 = P$.

■ **EXAMPLE 1.62**

Let S be again the subspace of $\mathbb{R}^2$ determined by the line $y = x$, and let $P = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Then, for any vector $x = [x_1 \; x_2]^T$, the vector Px lies on S; that is, ran$(P) = S$. However, P is not a projection because $P^2 \neq P$. In fact, if we take say $x = [0 \quad 1]^T$, then (see Figure 1.7)

$$Px = [1 \quad 1]^T \quad \text{but} \quad P^2(x) = P(Px) = [2 \quad 2]^T \neq [1 \quad 1]^T.$$

■ **EXAMPLE 1.63**

Let S be the subspace of $\mathbb{R}^3$ spanned by $u_1 = [2 \quad 1 \; -1]^T$ and $u_2 = [0 \; -1 \quad 1]^T$. Then the matrix

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

is a projection matrix onto S. One can verify that for an arbitrary vector $v \in \mathbb{R}^3$, the vector Pv can be written as a linear combination of u_1 and u_2.

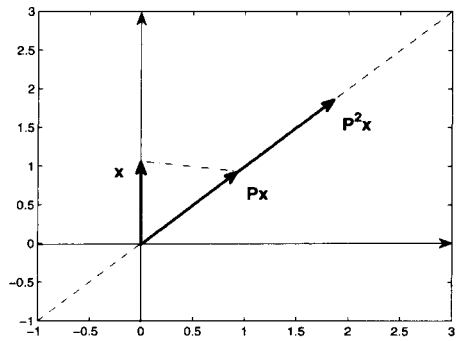


Figure 1.7 Example 1.62: P not a projection.

The projection matrices that probably have more applications are those that are also orthogonal.

Definition 1.54 *An orthogonal projection matrix is a projection matrix P for which $P^T = P$.*

■ **EXAMPLE 1.64**

The matrix

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

is an orthogonal projection. It clearly satisfies $P^2 = P$ and $P^T = P$. It is in fact an orthogonal projection onto the subspace S determined by the line $y = x$ (see Example 1.62).

Remark 1.55 *It is important to note that an orthogonal projection matrix is not necessarily an orthogonal matrix. However, we will use orthogonal matrices to construct orthogonal projection matrices.*

Explicit formulas for orthogonal projections. We want to see how to obtain an explicit expression for an orthogonal projection onto a subspace S, starting with the particular case when $S = \mathrm{col}(A)$.

Later we will see that for a general $m \times n$ matrix A, with $m > n$, the least squares solution to the system $Ax = b$ is given by the solution of the so-called *normal equations* $A^T A x = A^T b$. This is nothing else but a consequence of projecting b orthogonally onto $\mathrm{col}(A)$, as it will be explained in detail in Chapter 4. Observe also that from the normal equations, if the columns of A are linearly independent, we can write (see Exercise 1.54)

$$x = (A^T A)^{-1} A^T b.$$

The matrix

$$A^{\dagger} = (A^T A)^{-1} A^T \tag{1.52}$$

is called the **pseudoinverse matrix** of A.

Now, since Ax is in col(A), and $Ax = A(A^T A)^{-1} A^T b$, then the matrix

$$P = A\,(A^T A)^{-1} A^T \tag{1.53}$$

has taken the vector b onto col(A), so we suspect such a matrix is an orthogonal projection onto col(A). In fact,

$$P^2 = A\,(A^T A)^{-1} A^T\, A\,(A^T A)^{-1} A^T = A\,(A^T A)^{-1} A^T = P,$$

and since $A^T A$ is symmetric,

$$P^T = A(A^T A)^{-T} A^T = A(A^T A)^{-1} A^T = P.$$

■ **EXAMPLE 1.65**

Consider the matrix $A = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \\ 1 & 2 & 1 \end{bmatrix}$. Then, the matrix

$$P = A\,(A^T A)^{-1} A^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{bmatrix}$$

will project an arbitrary vector in $\mathbb{R}^4$ orthogonally onto the col(A).

Remark 1.56 *1) If the matrix A is square and nonsingular, then its pseudoinverse coincides with its inverse matrix. In fact, in this case, we have that $A^{\dagger} = (A^T A)^{-1} A^T = A^{-1} A^{-T} A^T = A^{-1}$.*

2) $A^{\dagger}$ satisfies $A^{\dagger} A = I$ and $A\,A^{\dagger} = P$.

- MATLAB command: pinv(A)

The definition (1.53) of the orthogonal projection onto col(A), although theoretically valid and illustrative, is not useful in practice due especially to the fact that an inverse has to be computed, which is something that must always be avoided, if possible. A more practical, and by far more efficient, way to compute a projection matrix onto col(A) is through orthogonal matrices.

The idea is to compute an orthonormal basis $\{q_1, \ldots, q_n\}$ of col(A) and define the orthogonal matrix $Q = [q_1 \cdots q_n]$. Then the projection is

$$P = Q Q^T. \tag{1.54}$$

Furthermore, this approach is in fact true for a projection matrix onto any vector subspace S. That is, the orthogonal projection matrix onto a vector subspace S can be defined as in (1.54), where Q is a matrix whose columns form an orthonormal basis of S. See Exercise 1.62.

■ **EXAMPLE 1.66**

Let S be the subspace spanned by $v_1 = [1 \quad 2 \quad 3]^T$, and $v_2 = [1 \quad 1 \quad 1]^T$. We can orthonormalize $\{v_1, v_2\}$, say, by using Gram–Schmidt (1.40) to obtain the vectors $q_1 = [1 \quad 2 \quad 3]^T/\sqrt{14}$, $q_2 = [4 \quad 1 \quad -2]^T/\sqrt{21}$. Then, we define the matrix Q with q_1 and q_2 as first column and second column respectively, so that the projection matrix onto S is given by

$$P = Q Q^T = \frac{1}{6}\begin{bmatrix} 5 & 2 & -1 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{bmatrix},$$

where $Q = [q_1 \quad q_2]$. Thus, given any vector $x \in R^3$, the vector Px will be in the subspace S.

■ **EXAMPLE 1.67**

Consider the matrix A of Example 1.65. The vectors

$$v_1 = [1 \quad 0 \quad 0 \quad 0]^T, \; v_2 = [0 \quad 1 \quad 0 \quad 1]^T, \; v_3 = [0 \quad 0 \quad 1 \quad 0]^T$$

form a basis of col(A). They can be orthonormalized using Gram–Schmidt to get

$$q_1 = [1 \quad 0 \quad 0 \quad 0]^T, \; q_2 = [0 \quad 1/\sqrt{2} \quad 0 \quad 1/\sqrt{2}]^T, \; q_3 = [0 \quad 0 \quad 1 \quad 0]^T.$$

Now define the matrix $Q = [q_1 \quad q_2 \quad q_3]$. Then, the projection matrix onto col(A) is

$$P = Q Q^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{bmatrix},$$

which is exactly what we had obtained before by using $P = A\,(A^T A)^{-1} A^T$.

Given an orthogonal projection P onto a subspace S, there is an obvious way of determining the orthogonal projection onto $S^{\perp}$.

Theorem 1.57 *If P is the orthogonal projection onto a subspace S, then $I - P$ is the orthogonal projection onto $S^{\perp}$.*

Through these two orthogonal projection matrices, an arbitrary vector $v \in \mathbb{R}^n$ gets effectively decomposed into its orthogonal components. In other words,

$$x = Px + (I - P)x.$$

The orthogonality property can also be observed by multiplying both matrices:

$$\begin{aligned} P(I-P) &= P - P^2 = P - P = 0, \quad \text{and} \\ (I-P)P &= P - P^2 = P - P = 0. \end{aligned} \tag{1.55}$$

We will find this property very useful when we study dynamical systems in Chapter 7.

A projection matrix onto a subspace S is in general not unique, but there is only one orthogonal projection.

Theorem 1.58 *Let V be a vector space, and let S be a vector subspace of V. Then, the orthogonal projection matrix P onto S is unique.*

Proof. Assume there are two orthogonal projections P_1 and P_2 onto the subspace S. Then, for any $x \in V$,

$$\begin{aligned} \|(P_1 - P_2)x\|^2 &= ((P_1x)^T - (P_2x)^T)(P_1x - P_2x) \\ &= x^T P_1 x - (P_1x)^T P_2 x - (P_2x)^T P_1 x + x^T P_2 x \\ &= (P_1x)^T (I - P_2)x + (P_2x)^T (I - P_1)x. \end{aligned}$$

But P_1x, $P_2x \in S$ and $(I - P_1)$, $(I - P_2) \in S^{\perp}$, so that the last expression above is zero and therefore $P_1 = P_2$.

□

1.10 EIGENVALUES AND EIGENVECTORS

Theorem 1.49 gives conditions for the existence and uniqueness of solutions of linear systems of equations of the form

$$Ax = b, \tag{1.56}$$

where A is a general rectangular matrix. We learned that (1.56) has a solution if b is in col(A). For the special case when A is square, which is the case we are interested in, we ask instead for conditions under which the system has a unique solution. The answer can be given in terms of the singularity of the square matrix A.

According to Theorem 1.49, a system with a square coefficient matrix A of order n has a unique solution if $\dim N(A) = 0$, or equivalently, $\dim \mathrm{col}(A) = n$; that is, all the columns of A must be linearly independent, and this is equivalent to saying that A is nonsingular. More precisely,

Theorem 1.59 *Let A be a square matrix of order n. The system $Ax = b$ has a unique solution if and only if the matrix A is nonsingular.*

Remark 1.60 *In particular, this theorem implies that if A is nonsingular, then the system $Ax = 0$ has the unique solution $x = 0$.*

Two essential concepts associated with every square matrix are eigenvalues and eigenvectors; they contain important information about the matrix, some of its associated subspaces, and about the structure of problems and phenomena whose modeling contain such a matrix. There are numerous applications of eigenvalues and eigenvectors, within and outside mathematics, e.g., differential equations, control theory, Markov chains, web page ranking, image compression, etc. We will study in detail some of these applications later on.

Definition 1.61 *Given a square matrix A of order n, we say that λ is an* **eigenvalue** *of A with associated* **eigenvector** *$v \neq 0$ if*

$$Av = \lambda v. \tag{1.57}$$

In general, eigenvalues are complex numbers, and eigenvectors are complex (n–dimensional) vectors, even though A is a real matrix.

One first and immediate geometric observation is that given an eigenvector v, equation (1.57) says that the vector Av is just a multiple of v, with larger or smaller magnitude. In fact, the line containing v is a one-dimensional vector space and Av lands on that same vector space. This is a simple instance of invariance, and we say that the space spanned by v is **invariant** with respect to A. We will return to this concept when we study matrix factorizations in Chapter 3 and dynamical systems in Chapter 7.

Observe that equation (1.57) can be written as

$$(A - \lambda I)v = 0. \tag{1.58}$$

Since we are looking for eigenvectors $v \neq 0$, we need the matrix $(A - \lambda I)$ to be singular (see Remark 1.60). In other words, we need to require that

$$\det(A - \lambda I) = 0. \tag{1.59}$$

Equation (1.59) is known as the *characteristic equation*. The left–hand side is a polynomial on λ of degree n, and the solutions to this equation are the eigenvalues of A. On the other hand, for each given eigenvalue λ, the system (1.58) is used to find a corresponding eigenvector.

- MATLAB command: eig(A)

Remark 1.62 *The eigenvalues are uniquely determined as solutions of the characteristic equation (1.59), but eigenvectors are not unique, as they are solutions of the singular system (1.58).*

■ **EXAMPLE 1.68**

Let $A = \begin{bmatrix} 1 & -2 \\ 0 & 3 \end{bmatrix}$. Then, the characteristic equation is

$$\det \begin{bmatrix} 1-\lambda & -2 \\ 0 & 3-\lambda \end{bmatrix} = (3-\lambda)(1-\lambda) = 0.$$

Then, the eigenvalues of A are $\lambda_1 = 3$ and $\lambda_2 = 1$. To find the corresponding eigenvectors, we use equation (1.58) for each eigenvalue. First let us consider λ =3. Then, equation (1.58) gives

$$\begin{bmatrix} 1-3 & -2 \\ 0 & 3-3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -2 & -2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

from which $v_2 = -v_1$. Thus, any nonzero vector of the form $v = [v_1 \quad -v_1]^T$ is an eigenvector of A corresponding to $\lambda = 3$. In particular, we may take $v = [1 \quad -1]^T$. In a similar way, any eigenvector associated with $\lambda = 1$ is of the form $[v_1 \quad 0]^T$, so in particular we may take $v = [1 \quad 0]^T$.

Remark 1.63 *In general, finding the eigenvalues of upper or lower triangular matrices like the one in Example 1.68 is immediate. The eigenvalues are nothing else but the diagonal entries.*

■ **EXAMPLE 1.69**

There is a particular matrix that appears very often, and its eigenvalues can be seen directly from the matrix itself:

$$\text{If} \quad A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}, \quad \text{then} \quad \lambda_{1,2} = a \pm i\,b. \tag{1.60}$$

■ **EXAMPLE 1.70**

Let $A = \begin{bmatrix} -2 & -1 & 0 \\ 1 & -2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$. Then, the characteristic equation is

$$\det \begin{bmatrix} -2-\lambda & -1 & 0 \\ 1 & -2-\lambda & 0 \\ 0 & 0 & 3-\lambda \end{bmatrix} = \lambda^3 + \lambda^2 - 7\lambda - 15 = 0.$$

The solution of this characteristic equation gives the eigenvalues: $\lambda_{1,2} = -2 \pm i$ and $\lambda_3 = 3$. To find the corresponding eigenvectors, let us consider first $\lambda_1 = -2 + i$. In this case, equation (1.58) gives

$$\begin{bmatrix} -2-(-2+i) & -1 & 0 \\ 1 & -2-(-2+i) & 0 \\ 0 & 0 & 3-(-2+i) \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

$$= \begin{bmatrix} -i & -1 & 0 \\ 1 & -i & 0 \\ 0 & 0 & 5-i \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

From the first or second row, we get that $v_1 = i\,v_2$, and from the third row, we get $v_3 = 0$. Then, any eigenvector associated with λ_1 is of the form $[\,i\,v_2 \quad v_2 \quad 0]^T$. Similarly, any eigenvector associated with λ_2 is for the form $[-i\,v_2 \quad v_2 \quad 0]^T$. Finally, it is easy to see that any eigenvector associated with $\lambda_3 = 3$ is of the form $[0 \quad 0 \quad v_3]^T$, and in particular, we can take $w = [0 \quad 0 \quad 1]^T$.

In some applications, like in the solutions of differential equations, it is necessary to extract real solutions by using eigenvectors (even if they are complex). In this case, using for instance the complex eigenvectors associated with $\lambda_{1,2} = -2 \pm i$, we can write

$$u \pm i\,v := \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \pm i \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

In this way, we can obtain two (linearly independent) real vectors u and v that span a 2-dimensional subspace.

Remark 1.64 *By using the fact that the determinant of any matrix is the same as that of its transpose, one can easily show that any matrix $A_{n\times n}$ and its transpose A^T have the same eigenvalues. However, their eigenvectors are not necessarily the same.*

The matrix A in Example 1.70 has a particular form. Observe that we can represent A as

$$A = \left[\begin{array}{cc|c} -2 & -1 & 0 \\ 1 & -2 & 0 \\ \hline 0 & 0 & 3 \end{array}\right] = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}.$$

Then, finding the eigenvalues of A reduces itself to finding the eigenvalues of A_1 and A_2 separately. In this case, according to (1.60), the eigenvalues of A_1 are $\lambda_{1,2} = -2 \pm \mathrm{i}$, and the eigenvalue of A_2 is $\lambda_3 = 3$.

This is actually a general result in linear algebra to compute the eigenvalues of *block diagonal matrices* : Let us denote with $\sigma(A)$ the set of eigenvalues of A (this set is known as the **spectrum** of A). Then, we have the following theorem.

Theorem 1.65 *Let* $A = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{bmatrix}$. *Then,*

$$\sigma(A) = \sigma(A_1) \cup \sigma(A_2) \cup \cdots \cup \sigma(A_k).$$

A similar result applies if the matrix is block upper triangular. We have the following

Theorem 1.66 *Consider the block upper triangular matrix*

$$A = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix},$$

where B and D are square matrices of order say p and q, respectively, for some positive integers p, q. Then, $\sigma(A) = \sigma(B) \cup \sigma(D)$, counting multiplicities.

Proof. Let λ be an eigenvalue of A with eigenvector $v = [v_1 \;\; v_2]^T$, where $v_1 \in \mathbb{R}^p$ and $v_2 \in \mathbb{R}^q$. Then,

$$Av = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

We either have $v_2 = 0$ or $v_2 \neq 0$. If $v_2 = 0$, then $v_1 \neq 0$ and $Bv_1 = \lambda v_1$, so that λ is an eigenvalue of B. If $v_2 \neq 0$, then $Dv_2 = \lambda v_2$ so that λ is an eigenvalue of D. This proves that $\sigma(A) \subseteq \sigma(B) \cup \sigma(D)$. In addition, the two sets $\sigma(A)$ and $\sigma(B) \cup \sigma(D)$ have the same cardinality. Therefore, they must be equal.

□

Remark. The result in Theorem 1.66 is also true if A is block lower triangular.

■ **EXAMPLE 1.71**

The eigenvalues of $A = \begin{bmatrix} 8 & -2 \\ 2 & 4 \end{bmatrix}$ are $\lambda = 6, 6$, and the eigenvalues of $A = \begin{bmatrix} 5 & 3 & 0 \\ -3 & 5 & 0 \\ 1 & -2 & 7 \end{bmatrix}$ are $\lambda = 7,\ 5 \pm 3\,i$. Therefore, the eigenvalues of

$$A = \left[\begin{array}{cc|ccc} 8 & -2 & 3 & 5 & 8 \\ 2 & 4 & 2 & -1 & 4 \\ \hline 0 & 0 & 5 & 3 & 0 \\ 0 & 0 & -3 & 5 & 0 \\ 0 & 0 & 1 & -2 & 7 \end{array}\right]$$

are $\lambda = 6, 6, 7,\ 5 \pm 3\,i$.

We have seen that in general the eigenvalues of a matrix A can be real or complex. However, if the matrix is symmetric, then its eigenvalues are real. Furthermore, its eigenvectors are orthogonal! The proof of the following theorem is left as an exercise.

Theorem 1.67 *Let $A_{n\times n}$ be a symmetric matrix. Then all its eigenvalues are real, and eigenvectors corresponding to different eigenvalues are orthogonal.*

■ **EXAMPLE 1.72**

Let $A = \begin{bmatrix} 3 & -1 & 4 \\ -1 & 0 & 1 \\ 4 & 1 & 3 \end{bmatrix}$. By using the characteristic equation, we find that the eigenvalues are $\lambda_1 = -2$, $\lambda_2 = 1$, and $\lambda_3 = 7$. Let us compute the eigenvectors. For $\lambda_1 = -2$, by letting $v = [x \quad y \quad z]^T$, equation (1.58) gives

$$5x - y + 4z = 0, \quad -x + 2y + z = 0, \quad 4x + y + 5z = 0.$$

From the second equation, $z = x - 2y$. By substituting this into the third equation, we get $x = y$. Then, any eigenvector associated with $\lambda_1 = -2$ is of the form $u = [x \quad x \quad -x]^T$.

For $\lambda_2 = 1$, equation (1.58) gives

$$2x - y + 4z = 0, \quad -x - y + z = 0, \quad 4x + y + 2z = 0.$$

From the second equation, $z = x + y$. Substituting this into the first equation, we get $y = -2x$. Then, any eigenvector associated with $\lambda_2 = 1$ is of the form $v = [x \quad -2x \quad -x]^T$.

For $\lambda_3 = 7$, equation (1.58) gives

$$-4x - y + 4z = 0, \quad -x - 7y + z = 0, \quad 4x + y - 4z = 0.$$

From the second equation, $z = x + 7y$. By substituting this into the third equation, we get $y = 0$. Then, any eigenvector associated with $\lambda_3 = 7$ is of the form $w = [x \quad 0 \quad x]^T$. Now, it is simple to verify that in fact

$$u^T v = 0, \quad u^T w = 0, \quad v^T w = 0,$$

which means that the eigenvectors are mutually orthogonal.

Note: In Example 1.72, a general fact is illustrated: since by definition, $\det(A - \lambda I) = 0$, in each set of three equations to determine the eigenvectors, each of the individual equations can be written in terms of the other two; that is why the solution is not unique (eigenvectors are not unique). In other words, in each of the sets of three equations, one is redundant and we can solve each system by using only two of its equations.

The next question is whether repeated eigenvalues of a general matrix can give linearly independent eigenvectors.

■ **EXAMPLE 1.73**

Let $A = \begin{bmatrix} 9 & 4 & 0 \\ -6 & -1 & 0 \\ 6 & 4 & 3 \end{bmatrix}$. By solving the corresponding characteristic equation, the eigenvalues are $\lambda_1 = 5$, $\lambda_2 = 3$, and $\lambda_3 = 3$. Thus, we have repeated eigenvalues. The eigenvector corresponding to λ_1 is $w_1 = [1 \quad -1 \quad 1]^T$. To find the eigenvector(s) corresponding to $\lambda_2 = \lambda_3$, equation (1.58) with $v = [x \quad y \quad z]^T$ gives

$$6x + 4y = 0, \quad -6x - 4y = 0, \quad 6x + 4y = 0.$$

The three equations are the same. From the first equation, we get $x = -\frac{2}{3}y$. The third entry z is free. Then, any eigenvector associated with $\lambda_2 = \lambda_3 = 3$ has the general form $[-\frac{2}{3}y \quad y \quad z]^T$. We have two free variables, and in particular, by taking $y = 0, z = 1$ and then $y = -3, z = 0$, respectively, we get $w_2 = [0 \quad 0 \quad 1]^T$, $w_3 = [2 \quad -3 \quad 0]^T$. Thus, even though there are repeated eigenvalues, we are still able to find a complete set of three linearly independent eigenvectors.

There is something very important to observe in Example 1.73. The eigenvectors w_1, w_2, and w_3 not only are different but also they are linearly independent. This is something that becomes very important in several applications, especially in the solution of differential equations. We will study this in more detail in Chapter 6.

In the examples above, we have seen that it was possible to find the same number of linearly independent eigenvectors and eigenvalues, even when the eigenvalues are repeated. However this is not always the case. There are cases when repeated eigenvalues give a smaller number of linearly independent eigenvectors, so that it is not possible to form a complete set of linearly independent eigenvectors. In this case, the repeated eigenvalue is

called **defective**, and we need to generalize the concept of eigenvector, so that we complete our set of (generalized) eigenvectors.

More formally, if an eigenvalue λ is repeated k times, we say λ has **algebraic multiplicity** k; e.g. in Example 1.73, $\lambda = 3$ has algebraic multiplicity two. The number of linearly independent eigenvectors associated with an eigenvalue λ is called the **geometric multiplicity** of λ; e.g., in Example 1.73, the geometric multiplicity of $\lambda = 3$ is two, so in this particular case, algebraic and geometric multiplicity coincide. In general, we have (see Exercise 3.60)

$$\text{geometric multiplicity} \leq \text{algebraic multiplicity}.$$

Thus, an eigenvalue is called defective if the above inequality is strict.

An eigenvalue with algebraic multiplicity 1 is called a simple eigenvalue.

For each eigenvalue λ of a matrix $A_{n\times n}$, we can define a special set: the set of all the eigenvectors associated with it. If we add to this set the zero vector, then it becomes a subspace of $\mathbb{R}^n$.

Definition 1.68 *Let A be a square matrix of order n, and λ an eigenvalue of A. The set*

$$E = \{v \in \mathbb{R}^n : Av = \lambda v\}$$

is called the **eigenspace** *of A with respect to λ.*

Several observations are in order.

1. We are considering all solutions of the equation $Av = \lambda v$. Thus, all eigenvectors of A associated with λ and also the zero vector are the elements of E.

2. The geometric multiplicity of λ is the dimension of its eigenspace E.

3. The linearly independent eigenvectors of A associated with λ form a basis of E.

4. The eigenspace of A can be defined as the nullspace of $A - \lambda I$.

When the geometric multiplicity of an eigenvector is strictly smaller than its algebraic multiplicity, that is, when the number of linearly independent eigenvectors is smaller than the number of times the corresponding eigenvalue is repeated, then the concept of an eigenvector needs to be generalized.

Definition 1.69 (Generalized Eigenvector) *If λ is an eigenvalue of a matrix A, then a rank-r generalized eigenvector associated with λ is a vector v such that*

$$(A - \lambda I)^r v = 0, \quad \text{and} \quad (A - \lambda I)^{r-1} v \neq 0. \tag{1.61}$$

Observe that a rank-1 generalized eigenvector is an ordinary eigenvector. That is,

$$(A - \lambda I)v = 0, \quad \text{and} \quad v \neq 0.$$

■ **EXAMPLE 1.74**

Let $A = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix}$. Then, according to Remark 1.63, the eigenvalues are $\lambda_{1,2} = 2$, $\lambda_3 = 3$. For $\lambda = 2$, equation (1.58) with $v = [x \quad y \quad z]^T$ gives $y = 0$, $z = 0$, and x is a free variable. This means that any eigenvector associated with $\lambda_{1,2} = 2$ is of the form $[x \quad 0 \quad 0]^T$. Thus, we can get only one (linearly independent) eigenvector. We need to find a generalized eigenvector associated with $\lambda = 2$. Following (1.61), the equation

$$(A - \lambda I)^2 v = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

gives $z = 0$. Thus, any rank-2 generalized eigenvector of A associated with $\lambda_1 = \lambda_2 = 2$ has the form $[x \quad y \quad 0]^T$, with two free variables. In particular, we can take $v_1 = [1 \quad 0 \quad 0]^T$ and $v_2 = [0 \quad 1 \quad 0]^T$. Observe that v_1 is the eigenvector that was determined before, and v_2 is a generalized eigenvector. Furthermore, we can verify that in this case

$$(A - \lambda I)v_2 = v_1 \neq 0,$$

thus satisfying the definition in (1.61). Finally, we can readily find out that any eigenvector associated with $\lambda_3 = 3$ is of the form $[0 \quad 0 \quad z]$, and in particular we can take $v_3 = [0 \quad 0 \quad 1]$.

We will see more about generalized eigenvectors in Chapter 6.

Left eigenvectors. It is possible to define eigenvectors when they are actually row vectors. Even though the eigenvalues of A would be the same, these new eigenvectors are in general different from the (right) eigenvectors introduced before. Suppose λ is an eigenvalue of A with associated eigenvector x, so that $Ax = \lambda x$. From Remark 1.64, we know that λ is also an eigenvalue of A^T, for some eigenvector v; that is,

$$A^T v = \lambda v.$$

By applying the transpose operation to both sides of this equation, we obtain the following definition.

Definition 1.70 *We say a nonzero vector $v \in \mathbb{R}^n$ is a* **left eigenvector** *of a matrix $A_{n\times n}$ associated with an eigenvalue λ if*

$$v^T A = \lambda v^T. \tag{1.62}$$

In other words, a left eigenvector of A associated with an eigenvalue λ is a (right) eigenvector of A^T corresponding to the same eigenvalue λ.

■ **EXAMPLE 1.75**

Let $A = \begin{bmatrix} 5 & 4 \\ 0 & 3 \end{bmatrix}$. For the eigenvalue $\lambda = 3$, any associated eigenvector x has the form $x = \begin{bmatrix} x_1 \\ -\frac{1}{2}x_1 \end{bmatrix}$; e.g., $x = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$. However, for the same eigenvalue $\lambda = 3$, any left eigenvector of A has the form $v = \begin{bmatrix} 0 \\ v_2 \end{bmatrix}$; e.g., $v = \begin{bmatrix} 0 \\ 5 \end{bmatrix}$ Thus, we have

$$\begin{bmatrix} 5 & 4 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = 3 \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \text{and}$$

$$[0 \quad 5] \begin{bmatrix} 5 & 4 \\ 0 & 3 \end{bmatrix} = 3\,[0 \quad 5].$$

1.11 SIMILARITY

We have seen that several sets of vectors can be chosen as a basis of a given vector space, and that arbitrary vectors in that space get different representations depending on the basis being used. In a similar way, given an arbitrary matrix A, we would like to find a simpler matrix B that is related or is equivalent to A in the sense that it shares some of its properties (e.g., same eigenvalues). Although ideally we would like B to be diagonal (the simplest possible), for most applications, it suffices if B is triangular, block diagonal, or block triangular.

We start this topic with the following definition.

Definition 1.71 *Two square matrices A and B are said to be* **similar** *if there exists a nonsingular matrix P such that*

$$B = P^{-1} A P. \tag{1.63}$$

Note: Equation (1.63) is sometimes written as $B = P A P^{-1}$.

The following theorem states one of the most important applications of similarity.

Theorem 1.72 *Similar matrices have the same eigenvalues, counting multiplicities.*

Proof. The idea is to show that characteristic polynomials of the similar matrices are equal, for if that is true, then they will have the same eigenvalues, counting multiplicities. Let A

and B be similar. Then,

$$\begin{aligned}
\det(B-\lambda I) &= \det(P^{-1}AP-\lambda I) = \det(P^{-1}AP-P^{-1}\lambda I P) \\
&= \det(P^{-1}(A-\lambda I)P) = \det(P^{-1})\det(A-\lambda I)\det P \\
&= \tfrac{1}{\det(P)}\det(A-\lambda I)\det(P) \\
&= \det(A-\lambda I).
\end{aligned}$$

□

■ **EXAMPLE 1.76**

The matrix $A = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}$ is similar to $B = \begin{bmatrix} 7 & 13 \\ -2 & -3 \end{bmatrix}$. In fact, you can verify that this is true by using the matrix $P = \begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}$. You can also verify that both A and B have the same eigenvalues $\lambda = 2 \pm \mathrm{i}$.

Ideally, we would like matrices to be similar to diagonal matrices, the simplest matrices possible. If this is true, the matrix has a special name.

Definition 1.73 *A matrix $A_{n\times n}$ is called* **diagonalizable** *if it is similar to a diagonal matrix.*

Given a matrix $A_{n\times n}$, how can we find the matrix P that performs the similarity transformation (1.63)? One answer to this is the eigenvectors. If A has n linearly independent eigenvectors, then the columns of the matrix P can be defined as the eigenvectors of A. Such a matrix P is nonsingular because its columns are linearly independent. The existence of such matrix P guarantees that A is diagonalizable. The following results come in handy.

Theorem 1.74 *Let A be a square matrix of order n. Then,*

$$A \text{ has distinct evalues} \implies A \text{ has } n \text{ l.i. evectors} \iff A \text{ is diagonalizable.}$$

Proof. To prove the first part of the theorem, assume by contradiction that A has n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ but that the corresponding set of eigenvectors $V = \{v_1, \ldots, v_n\}$ is not linearly independent. For some $m < n$, let $\{v_1, \ldots, v_m\} \subset V$ be the set with the maximum number of linearly independent eigenvectors of A. Then,

$$v_{m+1} = c_1 v_1 + \cdots + c_m v_m,$$

for some scalars $c_1, \ldots, c_m$. Using the general equation $(A-\lambda I)v = 0$ from (1.58), we have

$$\begin{aligned}
0 &= (A-\lambda_{m+1}I)v_{m+1} = (A-\lambda_{m+1}I)(c_1v_1+\cdots+c_mv_m) \\
&= (c_1\lambda_1 v_1+\cdots+c_m\lambda_m v_m) - (c_1\lambda_{m+1}v_1+\cdots+c_m\lambda_{m+1}v_m) \\
&= c_1(\lambda_1-\lambda_{m+1})v_1+\cdots+c_m(\lambda_m-\lambda_{m+1})v_m.
\end{aligned}$$

The linear independence of $v_1, \ldots, v_m$ implies that

$$c_1(\lambda_1 - \lambda_{m+1}) = \cdots = c_m(\lambda_m - \lambda_{m+1}) = 0,$$

and since the eigenvalues are distinct, we get $c_1 = \cdots = c_m = 0$. But then $v_{m+1} = 0$, which is not possible because v_{m+1} is an eigenvector.

The proof of the second part is left as an exercise.

□

■ **EXAMPLE 1.77**

Consider the matrix $A = \begin{bmatrix} 1 & 2 & 1 \\ 6 & -1 & 0 \\ -1 & -2 & -1 \end{bmatrix}$. The eigenvalues are $\lambda_1 = -4$, $\lambda_2 = 3$, $\lambda_3 = 0$. The eigenvectors are $v_1 = [1 \ -2 \ -1]^T$, $v_2 = [2 \ 3 \ -2]^T$, and $v_3 = [1 \ 6 \ -13]^T$, respectively, and they are linearly independent. Define

$$P = [v_1 \quad v_2 \quad v_3] = \begin{bmatrix} 1 & 2 & 1 \\ -2 & 3 & 6 \\ -1 & -2 & -13 \end{bmatrix}. \quad \text{Then,}$$

$$P^{-1}AP = \begin{bmatrix} -4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Thus, the matrix A is diagonalizable, and the diagonal matrix has the eigenvalues of A on its diagonal entries.

Invariant properties under similarity. We have seen that two similar matrices have the same eigenvalues. But not only eigenvalues are invariant under a similarity transformation. The following properties are also invariant (the proof of each is left as an exercise to the reader):

rank, determinant, trace, invertibility.

This means, e.g., that if A and B are similar matrices and A has rank k, then so does B and if A is invertible, so is B.

We finish this section by mentioning some more properties and applications of the eigenvalues of a matrix A.

1. The determinant of a matrix A can be defined as the product of its eigenvalues. That is, if $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then

$$\det A = \lambda_1 \cdot \lambda_2 \cdots \lambda_n. \tag{1.64}$$

2. From above, we see that A having one zero eigenvalue is necessary and sufficient to have a zero determinant. That is,

$$\det A = 0 \iff \lambda_i = 0, \quad \text{for some } i = 1, \ldots, n. \tag{1.65}$$

3. The matrix norm $\|\cdot\|_2$ given by (1.9) with $p = 2$, can also be defined through eigenvalues. More precisely, we compute the maximum eigenvalue of the symmetric matrix $A^T A$. That is,

$$\|A\|_2 = \sqrt{\max \lambda (A^T A)}. \tag{1.66}$$

1.12 BEZIER CURVES AND POSTSCRIPT FONTS

As we mentioned in Section 1.6, one powerful and useful concept in vector spaces is that of a basis. Given a vector space V of dimension n and a basis $\mathcal{B} = \{v_1, \ldots, v_n\}$ of V, then any element $x \in V$ can be uniquely expressed as

$$x = c_1 v_1 + \cdots + c_n v_n, \tag{1.67}$$

for some scalars $c_1, \ldots, c_n$. We say that x is written as a linear combination of the elements of the basis $\mathcal{B}$.

We want to see how this central concept in linear algebra can be used in a real-world application. Let V be the vector space of all real polynomials of degree at most n. This space has dimension $n + 1$, and one basis is given by the set

$$\{1, t, t^2, \ldots, t^n\}.$$

Several other bases can be formed for this vector space, but here we consider one that is very useful in several practical applications. In particular, it has proved to be essential in the construction of special curves, such as those needed in defining letter fonts, the ones we use everyday in text editing. The goal is to construct the so-called Bezier curves, which are currently used in computer font rendering technologies such as Adobe PDF and Postscript.

The special basis we are interested in can be formed using the so-called *Bernstein polynomials* of degree n, which can be defined recursively as

$$\begin{aligned} B_{0,0}(t) &= 1, \\ B_{i,n}(t) &= (1-t)B_{i,n-1}(t) + t\, B_{i-1,n-1}(t), \end{aligned}$$

where we take $B_{i,n}(t) = 0$, for $i < 0$, or $i > n$.

For example, the Bernstein polynomials of degrees 1, 2, and 3 are, respectively,

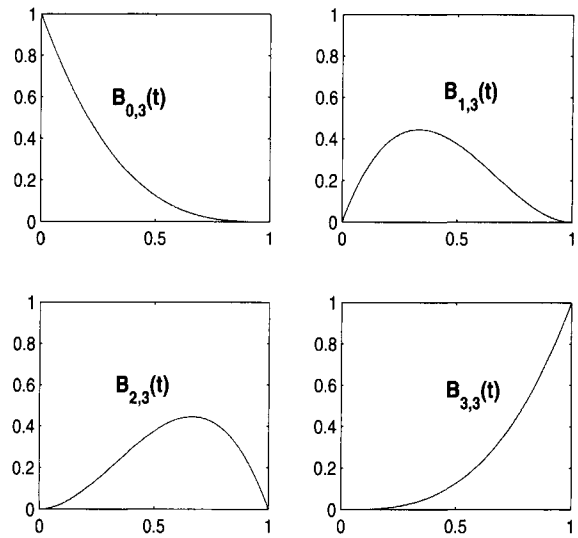


Figure 1.8 The cubic Bernstein polynomials.

$$
\begin{array}{lll}
B_{0,1}(t) = 1-t, & B_{1,1}(t) = t; & \\
B_{0,2}(t) = (1-t)^2, & B_{1,2}(t) = 2t(1-t), & B_{2,2}(t) = t^2; \\
B_{0,3}(t) = (1-t)^3, & B_{1,3}(t) = 3t(1-t)^2, & B_{2,3}(t) = 3t^2(1-t), \quad B_{3,3}(t) = t^3.
\end{array}
\tag{1.68}
$$

A much nicer way of obtaining the polynomials in (1.68) is by just expanding 1. For instance, for the Bernstein cubic polynomials:

$$1^3 = [(1-t)+t]^3 = (1-t)^3 + 3t(1-t)^2 + 3t^2(1-t) + t^3.$$

Observe that each term in this expansion is one of the cubic Bernstein polynomials $B_{i,3}$, $i = 0, \ldots, 3$. In general, we can write explicitly:

$$B_{i,n}(t) = \binom{n}{i} t^i (1-t)^{n-i}, \qquad i = 0, \ldots, n.$$

Here we are mostly interested in cubic Bernstein polynomials $B_{i,3}(t)$; these are shown in Figure 1.8. They form a basis of the space of polynomials of degree at most three. Again, the most important fact is that any polynomial of degree at most three can be expressed as a unique linear combination of those four cubic Bernstein polynomials.

As an example, let $P_3(t) = -5t^3 + 15t^2 - 9t + 2$. Then,

$$P_3(t) = 2B_{0,3}(t) - B_{1,3} + B_{2,3} + 3B_{3,3}.$$

Now let us see the usefulness of these special polynomials. Assume you are given four points

$$P_i = (x_i,\ y_i),\ i = 0, \ldots, 3,$$

henceforth called *control points*, and that you are looking for a polynomial of degree at most three that passes through the first and last point, and whose graph roughly follows the shape of the polygonal path determined by the four points.

This is a problem quite different from that of interpolation, where the curve must pass through all the given points (a general Bezier curve always passes through the first and the last point, but not necessarily through the other points, which are mostly used to control the shape of the curve). The problem here has another application in mind: to obtain an arbitrary plane curve and be able to modify its shape by redefining (moving) a few of the control points. This Bezier curve is to be defined using polynomials of degree at most three, and following (1.67), it will be expressed as a combination of basis functions

$$C(t) = \sum_{i=0}^{3} P_i B_{i,3}(t), \quad t \in [0,\ 1], \tag{1.69}$$

where $B_{i,3}$ are the cubic Bernstein polynomials, and P_i are the given control points.

For general linear combinations of the type (1.67), the main problem is to find the coefficients c_i, but in this present case, the coefficients P_i in (1.69) are the control points, which are always given.

An important fact is that we can express the Bezier curve (1.69) parametrically as

$$C(t) = (x(t),\ y(t)),$$

where

$$x(t) = \sum_{i=0}^{3} x_i B_{i,3}(t), \qquad y(t) = \sum_{i=0}^{3} y_i B_{i,3}(t), \qquad t \in [0,\ 1], \tag{1.70}$$

with $P_i = (x_i,\ y_i)$. Thus, a Bezier curve is a differentiable curve in the xy plane.

This is a very useful and illustrative application of the concept of basis of a vector space. Given four points P_i, $i = 0, 1, 2, 3$, we construct a smooth curve in parametric form (1.70), where the coefficients of the linear combination in (1.70); that is, x_i, and y_i, are the coordinates of the given points, and $B_{i,3}$ are the cubic Bernstein polynomials in (1.68), which as we know form a basis for the space of polynomials of degree ≤ 3.

1.12.1 Properties of Bezier curves

First and last point. An important property of the uniquely determined curve $C(t)$ is that it will pass through the first and last point. In fact, using (1.70), we have for the first coordinate of the curve at $t = 0$ and $t = 1$:

$$x(0) = x_0 B_{0,3}(0) + x_1 B_{1,3}(0) + x_2 B_{2,3}(0) + x_3 B_{3,3}(0)$$

$$= x_0 + 0 + 0 + 0 = x_0, \text{ and}$$

$$x(1) = x_0 B_{0,3}(1) + x_1 B_{1,3}(1) + x_2 B_{2,3}(1) + x_3 B_{3,3}(1)$$
$$= 0 + 0 + 0 + x_3 = x_3.$$

A similar result applies for the second coordinate $y(t)$ of $C(t)$ at $t = 0$ and $t = 1$.

Note: The fact that the curve passes through the first and the last points is true in general for Bernstein polynomials of arbitrary degree n, not only for cubic.

■ **EXAMPLE 1.78**

As an illustration, let us find the Bezier curve for the control points $P_0 = (4, 1), P_1 = (3, 2), P_2 = (5, 5)$, and $P_3 = (7, 3)$.

By substituting the first and second coordinates of the given control points into (1.70), we have

$$\begin{aligned} x(t) &= 4B_{0,3}(t) + 3B_{1,3}(t) + 5B_{2,3}(t) + 7B_{3,3}(t) \\ &= 4(1-t)^3 + 9t(1-t)^2 + 15t^2(1-t) + 7t^3 \\ &= 4 - 3t + 9t^2 - 3t^3. \end{aligned}$$

Similarly,

$$\begin{aligned} y(t) &= 1B_{0,3}(t) + 2B_{1,3}(t) + 5B_{2,3}(t) + 3B_{3,3}(t) \\ &= 1(1-t)^3 + 6t(1-t)^2 + 15t^2(1-t) + 3t^3 + \\ &= 1 + 3t + 6t^2 - 7t^3. \end{aligned}$$

Thus,

$$C(t) = (4 - 3t + 9t^2 - 3t^3, \quad 1 + 3t + 6t^2 - 7t^3), \quad t \in [0,\ 1].$$

The curve and the control points are shown in Figure 1.9. Observe how the curve somehow follows the polygonal path determined by the control points.

Tangency property. The main reason why a Bezier curve roughly follows the shape determined by the control points and the segments joining them is that the slope of the curve at P_0 is the same as that of the segment joining P_0 and P_1. Similarly, the slope of the curve at P_3 coincides with the slope of the segment joining P_2 and P_3. This is apparent in Figure 1.9.

In fact, with $P_0 = (x_0, y_0)$, and $P_1 = (x_1, y_1)$, the slope of the segment joining these two points is

$$m = \frac{y_1 - y_0}{x_1 - x_0}.$$

Now let us find the slope of the Bezier curve at P_0. In parametric form, the Bezier curve is

$$\begin{aligned} x(t) &= (1-t)^3 x_0 + 3t(1-t)^2 x_1 + 3t^2(1-t)x_2 + t^3 x_3, \\ y(t) &= (1-t)^3 y_0 + 3t(1-t)^2 y_1 + 3t^2(1-t)y_2 + t^3 y_3. \end{aligned} \tag{1.71}$$

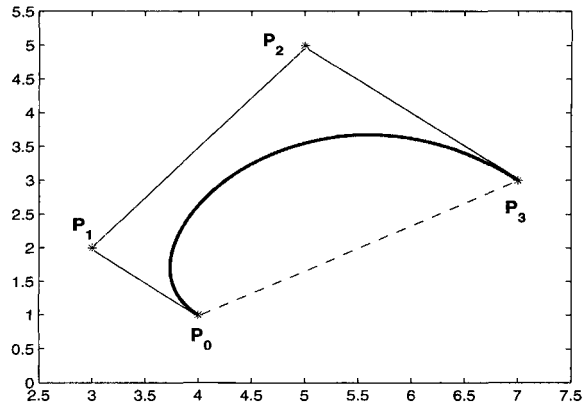


Figure 1.9 Bezier curve of Example 1.78.

The corresponding derivatives are

$$\frac{dx}{dt} = -3(1-t)^2x_0 + 3(1-t)^2x_1 - 6t(1-t)x_1 + 6t(1-t)x_2 - 3t^2x_2 + 3t^2x_3,$$

$$\frac{dy}{dt} = -3(1-t)^2y_0 + 3(1-t)^2y_1 - 6t(1-t)y_1 + 6t(1-t)y_2 - 3t^2y_2 + 3t^2y_3$$

Then, the slope of the curve at P_0 is

$$\left.\frac{dy}{dx}\right|_{t=0} = \left.\frac{dy/dt}{dx/dt}\right|_{t=0} = \frac{-3y_0 + 3y_1}{-3x_0 + 3x_1} = \frac{y_1 - y_0}{x_1 - x_0} = m.$$

Similarly, it can be proved that the slope of the Bezier curve at P_3 coincides with the slope of the segment joining P_2 and P_3.

Bezier curve and convex hull. The effectiveness and usefulness of Bezier curves lies in the great ease with which the shape of the curve can be modified, say by means of a mouse, by making adjustments to the control points. One good advantage is that the changes in the shape of the curve will be somewhat localized. The only term(s) in (1.71) that are modified are the ones involving the point(s) being moved.

■ **EXAMPLE 1.79**

Suppose that we change (or move) the control points P_1 and P_2 of Example 1.78 to $P_1 = (3.7, 3)$ and $P_2 = (6.5, 4.5)$. Then, the new Bezier curve will be

$$C(t) = (4 - 3t + 9t^2 - 3t^3, \quad 1 + 3t + 6t^2 - 7t^3), \quad t \in [0, 1].$$

In Figure 1.10, we show how by moving these two control points, the shape of the curve has been accordingly modified, the new curve being pulled toward the new control points.

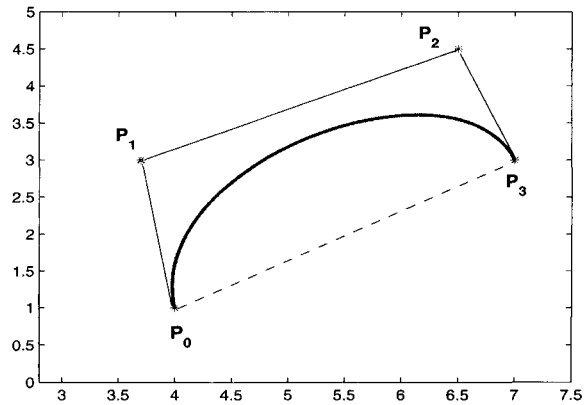


Figure 1.10 Modified Bezier curve.

The fact that the Bezier curve follows the shape of the polygonal path determined by the control points and stays within the quadrilateral is more than just a geometrical fact or coincidence. There is a concrete linear algebra concept behind this.

Definition 1.75 *Given a set of n points: $S = \{P_0, \ldots, P_{n-1}\}$, the* **convex hull** *of S is the smallest convex polygon that contains all the points in S.*

This definition is saying that any given point in the convex hull of S is either on the boundary of the polygon or in its interior. Intuitively, we can think of each point in S as a nail on a board; then, the convex hull is the shape formed by a tight rubber band that surrounds all the nails. Two clear examples are given in Figures 1.9 and 1.10, where the convex hull of the control points is the quadrilateral given by the lines and its interior. The relationship between convex hulls and Bezier curves, which is apparent in both Figures 1.9 and 1.10, is that

A Bezier curve always lies in the convex hull of its set of control points.

Midpoint property. We already know that $P_0 = (x(0),\ y(0))$ and that $P_3 = (x(1),\ y(1))$. All other points of the Bezier curve are $(x(t),\ y(t))$, for some $t \in (0,\ 1)$. One interesting question is whether we can characterize the point on the curve for which $t = \frac{1}{2}$.

If we substitute $t = \frac{1}{2}$ in the parametric equations (1.71), we obtain

$$\begin{aligned} x(\tfrac{1}{2}) &= \tfrac{1}{8}x_0 + \tfrac{3}{8}x_1 + \tfrac{3}{8}x_2 + \tfrac{1}{8}x_3, \\ y(\tfrac{1}{2}) &= \tfrac{1}{8}y_0 + \tfrac{3}{8}y_1 + \tfrac{3}{8}y_2 + \tfrac{1}{8}y_3. \end{aligned}$$

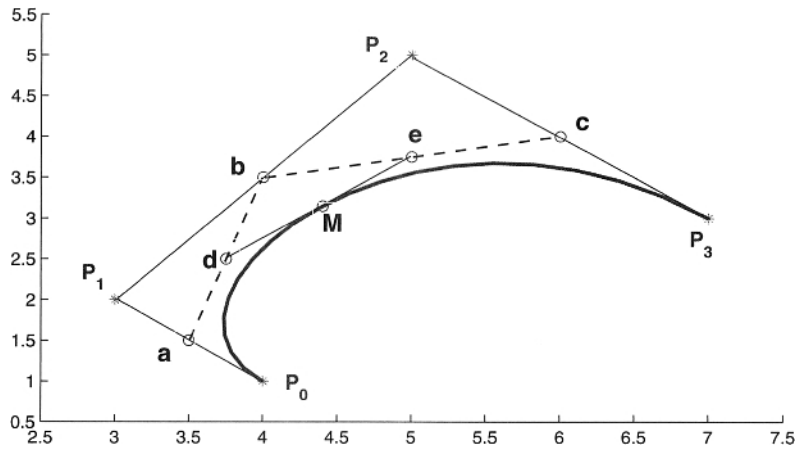


Figure 1.11 Midpoint property in Bezier curves.

These are the coordinates of the point M on the curve. However, the interesting thing is that this same point can be obtained in a different way, that geometrically is more intuitive (see Figure 1.11):

The point a is the midpoint between P_0 and P_1. The point b is the midpoint between P_1 and P_2, and the point c is the midpoint between P_2 and P_3. Similarly, the point d is the midpoint between a and b, and the point e is the midpoint between b and c. Finally, the sought point $M = (x(\frac{1}{2}),\ y(\frac{1}{2}))$ is the midpoint between d and e. We leave the verification of the statements above as an exercise.

Furthermore, it can be shown that the part of the Bezier curve that goes from P_0 to M can be defined using the control points, P_0, a, d, and M. Similarly, the part of the Bezier curve that goes from M to P_3 can be defined using the control points M, e, c, and P_3.

Bezier curve and center of mass. An interesting way of introducing Bezier curves is through the concept of center of mass of a set of point masses. Suppose we have four masses m_0, m_1, m_2 and m_3 which are located at points P_0, P_1, P_2, and P_3 respectively. Then, the center of mass of these four point masses is

$$P = \frac{m_0 P_0 + m_1 P_1 + m_2 P_2 + m_3 P_3}{m_0 + m_1 + m_2 + m_3}.$$

Assume also that the masses are not constant but that they vary as a function of a parameter t according to the equations

$$m_0 = (1-t)^3, \quad m_1 = 3t(1-t)^2, \quad m_2 = 3t^2(1-t), \quad m_3 = t^3,$$

for $t \in [0,\ 1]$.

Since for any value of t we always have that $m_0(t) + m_1(t) + m_2(t) + m_3(t) = 1$, the center of mass reduces to

$$P = m_0P_0 + m_1P_1 + m_2P_2 + m_3P_3.$$

Observe that for $t = 0$, the center of mass P is at P_0 and for $t = 1$ it is located at P_3. As t varies between 0 and 1, the center of mass P moves describing a curve, a cubic Bezier curve. The masses described above are the cubic Bernstein polynomials.

1.12.2 Composite Bezier curves

In practical applications, curves will be more sophisticated than a single Bezier curve, but these sophisticated curves can be produced by using a sequence of Bezier curves that share common end points and then are patched together ensuring the continuity of the final curve (but not necessarily differentiability). The final curve obtained is a composite Bezier curve, also known as **Bezier spline**.

Since a general Bezier curve always lies in the convex hull of its control points, oscillatory behavior will not be present. Also, changes in the curve will mean making just local changes of some control points, minimizing in this way the number of total modifications.

One more observation is in order. When performing interpolation, say with cubic splines, the resulting curve is smooth, with continuous first and second derivatives, so that sharp corners are out of the question. Composite Bezier curves are more flexible: A sharp corner will be well defined, because only continuity is required (in some applications, sharp cornes will be needed, such as in creating postscript fonts), and if we need to have smoothness at the points where two Bezier curves meet, it is sufficient to require (see Exercise 1.95) that the three control points (the point where they meet, the one before, and the one after) are collinear. We illustrate these ideas with the following example.

■ **EXAMPLE 1.80**

Let us find the composite Bezier curve for the following sets of control points:

$$\begin{aligned}
&\{(-5.0, 0.5), (-12, 1.5), (-12, 4.5), (-5.0, 5.0)\},\\
&\{(-5.0, 5.0), (-6.0, 3.0), (-2.0, 2.5), (-1.5, 5.0)\},\\
&\{(-1.5, 5.0), (-1.5, 5.0), (-1.0, 4.4), (-1.0, 4.4)\},\\
&\{(-1.0, 4.4), (-0.5, 4.6), (0.5, 4.6), (1.0, 4.4)\},\\
&\{(1.0, 4.4), (1.0, 4.4), (1.5, 5.0), (1.5, 5.0)\},\\
&\{(1.5, 5.0), (2.0, 2.5), (6.0, 3.0), (5.0, 5.0)\},\\
&\{(5.0, 5.0), (12, 4.5), (12, 1.5), (5.0, 5.0)\}.
\end{aligned}$$

Following the same process as in Example 1.78 to find one Bezier curve $C(t)$, we start by finding the following seven Bezier curves corresponding to the seven sets of

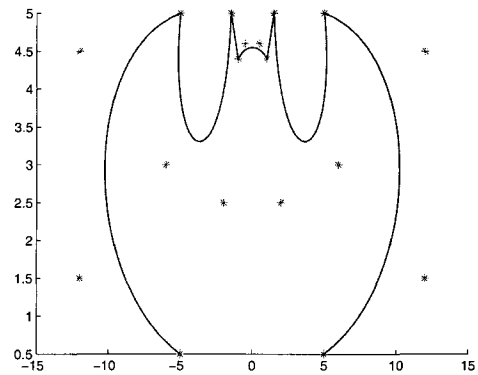


Figure 1.12 Composite Bezier curve of Example 1.80.

points above. This gives:

$$
\begin{aligned}
C_1(t) &= (21t^2 - 21t - 5, \quad -4.5t^3 + 6t^2 + 3t + 0.5), \\
C_2(t) &= (-8.5t^3 + 15t^2 - 3t - 5, \quad 1.5t^3 + 4.5t^2 - 6t + 5), \\
C_3(t) &= (-t^3 + 1.5t^2 - 1.5, \quad 1.2t^3 - 1.8t^2 + 5), \\
C_4(t) &= (0.6t^3 + 0.3t^2 + 2.7t - 1.4, \quad -0.6t^2 + 0.6t + 4.4), \\
C_5(t) &= (-t^3 + 1.5t^2 + 1, \quad -1.2t^3 + 1.8t^2 + 4.4), \\
C_6(t) &= (-8.5t^3 + 10.5t^2 + 1.5t + 1.5, \quad -1.5t^3 + 9t^2 - 7.5t + 5), \\
C_7(t) &= (-21t^2 + 21t + 5, \quad 9t^3 - 7.5t^2 - 1.5t + 5).
\end{aligned}
$$

The resulting composite Bezier curve is shown in Figure 1.12 along with the control points. Observe that we can have sharp corners wherever needed. Also observe that if a particular Bezier curve is to be a line segment, then the corresponding four points are listed by repeating them. For instance, for C_3, we have the control points

$$(-1.5, 5), (-1.5, 5), (-1, 4.4), (-1, 4.4).$$

One very important application of composite Bezier curves is the design of fonts. The polynomials used to create fonts could be quadratic or cubic (even linear), depending on the application. True Type typically uses quadratic composite Bezier curves, while Postscript uses cubic ones. In the next example, we want to illustrate this real-world application of Bezier curves (and hence of Bernstein polynomials and basis of a vector space) in creating a postscript font.

■ **EXAMPLE 1.81**

We can use the following set of points (listed without parenthesis for convenience) to generate the Times Roman character **R** as the composition of 22 Bezier curves. This

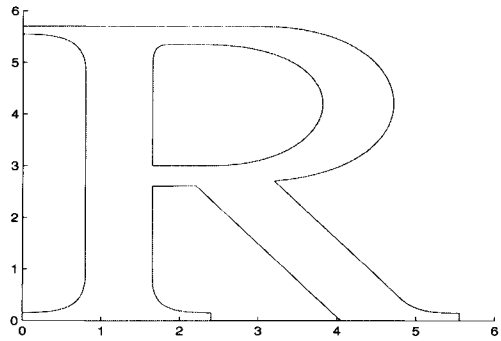


Figure 1.13 Times Roman character as a composite Bezier curve.

is illustrated in Figure 1.13.

0.00	5.70	0.00	5.70	0.00	5.55	0.00	5.55
0.00	5.55	0.60	5.55	0.80	5.35	0.80	4.80
0.80	4.80	0.80	4.80	0.80	0.90	0.80	0.90
0.80	0.90	0.80	0.30	0.60	0.15	0.00	0.15
0.00	0.15	0.00	0.15	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	2.40	0.00	2.40	0.00
2.40	0.00	2.40	0.00	2.40	0.15	2.40	0.15
2.40	0.15	1.80	0.15	1.65	0.30	1.65	0.90
1.65	0.90	1.65	0.90	1.65	2.60	1.65	2.60
1.65	2.60	1.65	2.60	2.65	2.62	2.65	2.62
2.65	2.62	2.65	2.62	4.10	0.00	4.10	0.00
4.10	0.00	4.10	0.00	5.50	0.00	5.50	0.00
5.50	0.00	5.50	0.00	5.50	0.15	5.50	0.15
5.50	0.15	5.15	0.15	4.95	0.20	4.76	0.50
4.76	0.50	4.80	0.50	3.50	2.74	3.50	2.74
3.50	2.74	5.38	3.00	5.10	5.70	3.05	5.70
3.05	5.70	3.05	5.70	0.00	5.70	0.00	5.70
1.65	3.00	1.65	3.00	1.65	5.00	1.65	5.00
1.65	5.00	1.65	5.30	1.75	5.35	1.84	5.35
1.84	5.35	1.84	5.35	2.30	5.35	2.30	5.35
2.30	5.35	4.40	5.20	4.20	3.00	2.40	3.00
2.40	3.00	2.40	3.00	1.65	3.00	1.65	3.00

1.13 FINAL REMARKS AND FURTHER READING

In this chapter, we have introduced those concepts and techniques of linear algebra and matrix analysis that are needed for some of the applications covered in this book. In particular, in Section 1.12, concepts like linear combination, basis of a vector space, convex

hull, and so on. have been illustrated as very useful tools for the construction of Bezier curves and postscript fonts. More applications are presented in the following chapters.

Linear algebra and matrix analysis offer several other very interesting concepts and techniques not covered here, but that are also very important in applications. An extensive study of matrix norms and their properties, matrix computations, and matrix analysis in general can be found in the classic books by Horn and Johnson [38] and Golub and Van Loan [25], the latter especially focused on the numerical and computational aspects of matrix algebra. An additional excellent reference for matrix computations, linear algebra, and their applications is the book by Meyer [49].

EXERCISES

1.1 Given two matrices A and B, show that

$$(A+B)^T = A^T + B^T, \quad \text{and} \quad (AB)^T = B^T A^T,$$

whenever the sum or the product, respectively, are well defined.

1.2 Show that for two matrices A and B whose product is well defined,

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

1.3 Find a nonzero vector $x = [x_1 \quad x_2]^T$ such that the sum, Euclidean, and maximum norm all coincide.

1.4 Show that for any n-dimensional vector x,

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1.$$

1.5 Prove the Cauchy–Schwarz inequality:

$$|x^T y| \leq \|x\|_2 \, \|y\|_2,$$

for any two vectors $x, y \in \mathbb{R}^n$.

1.6 Show that the Frobenius norm of a matrix is consistent, in the sense that

$$\|AB\|_F \leq \|A\|_F \|B\|_F,$$

for any two matrices A, B whose product is well–defined.

1.7 Let $A_{m\times n}$ be a matrix such that $\|Ax\|_2 \leq \|x\|_2$, for all vectors $x \in \mathbb{R}^n$. Show that also $\|A^T y\|_2 \leq \|y\|_2$, for all vectors $y \in \mathbb{R}^m$.

1.8 Let A be an $m \times n$ matrix, and $x \in \mathbb{R}^n$. Show that

$$\|Ax\|_2 = \|x\|_2 \qquad \text{if and only if} \qquad A^T A = I.$$

1.9 Show that for any matrix norm induced from a vector norm as in (1.9), we have

$$\|AB\| \leq \|A\| \, \|B\|, \tag{1.72}$$

for any two matrices A and B for which the product is well–defined.

1.10 Let $A = \begin{bmatrix} 3 & -2 & 5 \\ 0 & 8 & -1 \\ 2 & 2 & 7 \end{bmatrix}$. Find $\|A\|_F$ using the definitions given in (1.6), (1.7), and (1.8).

1.11 We may want to extend the vector norm $\|x\| = \max_i |x_i|$ into a matrix norm by defining

$$\|A\| = \max_{1 \leq i,j \leq n} |a_{ij}|.$$

However, this function cannot be considered a matrix norm because it does not satisfy the submultiplicative inequality (1.72). Provide a counterexample to confirm such statement.

1.12 Show that for $p = 1$ and $p = \infty$, the p-norms in (1.9) coincide with the corresponding norms in (1.6).

1.13 Show that the Frobenius norm in (1.6) can also be defined in the following two ways:

$$\text{(a)} \quad \|A\|_F^2 = \|a_1\|_2^2 + \cdots + \|a_n\|_2^2.$$

$$\text{(b)} \quad \|A\|_F^2 = \operatorname{tr}(A^T A).$$

Here, $a_1, \ldots, a_n$ denote the columns of $A_{m \times n}$.

1.14 Let A be an $m \times n$ matrix. Show that

$$\|A\|_2 \leq \sqrt{\|A\|_\infty \|A\|_1}.$$

1.15 Let $A_{n \times n}$ be an arbitrary matrix. We define its spectral radius as $\rho(A) = \max_{i=1,\ldots,n} |\lambda_i|$, where λ_i are the eigenvalues of A. Show that

$$\rho(A) \leq \|A\|,$$

for any matrix norm $\|\cdot\|$.

1.16 Show that the product of two square lower triangular matrices is lower triangular.

1.17 True or False? A symmetric Hessenberg matrix is tridiagonal, that is, a matrix whose nonzero elements are on the diagonal, the diagonal below and the diagonal above the main diagonal.

1.18 Show that the inverse of a nonsingular matrix $A_{n \times n}$ is always unique.

1.19 Let A be a nonsingular square matrix of order n. Show that if $AX = I$, for some matrix $X_{n \times n}$, then also $XA = I$.

1.20 Show that if $A_{n \times n}$ is nonsingular, then A^T is also nonsingular and

$$(A^T)^{-1} = (A^{-1})^T.$$

(We usually write this matrix as A^{-T}).

1.21 Let the matrices $A_1, A_2, \ldots, A_k$ be nonsingular. Show that

$$(A_1 A_2 \cdots A_k)^{-1} = A_k^{-1} A_{k-1}^{-1} \cdots A_1^{-1}.$$

1.22 Show that a matrix $A_{n\times n}$ is nonsingular if there is a matrix norm such that $\|I - A\| < 1$. Then show that in such a case,

$$A^{-1} = \sum_{k=0}^{\infty} (I - A)^k.$$

1.23 Let A be a square matrix of order n such that $\|I - A\| < 1$, for some matrix norm $\|\cdot\|$. Show that

$$\|A^{-1}\| \leq \frac{1}{1 - \|I - A\|}.$$

1.24 Let $A_{n\times n}$ be an arbitrary matrix that can be expressed as $A = M - N$, for some matrices M and N, with M nonsingular, and let $B = M^{-1}N$. Show that if $\|B)\| < 1$, for some matrix norm, then A is nonsingular.

1.25 Let $A = \begin{bmatrix} 1 & a & a \\ 0 & 1 & a \\ 0 & 0 & 1 \end{bmatrix}$. Show that its inverse is given by the matrix

$$A^{-1} = I + B + B^2,$$

where $B = I - A$. Find A^{-1} and B^3 explicitly.

1.26 Let A be a square matrix of order n that is strictly diagonal dominant; that is,

$$|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}|, \qquad i = 1, 2, \ldots, n. \tag{1.73}$$

Show that A is nonsingular.

Note: If the inequality in (1.73) is not strict, we say A is weakly diagonal dominant.

1.27 Give an example of a matrix that is symmetric but not hermitian and an example of matrix that is hermitian but not symmetric.

1.28 True or False? The diagonal entries of a Hermitian matrix must be real.

1.29 True or False? The sum of two symmetric matrices is a symmetric matrix.

1.30 Verify that the matrices in Example 1.13 are positive definite, and that the matrices in Example 1.11 are not.

1.31 Let $A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$ be symmetric positive definite (spd), with B and E square. Show that B and E are also spd.

1.32 Let A be a square matrix of order n. Show that

$$\frac{d}{dt} e^{At} = Ae^{At}.$$

1.33 Let $A_{n\times n}$ be nonsingular and assume that $e^{-At} \to 0$ as $t \to \infty$. Show that the inverse of A is given by the integral

$$\int_0^\infty e^{-At}dt.$$

1.34 Show that if a matrix $A_{m\times n}$ has rank n, then there exists a matrix X such that $XA = I$. Is the matrix X unique? (X is called a left inverse of A).

1.35 Let v and q be two n-dimensional vectors, with $\|q\| = 1$. We define the **projection** of v along q as

$$\mathrm{Proj}_q v = (v^T q)q,$$

and therefore, the **orthogonal projection** of v onto q is $v-(v^T q)q$, just as in Gram–Schimdt process. Let $v = [-1 \quad 0 \quad 3]^T$ and $u = [1 \quad 0 \quad 2]^T$. Find the orthogonal projection of v along $q = u/\|u\|$.

1.36 The vectors $v_1 = [-1 \quad 0 \quad 1]^T$, $v_2 = [0 \quad -1 \quad 1]^T$, and $v_3 = [1 \quad 1 \quad 0]^T$ are linearly independent. Apply the Gram-Schmidt process to these vectors, but using the following dot produt:

$$x^T y := \frac{1}{3}x_1y_1 + 2x_2y_2 + x_3y_3.$$

1.37 The same row and column permutation of the matrix A in Example 1.20 could have also been obtained by simply computing the product PAP. Is this true for any permutation matrix P?

1.38 Show that the columns of an orthogonal matrix $Q_{m\times n}$ are orthonormal vectors

1.39 Let $P_{n\times n}$ and $Q_{n\times n}$ be two orthogonal matrices. Show that the matrix PQ is also orthogonal. Show however that the sum $P + Q$ may fail to be orthogonal.

1.40 Let x, y be two arbitrary vectors in $\mathbb{R}^n$ and $Q_{n\times n}$ an orthogonal matrix. Let θ_1 be the angle between x and y, and θ_2 the angle between Qx and Qy. Show that

$$\cos\theta_1 = \cos\theta_2.$$

1.41 Let $Q_{m\times n} = [\,q_1 \quad q_2 \quad \cdots \quad q_n]$ be an orthogonal matrix. What do you obtain when Gram–Schmidt is applied to the columns of Q?

1.42 A matrix norm is called **orthogonally invariant** if for an arbitrary matrix $A_{m\times n}$ and orthogonal matrices Q and U of appropriate dimensions, we have

$$\|Q\,A\,U^T\| = \|A\|.$$

Show that the Frobenius and the 2-norm are orthogonally invariant.

1.43 Show that the orthogonal invariance of a norm implies that multiplication by an orthogonal matrix does not magnify errors, in the following sense: Let $A_{n\times n}$ be an arbitrary matrix and $Q_{n\times n}$ be orthogonal. If an error E is introduced in A, then the error in QAQ^T has the same norm as E. Also show that for a general nonsingular matrix P, the error in $PA\,P^{-1}$ is bounded by $\mathrm{cond}(P)\|E\|$.

1.44 Let $A_{n\times n}$ be a nonsingular matrix and let b be an arbitrary vector in $\mathbb{R}^n$. Let x be the solution of $Ax = b$, and suppose there is a perturbation δb on the vector b so that the solution of the system is perturbed to $x + \delta x$, that is, $A(x + \delta x) = b + \delta b$. Show that

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A)\frac{\|\delta b\|}{\|b\|}.$$

1.45 Let $Q_{n\times n}$ be an orthogonal matrix. Prove that $\text{cond}(Q) = 1$.

1.46 Let Q be an orthogonal matrix and A be a nonsingular matrix. Show that

$$\text{cond}\,(QA) = \text{cond}\,(A).$$

1.47 Show that no two vectors in $\mathbb{R}^3$ can span all of $\mathbb{R}^3$.

1.48 Find the subspace spanned by the three vectors $[2 \quad 3 \quad 1]^T$, $[2 \quad 1 \quad -5]^T$, and $[2 \quad 4 \quad 4]^T$.

1.49 Show that the set U in Example 1.30 is not a vector subspace of V.

1.50 Let V be a vector space and $S = \{v_1, \ldots, v_k\}$ be an arbitrary subset of V. Show that the set of all linear combinations of vectors from S is a subspace of V.

1.51 Let $\mathcal{B} = \{v_1, \ldots, v_n\}$ be a basis of vector space V. Show that any subset of V containing more than n vectors is linearly dependent (not linearly independent).

1.52 Let u and v be arbitrary linearly independent vectors in R^n. Show that for some values of the scalars c_1 and c_2, the vector $x = c_1u + c_2v$ has both positive and negative components.

1.53 Show that the set $\{\, f_1 = 2x^2 + 1,\ f_2 = x^2 + 4x,\ f_3 = x^2 - 4x + 1 \,\}$ is linearly dependent.

1.54 Show that if the columns of a matrix $A_{m\times n}$ are linearly independent, then the matrix A^TA has an inverse.

1.55 Let S be a subspace of a vector space V. Show that

$$(S^\perp)^\perp = S.$$

1.56 Let $\mathcal{B}_1 = \{u_1, \ldots, u_n\}$ and $\mathcal{B}_2 = \{w_1, \ldots, w_n\}$ be two orthonormal bases of $\mathbb{R}^n$, and let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be the coordinates of a vector $x \in \mathbb{R}^n$ in those bases, respectively. Show that

$$a_1^2 + \cdots + a_n^2 = b_1^2 + \cdots + b_n^2.$$

1.57 $A = \begin{bmatrix} 4 & 0 & 1 & -1 & 1 \\ 0 & -5 & 0 & 5 & 0 \\ 0 & 0 & 3 & 4 & 1 \end{bmatrix}$. Find $\dim \text{col}(A)$ and $\dim N(A)$.

1.58 Let A be an $m \times n$ matrix. Show that

$$N(A^TA) = N(A) \qquad \text{and} \qquad \text{rank}(A^TA) = \text{rank}(A).$$

1.59 Let V be the vector space of square real matrices of order 2, and let

$$U = \left\{ \begin{bmatrix} 0 & a \\ 0 & b \end{bmatrix} : a, b \in \mathbb{R} \right\}, \qquad W = \left\{ \begin{bmatrix} 0 & 0 \\ c & d \end{bmatrix} : c, d \in \mathbb{R} \right\}.$$

Find $\dim(U + W)$ and verify formula (1.38).

1.60 Consider the block matrix $A = [A_1 \quad A_2]$, for some matrices A_1 and A_2. Show that

$$\mathrm{col}(A) = \mathrm{col}(A_1) + \mathrm{col}(A_2).$$

1.61 Let P_1 and P_2 be orthogonal projection matrices onto two subspaces V_1 and V_2 of a vector space V. Show that

$$\mathrm{col}(P_1 + P_2) = \mathrm{col}(P_1) + \mathrm{col}(P_2).$$

1.62 Let the columns of a matrix $Q_{m\times n}$ form an orthonormal basis of a vector subspace S. Show that

$$P = QQ^T \tag{1.74}$$

is the orthogonal projection matrix onto S.

1.63 Let A be a square matrix of order n, and λ be an eigenvalue of A. Show that the set of all eigenvectors associated with λ, together with the zero vector, form a subspace of $\mathbb{R}^n$.

1.64 Show that the determinant of a matrix $A_{n\times n}$ is given by the product of its eigenvalues, that is,

$$\det A = \lambda_1 \cdot \lambda_2 \cdots \lambda_n,$$

where some of the eigenvalues may be repeated.

1.65 Consider the matrix

$$A = \begin{bmatrix} 2 & 1 & 0 & 3 \\ 0 & 3 & 3 & 0 \\ 4 & 1 & 0 & 1 \\ 1 & 0 & 0 & 5 \end{bmatrix}.$$

Without any computation of the characteristic equation, determine one of the eigenvalues of A as well as an associated eigenvector.

1.66 Suppose that the sum of the entries in each column of a square matrix A is 1. Show that 1 is an eigenvalue of A^T.

1.67 True or False? The eigenvalues of a positive definite matrix are positive numbers.

1.68 Let u and v be right and left eigenvectors, respectively, of a matrix $A_{n\times n}$ for a given eigenvalue λ of A, with $v^T u = 1$. Let $Q_{n\times(n-1)}$ be an orthogonal matrix whose columns form a basis of $v^\perp$. Show that the matrix $P = [u \quad Q]$ satisfies

$$P^{-1}AP = \begin{bmatrix} \lambda & 0 \\ 0 & B \end{bmatrix},$$

for some square matrix B of order $n - 1$.

1.69 Show by counterexample that the product of two symmetric matrices A and B is not necessarily symmetric. Show however that if A and B commute, then the product AB is also symmetric.

1.70 True or False? The inverse of a nonsingular symmetric matrix is also symmetric.

1.71 Let $A_{n\times n}$ be a symmetric matrix. Show that all its eigenvalues are real, and that eigenvectors corresponding to distinct eigenvalues are orthogonal.

1.72 Let $U = V \oplus W$, and let $\mathcal{B}_1, \mathcal{B}_2$ be bases of V and W, respectively. Prove that $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$ and that $\mathcal{B}_1 \bigcup \mathcal{B}_2$ is a basis for U.

1.73 For a matrix $A_{n\times n}$, define $r = Av - \lambda v \neq 0$, where v is a unit vector (we say v is an approximate eigenvector of A, with associated eigenvalue λ). Show that

$$(A + \delta A)v = \lambda v,$$

where $\delta A = -rv^T$.

1.74 Let $A_{n\times n}$ have n distinct eigenvalues with associated eigenvectors $v_1, \ldots, v_n$ and left eigenvectors $w_1, \ldots, w_n$. Show that $w_i^T v_j = 0$ if $i \neq j$, and $w_i^T v_j \neq 0$ if $i = j$.

1.75 **(Gershgorin)** Let A be a square matrix of order n, and for some nonsingular matrix P, let $P^{-1}AP = D + E$, with $D = \text{diag}(d_1, \ldots, d_n)$ and E has zero diagonal entries. Show that

$$\sigma(A) \subseteq \bigcup_{i=1}^{n} D_i,$$

where $D_i = \{\, z \in C : |z - d_i| \leq \sum_{j=1}^{n} |e_{ij}| \,\}$.

Hint: Show $(D - \lambda I) + E$ is singular, and use Exercise 1.23 with $p = \infty$.

1.76 Let D =diag$(d_1, \ldots, d_n)$ with all d_i distinct. Consider a scalar $c \neq 0$ and assume that all entries of a vector $u \in \mathbb{R}^n$ are nonzero. Define the matrix

$$A = D + cuu^T.$$

Show that D and A do not have any common eigenvalues.

1.77 Let S and K represent the subspaces of symmetric and skew-symmetric $n \times n$ matrices, respectively. Show that

$$R^{n\times n} = S \oplus K.$$

Hint: For any square matrix A of order n, the matrix $A + A^T$ is symmetric and $A - A^T$ is skew-symmetric.

1.78 Let $A_{n\times n}$ be a symmetric matrix. Show that the matrix iA is skew-hermitian.

1.79 Let $A_{n\times n}$ be a skew symmetric matrix. Show that A is singular when n is odd.

1.80 True or False? If A is a skew symmetric matrix, then e^A is orthogonal.

1.81 Let $P_{n\times n}$ be a projection matrix. Show that

$$\text{col}(P) \oplus N(P) = \mathbb{R}^n.$$

1.82 Let P be a projection matrix on a vector space V. Show that for $x \in V$,

$$x \in \mathrm{col}(P) \quad \text{if and only if} \quad x = Px.$$

1.83 Let x, y be two vectors in $\mathbb{R}^n$ such that $y^T x = 1$. Show that $P = xy^T$ is a projection matrix and that

$$\|P\|_2 = \|x\|_2 \, \|y\|_2.$$

1.84 Show that if P is a projection matrix, then each eigenvalue of P is either 0 or 1.

1.85 Let u and v be right and left eigenvectors, respectively, of a matrix $A_{n\times n}$, associated with a simple eigenvalue λ of A. Show that

$$P = \frac{u\,v^T}{v^T u}$$

is a projection matrix onto $N(A - \lambda I)$.

1.86 Show that if two square matrices A and B are similar, then they must have the same rank.

Hint. Use Theorem 1.44.

1.87 Let A and B be similar matrices. Show that for any positive integer k,

$$B^k = P^{-1} A^k P,$$

for some nonsingular matrix P.

1.88 True or False? Similar matrices have the same eigenvectors.

1.89 Let A be a square matrix of order n. Show that A has a complete set of n linearly independent eigenvectors if and only if A is diagonalizable.

1.90 Show that two matrices A and B can be simultaneously diagonalized if and only if $AB = BA$.

1.91 Let $A_{n\times n}$ be a symmetric matrix. Show that the algebraic multiplicity of any eigenvalue of A is equal to its geometric multiplicity.

1.92 Let A and B be two diagonalizable matrices such that $AB = BA$. Show that

$$e^{A+B} = e^A e^B.$$

1.93 Show that the matrix norm $\|\cdot\|_2$ in (1.9) with $p = 2$, can also be defined as

$$\|A\|_2 = \sqrt{\max \lambda\,(A^T A)}.$$

1.94 Is it possible to use quadratic Bernstein polynomials and still satisfy the tangential properties at both endpoints of a Bezier curve?

1.95 Show that to guarantee smoothness at a point P_i where two Bezier curves meet, the three points P_{i-1}, P_i, P_{i+1} must be collinear.

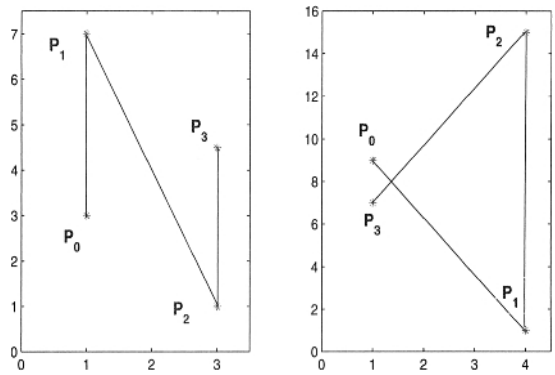


Figure 1.14 Control points for Exercise 1.97.

1.96 Verify the midpoint property of Bezier curves; that is, the point $M = (x(\frac{1}{2}, y(\frac{1}{2})))$ can be obtained by first computing five other midpoints. Refer to Figure 1.11.

1.97 Sketch the Bezier curves corresponding to the control points in Figure 1.14.

1.98 Show that the slope of the Bezier curve at the midpoint M coincides with the slope of the segment joining d and M.

1.99 True or False? It is possible for the control points P_1 and P_2 to lie on opposite sides of the corresponding Bezier curve.

1.100 True or False? It is not possible to draw an ellipse with one single Bezier curve.

CHAPTER 2

RANKING WEB PAGES

One relatively new application of eigenvalues, eigenvectors, and a number of results and techniques in matrix computations and numerical analysis is on the problem of ranking web pages or documents. When we enter a query, say on Google's famous search engine, a long list of pages or web documents are returned as relevant, that is, documents that match or are related to the query in some way (in Section 3.5, we study how relevant documents are matched with a given query). However, the order in which these pages appear on our screen is far from arbitrary; they follow a determined order according to a certain degree of importance. For instance, if we enter the query "news", and the first web pages on the list are cnn, msnbc, newsgoogle, foxnews, newsyahoo, and abcnews, that means that those news sources besides being related to the query, they are supposed to be more important than hundreds or thousands of other web pages related to news. But, how is this rank of importance determined? In particular, how is this done, taking into account the enormous size of the World Wide Web?

As we will see in this chapter, the mathematical modeling of the problem of ranking web pages leads to the problem of computing an eigenvector associated with the eigenvalue of largest magnitude. Obviously, this requires not only an efficient numerical method for this purpose, but also, given the special case of the huge matrix resulting from the modeling, it requires above all a feasible, simple, and hopefully fast numerical algorithm.

For reasons explained in detail later on, the numerical method of choice for this particular problem is the so–called *power method*, the simplest numerical method available for the

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

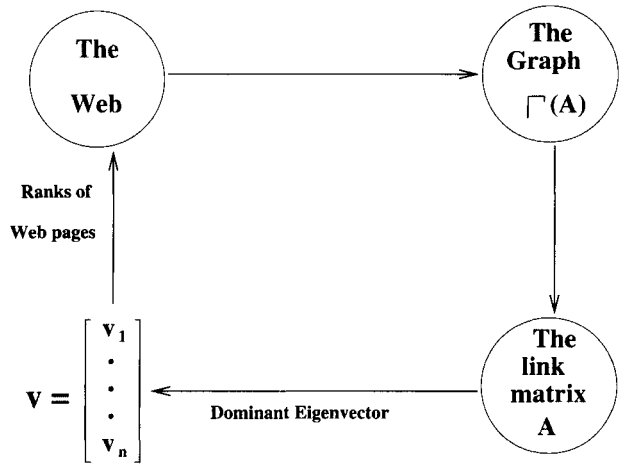


Figure 2.1 Modeling of Web page ranking

computation of eigenvectors. On our way to determine conditions for the uniqueness of solution and for convergence, we will encounter and apply several important concepts and methods from linear algebra and matrix computations.

After studying the eigenvector approach, we will introduce relatively recent results on how to compute such a vector containing the ranks of web pages by using approaches other than the power method itself, and we will show that such alternative methods can be more computationally efficient. We stress again that the problem of ranking web pages is an elegant, illustrative, and obviously very useful real-world application of theory and techniques in linear algebra and matrix analysis.

2.1 THE POWER METHOD

Recall that given an $n \times n$ matrix A, a vector $x \neq 0$ is called an eigenvector of A if

$$Ax = \lambda x,$$

where λ is the corresponding eigenvalue. Similarly, a vector $x \neq 0$ is called a left eigenvector of A if

$$x^T A = \lambda x^T.$$

Observe that a left eigenvector of A is in fact an eigenvector of A^T, for the same eigenvalue λ.

In Section 1.10, we showed how to find eigenvalues and eigenvectors analytically, by solving the associated characteristic equation and a singular linear system, respectively. This is feasible as long as n is small. For large values of n, we have to use numerical methods.

However, numerically approximating the eigenvalues and eigenvectors of a matrix A is not a very easy task; several sophisticated numerical methods have been developed for this task, the simplest of all of them is the so–called power method, which is however able to compute only the dominant eigenvector (the one associated with the eigenvalue of largest magnitude). For this reason, other methods are in general preferred over the power method, as most applications require the computation of more than just one of the eigenvalues or one of the eigenvectors. This had made the power method mostly just an illustrative introduction to the numerical computation of eigenvalues and eigenvectors.

Nevertheless, from the point of view of eigenvectors, the power method is now virtually the only method of choice for solving the problem of ranking web pages or documents, and search engines like Google make extensive use of this extremely simple but very effective numerical method, and it is in fact one of the main mathematical tools that Google and other search engines utilize. In the next sections, we explain in detail why this is so.

Let us start by describing how and why the power method works, and under what conditions. Assume that a matrix $A_{n\times n}$ has the following properties:

(a) it has a complete set of linearly independent eigenvectors $v_1, \ldots, v_n$, and

(b) the corresponding eigenvalues satisfy

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|. \tag{2.1}$$

Observe the strict inequality in (2.1). We are interested in computing v_1 – the dominant eigenvector, that is, the eigenvector associated with λ_1. Assume now that we choose an arbitrary initial guess vector x_0. Since the eigenvectors are linearly independent, then we can uniquely write

$$x_0 = c_1 v_1 + \cdots + c_n v_n, \tag{2.2}$$

for some scalars $c_1, \ldots, c_n$, with $c_1 \neq 0$.

We can multiply both sides of (2.2) by A repeatedly and use the fact that $Av_i = \lambda_i v_i$, for $i = 1, \ldots n$, to obtain

$$\begin{array}{rcl} Ax_0 & = & c_1\lambda_1 v_1 + c_2\lambda_2 v_2 + \cdots + c_n\lambda_n v_n \\ A^2x_0 & = & c_1\lambda_1^2 v_1 + c_2\lambda_2^2 v_2 + \cdots + c_n\lambda_n^2 v_n \\ & \vdots & \\ A^kx_0 & = & c_1\lambda_1^k v_1 + c_2\lambda_2^k v_2 + \cdots + c_n\lambda_n^k v_n. \end{array} \tag{2.3}$$

Now recall that $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Then, as k increases, the linear combination in A^kx_0 has more and more weight on the first term $c_1\lambda_1^k v_1$ compared with any of the other terms. In other words, as k increases, the vector A^kx_0 moves toward the dominant eigenvector direction, and we expect that as k gets larger and larger, the vector A^kx_0 will converge to v_1. That is exactly the main idea behind the power method.

Thus, we can define

$$\begin{aligned} x_1 &= Ax_0 \\ x_2 &= A^2x_0 = A(Ax_0) = Ax_1 \\ x_3 &= A^3x_0 = A(A^2x_0) = Ax_2 \\ &\vdots \\ x_{k+1} &= A^{k+1}x_0 = A(A^kx_0) = \cdots = Ax_k. \end{aligned} \tag{2.4}$$

This means that it is not necessary to explicitly compute the powers of A, and instead just perform a matrix-vector multiplication Ax_k at each step.

The eigenvalue λ_2 is called the **subdominant eigenvalue** and its corresponding eigenvector v_2, the **subdominant eigenvector**. There is an important fact to observe in (2.3), especially in its last line: The farther $|\lambda_1|$ is from $|\lambda_2|$, the faster the vector A^kx_0 will move toward the direction of the dominant eigenvector v_1. That is, the speed of convergence depends on the distance between the dominant and subdominant eigenvalue (in magnitude), or more properly, on how small the ratio $\frac{|\lambda_2|}{|\lambda_1|}$ is.

Now we collect the discussion above into the following

Theorem 2.1 *Suppose the eigenvalues of a matrix $A_{n\times n}$ have the property*

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|, \tag{2.5}$$

with corresponding linearly independent eigenvectors $v_1, \ldots, v_n$. Then, for an arbitrary x_0, the power iteration

$$x_{k+1} = A\,x_k, \qquad k = 0, 1, 2, \ldots \tag{2.6}$$

converges to the dominant eigenvector v_1, with rate of convergence

$$q = \left|\frac{\lambda_2}{\lambda_1}\right|. \tag{2.7}$$

■ **EXAMPLE 2.1**

Consider the matrix

$$A = \begin{bmatrix} 2 & -3 & 0 & 0 \\ 3 & 2 & 0 & 0 \\ 5 & 5 & 1 & 0 \\ 3 & -7 & 8 & -5 \end{bmatrix}.$$

The eigenvalues are

$$\lambda_1 = -5, \quad \lambda_{2,3} = 2 \pm 3, \mathrm{i}, \quad \lambda_4 = 1.$$

Thus, the relationship $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq |\lambda_4|$ is satisfied, and the power iteration $x_{k+1} = Ax_k$ will compute the dominant eigenvector

$$v = [0 \quad 0 \quad 0 \quad 1]^T.$$

Here we show some of the iterates:

$$
\begin{array}{rcrrrr}
x_0 & = & 1.00000000 & 1.00000000 & 1.00000000 & 1.00000000 \\
x_1 & = & 0.00127052 & -0.30365472 & -0.60857998 & 0.73309112 \\
x_5 & = & 0.02305952 & 0.00028946 & -0.02248059 & 0.99948126 \\
x_9 & = & 0.00032287 & -0.00030705 & -0.00093697 & 0.99999946 \\
x_{15} & = & 0.00000033 & -0.00000879 & -0.00001792 & 0.99999999 \\
ax_{25} & = & 0.00000008 & -0.00000003 & -0.00000014 & 0.99999999 \\
x_{39} & = & 0.00000000 & 0.00000000 & 0.00000000 & 1.00000000
\end{array}
$$

Computing the eigenvalue.

Although for the problem of ranking web pages we are not interested in finding the dominant eigenvalue (it will always be $\lambda_1 = 1$), in general, one simple way to compute the eigenvalue λ associated with an already computed eigenvector v is derived from least squares. We begin by writing the equation $Av = \lambda v$ as

$$v\,\lambda = Av. \tag{2.8}$$

Equation (2.8) can be understood as a linear system with the $(n \times 1)$ coefficient matrix v. This system is overdetermined, and therefore (see Chapter 4), we look for its least squares solution, which is determined by solving the normal equations

$$v^T v\,\lambda = v^T A v,$$

from which we get

$$\lambda = \frac{v^T A\, v}{v^T v}.$$

For instance, for the matrix A of Example 2.1, with $v = [0 \quad 0 \quad 0 \quad 1]^T$, we have that $Av = [0 \quad 0 \quad 0 \quad -5]^T$, and

$$\lambda = \frac{v^T(Av)}{1} = -5.$$

Some more details on this quotient are given in the Exercises section.

A different way to obtain the dominant eigenvalue is arrived at by first defining for a vector $x = [x_1 \cdots x_n]^T$ the linear function

$$f(x) = x_i, \quad \text{for some} \quad i = 1, \ldots, n.$$

Now, on the last line of (2.3), we can absorb the constants $c_1, \ldots, c_n$ into the corresponding vectors $v_1, \ldots, v_n$ to obtain

$$
\begin{aligned}
x_k &= \lambda_1^k v_1 + \lambda_2^k v_2 + \cdots + \lambda_n^k v_n \\
&= \lambda_1^k \left[v_1 + \left(\tfrac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \left(\tfrac{\lambda_n}{\lambda_1}\right)^k v_n \right].
\end{aligned}
$$

Then,

$$f(x_k) = \lambda_1^k \left[f(v_1) + f(\epsilon_k) \right],$$

where $\epsilon_k = \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^k v_n$.

Now let

$$\mu_k = \frac{f(x_{k+1})}{f(x_k)} = \lambda_1 \left[\frac{f(v_1) + f(\epsilon_{k+1})}{f(v_1) + f(\epsilon_k)} \right].$$

Since $\epsilon_k \to 0$ as $k \to \infty$ and f is linear, we have that

$$\mu_k \longrightarrow \lambda_1, \qquad \text{as} \quad k \longrightarrow \infty.$$

Thus, the ratio defined by μ_k approximates the dominant eigenvalue in an iterative process.

■ **EXAMPLE 2.2**

For the matrix of Example 2.1, we compute the dominant eigenvalue $\lambda_1 = -5$ by following the ratio $\mu = f(x_{k+1})/f(x_k)$, where we define f as $f(x) = x_4$. These are some of the iterates

$$\begin{aligned}
\mu_1 &= -4.54330708, \\
\mu_5 &= -4.69220022, \\
\mu_9 &= -4.99505589, \\
\mu_{15} &= -4.99998513, \\
\mu_{25} &= -4.99999975, \\
\mu_{39} &= -4.99999999.
\end{aligned}$$

Going back to our original goal, we remark that there is more than one way to compute the (dominant) vector that contains the rank of all web pages. We cover here the method used by Google, not only because of its obvious practical importance and unquestionable success, but also because of the interesting mathematics behind it. First, we need to introduce three special types of matrices that will be encountered when developing the so-called PageRank algorithm.

2.2 STOCHASTIC, IRREDUCIBLE, AND PRIMITIVE MATRICES

The entire web will be represented as a directed graph, where each node on the graph is a web page. In turn, a particular matrix will be associated with this graph, so that such matrix will represent the web pages and their links to each other (see Figure 2.1). After some modification, the original matrix will be **nonnegative**, that is, a matrix with nonnegative entries, and each entry will lie in the interval $[0, 1]$, with each row adding to one. It will be a stochastic matrix.

Definition 2.2 *A matrix $A_{n\times n}$ is called (row)* **stochastic** *if it is nonnegative and the sum of the entries in each row is 1.*

Note: In a similar way, we can define a *column stochastic* matrix as the one with the property that the sum of the entries in each column is 1. Observe that in any case, this means that all the entries of a stochastic matrix are numbers in the interval $[0,\ 1]$.

Due to their special form, stochastic matrices have all their eigenvalues lying in the unit circle. In fact, we have the following theorem.

Theorem 2.3 *If $A_{n\times n}$ is a stochastic matrix, then its eigenvalues have the property*

$$|\lambda_i| \leq 1, \qquad i = 1, \ldots, n,$$

and $\lambda = 1$ is always an eigenvalue of A.

Proof. Since the spectral radius of a matrix A never exceeds the norm of A (see Exercise 1.15), and since A is stochastic, then

$$\rho(A) \leq \|A\| = 1, \tag{2.9}$$

where $\|\cdot\|$ is the ∞ or 1- matrix norm, depending on whether A is row or column stochastic, respectively. On the other hand, let $e \in \mathbb{R}^n$ be a vector of ones. Because A is stochastic, we either have $Ae = e$ or $e^T A = e^T$ $(A^T e = e)$ depending whether A is row or column stochastic respectively. That is, $\lambda = 1$ is always an eigenvalue of A. But this also means that $\rho(A) \geq 1$, which combined with the inequality in (2.9) implies that $\rho(A) = 1$. Therefore, every eigenvalue of A must satisfy $|\lambda| \leq 1$.

□

■ **EXAMPLE 2.3**

The eigenvalues of the stochastic matrix

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

are (in order of magnitude):

$$\lambda_1 = 1,\ \lambda_2 = -0.9173,\ \lambda_{3,4} = -0.0413 \pm 0.2986\, i,\ \lambda_5 = 0.$$

■ **EXAMPLE 2.4**

This example shows that a stochastic matrix can have more than one eigenvalue of magnitude one.

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

The matrix A has eigenvalues $\lambda_1 = 1,\ \lambda_2 = -1,\ \lambda_3 = 0,\ \lambda_4 = 0,$ so two of them have magnitude 1.

Definition 2.4 *A matrix $A_{n\times n}$ is called* **reducible** *if there is a permutation matrix $P_{n\times n}$ and an integer $1 \le r \le n-1$ such that*

$$P^T A P = \begin{bmatrix} C & D \\ 0 & E \end{bmatrix}, \tag{2.10}$$

where 0 is an $(n-r)\times r$ block matrix. If a matrix A is not reducible, then it is called **irreducible**.

Remark 2.5 *Observe that equation* (2.10) *is a statement of similarity between the matrix A and a block upper triangular matrix.*

From Definition 2.4, we can conclude two things:

(a) A positive matrix (all its entries are positive) is irreducible.

(b) If A is reducible, it must have at least $(n-1)$ zero entries.

However, a matrix with $(n-1)$ or more zero entries is not necessarily reducible, as in the case of the matrix of Example 2.3, which is irreducible.

It is of course not simple to conclude that a given matrix is irreducible or not by just looking at it, or by trying to find the permutation matrix P that would transform it into the canonical form (2.10). However, if we have a web or a graph associated with the matrix, it is much simpler to deduce the reducibility property of the matrix, at least for small n.

Definition 2.6 *Given a matrix $A_{n\times n}$, the* **graph** $\Gamma(A)$ *of A is a directed graph with n nodes $P_1, P_2, \ldots, P_n$ such that*

$$P_i \text{ has a link to } P_j \iff a_{ij} \neq 0.$$

Note: Conversely, given a directed graph $\Gamma(A)$, the associated matrix A will be called the **link matrix**, or **adjacency matrix.**

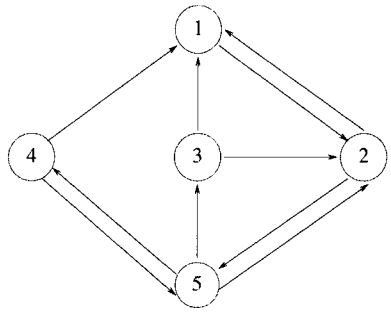


Figure 2.2 Web of Example 2.5.

Definition 2.7 *A directed graph is called* **strongly connected** *if it there exists a path from each node to every other node.*

This definition says that every node is reachable from any other node in a finite number of steps.

■ **EXAMPLE 2.5**

Consider the matrix of Example 2.3. The associated graph $\Gamma(A)$ is shown in Figure 2.2. We observe that every node in the graph is reachable from any other given node. Thus, $\Gamma(A)$ is strongly connected. The associated matrix A is irreducible.

The fact that the graph $\Gamma(A)$ in Example 2.5 is strongly connected and the associated matrix A in Example 2.3 is irreducible is a consequence of a general result that can be stated as follows.

Theorem 2.8 *A matrix $A_{n\times n}$ is irreducible $\iff$ $\Gamma(A)$ is strongly connected.*

Proof. Assume that $\Gamma(A)$ is not strongly connected. After a relabeling process, assume that $S_1 = \{P_1, \ldots, P_r\}$ is the set of all nodes that cannot be reached from node P_n. If we denote $S_2 = \{P_{r+1}, \ldots, P_{n-1}\}$, then no node in S_1 can be reached from any node in S_2 (Figure 2.3). Thus, if we denote with P the permutation matrix associated with the relabeling process, then we can write

$$P^T A P = \begin{bmatrix} C & D \\ 0 & E \end{bmatrix},$$

where C is square of order r and 0 is an $(n-r)\times r$ block matrix. C represents S_1 and E represents $S_2 \cup \{P_n\}$. This implies A is reducible.

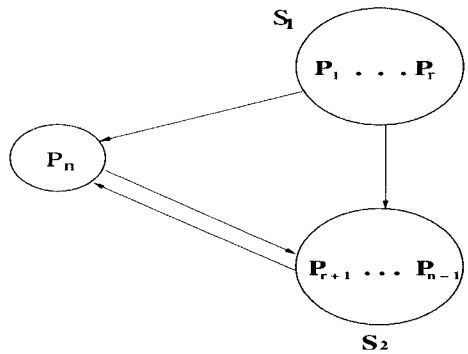


Figure 2.3 The graph $\Gamma(A)$ in Theorem 2.8.

Conversely, assume A is reducible. Then, from Definition 2.4, there is a permutation matrix $P_{n\times n}$ such that

$$B = P^T A P = \begin{bmatrix} C & D \\ 0 & E \end{bmatrix},$$

where 0 is a $(n-r)\times r$ block matrix, with $1 \le r \le n-1$. This means that the nodes $P_1, \ldots, P_r$ in $\Gamma(B)$ cannot be reached from any of the nodes $P_{r+1}, \ldots, P_n$, and therefore, $\Gamma(B)$ is not strongly connected. But, as noted before, because $\Gamma(A)$ is the same as $\Gamma(B)$ up to a renumbering of the nodes, (see Example 2.6), $\Gamma(A)$ is also not strongly connected.

□

■ **EXAMPLE 2.6**

Consider the matrix

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

The associated graph $\Gamma(A)$ is shown in Figure 2.4. We observe that the graph is not strongly connected. In particular, nodes 1 and 2 are not reachable from nodes 3 or 4. According to Theorem 2.8, this means the matrix A is reducible. In fact, we can transform A into the general reducible form (2.10) by permuting its rows and columns:

$$B = P^T A P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix}, \quad \text{with} \quad P = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}.$$

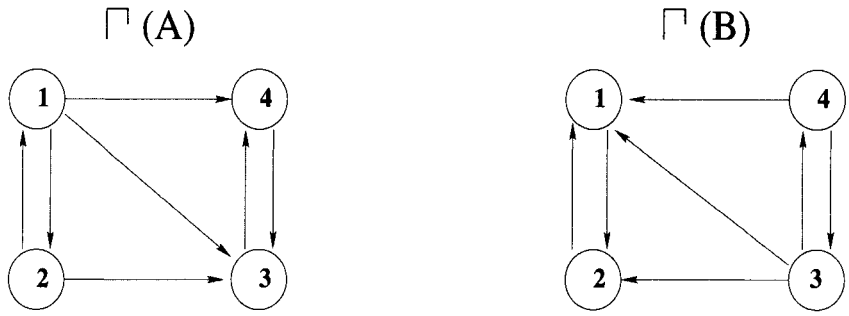


Figure 2.4 Web of Example 2.6

Note. Observe that the graph $\Gamma(B)$ is the same as $\Gamma(A)$ but with its nodes reordered according to the permutation assigned by P; that is:

$$1 \quad 2 \quad 3 \quad 4 \qquad \longrightarrow \qquad 3 \quad 4 \quad 1 \quad 2$$

Remark 2.9 *Even though the two matrices in Examples 2.5 and 2.6 are stochastic, in general, reducible or irreducible matrices need not be stochastic.*

When developing an algorithm to compute the rank of web pages, we will be interested in computing the dominant eigenvector of a link matrix. In general, for nonnegative matrices, there is an important theorem that provides conditions for the existence and uniqueness of such an eigenvector. This theorem is a generalization to nonnegative matrices of one originally proposed by Perron for positive matrices. The key property is that of irreducibility. Recall that $\rho(A)$ denotes the spectral radius of A; that is,

$$\rho(A) = \max\ |\lambda|, \qquad \lambda \text{ is eigenvalue of } A.$$

Theorem 2.10 (Perron–Frobenius) *Let $A_{n\times n}$ be a nonnegative irreducible matrix. Then,*

(a) $\rho(A)$ is a positive eigenvalue of A.

(b) There is a positive eigenvector v associated with $\rho(A)$.

(c) $\rho(A)$ has algebraic and geometric multiplicity 1.

In general, a matrix may have a spectral radius that is not an eigenvalue itself. For instance we may have the eigenvalues $\sigma(A) = \{-4, -2, -1, 3\}$. In this case, $\rho(A) = 4$, but 4 is not an eigenvalue of A. Also, $\rho(A)$ may be an eigenvalue but with algebraic multiplicity greater than one. For instance, we may have the case $\sigma(A) = \{2, 3, 4, 4\}$, where $\rho(A) = 4$ is an eigenvalue of algebraic multiplicity 2.

It is very important to stress that Theorem 2.10 guarantees that the vector space spanned by the dominant eigenvector is one-dimensional, which after normalization, it means that we will have a unique dominant eigenvector. For the purpose of ranking web pages, we will

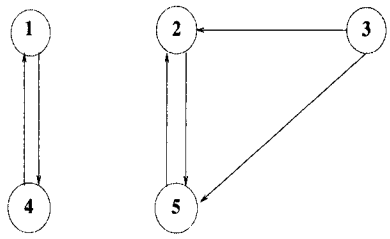


Figure 2.5 Web of Example 2.7.

be looking for an eigenvector that is a probabilistic vector, that is, a vector with entries between 0 and 1, and with the sum of its components equal to one. That is, the 1-norm of the dominant eigenvector $v = [\, v_1 \ \ldots \ v_n \,]^T$ is

$$\|v\|_1 = v_1 + \cdots + v_n = 1.$$

Let us illustrate what may happen when irreducibility fails and the Perron–Frobenius theorem cannot be applied.

■ **EXAMPLE 2.7**

The following nonnegative matrix is reducible

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

The graph $\Gamma(A)$ is shown in Figure 2.5. The eigenvalues of A are $\sigma(A) = \{1, 1, -1, -1, 0\}$. Thus, the spectral radius $\rho(A) = 1$ is an eigenvalue of A, with algebraic multiplicity 2. Furthermore, its associated dominant eigenvector is of the general form $v = [\, a \ b \ 0 \ a \ b \,]^T$. Thus, not only it is not positive, but the presence of two free variables a and b means that the eigenspace is two-dimensional. In fact, two linearly independent (normalized) eigenvectors can be chosen as $x = [\frac{1}{2} \ 0 \ 0 \ \frac{1}{2} \ 0]^T$ and $y = [0 \ \frac{1}{2} \ 0 \ 0 \ \frac{1}{2}]^T$.

From Example 2.7, we learn that if the matrix is reducible (the corresponding graph is not strongly connected), we may have a problem of uniqueness of the dominant eigenvector when computing web pages rankings. For instance, which of the two vectors in such example should we pick as the right PageRank vector? Furhtermore, since they are linearly independent, any combination of them will also be in the same eigenspace.

Remark 2.11 *If a web is considered as an indirected graph (links but no arrows), and it consists of k disconnected subgraphs, it can be proved that the corresponding dominant eigenspace has dimension at least k. See Exercise 2.25.*

A general nonnegative matrix that is irreducible is already very special according to Theorem 2.10, especially because $\rho(A)$ is an eigenvalue of algebraic and geometric multiplicity 1. But for the convergence of the power method, we will need the additional ingredient that ensures that the strict inequality in (2.5) is satisfied, that is, that the matrix in consideration should have only one eigenvalue of maximum magnitude.

■ **EXAMPLE 2.8**

The matrix

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

is stochastic and irreducible, and therefore, $\rho(A)$ has algebraic and geometric multiplicity one. However, it does not have a unique eigenvalue of largest magnitude. In fact, its eigenvalues are $\sigma(A) = \{1, -1, \pm \mathrm{i}\}$, all of them with magnitude 1. For this matrix, the strict inequality in (2.5) is not satisfied.

If we add the condition we need, then we get the following definition.

Definition 2.12 *A nonnegative matrix $A_{n\times n}$ is called* **primitive** *if it is irreducible and has only one eigenvalue of maximum magnitude.*

■ **EXAMPLE 2.9**

The matrix A of Example 2.3 is nonnegative and irreducible. In addition, its eigenvalues are $\lambda_1 = 1$, $\lambda_2 = -0.9173$, $\lambda_{3,4} = -0.0413 \pm 0.2986\,\mathrm{i}$, and $\lambda_5 = 0$. Therefore, A is primitive as it has only $\lambda_1 = 1$ as the dominant eigenvalue.

There is a useful characterization of primitive matrices that allows us to identify them without computing the eigenvalues.

Theorem 2.13 *A nonnegative matrix $A_{n\times n}$ is primitive if and only if*

$$A^k > 0, \qquad \text{for some} \quad k \geq 1. \tag{2.11}$$

Observe that according to this theorem, primitive matrices are just a generalization of positive matrices. In fact, a positive matrix is a primitive matrix with $k = 1$ in (2.11).

■ **EXAMPLE 2.10**

For the same matrix of Example 2.3, one can verify that A is primitive, with $k = 5$. In fact,

$$A^5 = \begin{bmatrix} 0.0417 & 0.4583 & 0.2708 & 0.4583 & 0.1111 \\ 0.6389 & 0.1250 & 0.3611 & 0.1250 & 0.5509 \\ 0.1389 & 0.0139 & 0.0694 & 0.0139 & 0.1204 \\ 0.1389 & 0.0139 & 0.0694 & 0.0139 & 0.1204 \\ 0.0417 & 0.3889 & 0.2292 & 0.3889 & 0.0972 \end{bmatrix} > 0.$$

2.3 GOOGLE'S PAGERANK ALGORITHM

After covering Sections 2.1 and 2.2, we have the basic background to introduce a well–defined algorithm for ranking web pages. The fundamental numerical method to compute the dominant eigenvector is the power method, and the matrix under consideration will be modified to be stochastic. After a second modification, it will be not only irreducible, but it will be even primitive, fulfilling all the conditions for convergence to a unique eigenvector, the PageRank vector. It all starts with the original ideas of Google's founders Larry Page and Sergey Brin, and their article [11].

Before Google was available, entering a query online in search of information meant obtaining several screens of web pages or documents which happened to match the search text, but the order in which they were shown was not efficient. To locate an important document out of the long list, one had to scroll down and first read a long list of pages of dubious source that besides matching the text were not necessarily reliable or useful. The main reason behind Google's great success is that its PageRank algorithm is able to give very reliable rates of importance to all web pages in the world, so that after a query, it presents the user with a list of web pages or documents starting with the most important and typically most relevant ones. In this chapter, we review the mathematical concepts and tools used to develop the PageRank algorithm and we discuss how a direct application of such mathematical concepts as well as some practical manipulation of the algorithm itself have led to its success.

Because web pages are supposed to be listed in order of importance, the first question to answer was how to define the importance of a web page. The reasoning is that if a page is important (e.g., Google, Yahoo, top universities, etc.), then a great number of other web pages will have links to it. Thus, a way to define the importance of a page is by counting the number of *inlinks* to it. In this context, a link from page P_1 to page P_2 can be understood as page P_1 "voting" for page P_2. However, to avoid one page exerting too much influence by casting multiple votes, these votes should be scaled in some way. At the same time, votes from important pages should be ranked higher than those from ordinary ones. Altogether, this should result on each page having a rank of importance.

The central ideas in the PageRank algorithm can be summarized as:

(a) The rank of a page P is determined by adding the ranks of the pages that have *outlinks* pointing to P, weighted by the number of outlinks of each of those pages. This means for instance that if pages P_1 and P_2 point to page P, and P_1 has 15 outlinks, while P_2 has 100 outlinks, then the link from P_1 will be given given more weight than P_2. Each vote or link from P_1 is worth 1/15, while each vote from P_2 is only worth 1/100.

(b) In addition, links from important pages (revered authors, top-rank colleges, etc.) will have a higher weight than ordinary web pages.

If we denote the rank of a given page P_j with $r(P_j)$, then part (a) can be expressed mathematically as

$$r(P_j) = \sum_{P \in B_{P_j}} \frac{r(P)}{|P|}, \tag{2.12}$$

where B_{p_j} is the set of all pages P pointing to P_j, and $|P|$ denotes the number of outlinks from P. This obviously means that some initial ranks have to be assigned to all web pages (otherwise formula (2.12) is self-referencing), and the ultimate rank has to be determined as an iterative process.

Thus, if we consider n web pages, the rank of page P_j at step k is defined as

$$r_k(P_j) = \sum_{P_i \in B_{P_j}} \frac{r_{k-1}(P_i)}{|P_i|}, \quad k = 1, 2, \ldots, \tag{2.13}$$

where $i, j = 1, \ldots, n$ and $i \neq j$.

If we denote $v_k = [r_k(P_1) \quad r_k(P_2) \quad \cdots \quad r_k(P_n)]^T$. Then, equation (2.13) can be expressed as

$$v_k^T = v_{k-1}^T H, \qquad \text{with} \quad H_{ij} = \begin{cases} \frac{1}{|P_i|} & \text{if } P_i \text{ has a link to } P_j \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

Thus, the equation $v_k^T = v_{k-1}^T H$ is nothing else but the power method, in this case to compute the dominant (left) eigenvector of H associated with the eigenvalue $\lambda = 1$. The associated (dominant) eigenvector

$$v = \lim_{k \longrightarrow \infty} v_k \tag{2.15}$$

is called the **rank vector** or **PageRank vector**.

Note. It is also possible to define the PageRank vector when the matrix H is column stochastic. See Exercise 2.63.

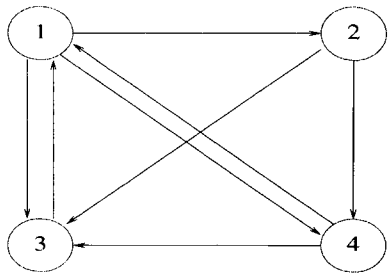


Figure 2.6 Web of Example 2.11.

■ **EXAMPLE 2.11**

Consider the web of Figure 2.6 with four pages (or nodes). By observing how the pages are linked to each other, and following the definition in (2.14), we obtain the *link* matrix

$$H = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix}.$$

Thus, the small web in Figure 2.6 consisting of four pages can be represented as a matrix H whose element h_{ij} can be defined as the probability of moving from node i to node j in one step (that is, in one mouse click).

In the matrix H of Example 2.11 we see that each row adds up to 1; in other words, H is a (row) stochastic matrix. Observe also that the third row has only one nonzero entry. This is because page 3 links only to one page; but this also implies that its vote for page 1 has the largest weight possible, a great benefit for page 1. The votes of the other pages are weighted so that all votes from one page have a total weight of one. In addition, H has a zero diagonal because the case of a page voting for itself is not considered. This is a good example of democracy.

■ **EXAMPLE 2.12**

Let us consider the web of Figure 2.7. The corresponding link matrix is given by

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}. \tag{2.16}$$

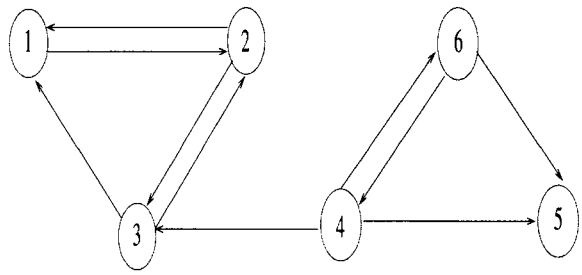


Figure 2.7 Web of Example 2.12.

The fact that page 5 does not have any outlinks is reflected in the matrix H having a corresponding row of zeros. This matrix is not stochastic. At the same time, H is reducible; in fact it can be written in the general reducible form (2.10), for a suitable permutation matrix P, and its graph is not strongly connected.

Example 2.12 is a simple illustration of what is to be expected in the real case of the World Wide Web: there will be several pages or documents with no outlinks (e.g., most PDF files), causing the corresponding link matrix to have a number of zero rows; in addition, because a given web page outlinks only to a few other pages, out of the several billions of pages available, the link matrix representing the entire World Wide Web will be very sparse and therefore, almost certainly reducible.

Here we discuss these two main issues, and how they can be resolved before one can proceed to apply power method to compute the pagerank vector v in (2.15).

(i) Making the matrix H stochastic.

As observed in Example 2.12, some of the rows in H may fail to add up to one because there is the possibility that some pages have no outlinks, and therefore, H will have some rows of zeros. Such a page or node is usually referred to as a **dangling node**. There is an easy solution to this problem: Just artificially add links by replacing the zero row with a nonzero vector. This is achieved by adding the matrix au^T to H, where u is an arbitrary probabilistic vector (say, $u = e/n$, with e is a vector of ones), and a is a vector with entry i equal to 1, if page i is a dangling node; all other entries are zero.

More precisely, we define

$$B = H + a\,u^T, \tag{2.17}$$

$$\text{where} \qquad a_i = \begin{cases} 1 & \text{if page } i \text{ is a dangling node} \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

This will modify only the zero rows, leaving the other rows untouched. Every zero row is effectively replaced by a probabilistic vector. This new matrix B is stochastic, and hence, the corresponding modified web has no dangling nodes.

■ **EXAMPLE 2.13**

The matrix H of Example 2.12 has one row of zeros. By using (2.17), with $u = e/6$, we define

$$B = H + a\,e^T/6, \qquad \text{with} \quad a = [0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0]^T$$

and e a vector of ones, to obtain

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}. \tag{2.19}$$

In the new web represented by B, page 5 now would link to all other pages and each of its votes is worth $1/6$.

Remark 2.14 *Observe that in Example 2.13, true democracy has been sacrificed because page 5 now has a vote for itself. However, in practice this does not represent a big concern. In general, due to the huge size of the web, each vote of a page that originally was a dangling node is actually worth almost nothing, e.g., $1/n$, where n is in the order of billions. Thus, for the World Wide Web practical purposes, B is still very close to the original model represented by H and basically, the relationship between the ranks of the pages is not much affected.*

Remark 2.15 *Defining the matrix au^T, with $u = e/n$ may be overkill by completely filling up rows of zeros. Recall that u is an arbitrary probabilistic vector, so that e.g., in Example 2.13 we could define*

$$u = [1/2 \quad 0 \quad 1/6 \quad 1/3 \quad 0 \quad 0]^T,$$

so that the the fifth row of the matrix B in (2.19) equals this vector u and some sparsity is preserved. Later we will see that this type of choice for u gives other additional advantages.

(ii) Making the matrix B irreducible.

Remember from Section 2.2 that if a matrix is irreducible, then (after normalization), there is a unique dominant eigenvector. If this condition fails, then we face a problem of uniqueness. The issue of reducibility happens when for example a web page P_i contains only a link to page P_j, and in turn P_j contains only a link to P_i, then a surfer hitting any of these two pages is virtually trapped bouncing between P_i and P_j forever. More generally, in the language of graphs, and as noted in Theorem 2.8, reducibility happens when the web is not strongly connected; that is, not every node or page is reachable from any given arbitrary node.

To ensure irreducibility, Google forces every page to be reachable from every other page. This is very reasonable from the practical point of view, because a surfer can eventually jump to any page entering a URL on the command line, no matter from which page.

The strategy is to add a perturbation matrix to B, and let

$$G = \alpha B + (1-\alpha)E, \tag{2.20}$$

where $\alpha \in (0 \quad 1)$. The matrix E can be defined as

$$E = e\,e^T/n \qquad \text{or} \qquad E = e\,u^T, \tag{2.21}$$

where again, e is a vector of ones, and the arbitrary probabilistic vector u is called the **personalization vector**, a vector that has a great role in the success of Google's search engine. The stochastic irreducible matrix G is what we can call the **"Google matrix"**.

The two choices of perturbation matrices in (2.21) are quite different. Every entry in the matrix $E = e\,e^T/n$ is given by $E_{ij} = 1/n$, so that every page will have its rank equally modified. This is the choice that Google used in its first years, but for several reasons that we explain below, now the second choice is being used. Also observe that the first choice of E is just a particular case of the second one, with $u = e/n$.

■ **EXAMPLE 2.14**

The matrix B of Example 2.13 is stochastic but reducible, because it can be expressed in the general form of (2.10) with $r = 3$. By taking the first choice in (2.21) we perturb B to make it irreducible, and for $\alpha = 0.8$ we define according to (2.20)

$$G = \alpha B + (1-\alpha)\,e\,e^T/6 = \frac{1}{30}\begin{bmatrix} 1 & 25 & 1 & 1 & 1 & 1 \\ 13 & 1 & 13 & 1 & 1 & 1 \\ 13 & 13 & 1 & 1 & 1 & 1 \\ 1 & 1 & 9 & 1 & 9 & 9 \\ 5 & 5 & 5 & 5 & 5 & 5 \\ 1 & 1 & 1 & 13 & 13 & 1 \end{bmatrix}. \tag{2.22}$$

All the entries are now positive, and that obviously makes G irreducible. In terms of the web it means every page is now reachable from any other page; that is, the associated graph is strongly connected.

The two major technical problems with the original matrix H have been solved above by making some modifications to such matrix so that it is stochastic and irreducible. But as observed in Example 2.14, even more has been obtained because the general Google matrix G in (2.20) has all its entries strictly positive. This also makes it primitive! Therefore, now we also know that $\lambda = 1$ is the only eigenvalue of G of largest modulus, thus satisfying the condition (2.5) for the convergence of the power method.

In fact, if we apply the power method to (2.22), it will uniquely converge to its normalized dominant (left) eigenvector (the PageRank vector):

$$v = [0.2537 \quad 0.3262 \quad 0.1948 \quad 0.0710 \quad 0.0900 \quad 0.0643]^T. \tag{2.23}$$

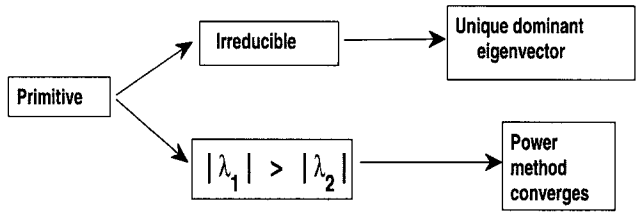


Figure 2.8 A primitive matrix satisfies all conditions needed

That is, this vector v satisfies

$$v^T = v^T G.$$

Suppose now that a query is entered, and the pages (or documents) 1, 2, 4, and 5 are returned as relevant; that is, they match the query. Then, these four pages will be shown to the user, but they will be arranged in the order given by the corresponding entries in the PageRank vector; that is,

$$\begin{array}{ll} P_2 & (\text{rank} = 0.3262), \\ P_1 & (\text{rank} = 0.2537), \\ P_5 & (\text{rank} = 0.0900), \\ P_4 & (\text{rank} = 0.0710). \end{array} \tag{2.24}$$

With the properties of the matrix G, convergence of the power method to a unique PageRank vector is now guaranteed, and one can proceed to compute such a vector (see Figure 2.8). However, there are a couple more issues we want to discuss.

First, besides a page being important in terms of how many inlinks it receives from other pages, it would be a great advantage to be able to manipulate some of the ranks so that really important pages are ensured to get a well–deserved high rank; at the same time the ranks of pages of financially generous businesses could also be pushed up, making sure their products will be listed first. These considerations have been implemented in PageRank and have put Google at the top of search engines. And second, nothing has been said about the speed of convergence, whether there could be some ways to improve the rate of convergence of the power method, and whether the sparsity of the original matrix H can still be exploited, although from (2.22) we can see that it has been lost.

It is remarkable to learn that all these issues can be solved by manipulating the entries of the personalization vector u in (2.21), by giving an appropriate value to α in (2.20,) and by rewriting the power method to avoid using the matrix G explicitly.

2.3.1 The personalization vector

Recall that we have modified the original link matrix H to make it stochastic and irreducible, and that the final matrix G in (2.20) we have obtained is even primitive, so that all conditions are given for uniqueness and convergence when applying the power method. Now we consider the second and more general choice in (2.21); that is, we take

$$E = e\,u^T, \tag{2.25}$$

where e is a vector of ones, and the vector u is in general an arbitrary probabilistic vector: its entries lie between 0 and 1, and they add up to 1.

It is this vector u that allows us to manipulate the PageRank values up or down according to several considerations. Thus, home pages of revered authors, top schools, or financially important businesses can be assigned higher ranks by appropriately fixing the entries of u. At the same time for reasons like spamming attempts or artificially trying to manipulate their own ranks, other pages can be pushed down to lower ranks.

To show how effective this approach can be, we will consider again the matrix B (2.19) from Example 2.13.

■ **EXAMPLE 2.15**

After modification of the original link matrix H, we have the matrix

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/3 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}.$$

To make this matrix reducible, we perturb as in (2.20), with $\alpha = 0.8$. Suppose now that we want to promote page 5 to a much higher rank. Then, we can define the probabilistic personalization vector

$$u = \frac{1}{50}[1 \quad 1 \quad 1 \quad 1 \quad 45 \quad 1]^T.$$

Now by taking the second choice in (2.21), we get

$$G = \alpha B + (1-\alpha)\, e\, u^T = \frac{1}{750}\begin{bmatrix} 3 & 603 & 3 & 3 & 135 & 3 \\ 303 & 3 & 303 & 3 & 135 & 3 \\ 303 & 303 & 3 & 3 & 135 & 3 \\ 3 & 3 & 203 & 3 & 335 & 203 \\ 103 & 103 & 103 & 103 & 235 & 103 \\ 3 & 3 & 3 & 303 & 435 & 3 \end{bmatrix}.$$

When we apply the power method to this matrix G, we get the PageRank vector

$$v = [0.2091 \quad 0.2688 \quad 0.1605 \quad 0.0585 \quad 0.2501 \quad 0.0530]^T. \tag{2.26}$$

Thus, assuming again that after entering a querry, pages 1,2 4, and 5 are the relevant ones, now they will be shown to the user in the following order:

$$\begin{array}{ll} P_2 & (\text{rank} = 0.2688), \\ P_5 & (\text{rank} = 0.2501), \\ P_1 & (\text{rank} = 0.2091), \\ P_4 & (\text{rank} = 0.0585). \end{array} \tag{2.27}$$

It is obvious in Example 2.15 that with that choice of a personalization vector, we force the rank of page 5 up from $v_5 = 0.0900$ to $v_5 = 0.2501$, a huge increase indeed (the new rank is almost three times the original one). This has moved page five to appear second in the list, with a rank very close to that of page 2. The ranks of all pages other than 5 have decreased.

This is how ranks of a few web pages can be accommodated so that they appear in a somehow prescribed order of importance. Nevertheless, this must be done carefully and applied to modify only a very few number of pages; otherwise, this notion of importance would be in clear conflict with that of relevance, that is, which pages actually are closer to the query.

2.3.2 Speed of convergence and sparsity

To ensure uniqueness of the PageRank vector associated with the dominant eigenvalue $\lambda = 1$, the original link matrix H representing the web has been modified according to (2.17) to obtain a stochastic matrix B, and then B itself has been modified according to (2.20) to obtain the stochastic irreducible (and in fact, primitive) matrix G. Also recall that with this matrix G, the condition (2.5) is satisfied and the convergence of power method is guaranteed.

One very important thing to note is that although the Google matrix G still has all of its eigenvalues lying inside the unit circle, they are in general different from those of B. What is surprising is that instead of causing any trouble, this has a positive effect on the algorithm itself and it can be effectively exploited.

Theorem 2.16 *Denote the spectra of B and G with*

$$\sigma(B) = \{1, \mu_2, \ldots, \mu_n\}, \quad \text{and} \quad \sigma(G) = \{1, \lambda_2, \ldots, \lambda_n\}. \tag{2.28}$$

Then,

$$\lambda_k = \alpha\, \mu_k, \qquad k = 2, 3, \ldots, n, \tag{2.29}$$

where α is as in (2.20).

Proof. Let $q \in \mathbb{R}^n$ so that $q_i = 1/\sqrt{n}$, for $i = 1, \ldots, n$. Since B is row stochastic, we have $Bq = q$. Now let X be a matrix so that

$$Q_{n\times n} = [q \quad X]$$

is orthogonal (see orthonormal extensions in Section 1.4.7). Then,

$$Q^T B Q = \begin{bmatrix} q^T \\ X^T \end{bmatrix} [Bq \quad BX] = \begin{bmatrix} 1 & q^T BX \\ 0 & X^T BX \end{bmatrix} =: \begin{bmatrix} 1 & v^T \\ 0 & C \end{bmatrix}.$$

On the other hand, for u a probabilistic vector and e a vector of ones,

$$u^T Q = u^T [q \quad X] = [u^T q \quad u^T X] = [\frac{1}{\sqrt{n}} \quad u^T X], \quad \text{and}$$

$$Q^T e = \begin{bmatrix} q^T \\ X \end{bmatrix} e = \begin{bmatrix} q^T e \\ X^T e \end{bmatrix} = \begin{bmatrix} \sqrt{n} \\ X^T(\sqrt{n}\, q) \end{bmatrix} = \begin{bmatrix} \sqrt{n} \\ 0 \end{bmatrix}.$$

Then,

$$\begin{aligned} Q^T G Q &= Q^T [\alpha B + (1-\alpha) e u^T] Q \\ &= \alpha \begin{bmatrix} 1 & v^T \\ 0 & C \end{bmatrix} + (1-\alpha) \begin{bmatrix} \sqrt{n} \\ 0 \end{bmatrix} [\tfrac{1}{\sqrt{n}} \quad u^T X] \\ &= \alpha \begin{bmatrix} 1 & v^T \\ 0 & C \end{bmatrix} + (1-\alpha) \begin{bmatrix} 1 & \sqrt{n}\, u^T X \\ 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 1 & * \\ 0 & \alpha C \end{bmatrix}. \end{aligned}$$

From the first similarity transformation $Q^T BQ$, we see that the eigenvalues of C are $\{\mu_2, \ldots, \mu_n\}$, and from the second one $Q^T GQ$, we conclude that indeed the eigenvalues of G satisfy $\lambda_k = \alpha\, \mu_k$, for $k = 2, 3, \ldots, n$.

□

■ **EXAMPLE 2.16**

Consider the matrices B and G of Example 2.15, where we set $\alpha = 0.8$. Their eigenvalues are

$$\sigma(B) = \{1,\ 0.6779,\ -0.0995,\ -0.4117,\ -0.5000,\ -0.5000\},$$

$$\sigma(G) = \{1,\ 0.5423,\ -0.0796,\ -0.3293,\ -0.4000,\ -0.4000\}.$$

Thus, we see that indeed $\lambda_k = \alpha\mu_k$, for $k = 2, \ldots, 6$.

Remark 2.17 *The proof of Theorem 2.16 can also be done by constructing a nonsingular (and not necessarily orthogonal) matrix $U = [q \quad X]$, where q is an eigenvector of U. See Exercise 2.46.*

As discussed in Section 2.1, the speed of convergence of the power method depends on the separation between the dominant eigenvalue and the subdominant eigenvalue. One problem with B was that the subdominant eigenvalue μ_2 may be too close or even equal to the dominant eigenvalue $\mu_1 = 1$, because B is not primitive. The usefulness of Theorem 2.16 is that when we have the matrix G, since $\lambda_2 = \alpha\,\mu_2$ and α lies in the interval $(0,\ 1)$, we can choose α away from 1, to increase the separation between $\lambda_1 = 1$ and λ_2 and, hence, speed up the convergence of the PageRank algorithm. It has been reported that Google uses a value of α between 0.8 and 0.85.

Observe that the choice of α affects how much the matrix B is modified in (2.20), and therefore it also affects the importance of the personalization vector u in (2.21). In other words, the parameter α and the vector u give Google the freedom and power to manipulate and modify page ranks, according to several considerations. Every time a surfer enters a query, there is the confidence that the pages returned are sorted in the right order of importance. At the same time, we just saw that the choice of α also gives control on the speed of convergence of the algorithm. Taking α far from 1 will cause the algorithm to converge faster and gives a larger weight to the perturbation matrix $E = e\,u^T$ and therefore to the personalization vector u. However, α cannot be taken too far from 1, because that will take the matrix G too far from B, and that would not be an accurate model of the original web.

As for sparsity, the original matrix H is very sparse because each page has only very few links to other pages (and remember there are billions of those pages). This sparsity is only barely affected by B. However, the matrix G in (2.20) is positive and has no zero entries anymore, so that in principle sparsity is completely lost. Fortunately this does not represent any problem in practice. Remember we need to apply the power method to find the dominant left eigenvector of G with eigenvalue $\lambda = 1$; that is, we are interested in the iteration

$$v_{k+1}^T = v_k^T G.$$

From (2.17) and (2.20), and using the fact that $v_k^T e = 1$, we have the following two cases, depending on whether $u = e/n$, or u is an arbitrary probabilistic vector.

If $u = e/n$, then

$$\begin{aligned} v_{k+1}^T = v_k^T G &= \alpha v_k^T B + (1-\alpha) v_k^T E \\ &= \alpha v_k^T H + \alpha\, v_k^T a e^T/n + (1-\alpha) v_k^T e\, e^T/n \qquad (2.30) \\ &= \alpha v_k^T H + [\,\alpha\, v_k^T a + (1-\alpha)\,]\, e^T/n. \end{aligned}$$

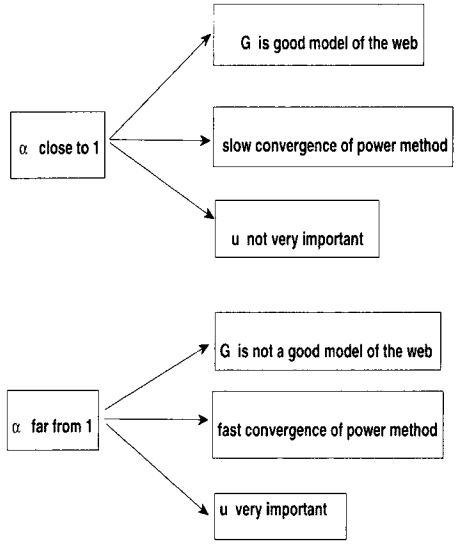


Figure 2.9 How α affects web page ranking.

If u is arbitrary, then

$$\begin{aligned} v_{k+1}^T = v_k^T G &= \alpha v_k^T B + (1-\alpha) v_k^T E \\ &= \alpha v_k^T H + \alpha\, v_k^T a u^T + (1-\alpha) v_k^T e\, u^T \\ &= \alpha v_k^T H + [\alpha\, v_k^T a + (1-\alpha)]\, u^T. \end{aligned} \tag{2.31}$$

Thus, in both cases, the only matrix-vector multiplication is $v_k^T H$, so that sparsity is preserved, something that can be fully exploited in practical terms. Also, the only vector-vector multiplication is $v_k^T a$, where a is a very sparse vector. Furthermore, the calculations above also show that the matrices B and E do not need to be formed explicitly. All this is possible because of the simplicity of the power iteration. In general this would not be possible for more sophisticated numerical methods, and that makes the power method just about the only feasible algorithm for computing the dominant eigenvector (the PageRank vector) of the very large Google matrix G.

Computing the vector a. Recall that the n-dimensional vector a contains the information about which pages are dangling nodes, or which is the same, which are the zero rows of H. See (2.18). For small n, the entries of a could found by hand; e.g., in Example 2.12, the vector a is simply $a = [0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0]^T$.

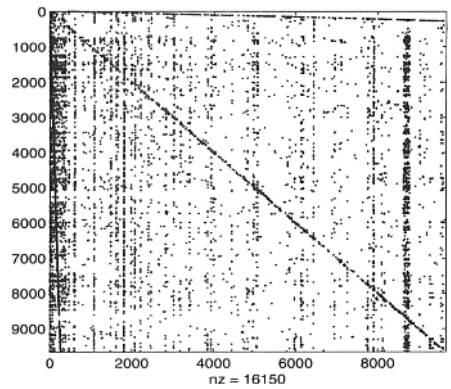


Figure 2.10 Matrix H for Example 2.17

In general, for a large sparse matrix H of order n, we can obtain a by using the following MATLAB commands:

$$\begin{aligned} a &= \text{sparse}\,(n, 1); \\ z &= \text{sum}\,(H, 2); \\ p &= \text{logical}\,(\sim z); \\ a(p) &= 1; \end{aligned}$$

This tells MATLAB to initially create a sparse vector a of zeros and then change the entry to 1 if the corresponding row sum of H is zero.

■ **EXAMPLE 2.17**

Datasets representing a given web are typically stored as an $n \times 3$ array. By instance, a web (or graph) generated after the query "California" can be stored as

$$\begin{array}{ccc} \vdots & \vdots & \vdots \\ 1 & 463 & 1/17 \\ 1 & 464 & 1/17 \\ 1 & 465 & 1/17 \\ 2 & 483 & 1 \\ 3 & 484 & 1/3 \\ 3 & 533 & 1/3 \\ 3 & 534 & 1/3 \\ 4 & 583 & 1/10 \\ \vdots & \vdots & \vdots \end{array}$$

(http://www.cs.cornell.edu/Courses/cs685/2002fa/data/gr0.California)

The first two columns represent the pages numbers, and the third is the weight of the link. In particular, page 1 has 17 outlinks, page 2 links only to page 483, page 3 has

exactly three outlinks: to pages 484, 533, and 583, and so on. There are $n = 9,664$ pages in total in this example.

Then, using MATLAB we can convert this dataset into an $n \times n$ link matrix H and visualize its sparsity pattern, where each blue dot represents a nonzero entry. See Figure 2.10.

- MATLAB commands: spconvert(H), spy(H)

Power method (2.30) with $\alpha = 0.83$ converges to the PageRank vector in 72 iterations, with an error 3.6502×10^{-9}. These are the ranks of pages P_1 to P_{10}:

$$
\begin{array}{l}
0.004054634831752 \\
0.001108812275000 \\
0.000100427746370 \\
0.001424756485048 \\
0.000106340487576 \\
0.000278621793647 \\
0.001664978614404 \\
0.000058109555930 \\
0.001373070568648 \\
0.002111426552892.
\end{array}
$$

Remark 2.18 *If in Example 2.17 we use $\alpha = 0.95$, then to achieve an error of about 10^{-9}, we need 261 iterations (more than three times the number of iterations needed with $\alpha = 0.83$). This illustrates how α determines the speed of convergence of the power method.*

2.3.3 Power method and reordering

As discussed before, the larger the gap between $|\lambda_1|$ and $|\lambda_2|$, the faster the power method will converge. Some theoretical results indicate that some reorderings speed up the convergence of numerical algorithms for the solution of linear systems. The question is whether some similar matrix reorderings can increase the separation between the dominant and the subdominant eigenvalue. For the particular case in mind, since $\lambda_1 = 1$, we would like to see whether some matrix reordering can take $|\lambda_2|$ farther from 1.

Extensive testing shows that by performing row and column permutations so as to bring most nonzero elements of the original matrix H say to the upper left corner, causes the gap between $|\lambda_1|$ and $|\lambda_2|$ to increase, and therefore, power method converges in fewer iterations.

- MATLAB commands: colperm(H), fliplr(H)

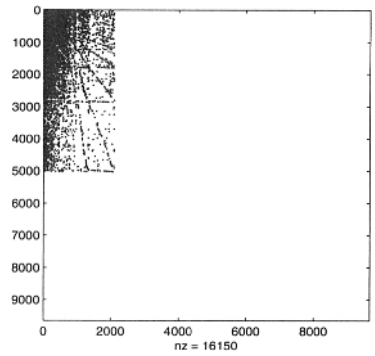


Figure 2.11 Reordered California matrix H.

When the power method with reordering is applied to the California dataset of Example 2.17 ($n = 9,664$), convergence is achieved in only 21 iterations, clearly better than the 72 iterations needed without permutation. As for clock time, including reordering, it is about 15% faster. Similarly, for a dataset Stanford, with $n = 281,903$, power iteration without reordering needs 66 iterations to converge, whereas with reordering only 19 iterations are needed. The total time for convergence, including reordering, is about 66% shorter than the power method without reordering.

In Figure 2.11, we show the reordered matrix H of Example 2.17. Compare this with Figure 2.10. If we want to visualize the original Stanford matrix we would mostly see a solid blue square, even though it is a sparse matrix. The reordered Stanford matrix is shown in Figure 2.12.

2.4 ALTERNATIVES TO THE POWER METHOD

The size of the link or adjacency matrix associated with the graph representing the entire World Wide Web has made the very simple power method and its modifications just about the only feasible method to compute the dominant eigenvector (the PageRank vector) of G. More sophisticated methods for computing eigenvectors turn out to be too expensive to implement. In recent years, however, researchers have been looking for alternative and hopefully faster methods to compute such PageRank vector. A first approach transforms the eigenvector problem into the problem of solving a linear system of equations. A second approach aims at reducing the size of the eigenvector problem by an appropriate blocking of the matrix G. We investigate these two different approaches and how they can be combined to build up more efficient algorithms.

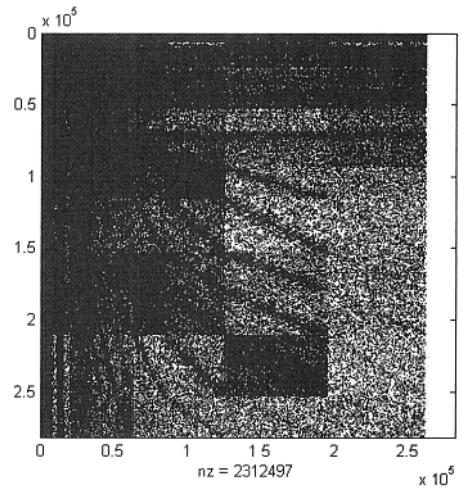


Figure 2.12 Reordered Stanford matrix H.

2.4.1 Linear system formulation

The motivation behind expressing the problem of ranking web pages as a linear system is the existence of a long list of well–studied methods that can be used to solve such systems. The central idea is to obtain a coefficient matrix that is very sparse; otherwise, the huge size of the matrix would make any linear system solver by far much slower than the power method itself.

We start by observing that for $0 < \alpha < 1$, the matrix $I - \alpha H^T$ is nonsingular. Indeed, the sum of each column of the matrix αH^T is strictly less than one and therefore $\|\alpha H^T\|_1 < 1$. Then, according to Exercise 1.22, the matrix $I - \alpha H^T$ must be nonsingular.

That the eigenvector problem to compute the PageRank vector is equivalent to the problem of solving a linear system is expressed in the following theorem.

Theorem 2.19 *Let $G = \alpha B + (1-\alpha)eu^T$, where B is as in (2.17) and $0 < \alpha < 1$. Then,*

$$v^T = v^T G \qquad \Longleftrightarrow \qquad \begin{cases} (I - \alpha H^T)\, x &= u \\ v = x/\|x\|_1 \,. \end{cases} \tag{2.32}$$

Proof. By substituting $G = \alpha(H + au^T) + (1-\alpha)eu^T$ into $v^T = v^T G$ and using the fact that $v^T e = 1$, we have that

$$v^T = v^T(\alpha H + \alpha a u^T) + (1-\alpha)u^T,$$

which can be written as $v^T(I - \alpha H - \alpha a u^T) = (1-\alpha)u^T$, or equivalently

$$(I - \alpha H^T - \alpha u a^T)\, v \;=\; (1-\alpha)\, u. \tag{2.33}$$

The coefficient matrix in the linear system (2.33) is a rank-1 modification of the matrix $I - \alpha H^T$, and therefore, we can apply the Sherman–Morrison formula (see Exercise 2.67) to get

$$(M - \alpha u\, a^T)^{-1} \;=\; M^{-1} + \frac{M^{-1} u\, a^T M^{-1}}{1/\alpha + a^T M^{-1} u},$$

where $M = I - \alpha H^T$. Thus, if x is the solution of $Mx = u$, then the solution of (2.33) can be computed explicitly as

$$\begin{aligned} v &= (1-\alpha)\left[M^{-1} + \tfrac{M^{-1} u\, a^T M^{-1}}{1/\alpha + a^T M^{-1} u}\right] u \\ &= (1-\alpha)\left[M^{-1} u + \tfrac{M^{-1} u\, a^T M^{-1} u}{1/\alpha + a^T M^{-1} u}\right] \\ &= (1-\alpha)\left[x + \tfrac{x\, a^T x}{1/\alpha + a^T x}\right] \\ &= (1-\alpha)\left[1 + \tfrac{a^T x}{1/\alpha + a^T x}\right] x. \end{aligned}$$

Thus, the PageRank vector v is just a multiple of x and we can use the constant $(1-\alpha)\left[1 + \frac{a^T x}{1/\alpha + a^T x}\right]$ to normalize v so that $\|v\|_1 = 1$.

□

■ **EXAMPLE 2.18**

Let us go back to the matrix H of Example 2.12. It was found later in Example 2.14 that the corresponding PageRank vector, that is the vector v that satisfies $v^T = v^T G$ was

$$v = [0.2537 \quad 0.3262 \quad 0.1948 \quad 0.0710 \quad 0.0900 \quad 0.0643]^T.$$

If instead we solve the system $(I - \alpha H^T)x = u$, with $u = e/n$ and e a vector of ones as usual, we get

$$x = [0.9328 \quad 1.1994 \quad 0.7161 \quad 0.2612 \quad 0.3308 \quad 0.2363]^T.$$

Then, after normalizing $x/\|x\|_1$ we get exactly the same PageRank vector v.

Remark 2.20 *The rescaling of the vector x to obtain v in Theorem 2.19 is not essential because the vector x itself already has the correct order of the web pages according to their ranks. Also, since in principle we only need to solve the linear system $(I - \alpha H^T)x = u$, the rank-1 matrix au^T (which was introduced to take care of the dangling nodes) does not play a direct role.*

Through Theorem 2.19, the eigenvector problem $v^T = v^T G$ has effectively been transformed into solving a linear system whose coefficient matrix $I - \alpha H^T$ is just about as sparse as the original matrix H. This gives us the opportunity to apply a variety of available numerical methods and algorithms for sparse linear systems and to look for efficient ways and strategies to solve such systems as a better alternative to finding a dominant eigenvector with the power method.

2.4.1.1 Reordering by dangling nodes The sparsity of the coefficient matrix in the linear system (2.32) is the main ingredient in the linear system formulation of the web page ranking problem. However, due to the enormous size of the web, we should look for additional ways to reduce the size of the problem or the computation time. One simple but very effective way to do this is by reordering the original matrix H so that zero rows (the ones that represent the dangling nodes) are moved to the bottom part of the matrix. This will allow us to reduce the size of the problem considerably. The more dangling nodes the web has, the more we can reduce the size of the problem through this approach.

By performing such reordering, the matrix H can be written as

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix}, \tag{2.34}$$

where H_{11} is a square matrix that represents the links from nondangling nodes to nondangling nodes and H_{12} represents the links from nondangling nodes to dangling nodes. In this case, we have

$$(I - \alpha H^T) = \begin{bmatrix} I - \alpha H_{11}^T & 0 \\ -\alpha H_{12}^T & I \end{bmatrix},$$

and therefore the system $(I - \alpha H^T)\, x = u$ is

$$\begin{bmatrix} I - \alpha H_{11}^T & 0 \\ -\alpha H_{12}^T & I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

which can be written as

$$\begin{aligned} (I - \alpha H_{11}^T) x_1 &= u_1 \\ -\alpha H_{12}^T x_1 + x_2 &= u_2 \end{aligned} . \tag{2.35}$$

Thus, we have an algorithm to compute the PageRank vector v:

1. Reorder H by dangling nodes.
2. Solve the system $(I - \alpha H_{11}^T) x_1 = u_1$.
3. Compute $x_2 = u_2 + \alpha H_{12}^T x_1$.
4. Normalize $v = x / \|x\|_1$.

Through this algorithm, the computation of the PageRank vector requires the solution of one sparse linear system $(I - \alpha H_{11}^T) x_1 = u_1$, which is much smaller than the original

one in (2.32), thus effectively reducing the size of the problem. The computational cost of steps 1, 3, and 4 of the algorithm is very small compared to that involved in the solution of the linear system itself.

■ **EXAMPLE 2.19**

Consider the matrix

$$H = \begin{bmatrix} 0 & 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 1/3 \end{bmatrix}.$$

When we reorder by dangling nodes, we get

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix} = \left[\begin{array}{ccc|ccc} 0 & 1/4 & 1/4 & 0 & 1/4 & 1/4 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 0 & 0 & 1/3 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]. \tag{2.36}$$

Let $u = [0.1 \quad 0.2 \quad 0.1 \mid 0.4 \quad 0 \quad 0.2]^T = [u_1 \quad u_2]^T$, and $\alpha = 0.8$. Following (2.35), the solution of the system $(I - \alpha H_{11}^T)x_1 = u_1$ is

$$x_1 = [0.1471 \quad 0.3824 \quad 0.1765]^T,$$

and $x_2 = u_2 + \alpha H_{12}^T x_1$ gives

$$x_2 = [0.4000 \quad 0.1824 \quad 0.2765]^T.$$

Then, by normalizing the vector $x = [x_1 \quad x_2]^T$, we get

$$v = \frac{x}{\|x\|_1} = [0.0940 \quad 0.2444 \quad 0.1128 \quad 0.2556 \quad 0.1165 \quad 0.1767]^T.$$

This vector v is the pagerank vector of the matrix G in Theorem 2.19, with H given in (2.36). In fact, that matrix G is given by

$$G = \begin{bmatrix} 0.0200 & 0.2400 & 0.2200 & 0.0800 & 0.2000 & 0.2400 \\ 0.0200 & 0.4400 & 0.0200 & 0.0800 & 0.4000 & 0.0400 \\ 0.2867 & 0.0400 & 0.2867 & 0.0800 & 0.0000 & 0.3067 \\ 0.1000 & 0.2000 & 0.1000 & 0.4000 & 0.0000 & 0.2000 \\ 0.1000 & 0.2000 & 0.1000 & 0.4000 & 0.0000 & 0.2000 \\ 0.1000 & 0.2000 & 0.1000 & 0.4000 & 0.0000 & 0.2000 \end{bmatrix}.$$

Its dominant eigenvector is

$$x = [0.2148 \quad 0.5585 \quad 0.2578 \quad 0.5843 \quad 0.2664 \quad 0.4039]^T,$$

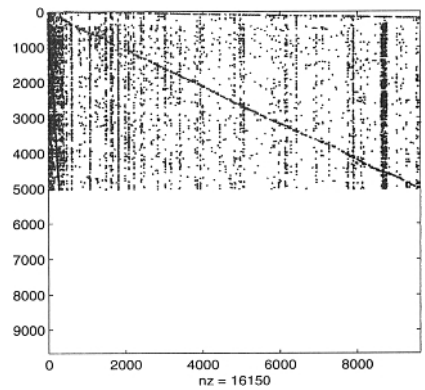


Figure 2.13 Reordered matrix H for Example 2.20.

which, after normalization in the 1-norm, gives again the vector v found above. Note that this vector x is different than the vector $x = [x_1 \quad x_2]^T$ above, but they are multiples of each other.

■ **EXAMPLE 2.20**

Let us consider the matrix $H_{n\times n}$ of Example 2.17, where $n = 9,664$, and let $u = e/n$, where e is a vector of ones. Instead of using the power method, we can solve the linear system (2.32), or first reorder such matrix H by dangling nodes (see Figure 2.13) and then apply (2.35). For the given example, since the number of dangling nodes of the original matrix H is $r = 4637$, the sparse block matrix H_{11} is square of order $n - r = 5027$. Once the matrix is reordered, the algorithm runs about 55% faster than the one without reordering.

2.4.2 Iterative aggregation/disaggregation (IAD)

Another attempt to outperform the power method in the computation of the PageRank vector comes from the theory of Markov chains. A **Markov chain** is a stochastic process describing a chain of events. Suppose we have a set of states $\mathcal{S} = \{s_1, \ldots, s_n\}$ and a process starts at one of these states and moves successively from one state to another. If the chain is currently at state s_i then it moves to state s_j in one step with probability p_{ij}, a probability that depends only on the current state. Thus, it is clear that such a chain can be represented by a stochastic matrix $G_{n\times n}$ (called the transition matrix), with entries p_{ij}.

If the transition matrix is irreducible, the corresponding Markov chain is also called irreducible. In such a case, a probabilistic vector v is called is called a **stationary distribution**

of the Markov chain if

$$v^T = v^T G.$$

This means that the PageRank vector (the dominant left eigenvector of G we are interested in) is also the stationary distribution of the Markov chain represented by the matrix G.

The main idea behind the IAD approach to compute the PageRank vector v is to block or partition the irreducible stochastic matrix G so that the size of the problem is reduced to about the size of one of the diagonal blocks. Thus, let

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \qquad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}. \tag{2.37}$$

It can be shown that $(I - G_{11})$ is nonsingular. Then, define

$$\begin{aligned} D &= \begin{bmatrix} I - G_{11} & 0 \\ 0 & I - S \end{bmatrix}, \qquad L = \begin{bmatrix} I & 0 \\ -G_{21}(I - G_{11})^{-1} & I \end{bmatrix}, \\ U &= \begin{bmatrix} I & -(I - G_{11})^{-1} G_{12} \\ 0 & I \end{bmatrix}, \end{aligned} \tag{2.38}$$

where

$$S = G_{22} + G_{21}(I - G_{11})^{-1} G_{12}. \tag{2.39}$$

S is called the **stochastic complement** of G_{22} in G, and it can be shown to be irreducible and stochastic if so is G. Then we have

$$I - G = LDU.$$

Thus,

$$v^T = v^T G \iff v^T(I - G) = 0 \iff v^T(LDU) = 0 \iff v^T LD = 0$$

because U is nonsingular. The last equation reduces to

$$[v_1^T \quad v_2^T] \begin{bmatrix} I - G_{11} & 0 \\ -G_{21} & I - S \end{bmatrix} = [0 \quad 0], \tag{2.40}$$

which in particular implies that

$$v_2^T = v_2^T S.$$

That is, v_2 is a stationary distribution of S. If u_2 is the unique stationary distribution of S with $\|u_2\|_1 = 1$, then we must have

$$v_2 = c u_2,$$

for some scalar c. It remains to find an expression for v_1.

Associated with G there is the so–called **aggregated matrix**, defined as

$$A = \begin{bmatrix} G_{11} & G_{12} e \\ u_2^T G_{21} & u_2^T G_{22} e \end{bmatrix}, \tag{2.41}$$

where e represents vectors of ones of appropriate size and u_2 is as defined above. First we want to find the stationary distribution of A.

Observe that from (2.40), we can get $v_1^T(I - G_{11}) - v_2^T G_{21} = 0$, which means

$$v_1^T = v_2^T G_{21}(I - G_{11})^{-1} \quad \text{and also} \quad v_1^T G_{11} = v_1^T - v_2^T G_{21}.$$

At the same time, the equation $v_2^T = v_2^T S$ implies that

$$v_2^T = v_2^T G_{22} + v_2^T G_{21}(I - G_{11})^{-1} G_{12}.$$

We then have

$$\begin{aligned}
[v_1^T \quad c]\,A &= [v_1^T \quad c]\begin{bmatrix} G_{11} & G_{12}e \\ u_2^T G_{21} & u_2^T G_{22}e \end{bmatrix} \\
&= [v_1^T G_{11} + c u_2^T G_{21} \qquad v_1^T G_{12} e + c u_2^T G_{22} e] \\
&= [(v_1^T - v_2^T G_{21}) + c u_2^T G_{21} \quad v_2^T G_{21}(I - G_{11})^{-1} G_{12} e + v_2^T G_{22} e] \\
&= [(v_1^T - v_2^T G_{21}) + v_2^T G_{21} \quad (\, v_2^T G_{22} + v_2^T G_{21}(I - G_{11})^{-1} G_{12}\,)\, e] \\
&= [v_1^T \quad v_2^T e] = [v_1^T \quad c u_2^T e] \\
&= [v_1^T \quad c].
\end{aligned}$$

We can now summarize the discussion above into the following theorem.

Theorem 2.21 (Exact aggregation/disaggregation)

If u_2 is the stationary distribution of $S = G_{22} + G_{21}(I - G_{11})^{-1} G_{12}$ and $\begin{bmatrix} v_1 \\ c \end{bmatrix}$ is the stationary distribution of $A = \begin{bmatrix} G_{11} & G_{12}e \\ u_2^T G_{21} & u_2^T G_{22} e \end{bmatrix}$, then the stationary distribution of

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$$

is given by $v = \begin{bmatrix} v_1 \\ c\,u_2 \end{bmatrix}$.

The computation of the matrices S and A and their corresponding stationary distributions is the aggregation step, while forming the vector $v = [v_1 \quad c u_2]^T$ is considered to be the disaggregation step.

■ **EXAMPLE 2.21**

Consider the stochastic irreducible matrix

$$G = \frac{1}{60}\left[\begin{array}{ccc|ccc} 15 & 10 & 10 & 5 & 10 & 10 \\ 5 & 5 & 15 & 15 & 5 & 15 \\ 8 & 8 & 8 & 6 & 10 & 20 \\ \hline 12 & 10 & 10 & 12 & 6 & 10 \\ 4 & 20 & 4 & 12 & 10 & 10 \\ 2 & 2 & 6 & 30 & 10 & 10 \end{array}\right] = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}.$$

The stationary distribution of $S = G_{22} + G_{21}(I - G_{11})^{-1}G_{12}$, with $\|u_2\|_1 = 1$ is

$$u_2 = \begin{bmatrix} 0.4112 \\ 0.2387 \\ 0.35 \end{bmatrix}.$$

The stationary distribution of $A = \begin{bmatrix} G_{11} & G_{12}e \\ u_2^T G_{21} & u_2^T G_{22} e \end{bmatrix}$, where the vector e is $e = [1 \quad 1 \quad 1]^T$, is given by

$$\begin{bmatrix} v_1 \\ c \end{bmatrix} := \left[\begin{array}{c} 0.1273 \\ 0.1457 \\ 0.1465 \\ \hline 0.5806 \end{array}\right], \quad \text{with} \quad \left\| \begin{bmatrix} v_1 \\ c \end{bmatrix} \right\|_1 = 1.$$

Therefore, the stationary distribution of G, with $\|v\|_1 = 1$ is

$$v = \begin{bmatrix} v_1 \\ c\,u_2 \end{bmatrix}^T = [0.1273 \quad 0.1457 \quad 0.1465 \quad 0.2388 \quad 0.1386 \quad 0.2032]^T.$$

That is, it satisfies $v^T = v^T G$, with $\|v\|_1 = 1$.

Through Theorem 2.21, the problem of computing the stationary distribution of G has been split into finding stationary distributions of two smaller matrices. Observe however that forming the matrix S and computing its stationary distribution u_2 would be very expensive and not efficient in practice. The idea is to avoid such computation and use instead an approximation.

More precisely, we define the **approximate aggregation matrix**:

$$\tilde{A} = \begin{bmatrix} G_{11} & G_{12}e \\ \tilde{u}_2^T G_{21} & \tilde{u}_2^T G_{22} e \end{bmatrix}, \tag{2.42}$$

where $\tilde{u}_2$ is an arbitrary probabilistic vector that plays the role of the exact stationary distribution u_2.

Such approximation is reasonable because the matrix $\tilde{A}$ differs from the original matrix A only in the last row, while typically, G_{11} is in the order of billions of rows for the PageRank

problem. It can be shown that $\tilde{A}$ is irreducible and stochastic if so is G; thus, we are guaranteed that $\tilde{A}$ has a unique stationary distribution.

In general, this approach does not give a very good approximation to the stationary distribution of the original matrix G, and to improve accuracy, a power method step is applied. More precisely:

To compute an approximate stationary distribution $\tilde{v} = [\tilde{v}_1 \quad \tilde{v}_2]^T$ of G, we apply the following **approximate aggregation/disaggregation** algorithm:

- Select an arbitrary probabilistic vector $\tilde{u}_2$.
- Compute the stationary distribution $\begin{bmatrix} w_1 \\ \tilde{c} \end{bmatrix}$ of $\tilde{A}$.
- Let $[\tilde{v}_1 \quad \tilde{v}_2]^T = [w_1 \quad \tilde{c}\,\tilde{u}_2]^T G$.

Typically, we will have $\tilde{v} \neq v$ so that the actual algorithm to be implemented consists of repeated applications of the algorithm above. That is, we start with an arbitrary $\tilde{u}_2$, compute the stationary distribution of $\tilde{A}$ and form $\tilde{v}$. For the next iteration, $\tilde{u}_2$ is taken as the second block of the current vector $\tilde{v}$. Then we repeat the process until convergence.

That is, we have the following **iterative aggregation/disaggregation** (IAD) algorithm:

1. Give an arbitrary probabilistic vector $\tilde{u}_2$ and a tolerance ϵ
2. For $k = 1, 2, \ldots$
   - Find the stationary distribution $\begin{bmatrix} w_1 \\ \tilde{c} \end{bmatrix}$ of $\tilde{A}$
   - Set $v_k^T = [w_1 \quad \tilde{c}\tilde{u}_2]^T$
   - Let $v_{k+1}^T = v_k^T G$
   - If $\|v_{k+1}^T - v_k^T\| < \epsilon$, stop. Otherwise, $\tilde{u}_2 = (v_{k+1})_y / \|(v_{k+1})_y\|_1$.

Here, $(v_{k+1})_y$ denotes the second block of the vector v_{k+1}.

2.4.2.1 IAD and power method Upon examining the IAD algorithm above, we see that we need to compute $\begin{bmatrix} w_1 \\ \tilde{c} \end{bmatrix}$, the stationary distribution of $\tilde{A}$. This vector satisfies

$$\begin{bmatrix} w_1 \\ \tilde{c} \end{bmatrix}^T = \begin{bmatrix} w_1 \\ \tilde{c} \end{bmatrix}^T \tilde{A}. \tag{2.43}$$

Thus, we can use the power method to solve for it. However, we are faced again with the same problem: the matrices G_{11}, G_{12}, G_{21}, and G_{22} that make up $\tilde{A}$ are full matrices and applying the power method directly would not be efficient.

The question is whether we can still apply a strategy similar to (2.30) and (2.31) so that we are able to exploit the sparsity of the original link matrix H.

We start by writing the matrix $\tilde{A}$ as (Exercise 2.80)

$$\tilde{A} = \begin{bmatrix} G_{11} & G_{12}e \\ \tilde{u}_2^T G_{21} & 1 - \tilde{u}_2^T G_{21} e \end{bmatrix},$$

so that we drop G_{22}. We can simplify further by setting $z^T = \tilde{u}_2^T G_{21}$, and thus $\tilde{A}$ becomes

$$\tilde{A} = \begin{bmatrix} G_{11} & G_{12}e \\ z^T & 1 - z^T e \end{bmatrix}.$$

From (2.43), we have

$$[w_1^T \quad \tilde{c}] = [w_1^T \quad \tilde{c}] \begin{bmatrix} G_{11} & G_{12}e \\ z^T & 1 - z^T e \end{bmatrix} = [w_1^T G_{11} + \tilde{c} z^T \qquad w_1^T G_{12} e + \tilde{c}(1 - z^T e)]. \tag{2.44}$$

To explicitly get some sparsity, first we write G in terms of H as before:

$$G = \alpha H + \alpha a u^T + (1 - \alpha) e u^T.$$

Then we block the matrices H, au^T and eu^T according to the blocking of G:

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} = \alpha \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} + \alpha \begin{bmatrix} A & B \\ C & D \end{bmatrix} + (1 - \alpha) \begin{bmatrix} E & F \\ J & K \end{bmatrix}, \tag{2.45}$$

for some matrices A, B, C, D, E, F, J, and K.

Now we see that

$$\begin{aligned} G_{11} &= \alpha H_{11} + \alpha a_x u_x^T + (1 - \alpha) e_x u_x^T, \\ G_{12} &= \alpha H_{12} + \alpha a_x u_y^T + (1 - \alpha) e_x u_y^T, \\ G_{21} &= \alpha H_{21} + \alpha a_y u_x^T + (1 - \alpha) e_y u_x^T, \end{aligned}$$

where the subindices x and y denote blocking of the vectors a, u, and e according to the blocking of G. Thus, e.g., a_x denotes the first block of entries of the vector a with size corresponding to the size of G_{11}.

Now we can use the sparsity of H_{11}, H_{12}, and H_{21} to make computations simpler and faster. The equations in (2.44) can now be written as

$$\begin{aligned} w_1^T &= \alpha w_1^T H_{11} + \alpha w_1^T a_x u_x^T + (1 - \alpha) w_1^T e_x u_x^T + \tilde{c} z^T, \\ \tilde{c} &= \alpha w_1^T H_{12} e + \alpha w_1^T a_x u_y^T e + (1 - \alpha) w_1^T e_x u_y^T e + \tilde{c}(1 - z^T e). \end{aligned} \tag{2.46}$$

For the iterative process of the power method within IAD, we give as usual an arbitrary initial guess $[w_1^T \quad \tilde{c}]$ and iterate according to (2.46) for the next approximation $[w_1^T \quad \tilde{c}]$ until certain tolerance is reached.

Exploiting dangling nodes. Although not obvious in (2.46), it is possible to exploit not only the sparsity of the matrices H_{11}, H_{12}, and H_{21} but also some reorderings applied to H. Fy instance, if H is reordered by dangling nodes, bringing all the zero rows to the bottom of the matrix, then the matrix H_{21}, which is used in $z = \tilde{u}_2^T G_{21}$ is a matrix of zeros. Also, by definition of the vector a (see (2.18)), we will have all the entries of a_x to be zeros. All this reduces power method computation in (2.46) to

$$\begin{aligned} w_1^T &= \alpha w_1^T H_{11} + (1-\alpha) w_1^T e_x u_x^T + \tilde{c} z^T, \\ \tilde{c} &= \alpha w_1^T H_{12} e + (1-\alpha) w_1^T e_x u_y^T e + \tilde{c}(1 - z^T e). \end{aligned} \tag{2.47}$$

Some other reorderings can even transform H_{12} into a zero matrix reducing computations even more.

2.4.3 IAD and linear systems

The two alternatives to the original power method we have presented above are a linear systems approach and IAD, both offering some clear advantages over the original algorithm. We want to illustrate how both approaches can be combined to provide an improved algorithm for the computation of the pagerank vector.

Recall that within the IAD algorithm we need to compute the stationary distribution of the approximate aggregation matrix $\tilde{A}$. This has been done by applying the power method to

$$[w_1^T \quad \tilde{c}] = [w_1^T \quad \tilde{c}]\tilde{A}.$$

Now we want to rewrite this equation so that we can arrive at the solution by solving some linear system. We have

$$[w_1^T \quad \tilde{c}] = [w_1^T \quad \tilde{c}] \begin{bmatrix} G_{11} & G_{12}e \\ \tilde{u}_2^T G_{21} & 1 - \tilde{u}_2^T G_{21} e \end{bmatrix} \tag{2.48}$$

After multiplication, we get

$$[w_1^T \quad \tilde{c}] = [w_1^T G_{11} + \tilde{c}\tilde{u}_2 G_{21} \qquad w_1^T G_{12} e + \tilde{c} - \tilde{c}\tilde{u}_2^T G_{21} e],$$

which can be written as

$$\begin{cases} w_1^T (I - G_{11}) &= \tilde{c}\tilde{u}_2^T G_{21} \\ w_1^T G_{12} e &= \tilde{c}\tilde{u}_2^T G_{21} e \end{cases} \tag{2.49}$$

Observe that the first equation in (2.49) is a linear system of the order of G_{11} and the second one is just a scalar equation. However, such a system with coefficient matrix $I - G_{11}$ cannot be solved because the right–hand side contains $\tilde{c}$, which is unknown.

Thus, the idea is initially to fix an arbitrary $\tilde{c}$ so that the system $w_1^T (I - G_{11}) = \tilde{c}\tilde{u}_2^T G_{21}$ can be solved, and once w_1 is computed, we can adjust $\tilde{c}$ using the second equation in (2.49). This approach is reasonable because, as remarked before, for the application in mind, G_{11} has billions of rows while $\tilde{c}$ is just a scalar.

More precisely, to compute the stationary distribution $[w_1 \quad \tilde{c}]^T$ of $\tilde{A}$, we apply the following algorithm.

1. Give an initial guess $\tilde{c}$ and a tolerance ϵ.
2. Repeat until $\|w_1 - w_1(\text{old})\| < \epsilon$:
   - Solve $(I - G_{11})^T w_1 = \tilde{c}\, G_{21}^T \tilde{u}_2$.
   - Adjust $\tilde{c} = \frac{w_1^T G_{12} e}{\tilde{u}_2^T G_{21} e}$.

Typically it takes just a couple of steps to reach a given tolerance ϵ so that this computation is relatively fast in terms of number of iterations. However, the matrices G_{11}, G_{12}, and G_{21} are full matrices so that computations at each step are in general very expensive. Thus, similarly to what it was done when studying IAD with the power method, we go back to the original matrix H to get some sparsity.

From equation (2.45), we can look at G_{11} in more depth to get

$$\begin{aligned} I - G_{11} &= I - \alpha H_{11} - \alpha A - (1-\alpha)E \\ &= I - \alpha H_{11} - \alpha a_x u_x^T - (1-\alpha) e_x u_x^T. \end{aligned} \tag{2.50}$$

The last term, $(1-\alpha)e_x u_x^T$ can disturb the sparsity of $I - G_{11}$, but we will use the fact that $w_1^T e_x = 1 - \tilde{c}$ to simplify. We know this to be true because

$$1 = [w_1^T \quad \tilde{c}]\, e = [w_1^T \quad \tilde{c}] \begin{bmatrix} e_x \\ 1 \end{bmatrix} = w_1^T e_x + \tilde{c}.$$

We also know that

$$G_{21} = \alpha H_{21} + \alpha a_y u_x^T + (1-\alpha) e_y u_x^T,$$

which can be used to keep the system sparse. With this knowledge, our linear system, $w_1^T(I - G_{11}) = c\tilde{u}_2^T G_{21}$, becomes

$$w_1^T(I - \alpha H_{11} - \alpha a_x u_x^T) = \tilde{c}\tilde{u}_2^T(\alpha H_{21} + \alpha a_y u_x^T + (1-\alpha) e_y u_x^T) + (1-\alpha)(1-\tilde{c}) u_x^T. \tag{2.51}$$

Thus, to compute w_1, we solve the linear system

$$(I - \alpha H_{11} - \alpha a_x u_x^T)^T w_1 = \tilde{c}(\alpha H_{21} + \alpha a_y u_x^T + (1-\alpha) e_y u_x^T)^T \tilde{u}_2 + (1-\alpha)(1-\tilde{c}) u_x. \tag{2.52}$$

Dangling nodes. Similarly as before, some reorderings of the matrix H can mean a reduction of the computational cost in the solution of the linear system. Thus, e.g., assume that H has been reordered by dangling nodes, i.e, the zero rows have been moved to the bottom part. In such a case, we have

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix},$$

so that H_{21} is a matrix of zeros. Then we have the following theorem.

Theorem 2.22 *Let the matrix H be reordered by dangling nodes. Then, the system (2.52) reduces to*

$$(I - \alpha H_{11})^T w_1 = (1 - \alpha(1 - \tilde{c}))\, u_x. \tag{2.53}$$

Proof. *Due to the reordering by dangling nodes, the vector a_x is a vector of zeros and the vector a_y is a vector of ones, of the same size as e_y, so that $a_y = e_y$. Since $\tilde{u}_2$ is a probabilistic vector, we also have $e_y^T \tilde{u}_2 = 1$. Then, we can simplify equation (2.52) to*

$$\begin{aligned}
(I - \alpha H_{11})^T w_1 &= \tilde{c}(\alpha a_y u_x^T + (1-\alpha) e_y u_x^T)^T \tilde{u}_2 + (1-\alpha)(1-\tilde{c}) u_x \\
&= \tilde{c}(e_y u_x^T)^T \tilde{u}_2 + (1-\alpha)(1-\tilde{c}) u_x \\
&= \tilde{c}\, u_x e_y^T \tilde{u}_2 + (1-\alpha)(1-\tilde{c}) u_x \\
&= \tilde{c}\, u_x + (1-\alpha)(1-\tilde{c}) u_x \\
&= (\tilde{c} + (1-\alpha)(1-\tilde{c}))\, u_x \\
&= (1 - \alpha(1-\tilde{c}))\, u_x.
\end{aligned} \tag{2.54}$$

□

Simply by reordering the matrix, in this case by dangling nodes, a computation that could have been time consuming has now been effectively simplified.

Following a similar reasoning, one can prove that the computation of the scalar $\tilde{c}$ can also be reduced in the case of a link matrix H that has been reordered by dangling nodes. More precisely, since

$$\tilde{c} = \frac{w_1^T G_{12} e}{\tilde{u}_2^T G_{21} e},$$

and $G_{12} = \alpha H_{12} + \alpha a_x u_y^T + (1-\alpha) e_x u_y^T$, calculations similar to those in Theorem 2.22 reveal that we can compute $\tilde{c}$ as

$$\tilde{c} = \frac{\alpha\, w_1^T H_{12} e + (1-\alpha)(1-\tilde{c}) u_y^T e}{u_x^T e}. \tag{2.55}$$

Thus, our algorithm that combines the ideas from IAD and linear systems, with the original link matrix H arranged by dangling nodes, can be expressed in a simplified form as

1. Give an initial guess $\tilde{c}$ and a tolerance ϵ.
2. Repeat until $\|w_1 - w_1(\text{old})\| < \epsilon$:

- Solve $(I - \alpha H_{11})^T w_1 = (1 - \alpha(1 - \tilde{c}))\, u_x$.

- Adjust $\tilde{c} = \frac{\alpha\, w_1^T H_{12} e + (1-\alpha)(1-\tilde{c}) u_y^T e}{u_x^T e}$.

Finally, we remark again that some particular reorderings of H can additionally give $H_{12} = 0$, simplifying computations even more.

2.5 FINAL REMARKS AND FURTHER READING

The problem of ranking web pages is a real-world application of some concepts and techniques from matrix algebra and numerical analysis. We have presented the minimum theoretical background necessary to build well-defined algorithms, and we have presented some examples with matrices of relatively small size. The starting and most basic algorithm is the power method, but recent studies have shown that a linear system approach and aggregation methods can also be applied and combined to achieve faster convergence. In all cases, permutations or reordering of the link matrices are important to develop more efficient algorithms.

The fundamental background needed is a set of results on nonnegative matrices. The book by Berman and Plemmons [4] offers a sound analysis of the most important properties of stochastic, irreducible, primitive, as well as M-matrices. Langville and Meyer [44] have studied the PageRank algorithm in detail and have reported their findings in several works including [44], [45], and [50]. Ipsen and Kirkland [39] provide with details on the convergence of an IAD algorithm developed by Langville and Meyer. A detailed report on the linear system approach for finding the PageRank vector can also be found in the work by del Corso et al. [16]. Osborne et al. [53] first combine power method with matrix reorderings and then combine the IAD approach with both, power method and linear systems, to obtain algorithms with fast convergence to the PageRank vector. As for other approaches toward improving the computation of webpage ranking, the work by Berkhin [3] proposes a bookmark coloring algorithm, with a predefined threshold value that induces sparsity and faster speed of convergence. In Andersen and Peres [2], an algorithm for PageRank computation is improved and further applied to develop PageRank-Nibble, a very efficient algorithm for local graph partitioning.

Research in this area of matrix computations applied to search engines is very active, and new and novel results are expected in the near future. More and more efficients algorithms are needed as the size of the World Wide Web increases by the millions every day. Thus, for realistic situations, the size of the matrices involved make the use of supercomputers and parallelism a must.

EXERCISES

2.1 Consider the matrix

$$A = \begin{bmatrix} 8 & -2 & 3 & 5 & 8 \\ 2 & 4 & 2 & -1 & 4 \\ 0 & 0 & 5 & 3 & 0 \\ 0 & 0 & -3 & 5 & 0 \\ 0 & 0 & 1 & -2 & 7 \end{bmatrix}.$$

Do the eigenvalues of A satisfy the condition (2.5)? If so, apply the power method to find the dominant eigenvector. What is the rate of convergence?

2.2 Find a matrix $A_{n\times n}$ and give an initial guess $x_0 \in \mathbb{R}^n$ so that the power method $x_{k+1} = Ax_k$ does not converge.

2.3 Rewrite the last equation in (2.3) as

$$A^k x_0 = \lambda_1^k \left[c_1 v_1 + c_2 \left(\frac{\lambda_2}{\lambda_1}\right)^k v_2 + \cdots + c_n \left(\frac{\lambda_2}{\lambda_1}\right)^k v_n \right],$$

where we assume again that $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Define the sequence $u_k = A^k x_0/\lambda_1^k$, $k = 0, 1, 2, \ldots$, and show that

$$\|u_k - c_1 v_1\| \leq C|\lambda_2/\lambda_1|^k,$$

for some constant C and for any vector norm $\|\cdot\|$. Thus, $u_k \to c_1 v_1$, as $k \to \infty$.

2.4 Assume a matrix $A_{n\times n}$ is nonsingular. Explain how you would modify the power method algorithm to compute an eigenvector associated with the smallest eigenvalue. What conditions must the eigenvalues of A satisfy?

2.5 Let u and v be nonorthogonal vectors. The quantity

$$\frac{v^T A u}{v^T u} \tag{2.56}$$

is called the generalized **Rayleigh quotient**. If u is an eigenvector of A, then the Rayleigh quotient gives the corresponding eigenvalue. Now consider a nonzero vector v and define the so–called residual $r(\mu) = Av - \mu v$. Show that $\|r(\mu)\|_2$ is minimized when

$$\mu = \frac{v^T A v}{v^T v}. \tag{2.57}$$

This μ is commonly known as the Rayleigh quotient.

2.6 Let x be a unit eigenvector of $A_{n\times n}$ with associated eigenvalue λ. Consider the Rayleigh quotient μ in (2.57), with $\|v\|_2 = 1$. Show that

$$|\lambda - \mu| \leq C\|x - v\|_2,$$

where $C = 2\|A\|_2$.

2.7 Let v be an eigenvector of a matrix A, with corresponding simple eigenvalue λ. Consider the iteration

$$\mu_{k+1} = v_k^T A v_k, \qquad (A - \mu_{k+1} I) v_{k+1} = v_k,$$

with $\|v_k\| = 1$, for all k, and assume that $v_k \to v$ as $k \to \infty$. Show that

$$\|v - v_{k+1}\| \leq C \|v - v_k\|^2,$$

for some constant C.

Hint. Use Exercise 2.6 and observe that the iteration above is the power method applied to $(A - \mu_{k+1} I)^{-1}$.

2.8 Consider the power iteration $x_{k+1} = Ax_k$, and suppose that for a given positive integer k, we have $x_{k-2} = v_1 + \alpha_2 v_2$, where v_1 and v_2 are the dominant and subdominant eigenvectors of $A_{n\times n}$, respectively. For $i = 1, \ldots, n$, define the entries of vectors b, c, and d respectively, by

$$\begin{aligned} b_i &= ((x_{k-1})_i - (x_{k-2})_i)^2, \\ c_i &= (x_k)_i - 2(x_{k-1})_i + (x_{k-2})_i, \\ d_i &= b_i / c_i. \end{aligned}$$

Show that $v_1 = x_{k-2} - d$.

2.9 Show that a matrix $A_{n\times n}$ is stochastic if and only if

(a) $Ax \geq 0$, for all vector $x \geq 0$, and

(b) $Ae = e$, where e is a vector of ones.

2.10 Show that a matrix $A_{n\times n}$ is stochastic if and only if the vector $w = u^T A$ is a probabilistic vector, for every probabilistic vector u.

2.11 Show that the convex combination of two stochastic matrices is a stochastic matrix. That is, if A and B are stochastic, then the matrix

$$C = \alpha A + (1 - \alpha) B, \qquad \text{with} \quad 0 < \alpha < 1$$

is also stochastic.

2.12 Show that if P is a stochastic matrix, then the matrix P^n is also stochastic for all integers $n > 0$.

2.13 Let $Q_{n\times n}$ be an orthogonal matrix. Show that the matrix A defined by $a_{ij} = q_{ij}^2$ is column and row stochastic.

2.14 True or False? Every stochastic matrix A is nonsingular.

2.15 Let $A_{n\times n}$ be an arbitrary nonsingular nonnegative matrix, and let $\tilde{A} = \frac{1}{\rho(A)} A$. Show that there exists a diagonal matrix D such that

$$B = D^{-1} \tilde{A} D$$

is stochastic.

2.16 Let $A_{n\times n}$ be a stochastic matrix and $\lambda \neq 1$ an eigenvalue of A. Show that if v is a left eigenvector of A, then it is also a left eigenvector of the matrix $B = (a_{ij} - c_j)$, where $c_j = \min_i a_{ij}$.

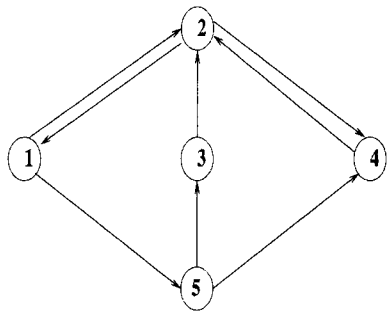


Figure 2.14 Graph of Exercise 2.22

2.17 Let $A_{n\times n}$ be a nonnegative matrix. Show that there exists a path of length k in $\Gamma(A)$ from a node P_i to a node P_j if and only if $(A^k)_{ij} \neq 0$.

Hint. Use induction, starting with $k = 2$.

2.18 An alternative way of defining the spectral radius of a square matrix A is

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}. \tag{2.58}$$

Use this definition to prove the following: If A and B are nonnegative matrices, then

$$A \geq B \quad \Longrightarrow \quad \rho(A) \geq \rho(B).$$

2.19 Let $A_{n\times n}$ be a nonnegative matrix. Show that

$$\rho(A) \;\geq\; \min_{1\leq i\leq n} \sum_{j=1}^{n} a_{ij}.$$

Hint: Construct a nonnegative matrix B that satisfies $A \geq B$ and whose row sums are all equal to $r = \min\limits_{1\leq i\leq n} \sum_{j=1}^{n} a_{ij}$.

2.20 Let $A_{n\times n}$ be a nonnegative matrix and $x \in \mathbb{R}^n$ a positive vector. Show that if $Ax > rx$ for some scalar $r > 0$, then $\rho(A) > r$.

2.21 Let $A_{n\times n}$ be a nonnegative matrix. Show that $\rho(A)$ is a positive eigenvalue of A, and that there is a positive eigenvector associated with it.

2.22 Is the link matrix associated with the graph in Figure 2.14 irreducible?

2.23 Show that the matrix

$$A = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1 \\ 1/4 & 1/4 & 1/4 & 0 & 1/4 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

is reducible by finding a permutation matrix P so that

$$B := P^T A P = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix},$$

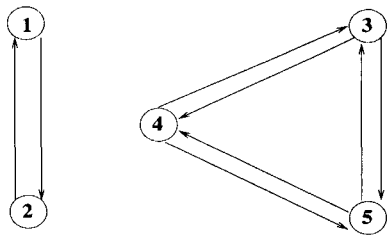


Figure 2.15 Graph of Exercise 2.28

for some matrices B, C, and D. Also draw the graphs $\Gamma(A)$ and $\Gamma(B)$.

2.24 Show that if A and B are nonnegative matrices and A is irreducible, then $A + B$ is irreducible.

2.25 Show that if an undirected graph modeling a web consists of k disconnected subgraphs, then the dimension of the dominant eigenspace (the one spanned by the dominant eigenvector) is at least k.

2.26 Let B be a stochastic positive matrix. Show that any eigenvector v in the dominant eigenspace has either all positive or all negative entries.

2.27 Let B be a stochastic positive matrix. Without using the Perron–Frobenius Theorem 2.10, show that the dominant eigenspace has dimension one.

2.28 Consider the graph of Figure 2.15. Find the corresponding link matrix and the dimension of the dominant eigenspace.

2.29 Let $A_{n\times n}$ be a positive column stochastic matrix, and let S be the subspace of $\mathbb{R}^n$ consisting of vectors v such that $\sum_{i=1}^{n} v_i = 0$. Show that S is invariant under A, that is, $AS \subset S$.

2.30 Find a nonnegative matrix $A_{n\times n}$, with at least one but not all entries equal zero, so that $\rho(A)$ is a zero eigenvalue with algebraic multiplicity greater than one.

2.31 Is the matrix $A = \begin{bmatrix} 0 & 5 & 1 & 6 \\ 0 & 0 & 4 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$ irreducible?

2.32 Let A be a reducible matrix; that is, $P^T AP = \begin{bmatrix} C & D \\ 0 & E \end{bmatrix}$, for some permutation matrix P. Show that solving a system $Ax = b$ can be reduced to solving the systems

$$E\tilde{x}_2 = \tilde{b}_2, \qquad C\tilde{x}_1 = \tilde{b}_1 - D\tilde{x}_2,$$

where $\tilde{b} = P^T b$, so that $x = P\tilde{x}$.

2.33 True or False? An irreducible matrix may have a row of zeros.

2.34 True or False? An irreducible matrix has at least one nonzero off-diagonal element in each row and column.

2.35 True or False? If a matrix A is reducible, then A^k is reducible for each $k \geq 2$.

2.36 Consider the matrix

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

Verify that its spectral radius $\rho(A)$ is an eigenvalue of algebraic and geometric multiplicity 1, yet the matrix A is reducible. Does this contradict Theorem 2.10?

2.37 Suppose that all row sums of a nonnegative matrix $A_{n\times n}$ are equal to a constant k. Show that $\rho(A) = k$.

2.38 Let $A_{n\times n}$ be a nonnegative matrix with row sums less than or equal to 1 (with at least one row sum strictly less than 1). Show that $\rho(A) \leq 1$.

2.39 Find a stochastic irreducible matrix A that does not have a unique eigenvalue of largest magnitude.

2.40 Verify and apply the Perron–Frobenius theorem to the matrix

$$A = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0 & 0 & 1 \\ 0.3 & 0.7 & 0 \end{bmatrix}.$$

2.41 Let $A_{n\times n}$ be a primitive stochastic matrix. Show that the rows of the matrix

$$S = \lim_{k\to\infty} A^k$$

are all identical.

Hint: First show that every row of S is a left eigenvector of A with eigenvalue $\lambda = 1$.

2.42 Consider the matrix

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

What is $\lim_{k\to\infty} A^k$?

2.43 Show that if A and B are nonnegative matrices and A is primitive, then $A + B$ is primitive.

2.44 True or False? The product of two primitive matrices is also a primitive matrix.

2.45 Let $A_{n\times n}$ be an irreducible stochastic matrix. Show that rank $(I - A) = n - 1$.

2.46 Give a slightly different proof of Theorem 2.16 by constructing a nonsingular matrix $U = [q \quad X]$, where q is an eigenvector of U.

2.47 Let $B_{n\times n}$ be a reducible matrix. Show that B is similar to a block upper triangular matrix. More precisely, show that there exists a permutation matrix P such that

$$P^T BP = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1k} \\ 0 & A_{22} & \cdots & A_{2k} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & A_{kk} \end{bmatrix},$$

where each A_{kk} is irreducible or a zero scalar.

2.48 Let $A_{n\times n}$ be an irreducible and weakly diagonal dominant matrix. Show that all its diagonal entries are nonzero.

2.49 The following matrix H is neither stochastic nor irreducible. Follow the equations (2.17) and (2.20) with $\alpha = 0.8$, $u = e/5$ and e a vector of ones to form a stochastic and irreducible matrix G.

$$H = \begin{bmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{bmatrix}.$$

2.50 The following matrix is stochastic but reducible. Apply equation (2.20) with $\alpha = 0.8$, $u = e/5$, where e is a vector of ones, to make it irreducible, and then apply the power method to find the PageRank vector, that is, the dominant eigenvector v with $\|v\|_1 = 1$.

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}.$$

What is the resulting order of the corresponding web pages in terms of importance?

2.51 Make the matrix B of Exercise 2.50 irreducible, with $\alpha = 0.8$, but now using the personalization vector

$$u = [0 \quad 0 \quad 0.1 \quad 0.9 \quad 0]^T.$$

What is the new order of importance of the web pages?

2.52 Given a stochastic matrix H representing a web, how would you quickly find a rough estimate of the order of the corresponding web pages in terms of importance?

2.53 Let $G_{n\times n} = \alpha B + (1-\alpha)eu^T$, where $u = e/n$ and e is a vector of ones, and let $\mu = \frac{1-\alpha}{\alpha-\lambda}$. Show that

$$\det(G - \lambda I) = \det(\alpha B - \lambda I)\det\left(I + \mu\, eu^T\right),$$

and that μ is an eigenvalue of $I + \mu\, eu^T$.

2.54 Find the eigenvalues of the matrices B and $G = \alpha B + (1-\alpha)eu^T$, where

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix},$$

$\alpha = 0.8$, $u = [0.3 \quad 0.1 \quad 0.2 \quad 0.4 \quad 0]^T$, e is a vector of ones, and verify Theorem 2.16.

2.55 Consider the matrix

$$H = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Let e be a vector of ones and $u = e/5$. Solve the system $(I - \alpha H^T)x = u$ with $\alpha = 0.85$ and then normalize $v = x/\|x\|_1$. Verify the statement in Theorem 2.19 that this vector v is the PageRank vector of the matrix $G = \alpha B + (1-\alpha)eu^T$, where B is as in (2.17).

2.56 Give a different proof of Theorem 2.19 by first proving that

$$(1-\alpha)x^T e + \alpha x^T a = 1,$$

and by observing that solving $v^T = v^T G$ is equivalent to solving $x^T(I - G) = 0$, where the PageRank vector v satisfies $v = x/\|x\|_1$.

2.57 Show that the system $(I - \alpha H^T)x = u$ in (2.32) has a unique solution.

2.58 We say a square matrix A is an M-matrix if and only if there exists a nonnegative matrix B and a real number k with $\rho(B) \le k$ such that $A = kI - B$. Show that the matrix $I - \alpha H$ is an M-matrix, where H is the usual link matrix and $0 < \alpha < 1$.

2.59 Let the matrix $G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$ be stochastic and irreducible. Show that the corresponding approximate aggregation matrix

$$\tilde{A} = \begin{bmatrix} G_{11} & G_{12}e \\ \tilde{u}_2^T G_{21} & \tilde{u}_2^T G_{22} e \end{bmatrix}$$

is also stochastic and irreducible.

2.60 Let $G_{n\times n}$ be an irreducible stochastic matrix. Show that $I - G$ is a singular M-matrix.

2.61 Let $A_{n\times n}$ be a singular irreducible M-matrix. First show that $\text{rank}(A) = n - 1$ and then show that there exists a positive vector x such that $Ax = 0$.

2.62 Let $A_{n\times n}$ be a matrix whose off-diagonal entries are nonpositive. Show that A is an M-matrix if and only if $A + \epsilon I$ is a nonsingular M-matrix for all $\epsilon > 0$.

2.63 Show that if the link matrix H is column stochastic, then the PageRank vector is $v = \lim_{k\to\infty} v_k$, with

$$v_{k+1} = G\, v_k,$$

where $G = \alpha(H + u\,a^T) + (1-\alpha)u\,e^T$, e is a vector of ones, u is an arbitrary probabilistic vector, and $0 < \alpha < 1$. How do you define the entries of H?

2.64 Let $u = e/7$, where $e \in \mathbb{R}^7$ is a vector of ones. Reorder by dangling nodes the following matrix:

$$H = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/3 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 & 0 \\ 0 & 0 & 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{bmatrix}$$

so that it takes the block form

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix}.$$

Then solve the equations (2.35) to find the vector x that solves $(I - \alpha H^T)x = u$. Finally, normalize $v = x/\|x\|_1$. Verify that this vector is the PageRank vector of the matrix G as given in Theorem 2.19.

2.65 Consider the original matrix H of Example 2.19. Compute the PageRank vector by directly applying (2.32), that is, without reordering by dangling nodes, and compare your answer with the one in the Example.

2.66 Suppose the link matrix H is column stochastic. Show that

$$v = G\,v \quad \Longleftrightarrow \quad \begin{cases} (I - \alpha H)\,x &= u \\ v = x/\|x\|_1, \end{cases}$$

where G is as in Exercise 2.63.

2.67 Let $A_{n\times n}$ be nonsingular, and let $x, y \in \mathbb{R}^n$ be two vectors such that $1 + y^T A^{-1} x \neq 0$. Show that $A + xy^T$ is nonsingular and that

$$(A + xy^T)^{-1} = A^{-1} - \frac{A^{-1}xy^TA^{-1}}{1 + y^TA^{-1}x}.$$

This is known as the **Sherman–Morrison formula**.

Hint: First show that $(I + uv^T)^{-1} = I - \frac{uv^T}{1+v^Tu}$, for any two nonzero vectors u, v such that $1 + v^Tu \neq 0$.

2.68 Suppose the link matrix H is column stochastic so that after reordering by dangling nodes, it can be expressed as

$$H = \begin{bmatrix} H_{11} & 0 \\ H_{21} & 0 \end{bmatrix}.$$

Show that the problem of solving the system $(I - \alpha H)x = u$ can be reduced in size in a similar way to (2.35).

2.69 Let $u = e/7$, where $e \in \mathbb{R}^7$ is a vector of ones. Reorder the following column stochastic matrix by dangling nodes

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 \\ 1/3 & 0 & 0 & 1/4 & 0 & 1/4 & 0 \\ 1/3 & 0 & 0 & 1/4 & 0 & 1/4 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 1/4 & 0 & 1/4 & 0 \end{bmatrix}$$

so that it takes the block form

$$H = \begin{bmatrix} H_{11} & 0 \\ H_{21} & 0 \end{bmatrix}.$$

Then solve the equations you found in Exercise 2.68 to find the vector x that solves $(I - \alpha H)x = u$. Finally, normalize $v = x/\|x\|_1$. Verify that this vector is the PageRank vector of the corresponding matrix G.

2.70 For the matrix G in Exercise 2.77, let S be the stochastic complement of G_{22}. Find the unique stationary distribution u_2 of S such that $\|u_2\|_1 = 1$.

2.71 Consider the matrix G in Exercise 2.77. Find the corresponding aggregated matrix A and its probabilistic stationary distribution, and then find the stationary distribution v of G with $\|v\|_1 = 1$.

2.72 Let $G_{n\times n} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$ be a row stochastic irreducible matrix with G_{11} a square block of order k. Let as usual e_i $(i = 1, \ldots, n)$ represent the i–th canonical vector. Show that

$$(e_{k+i} - e_{k+j})^T G = [\,(e_i - e_j)^T G_{21} \quad (e_i - e_j)^T G_{22}\,],$$

where $1 \leq i, j \leq n - k$.

2.73 Let G be a column stochastic irreducible matrix written as

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}. \tag{2.59}$$

Let S be defined as $S = G_{22} + G_{21}(I - G_{11})^{-1}G_{12}$. If u_2 satisfies $Su_2 = u_2$ and v_2 is the unique dominant eigenvector of S with $\|v_2\|_1 = 1$, that is, $v_2 = cu_2$, for some scalar c, show that the dominant eigenvector of G is given by $v = [v_1 \quad v_2]^T$, where

$$A \begin{bmatrix} v_1 \\ c \end{bmatrix} = \begin{bmatrix} v_1 \\ c \end{bmatrix},$$

for some matrix A (this is the corresponding aggregated matrix).

2.74 Let G be a column stochastic irreducible matrix written as in (2.59). Show that the equations (2.49) are now of the form

$$\begin{cases} (I - G_{11})w_1 &= \tilde{c}G_{12}\tilde{u}_2, \\ e^T G_{21} w_1 &= \tilde{c}(1 - e^T G_{22}\tilde{u}_2). \end{cases}$$

2.75 Let

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$$

be an irreducible stochastic matrix. Show that the matrix $I - G$ can be factored as $I - G = LDU$, where L, D, and U are given in (2.38).

2.76 Show that the stationary distribution v of the matrix G in the previous exercise is given by

$$v^T = c\,[x \quad u_2^T] \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} L^{-1},$$

where as usual u_2 is the unique stationary distribution of the stochastic complement S of G_{22}, $v_2 = cu_2$, L is given in (2.38), and x is an arbitrary vector of appropriate size.

2.77 Consider the stochastic irreducible matrix

$$G = \frac{1}{60} \left[\begin{array}{ccc|ccc} 10 & 10 & 15 & 10 & 10 & 5 \\ 5 & 10 & 10 & 15 & 5 & 15 \\ 8 & 8 & 8 & 10 & 6 & 20 \\ \hline 10 & 12 & 10 & 12 & 8 & 8 \\ 12 & 20 & 4 & 4 & 10 & 10 \\ 2 & 2 & 6 & 30 & 10 & 10 \end{array}\right] = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}.$$

Compute the factorization

$$I - G = LDU,$$

where L, D, U are given in (2.38).

2.78 Suppose that in the approximate aggregation disaggregation algorithm, the vector $\tilde{u}_2$ coincides with the unique stationary distribution of S; that is, $\tilde{u}_2 = u_2$. Show that the stationary distribution of G is exactly $[w_1 \quad \tilde{c}\,\tilde{u}_2]^T$.

2.79 Let $G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}$. Show that for an arbitrary vector x of appropriate size,

$$\|x^T(I - G_{11})^{-1}G_{12}\|_1 \le \|x\|_1.$$

2.80 Show that the aggregated matrix (2.41) can be written as

$$A = \begin{bmatrix} G_{11} & G_{12}e \\ u_2^T G_{21} & 1 - u_2^T G_{21} e \end{bmatrix}.$$

2.81 Prove formula (2.55).

2.82 Show that the vector $z = \tilde{u}_2^T G_{21}$ used in the definition of the approximate aggregated matrix $\tilde{A}$ can be computed as

$$z = \alpha\,\tilde{u}_2^T H_{21} + u_x^T.$$

CHAPTER 3

MATRIX FACTORIZATIONS

We are very familiar with factorizations of one-dimensional algebraic expressions such as

$$x^2 + 3x - 40 \;=\; (x+8)(x-5),$$

and in a very natural way we use such factorizations to solve algebraic equations like $x^2 + 3x - 40 = 0$. The idea behind this, is to transform the problem of solving one quadratic equation into the problem of solving two (much simpler) linear equations:

$$x^2 + 3x - 40 = 0 \quad \Longleftrightarrow \quad (x+8)(x-5) = 0 \quad \Longleftrightarrow \quad \begin{cases} x + 8 = 0 \\ x - 5 = 0. \end{cases}$$

It is very pleasant (and at this point maybe not surprising) to know that this idea generalizes to higher dimensions. More specifically, when we are faced with the problem of solving equations involving arbitrary matrices, that problem can be transformed into a simpler one via certain matrix factorization. For instance, as we will see in detail later on, we can use the LU factorization of a matrix A to transform the problem of solving a system of equations $Ax = b$ into solving two (much simpler) upper and lower triangular systems:

$$Ax = b \quad \Longleftrightarrow \quad LUx = b \quad \Longleftrightarrow \quad \begin{cases} Ly = b \\ Ux = y. \end{cases}$$

But there is more than that. Factorization of matrices is also useful because it uncovers several properties of the matrices and of special sets of vectors associated with them, which

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

are "invisible" without such factorization, and this has a great number of applications not only within mathematics, but especially in other fields like physics and computer sciences. For all these reasons and more, matrix factorizations is a topic of great importance in matrix algebra, and we have decided to devote a short chapter to study some of the most important matrix factorizations and some of their applications.

3.1 *LU* FACTORIZATION

One of the most common problems encountered in mathematics is that of solving a linear system of equations $Ax = b$. We encounter this problem in several instances like finding coefficients of linear combinations and linear independence, solution of boundary value problems, discretization of partial differential equations, solution of systems of nonlinear equations, interpolation, and so on.

If the matrix A (the coefficient matrix) is triangular, then the solution can be obtained by simple backward or forward substitution.

■ **EXAMPLE 3.1**

Solving $Ax = b$, with $A = \begin{bmatrix} 4 & 2 & 5 \\ 0 & 1 & 3 \\ 0 & 0 & 9 \end{bmatrix}$ and $b = \begin{bmatrix} 4 \\ -3 \\ 18 \end{bmatrix}$, means solving the system

$$\begin{aligned} 4x + 2y + 5z &= 4, \\ y + 3z &= -3, \\ 9z &= 18, \end{aligned}$$

and backward substitution gives

$$z = \frac{18}{9} = 2, \quad y = -3 - 3(2) = -9, \quad x = \frac{1}{4}\left(4 - 2(-9) - 5(2)\right) = 3.$$

Thus, for this system with an upper triangular coefficient matrix, we just need to perform backward substitution. Something similar happens with lower triangular matrices.

Given the simplicity of solving triangular systems, the idea for a general system is to transform the system of equations into two simpler systems, one upper triangular and the other lower triangular, and to solve these systems by backward and forward substitution, respectively.

Assume the matrix A can be factored as the product of a lower triangular matrix L and an upper triangular matrix U, $A = LU$. Then:

$$Ax = b \quad \rightarrow \quad LUx = b \quad \rightarrow \quad \begin{cases} Ly = b \\ Ux = y. \end{cases} \tag{3.1}$$

Thus, we first solve the system $Ly = b$ to find y, and then we solve the system $Ux = y$ to find x, the solution of $Ax = b$.

- MATLAB command: lu(A)

The LU factorization is related to the well-known **Gaussian elimination**, which is the process of solving $Ax = b$ by row reduction. We illustrate this process of elimination with an example.

■ **EXAMPLE 3.2**

Consider the linear system $Ax = b$:

$$\begin{aligned} x + 4y + 7z &= -9, \\ 2x + 5y + 8z &= 2, \\ 3x + 6y + 10z &= 7. \end{aligned}$$

Gaussian elimination consists of performing elementary row operations on the coefficient matrix A and on the right–hand side vector b. The purpose is to zero the elements below the diagonal.

If we denote the row i $(i = 1, 2, 3)$, with R_i and if we augment the vector b to the matrix A, we have:

$$[A \quad b] = \left[\begin{array}{ccc|c} 1 & 4 & 7 & -9 \\ 2 & 5 & 8 & 2 \\ 3 & 6 & 10 & 7 \end{array}\right] \xrightarrow[R_3 - 3R_1]{R_2 - 2R_1} \left[\begin{array}{ccc|c} 1 & 4 & 7 & -9 \\ 0 & -3 & -6 & 20 \\ 0 & -6 & -11 & 34 \end{array}\right].$$

Similarly, we have

$$\left[\begin{array}{ccc|c} 1 & 4 & 7 & -9 \\ 0 & -3 & -6 & 20 \\ 0 & -6 & -11 & 34 \end{array}\right] \xrightarrow{R_3 - 2R_2} \left[\begin{array}{ccc|c} 1 & 4 & 7 & -9 \\ 0 & -3 & -6 & 20 \\ 0 & 0 & 1 & -6 \end{array}\right].$$

The new system obtained is triangular:

$$\begin{aligned} x + 4y + 7z &= -9, \\ -3y - 6z &= 20, \\ z &= -6 \end{aligned}$$

and can be quickly solved by backward substitution, which gives

$$z = -6, \quad y = 16/3, \quad x = 35/3.$$

Note. Here we are using the fact that the elementary row operations of Gaussian elimination do not change the solution of the system.

We observe that the last matrix obtained through Gaussian elimination is upper triangular. This is precisely the matrix U in the LU factorization. Now let us see how we get L from this elimination process. The row operations performed above are:

$$R_2 - 2R_1, \quad R_3 - 3R_1 \qquad \text{and} \qquad R_3 - 2R_2. \tag{3.2}$$

The coefficients 2, 3, and 2 in (3.2), respectively are called *multipliers*. In matrix language, performing the first two row operations in (3.2) is equivalent to pre–multiplying the matrix A by a lower triangular matrix L_1 with 1's on the diagonal (we call such matrices unit lower triangular) to obtain the second matrix. Similarly, to obtain the third matrix, we pre–multiply the second one by another lower triangular matrix L_2 with 1's on the diagonal. For the example above, those matrices are

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}, \qquad L_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -2 & 1 \end{bmatrix}, \quad \text{so that}$$

$$L_1 A = \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & -6 & -11 \end{bmatrix} \quad \text{and} \quad L_2(L_1 A) = \begin{bmatrix} 1 & 4 & 7 \\ 0 & -3 & -6 \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus, we have

$$L_2 L_1 A = U. \quad \text{Then,} \quad A = (L_2 L_1)^{-1} U. \quad \text{Therefore, we define}$$

$$L := (L_2 L_1)^{-1} = L_1^{-1} L_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix},$$

so that we have $A = LU$.

Observe that the matrix L has 1's on the diagonal and that the elements below the diagonal are nothing else but the multipliers. Thus, in practice, the row operations in Gaussian elimination give an upper triangular matrix U, and L is just the collection of the multipliers. Also observe that the eigenvalues of L_1 and L_2 are all ones, and therefore, according to (1.65) in Chapter 1, their inverses are well defined.

In general, for an $n \times n$ matrix A, we have that

$$L_{n-1} \cdots L_2 L_1 A = U, \tag{3.3}$$

so that

$$A = LU, \qquad \text{with} \qquad L = (L_{n-1} \cdots L_2 L_1)^{-1}.$$

Computing inverses for this type of lower triangular matrices comes for free, because it just amounts to change the sign of the elements below the diagonal; e.g., in the example above we have

$$L_1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{bmatrix}. \quad \text{Then,} \quad L_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix}.$$

Not every square matrix A has an LU factorization as described above. The multipliers in the Gaussian elimination (the lower triangular elements in L) are in general defined as

$$l_{ik} = \frac{a_{ik}}{a_{kk}}, \quad i = k+1, \ldots, n, \quad (k = 1, \ldots, n-1),$$

which means that we must at least require that the diagonal entries a_{kk} are not zero at each step of the process. Although this is related to the nonsingularity of A, unfortunately, nonsingularity of A is not enough to guarantee the existence of the LU factorization; we also need to require some regularity of the principal submatrices. And with respect to uniqueness of the factorization, if we restrict L to have only 1's along the diagonal as above, the factorization is unique.

■ **EXAMPLE 3.3**

The following nonsingular matrix does not have an LU factorization, as described above.

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

■ **EXAMPLE 3.4**

The following singular matrix has an LU factorization. Here we do not set any restrictions on the diagonal elements of L.

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

■ **EXAMPLE 3.5**

The following matrix A has an LU factorization, but it is not unique if we do not restrict the diagonal elements of L to be 1.

$$A = \begin{bmatrix} 12 & 3 \\ 3 & 6 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/4 & 1 \end{bmatrix} \begin{bmatrix} 12 & 3 \\ 0 & 21/4 \end{bmatrix} = \begin{bmatrix} 12 & 0 \\ 3 & 21/4 \end{bmatrix} \begin{bmatrix} 1 & 1/4 \\ 0 & 1 \end{bmatrix}.$$

From Example 3.5, we see that if A is an arbitrary symmetric matrix, then the LU factorization is not unique if we do not fix the diagonal entries of L. Indeed, if $A = A^T$ and $A = LU$, then

$$A = A^T = (LU)^T = U^T L^T =: \tilde{L}\tilde{U},$$

which is a second LU factorization of A.

The following theorem gives conditions for the existence and uniqueness of an LU factorization. We use the notation introduced after Definition 1.17.

Theorem 3.1 *Let $A_{n\times n}$ be nonsingular. Then A can be uniquely factored as $A = LU$, with L unit lower triangular and U upper triangular, if and only if*

$$\det A(\{1,\ldots,k\}) \neq 0, \quad k = 1,\ldots,n. \tag{3.4}$$

(That is, each leading principal submatrix in A is nonsingular).

Proof. The elimination process of zeroing columns 1 through $k-1$ below the diagonal is given in matrix form as

$$L_{k-1}\cdots L_2 L_1 A = A^{(k-1)},$$

where A has been overwritten by $A^{(k-1)}$. For illustration purposes, suppose $k = 4$. Then

$$A^{(k-1)} = \left[\begin{array}{cccc|cccc} x & x & x & x & x & x & \cdots & x \\ 0 & x & x & x & x & x & \cdots & x \\ 0 & 0 & x & x & x & x & \cdots & x \\ 0 & 0 & 0 & x & x & x & \cdots & x \\ \hline 0 & 0 & 0 & x & x & x & \cdots & x \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & x & x & x & \cdots & x \end{array}\right].$$

Thus, in general, looking at the leading $k \times k$ block of $A^{(k-1)}$, we have

$$\det A(\{1,\ldots,k\}) = a_{11}^{(k-1)} a_{22}^{(k-1)} \cdots a_{kk}^{(k-1)}.$$

Thus, if $\det A(\{1,\ldots,k\}) \neq 0$ for $k = 1,\ldots,n$, then the pivot $a_{kk}^{(k-1)}$ is nonzero and the elimination process is guaranteed to exist.

Suppose now there are two LU factorizations $A = L_1U_1$ and $A = L_2U_2$ with unit lower triangular matrices L_1 and L_2. Then, from $L_1U_1 = L_2U_2$ we get that

$$L_2^{-1}L_1 = U_2U_1^{-1}.$$

The left–hand side is a lower triangular matrix, and the right–hand side is an upper triangular matrix, so both must be diagonal. But since $L_2^{-1}L_1$ has only 1's on its diagonal, both matrices are actually the identity matrix, that is, $L_2^{-1}L_1 = I = U_2U_1^{-1}$, from which we get $L_1 = L_2$ and $U_1 = U_2$.

□

Remark 3.2 *Unless otherwise stated, the matrix L in an LU factorization is a unit lower triangular (1's on the diagonal).*

3.1.1 The complex case

If the matrix $A_{n\times n}$ has complex entries, it also has an LU factorization as described above. As a matter of fact, Theorem 3.1 remains the same when A, and therefore, L and U are complex.

■ **EXAMPLE 3.6**

Let $A = \begin{bmatrix} 2 & 4+3i \\ i & 3 \end{bmatrix}$. Then,

$$A = LU = \begin{bmatrix} 1 & 0 \\ \frac{1}{2}i & 1 \end{bmatrix} \begin{bmatrix} 2 & 4+3i \\ 0 & \frac{9}{2} - 2i \end{bmatrix}.$$

3.1.2 Solving several systems

The LU factorization becomes especially useful when solving several systems of linear equations for the same matrix A but with different right–hand sides, a situation that often happens in applications. We consider solving the p systems

$$Ax^{(i)} = b^{(i)}, \quad i = 1, \ldots, p, \tag{3.5}$$

where each $b^{(i)}$ is an n-dimensional vector.

Following the idea in (3.1), here we use the LU factorization of A to obtain for each $i = 1, \ldots, p$

$$Ax^{(i)} = b^{(i)} \quad \rightarrow \quad LUx^{(i)} = b^{(i)} \quad \rightarrow \quad \begin{cases} Ly^{(i)} & = & b^{(i)} \\ Ux^{(i)} & = & y^{(i)}. \end{cases} \tag{3.6}$$

In practice when solving simultaneous systems of the form (3.5), we can perform the elimination process to A augmented by $b^{(1)}$, obtain U, solve for $x^{(1)}$, and then solve the remaining $2(p-1)$ triangular systems in (3.6).

■ **EXAMPLE 3.7**

Solve the systems (3.5) for

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}, \quad b^{(1)} = \begin{bmatrix} 4 \\ 10 \\ 6 \end{bmatrix}, \; b^{(2)} = \begin{bmatrix} 3 \\ 8 \\ 1 \end{bmatrix}, \; b^{(3)} = \begin{bmatrix} 1 \\ 12 \\ 4 \end{bmatrix}.$$

Elimination process gives

$$A = \left[\begin{array}{ccc|c} 1 & 0 & 1 & 4 \\ 0 & 2 & 2 & 10 \\ 1 & 1 & 1 & 6 \end{array}\right] \xrightarrow{R_3 - 1R_1} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 4 \\ 0 & 2 & 2 & 10 \\ 0 & 1 & 0 & 2 \end{array}\right].$$

Similarly, we have

$$\left[\begin{array}{ccc|c} 1 & 0 & 1 & 4 \\ 0 & 2 & 2 & 10 \\ 0 & 1 & 0 & 2 \end{array}\right] \xrightarrow{R_3 - \frac{1}{2}R_2} \left[\begin{array}{ccc|c} 1 & 0 & 1 & 4 \\ 0 & 2 & 2 & 10 \\ 0 & 0 & -1 & -3 \end{array}\right].$$

Thus, we have

$$U = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & -1 \end{bmatrix} \quad \text{and} \quad L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1/2 & 1 \end{bmatrix}.$$

So first we solve one triangular system to find $x^{(1)}$:

$$U\,x^{(1)} = \begin{bmatrix} 4 \\ 10 \\ -3 \end{bmatrix} \Longrightarrow x^{(1)} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix},$$

and then we solve the systems

$$Ly^{(2)} = b^{(2)}, \quad Ux^{(2)} = y^{(2)} \qquad \text{and} \qquad Ly^{(3)} = b^{(3)}, \quad Ux^{(3)} = y^{(3)}$$

to find $x^{(2)} = [-3 \;\; -2 \;\; 6]^T$ and $x^{(3)} = [-2 \;\; 3 \;\; 3]^T$.

The example above shows that once the LU factorization is computed, only triangular systems need to be solved to find the solutions to the multiple linear systems.

Computing A^{-1}. One instance where several linear systems need to be solved simultaneously, is the computation of the inverse of a matrix $A_{n\times n}$. If we denote with X the inverse of A, with columns $x_1, x_2, \ldots, x_n$, then we must have $AX = I$; that is,

$$AX = A[x_1 \; x_2 \; \cdots \; x_n] = [Ax_1 \; Ax_2 \; \cdots \; Ax_n] = [e_1 \; e_2 \; \cdots \; e_n],$$

where, as usual, e_i is the vector with its i-th component is equal to 1 and the rest of the entries are zero. Thus, we can compute the inverse of A by solving the n linear systems

$$Ax_1 = e_1, \quad Ax_2 = e_2, \quad \cdots, \quad Ax_n = e_n.$$

■ **EXAMPLE 3.8**

To find the inverse of the matrix $A = \begin{bmatrix} 1 & 4 & 0 \\ 1 & 6 & 2 \\ 3 & 0 & 2 \end{bmatrix}$, we can solve the three systems

$$Ax_1 = e_1, \quad Ax_2 = e_2, \quad Ax_3 = e_3,$$

which gives

$$x_1 = [0.4286 \;\; 0.1429 \;\; -0.6429]^T, \quad x_2 = [-0.2857 \;\; 0.0714 \;\; 0.4286]^T, \quad \text{and}$$

$x_3 = [0.2857 \;\; -0.0714 \;\; 0.0714]^T$. That is,

$$A^{-1} = \begin{bmatrix} 0.4286 & -0.2857 & 0.2857 \\ 0.1429 & 0.0714 & -0.0714 \\ -0.6429 & 0.4286 & 0.0714 \end{bmatrix}.$$

3.1.3 The $PA = LU$ factorization

From the discussion above, we see that we need to have nonzero pivots a_{kk} at each step of the elimination so that we are guaranteed to arrive at the triangular matrix U. However, in general, this condition may fail at some step k. If the matrix A is nonsingular, it is always possible (see Exercise 3.8) to interchange the row containing the zero pivot with one of the rows $k+1$ through n so that a nonzero pivot is available. But this does not give anymore the LU factorization of A, but instead we get the factorization of a permutation of A.

If no permutations are needed, we simply have as before

$$U = L_{n-1} \cdots L_2 L_1 A$$

and define $L = (L_{n-1} \cdots L_2 L_2)^{-1} = L_1^{-1} L_2^{-1} L_{n-1}^{-1}$, so that $A = LU$. If zero pivots are encountered, then we interchange rows via permutation matrices P_i before proceeding with the elimination process with the matrices L_i. More precisely, we have

$$\begin{aligned} A^{(1)} &= L_1 P_1 A \\ A^{(2)} &= L_2 P_2 A^{(1)} = L_2 P_2 L_1 P_1 A \\ &\vdots \\ U = A^{(n-1)} &= L_{n-1} P_{n-1} \cdots L_2 P_2 L_1 P_1 A \end{aligned}$$

The final matrix U is obviously upper triangular as desired. From the equation

$$U = L_{n-1} P_{n-1} \cdots L_2 P_2 L_1 P_1 A, \tag{3.7}$$

we would be tempted to define

$$\mathcal{L} = (L_{n-1} P_{n-1} \cdots L_2 P_2 L_1 P_1)^{-1}, \tag{3.8}$$

from which we would get $A = \mathcal{L}U$. Unfortunately this matrix $\mathcal{L}$ is not lower triangular, as a result of the permutations above (see also Remark 3.5).

Remark 3.3 *If no permutation is needed at step k, then $P_k = I$.*

To obtain an appropriate lower triangular matrix L, we use the fact that for any permutation matrix $P^{-1}P = P^T P = I$, so that

$$P_2^T P_3^T \cdots P_{n-1}^T P_{n-1} \cdots P_3 P_2 = I.$$

Then, we rewrite equation (3.7) as

$$\begin{aligned} U &= L_{n-1}P_{n-1}\cdots L_2P_2L_1(P_2^TP_3^T\cdots P_{n-1}^TP_{n-1}\cdots P_3P_2)P_1A \\ &= (L_{n-1}P_{n-1}\cdots L_2P_2L_1P_2^T\cdots P_{n-1}^T)P_{n-1}\cdots P_1A. \end{aligned}$$

If we set $L^{-1} = L_{n-1}P_{n-1}\cdots L_2P_2L_1P_2^T\cdots P_{n-1}^T$, then the matrix

$$L = P_{n-1}P_{n-2}\cdots P_2L_1^{-1}P_2^TL_2^{-1}\cdots P_{n-1}^TL_{n-1}^{-1} \tag{3.9}$$

is unit lower triangular (see Exercise 3.9). Then, we have $U = L^{-1}PA$, and therefore

$$PA = LU,$$

where the matrix

$$P = P_{n-1}\cdots P_2P_1$$

is the collection of all the permutations performed on A.

Thus, we have proved the following.

Theorem 3.4 *Let $A_{n\times n}$ be a nonsingular matrix. Then, there exist a permutation matrix P, a unit lower triangular matrix L, and an upper triangular matrix U such that*

$$PA = LU.$$

■ **EXAMPLE 3.9**

Let $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 1 & 2 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix}$. First we need to interchange the first and third rows to get

$$P_1A = \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix}, \quad \text{with} \quad P_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The elimination part gives

$$L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & -2 & 1 \end{bmatrix}.$$

Next we interchange the second and third rows:

$$P_2L_1P_1A = \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & -2 & 1 \end{bmatrix}, \quad \text{where} \quad P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The elimination part gives

$$L_2P_2L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & -2 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix}.$$

No row interchange is needed, so $P_3 = I$. The elimination part gives

$$L_3P_3L_2P_2L_1P_1A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = U.$$

Thus, following (3.7), we have obtained an upper triangular matrix U. If we define the matrix L as in (3.9) we get

$$L = P_3P_2L_1^{-1}P_2^TL_2^{-1}P_3^TL_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & -1 & -2 & 1 \end{bmatrix},$$

which again is a unit lower triangular matrix with the multipliers below the diagonal, and it is readily verified that in fact $PA = LU$, where $P = P_3P_2P_1$.

Remark 3.5 *Observe that if we had taken as in (3.8):*
$\mathcal{L} = (L_3P_3L_2P_2L_1P_1)^{-1}$, *then we would have obtained*

$$\mathcal{L} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & -1 & -2 & 1 \end{bmatrix},$$

which fails to be lower triangular. Also observe that this matrix $\mathcal{L}$ is nothing else than a permutation of the matrix L above; that is, $\mathcal{L} = P^TL$, or $P\mathcal{L} = L$.

Solving $Ax = b$. The idea of solving a linear system $Ax = b$ via the LU factorization with permutations is similar to the one where no permutations are needed. First we multiply the system $Ax = b$ by the permutation matrix $P = P_{n-1} \cdots P_2P_1$ (a permutation matrix is nonsingular), so that

$$Ax = b \quad \rightarrow \quad PAx = Pb \quad \rightarrow \quad LUx = Pb \quad \rightarrow \quad \begin{cases} Ly = Pb \\ Ux = y. \end{cases} \tag{3.10}$$

Thus, just as in (3.1), we simply solve two triangular systems to find the solution of the original system $Ax = b$, and the only real difference between (3.1) and (3.10) is that to

effectively apply the LU factorization in solving the system when row interchanges are needed, we permute some rows of A and the corresponding entries of the vector b, using the permutation matrix P.

■ **EXAMPLE 3.10**

Let us solve the system $Ax = b$, with

$$A = \begin{bmatrix} 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 3 \\ 2 & 2 & 1 & 0 \\ 3 & 1 & 0 & -1 \end{bmatrix}, \qquad b = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 2 \end{bmatrix}.$$

By proceeding as in Example 3.9, one obtains

$$L = P_3 P_2 L_1^{-1} P_2^T L_2^{-1} P_3^T L_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \frac{3}{2} & -1 & -\frac{1}{2} & 1 \end{bmatrix},$$

$$U = L_3 P_3 L_2 P_2 L_1 P_1 A = \begin{bmatrix} 2 & 2 & 1 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & \frac{1}{2} \end{bmatrix}, \quad P = P_3 P_2 P_1 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

so that solving $Ly = Pb$ gives $y = [1 \quad 2 \quad 3 \quad 4]^T$, and solving $Ux = y$ gives

$$x = [-\frac{1}{2} \quad \frac{23}{2} \quad -21 \quad 8].^T$$

3.2 QR FACTORIZATION

In this section, we introduce a special kind of matrix factorization, where one of the factors is an upper triangular matrix, but most importantly, the other factor is an orthogonal matrix. Several applications of this factorization are studied.

The main motivation and application of LU factorization in Section 3.1 was the solution of linear systems $Ax = b$, where A was a square matrix of order n. The central idea was that this factorization splits the problem into two simpler ones, with triangular matrices L and U. A clear improvement to such a method can be obtained by using orthogonal matrices instead of the matrices $L_1, \ldots, L_{n-1}$ in (3.3) to convert A into a triangular matrix. This is possible through a QR factorization of the matrix.

Since the QR factorization also exists for rectangular matrices, it can also be applied to the least squares solution of a rectangular overdetermined system and, surely, for the

computation of orthonormal bases. Finally, we will also illustrate how a QR factorization can be used to compute the eigenvalues of a matrix.

3.2.1 QR and Gram–Schmidt

We start by giving a theorem on the existence of the QR factorization of a rectangular matrix A of full rank, and relating it to the orthonormalization process of Gram–Schmidt. Recall that we call a matrix $A_{m\times n}$ full column rank, or simply **full rank**, if all its columns are linearly independent.

Theorem 3.6 *Let $A_{m\times n}$ be a matrix with linearly independent columns. Then there exists a factorization*

$$A = QR,$$

where $Q_{m\times n}$ is an orthogonal matrix and $R_{n\times n}$ is an invertible triangular matrix. The columns of Q form a Gram–Schmidt orthonormal basis of $col(A)$.

Proof. Recall that given n linearly independent vectors $\{v_1, \ldots, v_n\}$, an orthonormal set of vectors $\{q_1, \ldots, q_n\}$ can be obtained via Gram-Schmidt process by defining:

$$\begin{aligned}
& & & & q_1 &= \tfrac{v_1}{\|v_1\|}, \\
w_2 &= & v_2 &- (v_2^T q_1)\, q_1, & q_2 &= \tfrac{w_2}{\|w_2\|}, \\
w_3 &= & v_3 &- (v_3^T q_1)\, q_1 - (v_3^T q_2)\, q_2, & q_3 &= \tfrac{w_3}{\|w_3\|}, \\
& \vdots & & & \vdots & \\
w_n &= & v_n &- \sum_{k=1}^{n-1} (v_n^T q_k)\, q_k, & q_n &= \tfrac{w_m}{\|w_n\|}.
\end{aligned} \tag{3.11}$$

Denote the columns of A with $v_1, \ldots, v_n$. By following (3.11) and letting $r_{11} = q_1^T v_1 = v_1^T q_1$ and $r_{12} = q_1^T v_2 = v_2^T q_1$ we have that

$$\begin{aligned}
v_1 &= q_1\|v_1\| = q_1(q_1^T v_1) = q_1 r_{11}, \\
v_2 &= (v_2^T q_1)q_1 + w_2 = q_1(v_2^T q_1) + q_2\|w_2\| = q_1 r_{12} + q_2(q_2^T v_2) = q_1 r_{12} + q_2 r_{22}, \\
&\vdots
\end{aligned}$$

and in general, for each $j = 1, \ldots, n$, we have

$$v_j = \sum_{i=1}^{j} q_i\, r_{ij}, \tag{3.12}$$

where

$$r_{ij} = q_i^T v_j = v_j^T q_i. \tag{3.13}$$

In matrix form, equation (3.12) can be written as

$$[v_1 \cdots v_n] = [q_1 \cdots q_n] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}, \quad \text{or,}$$

$$A = QR,$$

where the matrix $Q = [q_1 \cdots q_n]$ is orthogonal. Also observe that for each $i = 1, \ldots, n$, we have $r_{ii} = \|w_i\| > 0$ and therefore R is invertible.

□

As stated in Theorem 3.6, since the QR factorization of a matrix A is related to the Gram–Schmidt process (applied to the columns of A), this factorization gives a reliable way to obtain an orthonormal basis for the col(A).

■ **EXAMPLE 3.11**

Let $A = \begin{bmatrix} 1 & 4 & -5 \\ 4 & 2 & 0 \\ 9 & -7 & 1 \\ 3 & 3 & -5 \end{bmatrix}$, with rank$(A) = 3$. Then, A can be factored as $A = QR$, where

$$Q = \begin{bmatrix} 0.0967 & 0.5601 & -0.4106 \\ 0.3867 & 0.4552 & 0.8004 \\ 0.8701 & -0.4421 & -0.1770 \\ 0.2900 & 0.5326 & -0.3993 \end{bmatrix}, \quad \text{with} \quad Q^TQ = I, \quad \text{and}$$

$R = \begin{bmatrix} 10.3441 & -4.0603 & -1.0634 \\ 0 & 7.8431 & -5.9056 \\ 0 & 0 & 3.8722 \end{bmatrix}$. The columns of the matrix Q form an orthonormal basis of col(A).

Remark 3.7 *In some cases, the QR factorization of Theorem 3.6 is called the* thin *or* economy size *QR factorization. This is because one can also factor $A_{m\times n}$ $(m \geq n)$ as $A = QR$, with $Q_{m\times m}$ and $R_{m\times n}$.*

- MATLAB commands: qr(A), qr(A, 0)

Theorem 3.6 states that the QR factorization and the Gram–Schmidt orthonormalization process are closely related. More precisely, for full rank matrices, the QR factorization can be simply thought of as the matrix version of the Gram–Schmidt process.

■ **EXAMPLE 3.12**

The columns of the matrix A in Example 3.11 are linearly independent. If we denote the three columns of A as v_1, v_2, and v_3 and apply the Gram–Schmidt process as in (1.40), then we obtain the orthonormal vectors

$$q_1 = [0.0967\ 0.3867\ 0.8701\ 0.2900]^T,\ q_2 = [0.5601\ 0.4552\ -0.4421\ 0.5326]^T,$$
$$q_3 = [-0.4106\ 0.8004\ -0.1770\ -0.3993]^T,$$

which are exactly the columns of Q in the QR factorization of A.

■ **EXAMPLE 3.13**

In Example 1.49, we applied Gram–Schmidt process to the vectors

$$v_1 = [2\ \ -1\ \ 0]^T, \quad v_2 = [1\ \ 0\ \ -1]^T, \quad v_3 = [3\ \ 7\ \ -1]^T$$

and obtained the orthonormal vectors

$$q_1 = \frac{1}{\sqrt{5}}[2\ \ -1\ \ 0]^T, \quad q_2 = \frac{1}{\sqrt{30}}[1\ \ 2\ \ -5]^T, \quad q_3 = \frac{1}{\sqrt{6}}[1\ \ 2\ \ 1]^T.$$

If we let $A = [v_1\ \ v_2\ \ v_3]$, then the matrix $Q = [q_1\ \ q_2\ \ q_3]$ is the orthogonal matrix in the factorization $A = QR$, and the entries of the triangular matrix R are defined by (3.13); that is (without computing the lower triangular part, which is zero):

$$r_{11} = q_1^T v_1 = 2.2361, \quad r_{12} = q_1^T v_2 = 0.8944, \quad r_{13} = q_1^T v_3 = -0.4472,$$
$$r_{22} = q_2^T v_2 = 1.0954, \quad r_{23} = q_2^T v_3 = 4.0166,$$
$$r_{33} = q_3^T v_3 = 6.5320.$$

Compare these calculations with the ones in Example 1.49.

Remark 3.8 *Using the fact that Q is orthogonal ($Q^TQ = I$), we can multiply both sides of the equation $A = QR$ by Q^T to obtain*

$$R = Q^T A. \tag{3.14}$$

Equation (3.14) is the matrix form of equation (3.13).

Remark 3.9 *If a matrix A does not have full rank, its QR factorization is still well defined. In this case, however, the matrix R will have one or more zero rows, and the columns of Q are not necessarily a basis for $col(A)$.*

■ **EXAMPLE 3.14**

The matrix

$$A = \begin{bmatrix} 1 & 4 & -5 \\ 2 & 8 & -10 \\ 4 & 16 & -20 \\ 8 & 2 & 40 \end{bmatrix}$$

has rank two. Its QR factorization, $A = QR$ gives

$$Q = \begin{bmatrix} 0.1085 & 0.1894 & -0.7695 \\ 0.2169 & 0.3787 & -0.4732 \\ 0.4339 & 0.7574 & 0.4290 \\ 0.8677 & -0.4971 & 0.0000 \end{bmatrix}, \quad R = \begin{bmatrix} 9.2195 & 10.8465 & 23.3200 \\ 0 & 14.9115 & -39.7640 \\ 0 & 0 & 0 \end{bmatrix}.$$

Since the columns of Q are linearly independent, they form a basis of a three-dimensional subspace S; however the col(A) is only two-dimensional. That is, the columns of Q do not form a basis of col(A).

In this Example 3.14, we also observe that if we block the matrices Q and R as

$$Q = [Q_1 \quad q_3], \quad R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix},$$

where Q_1 represents the first two columns of Q, and R_{11} is the upper left 2×2 block of R, then from $A = QR$, we get

$$A_1 = Q_1 R_{11},$$

which is the QR factorization (as in Theorem 3.6) of the matrix A_1 formed by the first two (linearly independent) columns of the original matrix A.

The following theorem establishes the existence of a QR factorization of a general matrix matrix A, which is not necessarily full-rank.

Theorem 3.10 *Let A be an $m \times n$ matrix $(m \geq n)$. Then, there exists an orthogonal matrix $Q_{m\times n}$ and an upper triangular matrix $R_{n\times n}$ such that*

$$A = QR.$$

Observe that this theorem does not say anything about the columns of Q forming a basis for col(A), nor anything about the nonsingularity of the matrix R. See Example 3.14.

We have seen before that the LU factorization with $l_{ii} = 1$ can be thought of as the matrix version of the Gaussian elimination. Similarly, we have just seen that for full rank matrices, the QR factorization with $r_{ii} > 0$ can be understood as the matrix version of the Gram–Schmidt orthonormalization process. That is,

$$\begin{aligned} \text{Gaussian elimination} &\longleftrightarrow LU \\ \text{Gram} - \text{Schmidt} &\longleftrightarrow QR \end{aligned} \tag{3.15}$$

Uniqueness. As explained above, the QR factorization of a full rank matrix A is closely related to the Gram–Schmidt process. This process is a source of uniqueness for such a factorization. Indeed, when we fix the diagonal entries of the matrix R to be positive, then the QR factorization of a full rank matrix A is exactly the Gram–Schmidt process. However, a QR factorization is not unique in general, even if the matrix is full rank, as the following example shows.

■ **EXAMPLE 3.15**

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

The given matrix A is full rank, but it has at least two QR factorizations. The first one is the one that corresponds to the Gram–Schmidt process.

Remark 3.11 *From Example 3.15, we observe a general fact: When $A_{m\times n}$ has full rank and we fix the diagonal entries of R to be positive, then R is unique and therefore so is Q.*

3.2.2 The complex case

Given a complex matrix A, its QR factorization exists just as described in Theorems 3.6 and 3.10 above, except that now the matrix Q is unitary. In particular, we have the following therorem.

Theorem 3.12 *Let $A_{m\times n}$ be a complex matrix with $m \geq n$ Then, there exists a unitary matrix $Q_{m\times n}$ and an upper triangular matrix $R_{n\times n}$ such that*

$$A = QR.$$

If A has full rank, then the factorization corresponds to the Gram–Schmidt process applied to the columns of A.

■ **EXAMPLE 3.16**

Consider the rank-2 matrix

$$A = \begin{bmatrix} 2-i & 8 \\ 9 & 2i \\ i & 3 \end{bmatrix}.$$

Its QR factorization gives

$$Q = \begin{bmatrix} -0.2144 + 0.1072i & 0.8936 - 0.0418i \\ -0.9649 & -0.2008 - 0.0460i \\ -0.1072i & 0.3959 - 0.0223i \end{bmatrix}, \qquad \text{and}$$

$$R = \begin{bmatrix} -9.3274 & -1.7154 - 2.4659i \\ 0 & 8.2448 \end{bmatrix}.$$

After seeing how the QR factorization can be used to compute orthonormal bases of the $\mathrm{col}(A)$, now we want to present some more applications of such a factorization, especially exploiting the presence of an orthogonal matrix as one of the factors.

3.2.3 QR and similarity

The QR factorization offers an alternative to a general similarity transformation of a matrix A. Remember we say a matrix A is similar to a matrix B if there exists a nonsingular matrix P such that

$$P^{-1}AP = B. \tag{3.16}$$

A clear improvement on the computation of such a similarity transformation would be to have a square orthogonal matrix Q available for use instead of the general nonsingular matrix P. In this case, the similarity transformation can be written as

$$Q^T A Q = B, \tag{3.17}$$

where explicit computation of an inverse matrix as in (3.16) has now been avoided.

On the other hand, let us assume that $P = QR$ is the QR factorization of the matrix P in (3.16) and that A is similar to a triangular matrix T; that is, $P^{-1}AP = T$. Then, using the QR factorization of P in this similarity transformation, we get $R^{-1}Q^T AQR = T$, and therefore

$$Q^T A Q = R^{-1} T R. \tag{3.18}$$

The matrix $R^{-1}TR$ is upper triangular and that makes (3.18) a similarity transformation of A into a triangular form by using orthogonal matrices. The inverse of R does not need to be computed; all that is needed is the QR factorization of P and the matrix product $Q^T AQ$.

■ **EXAMPLE 3.17**

Let $A = \begin{bmatrix} 5 & 15 & -41 \\ -12 & 20 & -4 \\ 1 & 11 & -21 \end{bmatrix}$. A has eigenvalues: $\lambda = -4, -8, 16$. With $P = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 3 \\ 1 & 0 & 1 \end{bmatrix}$, we have a similarity transformation:

$$P^{-1}AP = T = \begin{bmatrix} -8 & 12 & -4 \\ 0 & -4 & 12 \\ 0 & 0 & 16 \end{bmatrix}.$$

If $P = QR$, with $Q = \begin{bmatrix} -0.8165 & -0.0000 & -0.5774 \\ -0.4082 & -0.7071 & 0.5774 \\ -0.4082 & 0.7071 & 0.5774 \end{bmatrix}$, we get the similarity transformation

$$Q^TAQ = \begin{bmatrix} -8 & 48.4974 & 10.6066 \\ 0 & -4 & -15.9217 \\ 0 & 0 & 16 \end{bmatrix}.$$

Since similarity transformations preserve eigenvalues, the QR factorizations can be used in similarity transformations to develop an algorithm for the computation of eigenvalues. We give details next.

3.2.4 The QR algorithm

A very useful application of QR factorizations is the computation of eigenvalues. Finding eigenvalues λ and eigenvectors v of a square matrix of order n analytically by solving the corresponding equations

$$\det(A - \lambda \mathrm{I}) = 0, \qquad (A - \lambda \mathrm{I})\, v = 0,$$

as explained in Section 1.10, is only feasible for very small values of n. In general, iterative methods have to be used to approximate eigenvalues and eigenvectors. The simplest of all methods is the so–called power method, which is studied in some detail in Chapter 2. Here we show how by performing the QR factorization of the given matrix A, the eigenvalues are exposed by moving them to the the diagonal part of a triangular matrix. In the process, the eigenvectors can also be computed. Here we assume A has only real eigenvalues.

We start by recalling from Section 1.11 that two similar matrices A and B have the same eigenvalues, counting multiplicities. Let now $A = QR$ be the QR factorization of A. This implies that

$$Q^TAQ = Q^TQRQ = RQ. \tag{3.19}$$

That is,

If $A = QR$, then the matrix RQ is similar to A.

Therefore RQ has the same eigenvalues as A, counting multiplicities. The central idea of the algorithm is to perform such orthogonal similarity transformations to form a sequence of matrices A_k, $k = 1, 2, \ldots$, that will converge to a triangular form. That is,

$$\begin{array}{lll} \text{let} & A_1 & = A \\ & & \\ \text{factor} & A_1 & = Q_1 R_1 \\ \text{define} & A_2 & = R_1 Q_1 \\ & & \\ \text{factor} & A_2 & = Q_2 R_2 \\ \text{define} & A_3 & = R_2 Q_2 \\ & & \vdots \\ \text{factor} & A_k & = Q_k R_k \\ \text{define} & A_{k+1} & = R_k Q_k \\ & & \vdots \end{array} \tag{3.20}$$

repeat until convergence.

The sequence A_k will converge to an upper triangular matrix, and the (approximate) eigenvalues of A will be exposed along the diagonal.

In fact, observe that if we define the orthogonal matrix $P_k = Q_1 Q_2 \cdots Q_k$, then

$$\begin{aligned} P_{k-1}^T P_k &= (Q_1 Q_2 \cdots Q_{k-1})^T (Q_1 Q_2 \cdots Q_k) \\ &= Q_{k-1}^T \cdots Q_2^T Q_1^T Q_1 Q_2 \cdots Q_{k-1} Q_k \\ &= Q_{k-1}^T Q_{k-1} Q_k \\ &= Q_k. \end{aligned}$$

Now assuming that as $k \to \infty$ we have $P_k \to P$, for some orthogonal matrix P, we get

$$Q_k \to P^T P = I,$$

and therefore from the iteration $A_{k+1} = R_k Q_k$ in (3.20), we have

$$\lim_{k\to\infty} A_{k+1} = \lim_{k\to\infty} R_k Q_k = \lim_{k\to\infty} R_k = R,$$

an upper triangular matrix.

Remark 3.13 *The matrices A_k, $k = 2, 3, \ldots$, are all similar to the original matrix $A = A_1$. See Exercise 3.26.*

■ **EXAMPLE 3.18**

The eigenvalues of the matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 7 & 2 & 0 \\ 5 & 4 & 3 & 3 \\ 8 & 7 & 2 & 3 \end{bmatrix}$$

are $\sigma(A) = \{12.5351, -3.9610, 4.5105, 0.9154\}$. Two iterates of the QR algorithm are

$$A_{10} = \begin{bmatrix} 12.5392 & -4.8680 & -6.3580 & 0.0988 \\ 0.0139 & -3.9540 & -1.9643 & 1.1003 \\ 0.0001 & -0.0479 & 4.4994 & 2.4185 \\ 0.0000 & 0.0000 & 0.0000 & 0.9154 \end{bmatrix},$$

$$A_{30} = \begin{bmatrix} 12.5351 & -4.9178 & 6.3319 & -0.0997 \\ 0.0000 & -3.9610 & 1.9110 & -1.1139 \\ 0.0000 & 0.0000 & 4.5105 & 2.4122 \\ 0.0000 & 0.0000 & 0.0000 & 0.9154 \end{bmatrix}.$$

Observe that the diagonal entries of the matrices A_k, $k = 1, 2, \ldots$ converge to the eigenvalues of A.

In principle, the QR algorithm can be applied to any square matrix, but it could take too many steps to obtain a reasonable approximate of the eigenvalues. Thus, in practice, the preferable approach is first to reduce the given matrix A to Hessenberg form (see (1.18)) and then apply the QR algorithm to this new matrix. We can obtain the Hessenberg form of a matrix through orthogonal similarity transformations; if in addition, the original matrix is symmetric, then we obtain a tridiagonal matrix.

In Chapter 4, we will study the application of the QR factorization of a rectangular matrix $A_{m\times n}$ to the least squares solution of systems $Ax = b$. Here we study another very important applications of QR factorization, which is associated with Householder matrices.

3.2.5 QR and LU

In Section 3.1, we used the LU factorization to reduce a given square matrix A of order n to a triangular matrix U, not by a similarity transformation but by a series of elementary operations (Gaussian elimination) that in matrix language amounts to performing matrix multiplications

$$L_{n-1} \cdots L_2 L_1 A = U. \tag{3.21}$$

As mentioned before, a more efficient way to do this is by using orthogonal matrices instead of the triangular matrices L_i. Here we can use the Householder matrices introduced in

Chapter 1: For an arbitrary vector u we define

$$H = I - \frac{2\,u\,u^T}{u^T u}. \tag{3.22}$$

This Householder matrix H is orthogonal and symmetric. Recall that multiplying A by L_1 from the left will zero all the entries of A on the first column below the diagonal. We would like to have a Householder matrix H_1 that has the same effect on A.

If we denote the first column of A with a_1, then it can be proved (see Exercise 3.28) that the required Householder matrix H_1 is the one in (3.22) for which

$$u = \begin{bmatrix} a_{11} + \text{sign}(a_{11})\|a_1\| \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}. \tag{3.23}$$

With this definition, we get

$$H_1 A = \begin{bmatrix} \alpha & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{bmatrix},$$

where $|\alpha| = \|a_1\|$.

Thus, the first column has been reduced as required. In a similar way, Householder matrices $H_2, \ldots H_{n-1}$ can be defined (see Example 3.19 below) to reduce the other columns so that we have

$$H_{n-1} \cdots H_2 H_1\, A = R, \tag{3.24}$$

with R upper triangular. Compare this with (3.21).

How is this related to QR factorization?

Since $H_i^T H_i = I$ for all $i = 1, \ldots n-1$, we can multiply both sides of (3.24) by $H_1^T \cdots H_{n-1}^T$ to get

$$\begin{aligned}
(H_1^T \cdots H_{n-1}^T)\,(H_{n-1} \cdots H_2 H_1)\,A &= (H_1^T \cdots H_{n-1}^T)\,R \\
(H_1^T \cdots H_{n-2}^T)\,(H_{n-2} \cdots H_2 H_1)\,A &= (H_1^T \cdots H_{n-1}^T)\,R \\
(H_1^T \cdots H_{n-3})\,(H_{n-3} \cdots H_2 H_1)\,A &= (H_1^T \cdots H_{n-1}^T)\,R \\
&\vdots \\
H_1^T H_1\, A &= (H_1^T \cdots H_{n-1}^T)\,R \\
A &= (H_1^T \cdots H_{n-1}^T)\,R.
\end{aligned}$$

In other words, we have

$$A = Q\,R, \tag{3.25}$$

with $Q = H_1^T \cdots H_{n-1}^T$.

We have just showed how the Householder matrices can be applied to the problem of reducing a matrix A to a triangular form R through the matrix operations in (3.24). Since we are using orthogonal matrices, this is clearly a better approach than that in (3.21). At the same time, this has provided us with a way of proving the existence of a **(Householder) QR factorization** of the given square matrix A.

Remark 3.14 *Neither (3.21) nor (3.24) are similarity transformations, so that even though U and R are upper triangular matrices, their diagonal entries are not necessarily the eigenvalues of A.*

■ **EXAMPLE 3.19**

Let $A = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 10 \end{bmatrix}$. This is the same matrix of Example 3.2.

To zero the elements of the first column below the diagonal, we use (3.23) so that $u = \begin{bmatrix} 4.7417 \\ 2 \\ 3 \end{bmatrix}$. According to (3.22), this gives

$$H_1 = \begin{bmatrix} -0.2673 & -0.5345 & -0.8018 \\ -0.5345 & 0.7745 & -0.3382 \\ -0.8018 & -0.3382 & 0.4927 \end{bmatrix},$$

and we get

$$H_1 A = \begin{bmatrix} -3.7417 & -8.5524 & -14.1648 \\ 0.0000 & -0.2945 & -0.9272 \\ 0.0000 & -1.9418 & -3.3908 \end{bmatrix}.$$

We now repeat the process but only for the 2×2 lower right block of $H_1 A$.

Thus, using again (3.22) and (3.23), our new vector u is $u = \begin{bmatrix} -2.2585 \\ -1.9418 \end{bmatrix}$ and

$\tilde{H}_2 = \begin{bmatrix} -0.1500 & -0.9887 \\ -0.9887 & 0.1500 \end{bmatrix}$, and we define

$$H_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.1500 & -0.9887 \\ 0 & -0.9887 & 0.1500 \end{bmatrix},$$

so that

$$H_2 H_1 A = \begin{bmatrix} -3.7417 & -8.5524 & -14.1648 \\ 0.0000 & 1.9640 & 3.4915 \\ 0.0000 & 0.0000 & 0.4082 \end{bmatrix} = R.$$

This is the upper triangular matrix obtained with the Householder QR factorization of A, with

$$Q = H_1^T H_2^T = \begin{bmatrix} -0.2673 & 0.8729 & 0.4082 \\ -0.5345 & 0.2182 & -0.8165 \\ -0.8018 & -0.4364 & 0.4082 \end{bmatrix}.$$

QR and linear systems. We know that with an LU factorization $A = LU$, we are applying Gaussian elimination to row-reduce A into a triangular form. This allows us to solve $Ax = b$ by solving instead two triangular systems, that is,

$$Ax = b \qquad \Longleftrightarrow \qquad \begin{cases} Ly &= b \\ Ux &= y. \end{cases}$$

With a QR factorization, $A = QR$, we also transform A into a triangular matrix, but this time in a more efficient way, by using orthogonal matrices. Using this factorization, we can rewrite $Ax = b$ as $QRx = b$ and then we multiply both sides of the equation by Q^T so that by using orthogonality, we have $Rx = Q^T b$. Thus, we solve instead one triangular system; that is,

$$Ax = b \qquad \Longleftrightarrow \qquad Rx = Q^T b. \tag{3.26}$$

■ **EXAMPLE 3.20**

Consider again the matrix A of Example 3.19. We have $A = QR$, with

$$Q = \begin{bmatrix} -0.2673 & 0.8729 & 0.4082 \\ -0.5345 & 0.2182 & -0.8165 \\ -0.8018 & -0.4364 & 0.4082 \end{bmatrix}, \; R = \begin{bmatrix} -3.7417 & -8.5524 & -14.1648 \\ 0.0000 & 1.9640 & 3.4915 \\ 0.0000 & 0.0000 & 0.4082 \end{bmatrix}.$$

Using the same vector b of Example 3.2, we have that

$$Q^T b = [-4.2762 \quad -10.4745 \quad -2.4495]^T.$$

Then, by solving the system $Rx = Q^T b$, we get $x = [35/3 \quad 16/3 \quad -6]^T$.

The following example illustrates why a QR factorization, which uses orthogonal matrices, is more efficient than an LU factorization of a given matrix A. In particular, one way to measure this is to compare the condition number of the original matrix A with those of the obtained matrices U and R. Recall that a large condition number indicates a potential loss of accuracy in the numerical computations.

■ **EXAMPLE 3.21**

Let $A = \begin{bmatrix} 0.001 & 2 \\ 1 & 1 \end{bmatrix}$. Then, $\text{cond}(A) = 2.6198$, and

$$A = LU = \begin{bmatrix} 1 & 0 \\ 1000 & 1 \end{bmatrix} \begin{bmatrix} 0.001 & 2 \\ 0 & -1999 \end{bmatrix},$$

where $\text{cond}(U) = 1.999 \times 10^6$.

On the other hand,

$$A = QR = \begin{bmatrix} 0.001 & 1 \\ 1 & -0.001 \end{bmatrix} \begin{bmatrix} 1 & 1.002 \\ 0 & 1.999 \end{bmatrix},$$

and $\text{cond}(R) = 2.6198$, which is the same as $\text{cond}(A)$.

3.3 SINGULAR VALUE DECOMPOSITION (SVD)

The singular value decomposition of a matrix A is relatively simple, it has a wide range of applications, and it is an excellent showcase for several concepts in linear algebra and matrix computations. Nevertheless, it has been traditionally skipped in undergraduate and sometimes even in graduate courses in linear algebra. Its importance cannot be overstressed. The SVD allows us to decompose a matrix A into several components, exposing many of its properties that are invisible otherwise, such as the rank, spectral norm, orthonormal bases, the sensitivity of a matrix to numerical error in a linear system, lower-rank approximations to A, and related lower-dimensional subspaces. All these nice properties of the SVD factorization of a matrix will allow for several applications, some of which we will cover in detail.

Again, orthogonality is the key ingredient.

Theorem 3.15 *Given any matrix $A_{m\times n}$, there exist orthogonal matrices $U_{m\times m}$, $V_{n\times n}$ and a diagonal matrix $\Sigma_{m\times n}$ such that*

$$A = U\Sigma V^T, \tag{3.27}$$

and the following is true:

1. *The columns of U are eigenvectors of the matrix $A\,A^T$.*
2. *The columns of V are eigenvectors of the matrix $A^T A$.*
3. *The diagonal entries of Σ, the* **singular values** *of A, are the nonnegative square roots of the eigenvalues λ_i of $A^T A$ (or $A\,A^T$); that is,*

$$\sigma_i = \sqrt{\lambda_i}, \quad i = 1, \ldots p, \quad p = \min\{m, n\},$$

arranged in descending order.

Proof. The proof of existence goes by induction. Let $\sigma_1 = \|A\|_2$, and consider two unit vectors $u_1 \in \mathbb{R}^m$, $v_1 \in \mathbb{R}^n$ that satisfy $Av_1 = \sigma_1 u_1$. The vectors u_1 and v_1 can be extended to orthonormal bases of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively (see orthonormal extensions in section 1.4.7). That is, there exist $X \in \mathbb{R}^{m\times(m-1)}$ and $Y \in \mathbb{R}^{n\times(n-1)}$ such that $U_1 = [u_1 \;\; X]_{m\times m}$ and $V_1 = [v_1 \;\; Y]_{n\times n}$ are orthogonal. Then,

$$U_1^T A V_1 = \begin{bmatrix} u_1^T \\ X^T \end{bmatrix} A \, [v_1 \quad Y] = \begin{bmatrix} u_1^T A v_1 & u_1^T A Y \\ X^T A v_1 & X^T A Y \end{bmatrix} = \begin{bmatrix} \sigma_1 & z^T \\ 0 & A_2 \end{bmatrix} =: S_1,$$

where $z \in \mathbb{R}^{n-1}$ and $A_2 \in \mathbb{R}^{(m-1)\times(n-1)}$. We need to show $z = 0$. To this end, we see that

$$\left\| S_1 \begin{bmatrix} \sigma_1 \\ z \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \sigma_1^2 + z^T z \\ A_2 z \end{bmatrix} \right\|_2^2 \geq (\sigma_1^2 + z^T z)^2.$$

Then, by the definition (1.11) of the 2-norm for matrices, we have that

$$\|S_1\|_2^2 \;\geq\; \sigma_1^2 + z^T z.$$

However, by orthogonal invariance (Exercise 1.42), A and S_1 must have the same norm; that is, $\|S_1\|_2^2 = \sigma_1^2$ and therefore $z = 0$.

By applying the same process on A_2, we can conclude that there exist orthogonal matrices U_2, V_2 such that $U_2^T A_2 V_2 = S_2 = \begin{bmatrix} \sigma_2 & d^T \\ 0 & A_3 \end{bmatrix}$, with $d = 0$. Following by induction, the existence statement is proved.

Let us now see why in fact the columns of U are eigenvectors of the matrix AA^T. First observe that we have $A^T = V\Sigma^T U^T$. Combining this with the factorization (3.27), we have

$$AA^T = U\Sigma V^T V \Sigma^T U^T = U \Sigma \Sigma^T U^T.$$

This means

$$AA^T U = U \Sigma \Sigma^T, \tag{3.28}$$

where $\Sigma\Sigma^T = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$.

Thus, the columns of U are eigenvectors of AA^T with eigenvalues $\lambda_i = \sigma_i^2$. The proof of the other statements in the theorem is left as exercise.

□

Terminology: The columns of U are called the **left singular vectors** of A and the columns of V the **right singular vectors** of A.

We develop an example below, following what would be a simple algorithm to compute the SVD of a given matrix A.

■ **EXAMPLE 3.22**

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

(1) Find the eigenvalues of the matrix $A^T A$, arrange them in descending order, and determine the number r of positive eigenvalues:

$$A^T A = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}. \quad \lambda_1 = 12,\ \lambda_2 = 10,\ \lambda_3 = 0. \qquad r = 2.$$

(2) Find orthogonal eigenvectors of $A^T A$ and form with them the matrix V:

$$v_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad v_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \quad v_3 = \frac{1}{\sqrt{30}} \begin{bmatrix} 1 \\ 2 \\ -5 \end{bmatrix} \quad V = [v_1 \quad v_2 \quad v_3].$$

(3) Form the matrix Σ, with diagonal elements in descending order:

$$\Sigma = \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix}.$$

(4) Find the first r columns of the matrix U using $u_i = \frac{1}{\sigma_i} A v_i, \quad i = 1, \ldots, r$:

$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad u_2 = \frac{1}{\sigma_2} A v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

(5) If $r < m$, compute the remaining $m - r$ columns of U using Gram–Schmidt.
This step is not necessary for this example, since $m = 2 = r$.

Thus, the SVD of the matrix A is

$$A = U \Sigma V^T = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & -\frac{5}{\sqrt{30}} \end{bmatrix}^T.$$

Several observations are in order.

Remark 3.16 *(1) The eigenvalues of $A^T A$ are real (and nonnegative). Thus, no complex arithmetic is involved.*

(2) From $A = U \Sigma V^T$, we can write $AV = U\Sigma$, which is used in step (4) of Example 3.22.

(3) Since the matrix $A^T A$ is symmetric, the existence of orthogonal eigenvectors is guaranteed (Exercise 1.71).

Through the SVD of an arbitrary matrix, we have computed not just any orthogonal matrices U and V. The following theorem states one of the main reasons why these matrices become very important in applications.

Theorem 3.17 *Let $A = U\Sigma V^T$ be a singular value decomposition of a matrix A of rank r. Then,*

- *the first r columns of U form an orthonormal basis of col(A), and*
- *the first r columns of V form an orthonormal basis of row(A).*

Proof. We have that rank$(A) = r$, and that $U^T AV = \Sigma$. The rank of the diagonal matrix Σ is the number of nonzero diagonal entries σ_i. But since orthogonal transformations preserve linear independence, then the first r singular values $\sigma_1, \ldots, \sigma_r$ are nonzero. From $AV = U\Sigma$, we have that

$$Av_i = \sigma_i u_i, \quad i = 1, \ldots, p, \quad p = \min\{m, n\}.$$

Thus, for $i = 1, \ldots, r$ $(r \leq p)$, we have $Av_i = \sigma_i u_i \neq 0$ and therefore $u_i \in \text{col}(A)$, which is r-dimensional. Hence, the vectors $\{u_1, \ldots, u_r\}$ form an orthonormal basis of col(A). Using the same idea for A^T, we get from $A^T U = V\Sigma^T$:

$$A^T u_i = \sigma_i v_i, \quad i = 1, \ldots, p,$$

so that for $i = 1, \ldots, r$, we have $A^T u_i = \sigma_i v_i \neq 0$ and $v_i \in \text{col}(A^T)$, which is also r-dimensional. Then, $\{v_1, \ldots, v_r\}$ form an orthonormal basis of col(A^T) =row(A).

□

Remark 3.18 *Theorem 3.17 implies that the last columns $u_{r+1}, \ldots, u_m$ of U form an orthonormal basis of $N(A^T)$, and that the columns $v_{r+1}, \ldots, v_n$ of V form an orthonormal basis of $N(A)$. See Exercise 3.33.*

Uniqueness. The singular values σ_i of A are uniquely determined as they are the p largest eigenvalues of AA^T or $A^T A$. However, the orthogonal matrices U and V are clearly not unique. Recall that the columns of V are eigenvectors of $A^T A$ and that these eigenvectors can be placed in any order to form the matrix V. This also implies the nonuniqueness of U, since $u_i = \frac{1}{\sigma_i} Av_i$, for $i = 1, \ldots, r$. Furthermore, if $r < m$, the remaining $m - r$ columns of U can be chosen arbitrarily to complete an orthogonal matrix U.

■ **EXAMPLE 3.23**

Consider again the matrix $A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$ of Example 3.22. By permuting the first and second eigenvectors of $A^T A$, we now can form a new orthogonal matrix $V = [v_1 \quad v_2 \quad v_3]$, where

$$v_1 = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \quad v_2 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad v_3 = \frac{1}{\sqrt{30}} \begin{bmatrix} 1 \\ 2 \\ -5 \end{bmatrix}.$$

The columns of U are determined by $u_i = \frac{1}{\sigma_i} A v_i$:

$$u_1 = \frac{1}{\sigma_1} A\, v_1 = \frac{1}{\sqrt{60}} \begin{bmatrix} 5 \\ -5 \end{bmatrix}, \quad u_2 = \frac{1}{\sigma_2} A\, v_1 = \frac{1}{\sqrt{60}} \begin{bmatrix} 6 \\ 6 \end{bmatrix}.$$

Then, a second SVD of A is

$$A = U \Sigma V^T = \begin{bmatrix} \frac{5}{\sqrt{60}} & \frac{6}{\sqrt{60}} \\ \frac{-5}{\sqrt{60}} & \frac{6}{\sqrt{60}} \end{bmatrix} \begin{bmatrix} \sqrt{12} & 0 & 0 \\ 0 & \sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{30}} \\ \frac{-1}{\sqrt{5}} & \frac{2}{\sqrt{6}} & \frac{2}{\sqrt{30}} \\ 0 & \frac{1}{\sqrt{6}} & -\frac{5}{\sqrt{30}} \end{bmatrix}^T.$$

We next state an important theorem that we will use later.

Theorem 3.19 *Let A be an $m \times n$ matrix. Then, rank$(A) = 1$ if and only if there exist two nonzero vectors $x \in \mathbb{R}^m$ and $y \in \mathbb{R}^n$ such that*

$$A = x\, y^T. \tag{3.29}$$

Proof. If rank$(A) = 1$, then the Σ matrix in the SVD of A has only one nonzero entry. That is,

$$\begin{aligned} A &= U \Sigma V^T = [u_1 \ \cdots \ u_m] \begin{bmatrix} \sigma_1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} [v_1 \ \cdots \ v_n]^T \\ &= [u_1 \ \cdots \ u_m] \begin{bmatrix} \sigma_1 v_{11} & \sigma_1 v_{21} & \cdots & \sigma_1 v_{n1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \\ &= [v_{11}(\sigma_1 u_1) \ \cdots \ v_{n1}(\sigma_1 u_1)] \\ &= (\sigma_1 u_1)\, v_1^T =: x\, y^T. \end{aligned}$$

Conversely, assume that $A = xy^T$. We can define two Householder matrices H_1 and H_2 as in (3.22) and (3.23) to zero all the entries but the first of the vectors x and y respectively. Thus, the vectors $H_1 x$ and $H_2 y$ are just multiples of the canonical vector e_1, and therefore the matrix $(H_1 x)(H_2 y)^T$ is obviously of rank 1. We also have that

$$(H_1 x)(H_2 y)^T = H_1 (xy^T) H_2^T = H_1 A\, H_2^T,$$

so that using Theorem 1.44, we conclude

$$1 = \text{rank}(H_1 A H_2^T) = \text{rank}(A H_2^T) = \text{rank}(A).$$

□

■ **EXAMPLE 3.24**

Let $A = \begin{bmatrix} 2 & 3 & 2 \\ 4 & 6 & 4 \\ 6 & 9 & 6 \end{bmatrix}$. This is clearly a rank-1 matrix. If we define $x = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $y = \begin{bmatrix} 2 \\ 3 \\ 2 \end{bmatrix}$. Then, $A = xy^T$.

From the proof of Theorem 3.19, we can also get two other vectors x and y that do the work. Indeed, let

$$x = \sigma_1 u_1 = \begin{bmatrix} -4.1231 \\ -8.2462 \\ -12.3693 \end{bmatrix}, \qquad y = v_1 = \begin{bmatrix} -0.4851 \\ -0.7276 \\ -0.4851 \end{bmatrix},$$

where u_1, v_1 are the first column of U and V, respectively, and σ_1 is the first entry of Σ in the SVD factorization $A = U\Sigma V^T$. Then, the reader can verify that again $A = xy^T$.

3.3.1 The complex case

The singular value decomposition of a matrix A exists in general for complex matrices. The statements of Theorem 3.15 remain for the most part unchanged. More precisely, we have the following theorem.

Theorem 3.20 *Given a complex matrix $A_{m\times n}$, there exist unitary matrices $U_{m\times m}$, $V_{n\times n}$ and a diagonal matrix $\Sigma_{m\times n}$ such that*

$$A = U\Sigma V^*, \tag{3.30}$$

and the following is true:

1. *The columns of U are eigenvectors of the matrix AA^*.*
2. *The columns of V are eigenvectors of the matrix A^*A.*
3. *The diagonal entries of Σ, the* **singular values** *of A, are the nonnegative square roots of the eigenvalues λ_i of A^*A (or AA^*); that is,*

$$\sigma_i = \sqrt{\lambda_i}, \quad i = 1, \ldots p, \quad p = \min\{m, n\},$$

arranged in descending order.

■ **EXAMPLE 3.25**

Let $A = \begin{bmatrix} 2 & 3i & 4 \\ 1+2i & -1 & 4 \end{bmatrix}$. Then its singular value decomposition $A = U\Sigma V^*$ gives

$$U = \begin{bmatrix} -0.7676 & 0.6410 \\ -0.5974 - 0.2323i & -0.7154 - 0.2782i \end{bmatrix}, \Sigma = \begin{bmatrix} 6.7177 & 0 & 0 \\ 0 & 2.4233 & 0 \end{bmatrix},$$

$$V = \begin{bmatrix} -0.3866 + 0.1433i & 0.0042 + 0.4756i & -0.4667 + 0.6212i \\ 0.0889 + 0.3774i & 0.2952 - 0.6787i & -0.5440 + 0.0773i \\ -0.8127 - 0.1383i & -0.1228 - 0.4592i & 0.2913 + 0.0974i \end{bmatrix}.$$

3.3.2 Low–rank approximations

A group of important and exciting applications of the SVD factorization of a matrix A originates from the fact that with such a factorization, it is possible to construct matrices A_k of rank k that approximate the original matrix A. And the best part is that among all possible rank-k approximations to A, the one provided by the SVD is the one with the smallest error, in the Frobenius matrix norm. There is a remarkable classical theorem that formalizes these ideas.

Theorem 3.21 *Let $A = U\Sigma V^T$ be the SVD of $A_{m\times n}$, and let*

$$A_k = U_k \Sigma_k V_k^T, \tag{3.31}$$

where U_k is the $m \times k$ matrix whose columns are the first k columns of U, V_k is the $n \times k$ matrix whose columns are the first k columns of V, and Σ_k is the $k \times k$ diagonal matrix whose diagonal elements are the k largest singular values of A. Then, A_k is a rank-k approximation of A and

$$\|A - A_k\|_F = \min_{\text{rank}(X)=k} \|A - X\|_F.$$

Remark 3.22 *Observe that given $A = U\Sigma V^T$, to find any rank k approximation A_k of A, no additional computation or factorization is needed; we just simply take the first k columns of U and V to get U_k and V_k, and the k largest singular values σ_i to get Σ_k.*

■ **EXAMPLE 3.26**

The matrix $A = \begin{bmatrix} 1 & 4 & 0 \\ 1 & 6 & 2 \\ 3 & 0 & 2 \\ 2 & 2 & 6 \end{bmatrix}$ has rank$(A) = 3$, and its SVD is $A = U\Sigma V^T$, where

$$U = \begin{bmatrix} -0.3519 & -0.4762 & 0.3338 & -0.7335 \\ -0.6444 & -0.4886 & -0.1038 & 0.5791 \\ -0.2430 & 0.4154 & 0.8454 & 0.2316 \\ -0.6340 & 0.6016 & -0.4038 & -0.2702 \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} 9.1035 & 0 & 0 \\ 0 & 5.2344 & 0 \\ 0 & 0 & 2.1745 \\ 0 & 0 & 0 \end{bmatrix}, V = \begin{bmatrix} -0.3288 & 0.2836 & 0.9008 \\ -0.7186 & -0.6941 & -0.0437 \\ -0.6128 & 0.6617 & -0.4320 \end{bmatrix}.$$

Then, the rank-2 approximation (3.31) is given by

$$A_2 = \begin{bmatrix} -0.3519 & -0.4762 \\ -0.6444 & -0.4886 \\ -0.2430 & 0.4154 \\ -0.6340 & 0.6016 \end{bmatrix} \begin{bmatrix} 9.1035 & 0 \\ 0 & 5.2344 \end{bmatrix} \begin{bmatrix} -0.3288 & 0.2836 \\ -0.7186 & -0.6941 \\ -0.6128 & 0.6617 \end{bmatrix}^T,$$

which gives

$$A_2 = \begin{bmatrix} 0.3462 & 4.0317 & 0.3136 \\ 1.2033 & 5.9901 & 1.9025 \\ 1.3440 & 0.0804 & 2.7942 \\ 2.7909 & 1.9616 & 5.6207 \end{bmatrix}.$$

From Theorem 3.21 we know that among all approximations to A of rank 2, this matrix A_2 is the closest to A in the Frobenius norm.

With an eye on some applications of low-rank approximations, e.g. ,compression, we introduce another way of writing the SVD of a given matrix A, in a form that is very illustrative about the importance and usefulness of such a factorization. The low-rank approximations given in (3.31) suggest this new approach.

If we denote each column of U as u_i, each column of V as v_i, and each singular value as σ_i, then we can have an alternative and extremely useful way to write the SVD factorization (3.27). More precisely, if rank$(A) = r$, we know that the last singular values satisfy

$$\sigma_{r+1} = \cdots = \sigma_p = 0,$$

where as always, $p = \min(m, n)$, and we have the following theorem.

Theorem 3.23 *The SVD factorization $A = U\Sigma V^T$ of a matrix $A_{m\times n}$ of rank r is given by*

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_r u_r v_r^T. \tag{3.32}$$

Proof. The factorization $A = U\Sigma V^T$ can be written as

$$\begin{aligned} A &= [u_1 \ \cdots \ u_m] \begin{bmatrix} \sigma_1 & & & & & \\ & \ddots & & & & \\ & & \sigma_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} \\ &= [u_1 \ \cdots \ u_m] \begin{bmatrix} \sigma_1 v_1^T \\ \vdots \\ \sigma_r v_r^T \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T. \end{aligned}$$

□

An immediate consequence of this new way of writing the SVD of a matrix A, is that the rank-k approximation (3.31) can now be expressed as

$$A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T, \qquad k \le r. \tag{3.33}$$

Thus, by comparing with (3.32), we see that (3.31) is just a truncation of the original SVD factorization of A and that the more terms we take (that is, the higher the rank of A_k), the better the approximation to A. At the same time, we also observe that if the singular values σ_i decrease rapidly as i increases, then an accurate approximation of A can be obtained by truncating the expansion after only a few terms. As a matter of fact, since the singular values satisfy

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0,$$

they are indicators of the importance of each term in the low–rank approximation A_k. The most important low rank approximations are obtained with the first few terms in (3.33) since they have the larger weights. As the index k increases, the corresponding terms in the expansion (3.33) become less and less important. This is of great utility for the applications we have in mind.

■ **EXAMPLE 3.27**

Consider the matrix A of Example 3.26. According to (3.33), its SVD rank-2 approximation can be obtained as $A_2 = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T$. That is,

$$A_2 = \sigma_1 \begin{bmatrix} -0.3519 \\ -0.6444 \\ -0.2430 \\ -0.6340 \end{bmatrix} \begin{bmatrix} -0.3288 \\ -0.7186 \\ -0.6128 \end{bmatrix}^T + \sigma_2 \begin{bmatrix} -0.4762 \\ -0.4886 \\ 0.4154 \\ 0.6016 \end{bmatrix} \begin{bmatrix} 0.2836 \\ -0.6941 \\ 0.6617 \end{bmatrix}^T,$$

where $\sigma_1 = 9.1035$ and $\sigma_2 = 5.2344$. This gives

$$A_2 = \begin{bmatrix} 0.3462 & 4.0317 & 0.3136 \\ 1.2033 & 5.9901 & 1.9025 \\ 1.3440 & 0.0804 & 2.7942 \\ 2.7909 & 1.9616 & 5.6207 \end{bmatrix},$$

which is of course what we obtained in Example 3.26.

3.3.3 SVD and spectral norm

Besides providing orthonormal bases for the subspaces

$$\text{col}(A),\ N(A),\ N(A^T) \text{ and } \text{col}(A^T),$$

the SVD of a matrix $A_{m\times n}$ also displays a matrix norm. Recall that for $p = 2$, the matrix norm (1.9) is

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2.$$

This norm coincides with the largest singular value of A. More precisely,

Theorem 3.24 *If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are the singular values of a matrix $A_{m\times n}$, then*

$$\sigma_1 = \|A\|_2.$$

Proof. Let $A = U\Sigma V^T$, and let $\text{rank}(A) = r$. Since the 2-norm is orthogonally invariant (Exercise 1.42), we have

$$\begin{aligned}
\|A\|_2 &= \|U\Sigma V^T\|_2 = \|\Sigma\|_2 = \max_{\|x\|_2=1} \|\Sigma x\|_2 = \max_{\|x\|_2=1} \left\| \begin{bmatrix} \sigma_1 x_1 \\ \vdots \\ \sigma_r x_r \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right\|_2 \\
&= \max_{\|x\|_2=1} \sqrt{\sigma_1^2 x_1^2 + \cdots + \sigma_r^2 x_r^2} \\
&\leq \max_{\|x\|_2=1} \sqrt{\sigma_1^2 (x_1^2 + \cdots + x_r^2)} \leq \sigma_1.
\end{aligned}$$

On the other hand, from $AV = U\Sigma$, we have in particular that $Av_1 = \sigma_1 u_1$, where as usual, u_1 and v_1 are the first column of U and V, respectively. Then,

$$\|Av_1\|_2 = \sigma_1 \|u_1\|_2 = \sigma_1.$$

This means that $\|A\|_2 \geq \sigma_1$, and the theorem is proved.

□

Let us recall that the singular values of a matrix A are the square roots of the eigenvalues of the symmetric matrix $A^T A$; that is, $\sigma_i = \sqrt{\lambda_i}$, where λ_i are the eigenvalues of $A^T A$. Therefore, the 2-norm of a matrix $A_{m\times n}$ can also be defined as

$$\|A\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^T A\}. \tag{3.34}$$

In this context, the 2-norm is also known as the **spectral norm** of A.

■ **EXAMPLE 3.28**

Let $A = \begin{bmatrix} 2 & 5 \\ 3 & 9 \\ 1 & 6 \end{bmatrix}$. Then, $\|A\|_2 = 12.4541$. At the same time, the matrix Σ in the factorization $A = U\Sigma V^T$ is

$$\Sigma = \begin{bmatrix} 12.4541 & 0 \\ 0 & 0.9467 \\ 0 & 0 \end{bmatrix}.$$

Therefore, $\sigma_1 = \|A\|_2$. Similarly, $A^T A = \begin{bmatrix} 14 & 43 \\ 43 & 142 \end{bmatrix}$. Its eigenvalues are $\lambda_1 = 155.1038$ and $\lambda_2 = 0.8962$. Thus, $\sqrt{\lambda_1} = 12.4541$, and $\sqrt{\lambda_2} = 0.9467$.

3.4 SCHUR FACTORIZATION

Although not as well–known as the previous matrix decompositions, Schur factorization is a very powerful and useful tool, which provides a way to compute and associate eigenvalues with orthogonal sets of vectors that span the corresponding invariant subspaces. This is, for example, an essential tool in boundary value problems and dynamical systems in general, when we are trying to compute orthonormal sets of vectors that span stable or unstable subspaces of an equilibrium point or a periodic orbit (see Section 7.2.3). Schur factorizations are also useful for solving algebraic Ricatti equations, in systems and control theory, and in general linear algebra and matrix computations. The Schur factorization exists for any square matrix A.

First, we give an important definition and a result that we will need to establish the Schur factorization of a matrix.

Definition 3.25 *We say that a subspace S of $\mathbb{R}^n$ is* **invariant** *under $A_{n\times n}$, or* **A-invariant** *if*

$$x \in S \quad \Longrightarrow \quad Ax \in S,$$

or equivalently, if $AS \subseteq S$.

■ **EXAMPLE 3.29**

The simplest examples of A-invariant subspaces are given by sets spanned by eigenvectors associated with given eigenvalues. For instance, if v is an eigenvector associated with a simple eigenvalue λ, then $Av = \lambda v$. If S is the line spanned by v, this means that Av is just a multiple of v and therefore also lies in S. More details on this are given in Section 3.4.2.

■ **EXAMPLE 3.30**

Let λ be an eigenvalue of A. For $k = 1, 2, \ldots$, the following sets are A-invariant

$$N(A - \lambda I)^k \qquad \text{and} \qquad \text{col}(A - \lambda I)^k.$$

In fact, from the definition of nullspace (1.45), if $x \in N(A - \lambda I)^k$, then we have $(A - \lambda I)^k x = 0$. Therefore, $(A - \lambda I)^k Ax = A(A - \lambda I)^k x = 0$, which means that $Ax \in N(A - \lambda I)^k$ and hence that $N(A - \lambda I)^k$ is A-invariant. The proof for the second set is left as an exercise.

The next theorem says that given an invariant subspace S of A, it is always possible to obtain a nonsingular matrix so that A is similar to a block upper triangular matrix, and

the leading diagonal block is associated with S. More precisely, we have the following theorem.

Theorem 3.26 *Let S be invariant under $A_{n\times n}$, with $dim(S) = r$. Then, there exists a nonsingular matrix $U_{n\times n}$ such that*

$$U^{-1}AU = T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix},$$

where T_{11} is square of order r.

Proof. Let $\mathcal{B} = \{u_1, \ldots, u_r\}$ be a basis of the subspace S. Since S is A-invariant, the vector Au_j lies in S, for any $j = 1, \ldots, r$, and therefore we can write it as the unique combination

$$Au_j = c_1u_1 + \cdots + c_ru_r,$$

for some scalars $c_1, \ldots, c_r$. On the other hand, we can extend the basis $\mathcal{B}$ with vectors $u_{r+1}, \ldots, u_n$ to obtain a basis of $\mathbb{R}^n$. Now define the matrix $U = [u_1 \cdots u_n]$ and $T = U^{-1}AU = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix}$, with T_{11} of order $r \times r$.

The goal is now to show that T_{21} is a zero matrix. We have that $AU = UT$. Columnwise, this equation can be written as

$$Au_j = \sum_{i=1}^{n} t_{ij}u_i,$$

which is also a unique combination. Therefore, we must have $c_i = t_{ij}$, for $i = 1, \ldots, r$, and consequently, $t_{ij} = 0$ for $i = r+1, \ldots, n$ and $j = 1, \ldots, r$. This implies the block T_{21} is a matrix of zeros.

□

■ **EXAMPLE 3.31**

Let $A = \begin{bmatrix} 79 & -64 & -2 \\ 8 & 13 & 35 \\ -20 & 8 & 7 \end{bmatrix}$, and denote with S the subspace spanned by the vector $[-0.9645 \quad 0.0219 \quad 0.2631]^T$. Then, S is A-invariant. Define the matrix $U = \begin{bmatrix} -0.945 & 1 & 0 \\ 0.0219 & 1 & 1 \\ 0.2631 & 0 & 0 \end{bmatrix}$. Then, we get

$$U^{-1}AU = \left[\begin{array}{c|cc} 81 & -45.6180 & 30.4120 \\ \hline 0 & -29 & -34.6667 \\ 0 & 51 & 42 \end{array}\right].$$

With the help of Theorem 3.26, we can now prove the existence of the Schur factorization of a real matrix.

Theorem 3.27 (Real Schur Form) *Let $A_{n\times n}$ be an arbitrary real matrix. Then, there exists an orthogonal matrix $Q_{n\times n}$, and a block upper triangular matrix T such that*

$$Q^T A Q = T = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1m} \\ 0 & T_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & T_{mm} \end{bmatrix}, \tag{3.35}$$

where each diagonal block T_{ii} is either a 1×1 or a 2×2 real matrix, the latter with a pair of complex conjugate eigenvalues. The diagonal blocks can be arranged in any prescribed order.

Proof. Let $\lambda_1, \ldots, \lambda_n$ be any ordering of the eigenvalues of A. The theorem is obviously true if $n = 1$ or if $n = 2$ and A has a pair of complex conjugate eigenvalues. To use induction, we assume that the statement is true for any $m < n$, and we have to show it is also true for $m = n$.

Let $\lambda = \lambda_1$ be real (and $m \geq 2$) with unit eigenvector u_1. Using orthonormal extensions, add vectors $u_2, \ldots, u_m$ to form an orthogonal matrix $U = [u_1 \ \cdots \ u_m]$. Since span$\{u_1\}$ is A-invariant, by Theorem 3.26, we have $B := U^T AU = \begin{bmatrix} b_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$, where $b_{11} = u_1^T A u_1 = u_1^T \lambda u_1 = \lambda(u_1^T u_1) = \lambda$. The eigenvalues of B_{22} are $\lambda_2, \ldots, \lambda_n$. By the induction hypothesis, we also know there exists an orthogonal matrix W of order $(m-1)\times(m-1)$ such that $W^T B_{22} W$ is block upper triangular, with its diagonal blocks in the prescribed order.

If we define $V =$diag$(1, W)$ and $Q = UV$, then we have

$$Q^T A Q = V^T U^T A U V = V^T B V = \left[\begin{array}{c|c} \lambda & * \\ \hline 0 & W^T B_{22} W \end{array}\right] =: T,$$

and T is block upper triangular.

We also need to consider the case of a complex eigenvalue.

Let $\lambda = a + ib$ be an eigenvalue of A (and $m \geq 3$), with eigenvector $v = v_1 + i\, v_2$. We can apply Gram–Schmidt to v_1, v_2 so that we can get two orthonormal vectors u_1, u_2 that span the same invariant subspace spanned by v_1 and v_2. As before, add vectors $u_3, \ldots, u_m$ to form an orthogonal matrix $U = [u_1 \ \cdots \ u_m]$. By Theorem 3.26, we have $B := U^T AU = \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix}$, where now B_{11} is 2×2. First let us show B_{11} has eigenvalues λ and $\bar{\lambda}$.

From $Av = \lambda v$, we get $Av_1 = av_1 - bv_2$ and $Av_2 = bv_1 + av_2$, or $AV = VS$, where $V = [v_1 \ v_2]$ and $S = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$. From the relationship between Gram–Schmidt and

QR factorization, we have $V = U_1 R$, with $U_1 = [u_1 \ u_2]$ and $R_{2\times 2}$ invertible, so that $AU_1 = AVR^{-1}$. Then,

$$B_{11} = U_1^T AU_1 = U_1^T AVR^{-1} = U_1^T VSR^{-1} = U_1^T U_1 RSR^{-1} = RSR^{-1}.$$

Thus, B_{11} has the same eigenvalues as S, that is, λ and $\bar{\lambda}$.

Now repeat the induction process as above; that is, there exists an orthogonal matrix W of order $(m-2)\times(m-2)$ such that $W^T B_{22} W$ is block upper triangular. If we define $X = \text{diag}(I_{2\times 2}, W)$ and $Q = UX$, then we have

$$Q^T AQ = X^T U^T AUX = X^T BX = \left[\begin{array}{c|c} B_{11} & * \\ \hline 0 & W^T B_{22} W \end{array}\right] =: T,$$

where T is now block upper triangular.

□

Remark 3.28 *An equivalent and useful way of writing the Schur factorization in (3.35) is*

$$AQ = QT. \tag{3.36}$$

Obviously, in the particular case when all the eigenvalues of the matrix A are real, we obtain an upper triangular matrix, but in general what we can get is a so-called upper quasi–triangular or block upper triangular matrix as in (3.35). In any case, we immediately see that (3.35) is a similarity transformation obtained by using orthogonal matrices and that hence a Schur factorization is an efficient way to obtain the eigenvalues of a matrix A.

■ **EXAMPLE 3.32**

Let $A = \begin{bmatrix} -2 & 4 & 0 & -1 \\ 8 & -1 & 0 & 4 \\ -2 & 1 & 1 & 9 \\ 7 & 7 & -1 & -1 \end{bmatrix}$. The matrix A has two real and a pair of complex conjugate eigenvalues:

$$\lambda_1 = 6.1429,\ \lambda_2 = -7.9078,\ \lambda_{3,4} = -0.6175 \pm 1.7365\text{i}.$$

The Schur factorization gives

$$Q^T AQ = T = \left[\begin{array}{r|rr|r} 6.1429 & 2.3027 & 7.1693 & -2.8451 \\ \hline 0 & -0.6175 & -0.2731 & 0.9995 \\ 0 & 11.0418 & -0.6175 & 0.4232 \\ \hline 0 & 0 & 0 & -7.9078 \end{array}\right], \tag{3.37}$$

with $Q = \begin{bmatrix} -0.1479 & -0.8513 & -0.0239 & 0.5029 \\ -0.4103 & -0.3567 & -0.3903 & -0.7430 \\ -0.7867 & 0.1490 & 0.5970 & 0.0493 \\ -0.4368 & 0.3548 & -0.7005 & 0.4389 \end{bmatrix}$.

Observe that the first and last diagonal blocks of R in (3.37) are 1×1 blocks containing the real eigenvalues, whereas the eigenvalues of the second diagonal block $T_{22} := \begin{bmatrix} -0.6175 & -0.2731 \\ 11.0418 & -0.6175 \end{bmatrix}$ are the complex conjugate eigenvalues of A. Also observe in this example that the first column q_1 of Q is a unit eigenvector of A corresponding to the eigenvalue $\lambda_1 = 6.1429$; that is, $Aq_1 = \lambda_1 q_1$.

■ **EXAMPLE 3.33**

The matrix $A = \begin{bmatrix} 9 & -1 & 4 \\ 7 & 6 & 8 \\ -5 & 5 & -2 \end{bmatrix}$ has only real eigenvalues: $\lambda_1 = -3.0428$, $\lambda_2 = 5.8934$, and $\lambda_3 = 10.1494$. In this case, the Schur factorization is

$$Q^T A Q = T = \begin{bmatrix} -3.0428 & 6.9273 & 9.8058 \\ 0 & 5.8934 & -3.1398 \\ 0 & 0 & 10.1494 \end{bmatrix}, \tag{3.38}$$

with $Q = \begin{bmatrix} 0.3119 & 0.6770 & 0.6666 \\ 0.4826 & -0.7173 & 0.5026 \\ -0.8184 & -0.1649 & 0.5504 \end{bmatrix}$. Thus, in this case, we get an upper triangular matrix T, with the eigenvalues of A on its diagonal.

Remark 3.29 *A direct consequence of the Schur factorization of a symmetric matrix $A_{n\times n}$ is that A is similar to a diagonal matrix. See Exercise 3.57.*

Uniqueness. It should be clear from the presentation above, and in particular from the proof of Theorem 3.27, that a Schur factorization of a matrix A is not unique. First, given a unit eigenvector u_1 corresponding to a real eigenvalue λ, there is more than one way to extend u_1 to form an orthogonal matrix U. And second, the diagonal blocks in (3.35) can be arranged in any prescribed order.

It is also possible to arrange the blocks in (3.35) to form a compact form of the factorization, to obtain what is known as **ordered block Schur factorization**:

Corollary 3.4.1 *Let $A_{n\times n}$ be an arbitrary real matrix. Then, there exists an orthogonal matrix $Q_{n\times n}$, and a block upper triangular matrix T such that*

$$Q^T A Q = T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \tag{3.39}$$

where T_{11} is $m \times m$ and T_{22} is $(n-m) \times (n-m)$, for some positive integer m. The diagonal blocks can be arranged in any prescribed order.

■ **EXAMPLE 3.34**

$$\text{Let } A = \begin{bmatrix} 3 & 5 & 5 & -5 & 3 \\ -19 & -25 & -15 & 3 & -13 \\ 41 & 41 & 19 & -9 & 13 \\ 9 & 19 & 5 & 3 & 17 \\ 14 & 18 & 10 & 2 & 8 \end{bmatrix}. \text{ Then, with}$$

$$Q = \begin{bmatrix} -0.5970 & -0.2784 & -0.6892 & 0.2441 & -0.1775 \\ 0.6692 & 0.0230 & -0.3656 & 0.1792 & -0.6211 \\ 0.0753 & -0.1647 & -0.3272 & -0.9274 & -0.0000 \\ -0.1062 & -0.7421 & 0.4886 & -0.0492 & -0.4437 \\ -0.4228 & 0.5866 & 0.2135 & -0.2138 & -0.6211 \end{bmatrix}, \text{ we get}$$

$$Q^T A Q = \left[\begin{array}{cc|ccc} -2 & -3.7790 & 25.9596 & 8.9818 & 27.2895 \\ 0 & -6 & 8.3761 & 1.0333 & 14.8848 \\ \hline 0 & 0 & 4 & -1.8420 & -11.5886 \\ 0 & 0 & 54.2893 & 4 & 43.3662 \\ 0 & 0 & 0 & 0 & 8 \end{array}\right] = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}.$$

The eigenvalues of A are $\lambda = 8, -2, -6, 4 \pm 10i$. With this ordered block Schur factorization, we have obtained a block triangular matrix, where the block T_{11} contains the eigenvalues of A with negative real part, and the block T_{22} contains the eigenvalues of A with positive real part.

In addition (as we will see in general later on), the first two columns of Q form an orthonormal basis of the vector space spanned by the eigenvectors associated with the two eigenvalues of A with negative real part.

For most applications, in particular within the frame of numerical computations, the block upper triangular form (3.39) is not just good enough, but it is in fact a more efficient and useful arrangement of eigenvalues than that in (3.35), as we will see in Section 7.2.3.

- MATLAB commands: schur(A), ordschur(A, T, select)

3.4.1 The complex case

In a certain sense, the complex version of a Schur factorization of a matrix is simpler than the real case, for if complex objects are allowed, then there is no need for 2×2 diagonal blocks in (3.35). We have the following theorem.

Theorem 3.30 *Let $A_{n\times n}$ be an arbitrary complex matrix. Then, there exists a unitary matrix $Q_{n\times n}$, and an upper triangular complex matrix T such that*

$$Q^*AQ = T = \begin{bmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ 0 & \lambda_2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}, \tag{3.40}$$

where the eigenvalues of A, $\lambda_1, \ldots, \lambda_n$ can be arranged in any prescribed order.

■ **EXAMPLE 3.35**

The matrix $A = \begin{bmatrix} 1 & 3 & 2+i \\ 1-2i & 3-i & 2 \\ 4 & 1 & -1 \end{bmatrix}$ has eigenvalues: $\lambda_1 = -2.6974 - 0.5723\,i$, $\lambda_2 = 0.1260 + 0.7641i$, $\lambda_3 = 5.5713 - 1.1918i$. The Schur factorization gives

$$Q^*AQ = \begin{bmatrix} -2.6974 - 0.5723i & 0.8207 - 2.4487i & 0.5511 + 1.5238i \\ 0 & 0.1260 + 0.7641i & -0.7153 + 1.2368i \\ 0 & 0 & 5.5713 - 1.1918i \end{bmatrix}, \tag{3.41}$$

with $Q = \begin{bmatrix} 0.3453 + 0.0116i & -0.0647 + 0.7744i & -0.5170 + 0.0971i \\ 0.2383 + 0.0544i & -0.0139 - 0.6119i & -0.7054 + 0.2610i \\ -0.8749 + 0.2355i & -0.0179 + 0.1456i & -0.3129 + 0.2444i \end{bmatrix}$.

As usual, the first column of Q is a unit eigenvector of A associated with the complex eigenvalue λ_1.

3.4.2 Schur factorization and invariant subspaces

We now turn back to eigenvalues and eigenvectors to find the simplest examples of subspaces that are invariant under a matrix $A_{n\times n}$. The equation that defines an eigenvalue λ of a matrix A, with corresponding eigenvector $v \neq 0$, is

$$Av = \lambda v. \tag{3.42}$$

What this definition says is that when v is multiplied by A, what we get is a new vector, λv, which is just a multiple of v. More precisely, if v lies on a certain line, then $Av = \lambda v$ lies on the same line. That line is actually a one-dimensional vector subspace W spanned by v, and we say that this subspace W is *invariant* under A.

We have also seen in the proof of Theorem 3.27 that if $\lambda = a \pm ib$ are complex conjugate eigenvalues of A with eigenvectors $v = v_1 \pm iv_2$, then the two-dimensional set spanned by

v_1, v_2 is invariant under A. More precisely, we have (see the equation $AV = VS$ in the proof of Theorem 3.27):

$$A[v_1 \quad v_2] = [v_1 \quad v_2] \begin{bmatrix} a & b \\ -b & a \end{bmatrix}. \tag{3.43}$$

Since the product on the right–hand side of (3.43) is just combinations of the vectors v_1 and v_2, we can conclude that the set spanned by v_1 and v_2 is invariant under A.

On the other hand, consider any two linearly independent vectors q_1, q_2 that span a two-dimensional subspace W, and assume that this subspace is invariant under A. Then, there exist scalars $r_{11}, r_{12}, r_{21}, r_{22}$ such that

$$\begin{aligned} Aq_1 &= r_{11}q_1 + r_{21}q_2, \\ Aq_2 &= r_{12}q_1 + r_{22}q_2. \end{aligned} \tag{3.44}$$

Equation (3.44) can also be written as

$$[Aq_1 \quad Aq_2] = [q_1 \quad q_2] \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix},$$

or even better, as

$$AQ = QR, \tag{3.45}$$

where $Q = [q_1 \quad q_2]$ and $R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$.

Thus, from (3.42), (3.43) and (3.45) we realize that eigenvalues, invariant subspaces, and Schur factorization are all closely related. We want to give some more details on this important fact. We start by giving a theorem that generalizes the present discussion.

Theorem 3.31 *Let W be an invariant subspace under a matrix $A_{n\times n}$, and let $\mathcal{B} = \{v_1, \ldots, v_k\}$ be a basis for W. Then, there is a* unique *matrix R such that*

$$AV = VR, \tag{3.46}$$

where $V = [v_1 \ \cdots \ v_k]$.

Proof. First define the matrix $Y = AV$, partitioned as $Y = [y_1 \ \cdots \ y_k]$. Obviously we have that for each $i = 1, \ldots, k$, the vector y_i lies in W, and since $\mathcal{B}$ is a basis of W, we can write each y_i as a unique linear combination

$$y_i = r_{i1}v_1 + r_{i2}v_2 + \cdots + r_{ik}v_k = [v_1 \ v_2 \ \cdots \ v_k] \begin{bmatrix} r_{i1} \\ r_{i2} \\ \vdots \\ r_{ik} \end{bmatrix}.$$

If we define the matrix $R = \begin{bmatrix} r_{11} & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & r_{kk} \end{bmatrix}$, then we have that $Y = VR$, and therefore, $AV = VR$.

□

The following example illustrates the previous theorem.

■ **EXAMPLE 3.36**

Let $A = \begin{bmatrix} 3 & -2 & 5 \\ 4 & -1 & 4 \\ 0 & 0 & 9 \end{bmatrix}$. For the eigenvalue $\lambda = 1 + 2i$, the associated eigenvector is $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} i$. This indicates that the two vectors $v_1 = [1 \quad 1 \quad 0]^T$ and $v_2 = [0 \quad 1 \quad 0]^T$ form a basis of the corresponding two-dimensional A-invariant subspace. Now define $V = [v_1 \quad v_2]$. Then we get $AV = \begin{bmatrix} 1 & -2 \\ 3 & -1 \\ 0 & 0 \end{bmatrix} = [y_1 \quad y_2]$. Now we can write the unique linear combinations

$$y_1 = v_1 + 2v_2, \qquad y_2 = -2v_1 + v_2.$$

If we let $R = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$, then, in matrix form, the two equations above can be written as $Y = VR$ and therefore $AV = VR$. The matrix R is uniquely determined.

Now let us go back to equation (3.39) of Corollary 3.4.1. If we block the orthogonal matrix Q according to the dimensions of T; that is, if $Q = [Q_1 \quad Q_2]$, where Q_1 is $n \times m$ and Q_2 is $n \times (n - m)$, then from the block Schur factorization $Q^T A Q = T$, we can write $AQ = QT$, and this in turn can now be written as

$$[AQ_1 \quad AQ_2] = [Q_1 \quad Q_2] \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}.$$

Thus, by equating the first block, we have

$$A Q_1 = Q_1 T_{11}. \tag{3.47}$$

Equation (3.47) makes Schur a very useful matrix factorization. Let us explain why. First, this equation is of the same type as (3.45), that is, it expresses the fact that the subspace spanned by the columns of Q_1, that is, $W = \text{span}\{q_1, \ldots, q_m\}$, is invariant under A. Second, recall from Remark 1.66 in Chapter 1 that the eigenvalues of R (and by similarity, those of A) are partitioned in two disjoint sets, that is,

$$\sigma(A) = \sigma(T_{11}) \cup \sigma(T_{22}).$$

Referring back to (3.47), this means that the invariant subspace S spanned by the columns of Q_1 is associated with the eigenvalues of A that are clustered as the eigenvalues of T_{11}.

And since any of the blocks T_{11} or T_{22}, containing the eigenvalues of A, can be placed in the upper left corner (Theorem 3.36), we conclude that

The Schur factorization provides with a very efficient way to compute a set of orthonormal vectors that span the eigenspace associated with a predetermined set of eigenvalues.

This is of great importance in applications. More formally, we have just proved the following theorem.

Theorem 3.32 *Given the ordered block Schur factorization*

$$Q^T A Q = T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \tag{3.48}$$

and given the blocking of $Q = [Q_1 \quad Q_2]$ *according to the dimensions of* T, *the columns of* Q_1 *form an orthonormal basis of the subspace associated with the eigenvalues of* A, *which are the eigenvalues of* T_{11}.

Remark 3.33 *We cannot overstress the importance of* (3.48) *in Theorem 3.32. Given a set of eigenvalues, the associated eigenvectors span the same subspace as the one spanned by the columns of* Q_1. *The main point is that such eigenvectors could be almost linearly dependent, while the columns of* Q_1 *are orthonormal. This provides us with a basis of such subspace that is by far more stable, and can be reliably used in numerical computations.*

■ **EXAMPLE 3.37**

The eigenvalues of the matrix

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

are $\sigma(A) = \{3,\ 6.4828,\ -3.9343,\ -2.5485\}$. Suppose we are interested in the negative eigenvalues of A. The corresponding (ordered) block Schur factorization (3.48) gives

$$Q^T A Q = T = \left[\begin{array}{rr|rr} -3.9343 & 1.5086 & -2.7042 & -4.3433 \\ 0 & -2.5485 & -3.4171 & -0.7199 \\ \hline 0 & 0 & 6.4828 & -0.5932 \\ 0 & 0 & 0 & 3.0000 \end{array}\right], \tag{3.49}$$

$$\text{with } Q = [Q_1 \quad Q_2] = \left[\begin{array}{rr|rr} 0.1854 & 0.2744 & -0.9311 & 0.1533 \\ -0.9470 & 0.0217 & -0.1342 & 0.2910 \\ -0.2599 & 0.2449 & -0.1318 & -0.9247 \\ 0.0359 & 0.9297 & 0.3126 & 0.1916 \end{array}\right].$$

The negative eigenvalues of A are the eigenvalues of the upper left block of T, and the subspace associated with such eigenvalues is spanned by the (orthonormal) columns of Q_1.

Remark 3.34 *In general, the matrix T in a Schur factorization $Q^T AQ = T$ of a real matrix A is triangular only when the all eigenvalues of A are real. This is the case of Example 3.37. This is also why the blocks T_{11} and T_{22} are triangular.*

■ **EXAMPLE 3.38**

The eigenvalues of the matrix

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

are $\sigma(A) = \{7.6651,\ -0.9137,\ -3.4587,\ 3.3537 \pm 4.2453\,\mathrm{i}\}$. Its (ordered) block Schur factorization (3.48) is

$$Q^T A Q = T = \left[\begin{array}{rrr|rr} 7.6651 & -1.8265 & -1.2204 & 1.0638 & -0.7142 \\ 0 & 3.3537 & 10.7330 & 1.7211 & 2.7507 \\ 0 & -1.6792 & 3.3537 & -2.8273 & 1.6277 \\ \hline 0 & 0 & 0 & -0.9137 & -0.5460 \\ 0 & 0 & 0 & 0 & -3.4587 \end{array}\right], \tag{3.50}$$

$$\text{with } Q = [Q_1 \quad Q_2] = \left[\begin{array}{rrr|rr} 0.7306 & 0.5790 & -0.1721 & -0.3178 & 0.0188 \\ -0.1312 & -0.2196 & -0.8666 & -0.2528 & -0.3459 \\ 0.2673 & -0.5592 & 0.3888 & -0.6302 & -0.2600 \\ 0.1768 & -0.3795 & -0.2480 & -0.0994 & 0.8680 \\ 0.5884 & -0.3998 & -0.0816 & 0.6543 & -0.2431 \end{array}\right].$$

This Schur factorization has been ordered in such a way to put the eigenvalues of A with a positive real part as the eigenvalues of the 3×3 upper left block of T. The corresponding subspace is spanned by the columns of Q_1.

Next we want to see how (1×1) and (2×2) blocks along the diagonal of matrix can be permuted through orthogonal transformations, without affecting the eigenvalues of the given matrix. As a matter of fact, in the Schur factorization of a matrix, we will be interested

in exchanging the positions of the eigenvalues, which will be located on the diagonal blocks of a given matrix.

3.4.3 Exchanging eigenblocks

Here we present a general technique to exchange diagonal blocks of a given matrix. This algorithm has proved very useful in matrix computations, for instance, in the numerical computation of invariant subspaces. More than the eigenblocks themselves, it is the eigenvalues they contain that we want to move around. We start with the following definition.

Definition 3.35 *If λ is an eigenvalue of a matrix $A_{n\times n}$ with corresponding eigenvector v, we say that the pair (λ, v) is an* **eigenpair** *of A. More generally, for $R_{k\times k}$ and $V_{n\times k}$, we say that (R, V) is an eigenpair of A of order k if*

(i) V has full rank k, and
(ii) $AV = VR$.

The matrix V is called an **eigenbasis** *and the matrix R an* **eigenblock**.

Theorem 3.36 *Consider the matrix*

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & r_{22} \end{bmatrix},$$

where R_{11} is either 1×1 or 2×2. Let q be a unit eigenvector of R corresponding to r_{22}, and let $Q = [q \quad Y]$ be orthogonal. Then

$$Q^T R Q = \begin{bmatrix} r_{22} & \tilde{R}_{12}^T \\ 0 & \tilde{R}_{11} \end{bmatrix},$$

where $\tilde{R}_{11}$ has the same eigenvalues as R_{11}.

Proof. First observe that the matrix R has one of the two forms

$$\begin{bmatrix} * & * \\ 0 & r_{22} \end{bmatrix}, \qquad \text{or} \qquad \left[\begin{array}{cc|c} * & * & * \\ * & * & * \\ \hline 0 & 0 & r_{22} \end{array}\right].$$

Performing matrix multiplication and using orthogonality, we get

$$\begin{aligned} Q^T R Q &= \begin{bmatrix} q^T \\ Y^T \end{bmatrix} [Rq \quad RY] = \begin{bmatrix} q^T R q & q^T R Y \\ Y^T R q & Y^T R Y \end{bmatrix} \\ &= \begin{bmatrix} r_{22} q^T q & q^T R Y \\ r_{22} Y^T q & Y^T R Y \end{bmatrix} = \begin{bmatrix} r_{22} & q^T R Y \\ 0 & Y^T R Y \end{bmatrix}. \end{aligned}$$

Because of similarity, the matrix $Q^T RQ$ has the same eigenvalues as R, and therefore the matrix $\tilde{R}_{11} = Y^T RY$ has the same eigenvalues as R_{11}.

□

■ **EXAMPLE 3.39**

Let

$$R = \left[\begin{array}{cc|c} 4 & -2 & 1 \\ 9 & -2 & 5 \\ \hline 0 & 0 & -6 \end{array}\right] = \begin{bmatrix} R_{11} & R_{12} \\ 0 & r_{22} \end{bmatrix}.$$

The eigenvalues of R are $\lambda = 1 \pm 3i$, -6, and $q = [0.1934 \quad 0.5663 \quad -0.8012]^T$ is a unit eigenvector of R associated to $\lambda = -6$. By using orthonormal extensions, we can define the orthogonal matrix

$$Q = \left[\begin{array}{r|rr} 0.1934 & 0.5663 & -0.8012 \\ 0.5663 & 0.6024 & 0.5625 \\ -0.8012 & 0.5625 & 0.2043 \end{array}\right].$$

Then, we get

$$Q^T RQ = \left[\begin{array}{r|rr} -6 & 6.8151 & -3.9583 \\ \hline 0 & 3.0595 & -7.4316 \\ 0 & 1.7818 & -1.0595 \end{array}\right],$$

where the lower right block has eigenvalues $\lambda = 1 \pm 3i$. Thus, the eigenvalues of R have effectively interchanged positions on the diagonal blocks.

The theorem above shows how to perform orthogonal similarity transformations so that the eigenvalues contained in the diagonal blocks interchange positions.

Now we want to prove a similar result but with the dimension of the lower right block being 2×2.

Theorem 3.37 *Consider the matrix*

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix},$$

where R_{11} is either 1×1 or 2×2 and $R_{2\times 2}$ is order 2×2. The diagonal eigenblocks can be interchanged by orthogonal transformations.

Proof. We have a matrix R in one of these two cases

$$\left[\begin{array}{c|cc} r_{11} & * & * \\ \hline 0 & * & * \\ 0 & * & * \end{array}\right], \qquad \text{or} \qquad \left[\begin{array}{cc|cc} * & * & * & * \\ * & * & * & * \\ \hline 0 & 0 & * & * \\ 0 & 0 & * & * \end{array}\right].$$

The first case is treated similarly as in Theorem 3.36: Let q be a unit left eigenvector of R corresponding to r_{11}, that is, $q^T R = r_{11} q^T$, and construct the matrix $Q = [X \quad q]$ to be orthogonal. Then, using orthogonality, we get

$$\begin{aligned} Q^T R Q &= \begin{bmatrix} X^T \\ q^T \end{bmatrix} [RX \quad Rq] = \begin{bmatrix} X^T RX & X^T Rq \\ q^T RX & q^T Rq \end{bmatrix} \\ &= \begin{bmatrix} X^T RX & X^T Rq \\ r_{11} q^T X & r_{11} q^T q \end{bmatrix} = \begin{bmatrix} X^T RX & X^T Rq \\ 0 & r_{11} \end{bmatrix}. \end{aligned}$$

By similarity, the matrix $\tilde{R}_{22} = X^T RX$ has the same eigenvalues as R_{22}.

The second case is treated a little differently. Let (R_{22}, X) be an eigenpair of R, with X orthogonal. That is, $RX = XR_{22}$. As before, extend X so that the matrix $Q = [X \quad Y]$ is also orthogonal. Then,

$$\begin{aligned} Q^T R Q &= \begin{bmatrix} X^T \\ Y^T \end{bmatrix} [RX \quad RY] = \begin{bmatrix} X^T RX & X^T RY \\ Y^T RX & Y^T RY \end{bmatrix} \\ &= \begin{bmatrix} X^T X R_{22} & X^T RY \\ Y^T X R_{22} & Y^T RY \end{bmatrix} = \begin{bmatrix} R_{22} & X^T RY \\ 0 & Y^T RY \end{bmatrix}. \end{aligned}$$

Again, because of similarity, the matrix $\tilde{R}_{11} = Y^T RY$ has the same eigenvalues as R_{11}.

□

■ **EXAMPLE 3.40**

Consider the matrix

$$R = \left[\begin{array}{cc|cc} -4 & 1 & 3 & 5 \\ -1 & -4 & 6 & 1 \\ \hline 0 & 0 & 1 & 0.6612 \\ 0 & 0 & -6.0495 & 1 \end{array}\right].$$

This matrix has eigenvalues $\lambda_{1,2} = -4 \pm i$ and $\lambda_{3,4} = 1 \pm 2i$. The pair (R_{22}, X), where

$$R_{22} = \begin{bmatrix} 1 & 0.6612 \\ -6.0495 & 1 \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} -0.4531 & -0.7109 \\ 0.4939 & -0.3698 \\ 0.7306 & -0.2846 \\ -0.1301 & -0.5261 \end{bmatrix},$$

is an eigenpair of R, with X orthogonal. By using e.g., Gram–Schmidt, we extend X to form an orthogonal matrix $Q = [X \quad Y]$, where

$$Y = \begin{bmatrix} 0.4330 & -0.3190 \\ -0.5234 & -0.5877 \\ 0.5324 & 0.3190 \\ -0.5051 & 0.6716 \end{bmatrix}.$$

Then, we get

$$Q^T R Q = \left[\begin{array}{cc|cc} 1.0000 & 0.6612 & 5.1531 & -0.0323 \\ -6.0495 & 1.0000 & -1.4731 & -5.7091 \\ \hline 0.0000 & -0.0000 & -3.5980 & -1.0949 \\ 0.0000 & -0.0000 & 1.0610 & -4.4020 \end{array}\right].$$

The lower right block of R has been moved to the upper left position. The eigenvalues of the lower right block of $Q^T R Q$ are $-4 \pm i$.

Remark 3.38 *One way to obtain an eigenpair (R_{22}, X) is to solve a certain Sylvester equation for X (see Exercise 3.73). The matrix X can then be orthonormalized.*

In all the similarity operations performed in the theorems above, we have used orthogonal matrices. The transformations had the form

$$Q^T R Q = T,$$

which is in fact, an ordered Schur factorization of the matrix R. The applicability of these factorizations and the ability to move eigenblocks will be crucial in some problems in dynamical systems studied in Chapter 7.

3.4.4 Block diagonalization

In the preceding discussion we have been able to transform a given matrix A into a block triangular matrix

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \tag{3.51}$$

using a Schur factorization of A. Now we would like to take it one step further and transform T into a block diagonal matrix.

A central tool for such a transformation is a very important matrix equation that is useful in several other instances, such as finding eigenpairs of a given matrix.

Given the matrices $A_{m\times m}$, $B_{n\times n}$, and $C_{m\times n}$, the equation

$$AX - XB = C \tag{3.52}$$

is known as the **Sylvester's equation**.

Theorem 3.39 *The Sylvester's equation (3.52) has a unique solution $X_{m\times n}$ if and only if the matrices A and B have no eigenvalues in common; that is*

$$\sigma(A) \cap \sigma(B) = \emptyset.$$

■ **EXAMPLE 3.41**

Let

$$A = \begin{bmatrix} 4 & 5 \\ 3 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & -2 & 1 \\ 9 & -2 & 5 \\ 0 & 0 & -6 \end{bmatrix}, \quad C = \begin{bmatrix} -10 & 7 & 24 \\ -2 & 0 & 11 \end{bmatrix}.$$

The eigenvalues of A are $\lambda = 7, -1$, and the eigenvalues of B are $\lambda = -6, 1 \pm 3i$. Then, the Sylvester equation $AX - XB = C$ has the unique solution

$$X = \begin{bmatrix} 1 & 0 & 2 \\ -2 & 1 & 1 \end{bmatrix}.$$

Remark 3.40 *It is very important to note that Theorem 3.39 generalizes the uniqueness of solutions of linear systems of equations $Ax = b$. In fact, in this latter case, B is the scalar zero ($n = 1$), and $C = b$. Theorem 3.39 says that the system $Ax = b$ has a unique solution if and only if A does not have any zero eigenvalues.*

Now suppose we want to transform the matrix T in (3.51) into block diagonal form. Since we must preserve the zeros in the lower left block, we define the square matrix

$$U = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix},$$

for some $X_{m \times n}$. This matrix U is obviously nonsingular. Then,

$$\begin{aligned} U^{-1}TU &= \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} I & X \\ 0 & I \end{bmatrix} \\ &= \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix} \begin{bmatrix} T_{11} & T_{11}X + T_{12} \\ 0 & T_{22} \end{bmatrix} \\ &= \begin{bmatrix} T_{11} & T_{11}X - XT_{22} + T_{12} \\ 0 & T_{22} \end{bmatrix}. \end{aligned}$$

Thus, the matrix $U^{-1}TU$ is block diagonal if

$$T_{11}X - XT_{22} = -T_{12},$$

which is a Sylvester's equation on X.

Hence, we have just proved the following theorem.

Theorem 3.41 *Consider the matrix T in (3.51), and assume that*

$$\sigma(T_{11}) \cap \sigma(T_{22}) = \emptyset.$$

Then, there exists a unique matrix $X_{m\times n}$ that satisfies

$$T_{11}X - XT_{22} = -T_{12}, \tag{3.53}$$

such that

$$U^{-1}TU = \begin{bmatrix} T_{11} & 0 \\ 0 & T_{22} \end{bmatrix},$$

where $U = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$.

■ **EXAMPLE 3.42**

Consider the matrix

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} = \begin{bmatrix} 3 & -2 & 8 & 1 \\ 4 & -1 & 3 & 2 \\ 0 & 0 & -5 & 1 \\ 0 & 0 & -1 & -5 \end{bmatrix}.$$

Since the eigenvalues of T_{11} are $\lambda_{1,2} = 1 \pm 2i$, and the eigenvalues of T_{22} are $\lambda_{3,4} = -5 \pm i$, we have $\sigma(T_{11}) \cap \sigma(T_{22}) = \emptyset$. Then the Sylvester's equation (3.53) has the unique solution

$$X = \begin{bmatrix} -0.9135 & -0.2811 \\ 0.2054 & -0.1676 \end{bmatrix}.$$

If we define $U = \begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$, then,

$$U^{-1}TU = \begin{bmatrix} 3 & -2 & 0 & 0 \\ 4 & -1 & 0 & 0 \\ 0 & 0 & -5 & 1 \\ 0 & 0 & -1 & -5 \end{bmatrix}.$$

The central requirement in Theorem 3.41 is to have two disjoint sets of eigenvalues of T. The good news is that this can be generalized to the case when the matrix has more than two disjoint sets of eigenvalues.

Suppose that the eigenvalues of A can be partitioned into the union

$$\sigma_1 \cup \sigma_2 \cup \cdots \cup \sigma_k$$

of disjoint sets, each containing a number of eigenvalues. Using Schur factorization, we can get

$$Q^TAQ = T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & B \end{bmatrix}, \tag{3.54}$$

where $\sigma(T_{11}) = \sigma_1$. By using Theorem 3.41, we can block diagonalize T to get $\begin{bmatrix} T_{11} & 0 \\ 0 & B \end{bmatrix}$. We can apply the same process to the matrix B to pull out a block T_{22} so that $\sigma(T_{22}) = \sigma_2$. Following this in a recursive way, we obtain the following result.

Theorem 3.42 *Suppose the eigenvalues of a matrix $A_{n\times n}$ can be partitioned into the union $\sigma(A) = \sigma_1 \cup \sigma_2 \cup \cdots \cup \sigma_k$ of disjoint multisets. Then, there is a nonsingular matrix U such that*

$$U^{-1}AU = \begin{bmatrix} T_{11} & & & \\ & T_{22} & & \\ & & \ddots & \\ & & & T_{kk} \end{bmatrix}, \tag{3.55}$$

where $\sigma(T_{ii}) = \sigma_i$.

■ **EXAMPLE 3.43**

Consider the matrix

$$A = \frac{1}{15}\begin{bmatrix} -3 & 24 & -24 & -1 & -24 \\ 0 & 12 & 36 & 6 & 36 \\ -78 & 48 & -15 & -40 & -51 \\ -33 & 24 & 0 & -3 & 0 \\ 45 & -24 & 0 & 39 & 36 \\ 39 & -24 & 24 & 1 & 0 \\ 33 & -24 & 24 & 1 & 24 \end{bmatrix}.$$

We apply an ordered Schur factorization

$$Q_1^T A Q_1 = T_1 = \begin{bmatrix} T_{11} & T_{12} \\ 0 & B \end{bmatrix}$$

as in (3.54) to locate an upper left block containing, say, the eigenvalues of A with negative real part:

$$T_1 = \left[\begin{array}{cc|ccccc} -0.2000 & 2.0443 & 3.7672 & -6.2350 & 4.4535 & -2.2979 & 7.6028 \\ 0 & -0.1333 & -1.2975 & 3.0465 & -0.2244 & -0.5976 & -1.5779 \\ \hline 0 & 0 & 2.0000 & -3.1500 & 0.3116 & -0.6252 & 2.0081 \\ 0 & 0 & 0 & 1.6000 & -0.3553 & -0.6672 & 0.8544 \\ 0 & 0 & 0 & 1.8012 & 1.6000 & 0.3548 & -1.1839 \\ 0 & 0 & 0 & 0 & 0 & 0.6000 & 4.7198 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.8000 \end{array}\right].$$

Using Theorem 3.41, we can block diagonalize the matrix T_1 above:

$$U_1^{-1}Q_1^T A Q_1 U_1 = U_1^{-1}T_1 U_1 = \begin{bmatrix} T_{11} & 0 \\ 0 & B \end{bmatrix}.$$

Assume that we want to apply the same process to B, but now to pull out a block containing the eigenvalues of the original matrix A with magnitude less than one. A

corresponding ordered Schur factorization will give:

$$\tilde{Q}_2^T B \tilde{Q}_2 = \tilde{T}_2 = \begin{bmatrix} T_{22} & R \\ 0 & T_{33} \end{bmatrix},$$

where

$$\tilde{T}_2 = \left[\begin{array}{rr|rrr} 0.6000 & -3.6150 & 0.0392 & 2.6913 & 3.1359 \\ 0 & 0.8000 & -0.4816 & -1.3995 & 2.7681 \\ \hline 0 & 0 & 2.0000 & 0.7354 & -0.7406 \\ 0 & 0 & 0 & 1.6000 & 1.1891 \\ 0 & 0 & 0 & -0.5382 & 1.6000 \end{array}\right].$$

Again, using Theorem 3.41, we can block diagonalize $\tilde{T}_2$:

$$\tilde{U}_2^{-1} \tilde{T}_2 \tilde{U}_2 = \begin{bmatrix} T_{22} & 0 \\ 0 & T_{33} \end{bmatrix}.$$

If we let $U_2 = \begin{bmatrix} I & 0 \\ 0 & \tilde{U}_2 \end{bmatrix}$ and $Q_2 = \begin{bmatrix} I & 0 \\ 0 & \tilde{Q}_2 \end{bmatrix}$, where I is the 2×2 identity matrix, then we have

$$U^{-1} A U := \begin{bmatrix} T_{11} & 0 & 0 \\ 0 & T_{22} & 0 \\ 0 & 0 & T_{33} \end{bmatrix},$$

where $U = Q_1 U_1 Q_2 U_2$. The eigenvalues of A are split:

$$\begin{array}{lll} \sigma(T_{11}) = -0.2,\ -0.13333 & \rightarrow & \mathrm{Re}(\lambda) < 0, \\ \sigma(T_{22}) = \ \ 0.6,\ 0.8 & \rightarrow & |\lambda| < 1, \\ \sigma(T_{33}) = \ \ 2,\ 1.6 \pm 0.8i & \rightarrow & |\lambda| > 1. \end{array}$$

The residual. Recall from Definition 3.35 that a matrix X is an eigenbasis of A if $AX = XL$. In practice this eigenbasis needs to be computed numerically as the limit of a sequence X_k of bases whose span converges to an eigenspace of A. To determine convergence we monitor the so-called residual

$$R_k = AX_k - X_k L_k$$

and hope it will converge to zero. The natural way to choose L would be the one that minimizes the norm of R. More formally, we have the following theorem.

Theorem 3.43 *Let $[X \quad Y]$ be a square orthogonal matrix, and let $R = AX - XL$. Then the matrix*

$$L = X^T A X$$

minimizes the norm $\|R\|_2$.

Proof.

$$\begin{bmatrix} X^T \\ Y^T \end{bmatrix} A [X \quad Y] = \begin{bmatrix} X^TAX & X^TAY \\ Y^TAX & Y^TAY \end{bmatrix} =: \begin{bmatrix} B & C \\ D & E \end{bmatrix}, \textit{ and}$$

$$\begin{bmatrix} X^T \\ Y^T \end{bmatrix} A [X \quad Y] \begin{bmatrix} X^T \\ Y^T \end{bmatrix} X = \begin{bmatrix} X^T \\ Y^T \end{bmatrix} A [X \quad Y] \begin{bmatrix} I \\ 0 \end{bmatrix} = \begin{bmatrix} X^T \\ Y^T \end{bmatrix} AX.$$

Hence,

$$\begin{aligned} \begin{bmatrix} X^T \\ Y^T \end{bmatrix} R &= \begin{bmatrix} X^T \\ Y^T \end{bmatrix} (AX - XL) \\ &= \begin{bmatrix} X^T \\ Y^T \end{bmatrix} A [X \quad Y] \begin{bmatrix} X^T \\ Y^T \end{bmatrix} X - \begin{bmatrix} X^T \\ Y^T \end{bmatrix} XL \\ &= \begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} X^T \\ Y^T \end{bmatrix} X - \begin{bmatrix} X^T \\ Y^T \end{bmatrix} XL \\ &= \begin{bmatrix} B & C \\ D & E \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} - \begin{bmatrix} L \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} B - L \\ D \end{bmatrix}. \end{aligned}$$

Since the norm $\|\cdot\|_2$ *is orthogonally invariant,*

$$\|R\|_2 = \left\| \begin{bmatrix} X^T \\ Y^T \end{bmatrix} R \right\|_2 = \left\| \begin{bmatrix} B - L \\ D \end{bmatrix} \right\|_2, \textit{ which is minimum if}$$

$$L = B = X^TAX.$$

□

■ **EXAMPLE 3.44**

The matrix

$$[X \quad Y] = \begin{bmatrix} -0.4531 & -0.7109 & 0.4330 & -0.3190 \\ 0.4939 & -0.3698 & -0.5234 & -0.5877 \\ 0.7306 & -0.2846 & 0.5324 & 0.3190 \\ -0.1301 & -0.5261 & -0.5051 & 0.6716 \end{bmatrix} \text{ is orthogonal.}$$

Now let $A = \begin{bmatrix} -4 & 1 & 3 & 5 \\ -1 & -4 & 6 & 1 \\ 0.5 & 0 & 3 & -2 \\ 0 & 0 & 4 & -1 \end{bmatrix}$. If we let $R = AX - XL$, where

$L = X^TAX = \begin{bmatrix} 0.8345 & 0.4015 \\ -5.9850 & 1.1012 \end{bmatrix}$, then the matrix R has the minimum norm $\|R\|_2 = 0.2616$.

(Observe that if the (3,1) entry of A had been zero, then we would have obtained $\|R\|_2 = 0$.)

Remark 3.44 *It is clear that the theorem above is also true for the Frobenius norm and for any other orthogonally invariant norm.*

3.5 INFORMATION RETRIEVAL

One area where linear algebra and matrix computations have lately found several important applications is that of information retrieval (IR) in digital libraries and the World Wide Web. Think about a digital library and you are searching for a given book or article. Is there any linear algebra behind the search engine you are using?

Large databases of books, journals, and so on, can be modeled and represented as matrices, so that when a user enters a query in the computer through a search engine, that query is transformed into a vector and compared with the columns of the database matrix to pull the relevant documents associated with the query. The columns of such a (very large) matrix represent in some way each entry of the database. For instance, a book, indexed e.g., by words on its title, can be represented as a vector where the nonzero entries of such vector correspond to words on the title. We want to explain and illustrate the mathematics behind this process of entering a query and matching with one or more items on a digital library or the World Wide Web (WWW).

In a vector space IR model, a vector is used to represent each item or document in a collection. Each component of the vector reflects a particular key word or term associated with the given document. The value assigned to that component could be, for instance, the (scaled) frequency with which the term occurs in the document.

A database containing n documents and described by a total of m terms or words is represented as an $m \times n$ term-document matrix A. Thus, the matrix element a_{ij} could be interpreted as the weighted frequency at which term i occurs in document j. The important fact is that

The semantic content of the database is completely contained in

the column space of A.

In other words, the document column vectors span the content.

The relationship between the number of rows and columns of A depends on the data it represents. For contexts such as newspapers or encyclopedias, the number of terms is typically larger than the number of documents. In the case of the WWW, the situation is reversed: The number of terms is the number of words in the largest dictionary, and the number of columns is the the number of web pages, which gives a matrix of about 300,000 rows by several billions of columns.

With all the considerations above, we are moving the problem of information retrieval to the area of matrix analysis and vector spaces, and we can now take advantage of the available mathematical tools.

3.5.1 Query matching

Query matching is the process of finding the documents most similar or closely related to a given query. Since the query is transformed into a vector, and it has to be compared with the columns of the term-document matrix, the documents selected are those geometrically closest to the query. One common measure of similarity is the cosine of the angle between the query and the document vectors.

If the term-document matrix A has columns a_j, $j = 1, \ldots, n$, the cosines of the angles between the column a_j and the query vector q are defined by

$$\cos\theta_j = \frac{a_j^T q}{\|a_j\| \, \|q\|}, \quad j = 1, \ldots, n. \tag{3.56}$$

The computations in (3.56) are not expensive because the vectors a_j and q are typically very sparse (think about the total terms of a dictionary and how many of them you would use on a query, or how many would be used as key words to index a document). Also observe that the norm $\|a_j\|$ is not exactly part of the computation in (3.56) for every query q, because it is computed just once for every term-document matrix. It is independent of the query.

As an illustration, suppose our list of terms has only seven elements (identified by their word stems):

T1:	python	T5:	formula(s)
T2:	apply(ied)	T6:	language(s)
T3:	spam(s)	T7:	programming.
T4:	mathematics		

Also suppose that our collection contains only five documents, indexed by some title words

D1: Introduction to <u>applied</u> <u>mathematics</u> in <u>python</u>

D2: The <u>python</u> <u>programming</u> <u>language</u>

D3: Monty <u>python</u> skids about <u>spam</u>

D4: Some <u>formulas</u> behind a <u>spam</u>

D5: Understanding <u>mathematics</u> with <u>python</u>.

Then, the term-by-document matrix $A_{7\times 5}$, with a_{ij} representing the number of times term T_i appears in the title of document D_j, is

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

After normalization, the matrix is

$$A = \begin{bmatrix} 0.5774 & 0.5774 & 0.7071 & 0 & 0.7071 \\ 0.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.7071 & 0.7071 & 0 \\ 0.5774 & 0 & 0 & 0 & 0.7071 \\ 0 & 0 & 0 & 0.7071 & 0 \\ 0 & 0.5774 & 0 & 0 & 0 \\ 0 & 0.5774 & 0 & 0 & 0 \end{bmatrix}. \tag{3.57}$$

We do this normalization because we are not exactly interested in the total number of times terms appear, but instead in their relative frequencies. At the same time, we now have

$$\|a_j\| = 1, \qquad \text{for all} \quad j = 1, \ldots, n.$$

■ **EXAMPLE 3.45**

Assume that a user enters the query python mathematics. By comparing this query with the seven terms above, the query vector is

$$q = [1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0]^T,$$

with nonzero entries for the terms *python* and *mathematics*. Consequently, the search for relevant documents is performed by computing the cosines of the angles between this query vector q and the document vectors represented by the columns a_j of the matrix A, according to (3.56).

The closer the cosine is to one (the smaller the angle between the q and a_j is), the more relevant the document is to the query. But then a threshold or cutoff value has to be predetermined, so that only the documents whose cosines with the query vector pass this value are pulled as relevant. Otherwise, too many documents could be retrieved as relevant, some of them with almost no connection to the query. Let us assume this cutoff value is equal to 0.55. The cosines are

$$\cos\theta_1 = 0.8165, \quad \cos\theta_2 = 0.4082, \quad \cos\theta_3 = 0.5, \quad \cos\theta_4 = 0, \quad \cos\theta_5 = 1.$$

Thus, documents D1 and D5 will be accurately returned as relevant, while the other three will be correctly ignored!

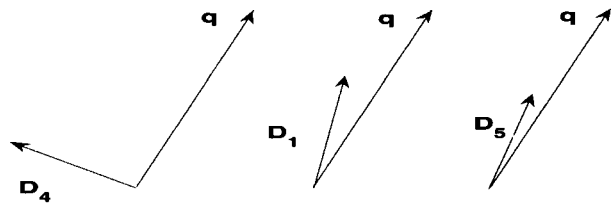


Figure 3.1 Schematic representation of query matching.

What do matrix factorizations have to do with IR? The main idea is to use matrix factorizations to identify and remove redundant information or noise in the matrix representation of the database, using a low-rank approximation of the original term-by-document matrix, contributing in this form to reducing the size of the problem and also to help dealing with problems of synonymy and polysemy.

Two factorizations can be used: QR and SVD, but the latter is more powerful, first, because as Theorem 3.17 states, the SVD provides a basis for both, col(A) and row(A), and second, from Theorem 3.21, we know that among all possible rank-k approximations of the matrix A, the one provided by the SVD is the closest to A.

■ **EXAMPLE 3.46**

Consider the matrix A in (3.57). The SVD of that matrix is $A = U\Sigma V^T$, where

$$U = \begin{bmatrix} -0.8005 & 0.1015 & 0.1507 & -0.3216 & 0.0229 & -0.3333 & -0.3333 \\ -0.1901 & 0.1484 & -0.2624 & 0.3350 & -0.8722 & -0.0000 & -0.0000 \\ -0.2903 & -0.7772 & -0.0613 & -0.2528 & -0.1477 & 0.3333 & 0.3333 \\ -0.4374 & 0.3123 & -0.4820 & 0.2994 & 0.4086 & 0.3333 & 0.3333 \\ -0.0725 & -0.5096 & -0.1206 & 0.6699 & 0.2227 & -0.3333 & -0.3333 \\ -0.1453 & 0.0568 & 0.5735 & 0.3017 & -0.0153 & 0.6667 & -0.3333 \\ -0.1453 & 0.0568 & 0.5735 & 0.3017 & -0.0153 & -0.3333 & 0.6667 \end{bmatrix},$$

$$\Sigma = \begin{bmatrix} 1.5822 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.1236 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.8685 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.5579 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.4104 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \end{bmatrix},$$

$$V = \begin{bmatrix} -0.5211 & 0.2889 & -0.3947 & 0.3238 & -0.6200 \\ -0.3981 & 0.1105 & 0.8627 & 0.2915 & -0.0109 \\ -0.4875 & -0.4252 & 0.0728 & -0.7280 & -0.2149 \\ -0.1621 & -0.8098 & -0.1481 & 0.5285 & 0.1292 \\ -0.5532 & 0.2604 & -0.2698 & -0.0281 & 0.7434 \end{bmatrix}.$$

By looking at Σ, we realize that $\text{rank}(A) = 5$. Since the SVD of A is available, we can follow the rank-k approximation given by Theorem 3.21:

$$A_k = U_k \Sigma_k V_k^T \tag{3.58}$$

to obtain A_4 and A_3:

$$A_4 = \begin{bmatrix} 0.5832 & 0.5775 & 0.7091 & -0.0012 & 0.7001 \\ 0.3554 & -0.0039 & -0.0769 & 0.0463 & 0.2661 \\ -0.0376 & -0.0007 & 0.6941 & 0.7149 & 0.0450 \\ 0.6813 & 0.0018 & 0.0360 & -0.0217 & 0.5825 \\ 0.0567 & 0.0010 & 0.0196 & 0.6953 & -0.0679 \\ -0.0039 & 0.5773 & -0.0014 & 0.0008 & 0.0047 \\ -0.0039 & 0.5773 & -0.0014 & 0.0008 & 0.0047 \end{bmatrix}, \tag{3.59}$$

$$A_3 = \begin{bmatrix} 0.6413 & 0.6298 & 0.5785 & 0.0936 & 0.6951 \\ 0.2949 & -0.0584 & 0.0592 & -0.0525 & 0.2713 \\ 0.0081 & 0.0405 & 0.5914 & 0.7895 & 0.0411 \\ 0.6272 & -0.0469 & 0.1577 & -0.1100 & 0.5872 \\ -0.0643 & -0.1080 & 0.2917 & 0.4978 & -0.0574 \\ -0.0584 & 0.5282 & 0.1212 & -0.0881 & 0.0094 \\ -0.0584 & 0.5282 & 0.1212 & -0.0881 & 0.0094 \end{bmatrix}. \tag{3.60}$$

Then, we can compute the relative errors:

$$\frac{\|A - A_4\|_F}{\|A\|_F} \approx 0.1835, \qquad \frac{\|A - A_2\|_F}{\|A\|_F} \approx 0.3097.$$

This means that approximating A with a rank-4 matrix A_4 gives a relative error of 18.35%, which can be considered as reasonably small, for the application in mind, but approximating A with a rank-3 matrix A_3 gives a relative error of 30.97%, which may be too large. That is, we can expect to get accurate results if we use A_4, but not necessarily accurate results if we use A_3 as an approximation to the matrix A.

3.5.2 Low-rank query matching

When we use low-rank query matching, given a query vector q, we compare this vector not with the columns of A anymore, but with the columns of A_k. But remember from Remark 3.22 that this matrix A_k does not need to be computed explicitly, thanks to the way it is

defined in (3.31). We are mostly interested in the cosines between the query vector and the columns of A_k, for which we will only need the matrices U_k and V_k.

Recall that the j–th column of any matrix A is given by Ae_j, where e_j is the canonical vector with a one in the j–th entry and zeros elsewhere. In general, we compute the cosines of the angles between q and the j-th column of A_k as we did in (3.56):

$$\cos\theta_j = \frac{(A_k e_j)^T q}{\|A_k e_j\|\,\|q\|} = \frac{(U_k \Sigma_k V_k^T e_j)^T q}{\|U_k \Sigma_k V_k^T e_j\|\,\|q\|} = \frac{e_j^T V_k \Sigma_k^T (U_k^T q)}{\|\Sigma_k V_k^T e_j\|\,\|q\|}, \quad j = 1, \ldots, n.$$

If we denote $p_j = \Sigma_k V_k^T e_j$, then we can write

$$\cos\theta_j = \frac{p_j^T (U_k^T q)}{\|p_j\|\,\|q\|}, \quad j = 1, \ldots, n. \tag{3.61}$$

Now we can compare (3.61) with the original formula (3.56). The expression in (3.61) requires only operations with k-dimensional vectors, except the norm of q (which is already very sparse, and it is computed only once), whereas (3.56) involves m-dimensional vector computations. Thus, lowering the rank also lowers the cost of query matching. Furthermore, it is never necessary to compute the full SVD of A; it is sufficient to compute only the first singular values and vectors that make up Σ_k, U_k and V_k. In addition, the norms $\|p_j\|$ are computed only once for each term-by-document matrix. They do not depend on the query vector q.

■ **EXAMPLE 3.47**

Let us apply this rank-reduction approach to the matrix (3.57), with the same query python mathematics, and

$$q = [1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0]^T.$$

By using the rank-4 approximation A_4 in (3.59), that is, $k = 4$, the cosine formula in (3.61) gives

$$\cos\theta_1 = 0.9246, \quad \cos\theta_2 = 0.4096, \quad \cos\theta_3 = 0.5290,$$

$$\cos\theta_4 = -0.0121, \quad \cos\theta_5 = 0.9522.$$

For the same cutoff value of 0.55, documents 1 and 5 are again correctly identified, and the other three are correctly ignored.

If instead we had used a rank-3 approximation A_3, we would have obtained

$$\cos\theta_1 = 0.9441, \quad \cos\theta_2 = 0.4177, \quad \cos\theta_3 = 0.5723,$$

$$\cos\theta_4 = -0.162, \quad \cos\theta_5 = 0.9523.$$

Thus, in this case documents 1 and 5 are still correctly chosen as relevant. However, also document 3 (*Monty python skids about spam*) passes the cutoff value 0.55 as matching the

query python mathematics, which is not exactly right. This confirms what we observed about the relationship between the relative error and the accuracy of the approximation. (Clearly, a little higher cutoff value would solve this problem).

We need to remark that the results of low-rank query matching depend on the query and its relationship with the total list of terms or dictionary. They also depend on the prescribed cutoff value, and on the size of the term-document matrix (that is, the number of terms available and the number of documents). In a few cases, the accuracy of the approximation may not improve as we increase the the value of k. See Exercise 3.79.

3.5.3 Term–term comparison

As mentioned before, the SVD of a matrix A besides giving a basis for col(A), it also provides us with a basis for row(A). A basis for row(A) can be used to compare a pool of relevant documents and to group them according to the different meanings of a key word in the query. This word is present in all those documents, but it may have a totally different meaning on each one of them. This process of grouping documents according to the different meanings of a common key word is known as **clustering**, a technique very useful for dealing with polysemy. We will see how the SVD of the term-document matrix magically clusters documents and terms into concepts.

As an illustration, consider the following terms:

T1:	ski(ing)	T5:	mountain(s)
T2:	boat(s)	T6:	snow
T3:	lake(s)	T7:	sand,
T4:	water(s)		

and assume that the following documents have been selected as relevant, after the query *ski*:

D1: Manual of <u>snow</u> <u>skiing</u> in the <u>mountains</u>

D2: <u>Water</u> <u>skiing</u> at the <u>lakes</u> of Canada.

D3: <u>Boating</u> and <u>skiing</u> on the cold <u>waters</u> of the <u>lake</u>.

D4: <u>Sand</u> <u>skiing</u> in Peru.

D5: <u>Snow</u> on the Rocky <u>mountains</u>.

The normalized term-by-document matrix is

$$A = \begin{bmatrix} 0.5774 & 0.5774 & 0.5 & 0.7071 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0.5774 & 0.5 & 0 & 0 \\ 0 & 0.5774 & 0.5 & 0 & 0 \\ 0.5774 & 0 & 0 & 0 & 0.7071 \\ 0.5774 & 0 & 0 & 0 & 0.7071 \\ 0 & 0 & 0 & 0.7071 & 0 \end{bmatrix}.$$

Suppose then that we want to refine our search by grouping the documents D_1 through D_5 according to common meanings. For this to be achieved, we now look at the vector space spanned by the terms under consideration, that is, we look at the row(A). This means we need to perform a term–term comparison by computing the cosines of the angles between the term vectors (the rows of A).

Recall that in general, the vector $A^T e_j$, where e_j is the usual canonical vector, is the j–th row of A. Thus, we compute

$$\cos\theta_{ij} = \frac{(e_i^T A)(A^T e_j)}{\|A^T e_i\| \, \|A^T e_j\|}, \qquad i, j = 1 \ldots, 7. \tag{3.62}$$

These cosines are shown below as the entries of the symmetric matrix C below (we do not show the lower triangular part), with $c_{ij} = \cos\theta_{ij}$ representing how closely term i is associated with term j. The closer the entry is to 1, the more closely related the terms are; likewise, the closer the entry is to zero, the less related the terms are.

$$C = \left[\begin{array}{c|cccc|cc|c}
1 & 0.4201 & 0.6417 & 0.6417 & & 0.3068 & 0.3068 & 0.5941 \\
\hline
 & 1 & 0.6547 & 0.6547 & & 0 & 0 & 0 \\
 & & 1 & 1 & & 0 & 0 & 0 \\
 & & & 1 & & 0 & 0 & 0 \\
\hline
 & & & & & 1 & 1 & 0 \\
 & & & & & & 1 & 0 \\
\hline
 & & & & & & & 1
\end{array}\right]. \tag{3.63}$$

For instance, we expect T_2 (boat) to be related to T_3 (lake) and T_4 (water), but not to T_5 (mountain), T_6 (snow), or T_7 (sand). This means that on the second row of the matrix C, besides the second entry, which compares T_2 with itself, only the third and fourth entries should be away from zero and closer to one.

We are interested in the cosines of all the terms other than ski (rows 2 through 7 in the matrix C). Observe that the computation of the cosines between the terms has resulted in a partition of the terms into three groups that are semantically independent:

Group 1: { boat, lake, water}, Group 2: { mountain, snow},

Group 3: { sand}.

This means that we can group the documents into the sets

$$\{D_1, D_3, D_5\}, \ \{D_4\}, \quad \text{and} \quad \{D_2\}.$$

Low-rank computation. Since the SVD of A provides us with a basis for row(A), we can also use the SVD low-rank approach in a term–term comparison, just as we did for the regular query matching problem. Thus, as before, let $A_k = U_k \Sigma_k V_k^T$ be an approximation

of rank k to the term-by-document matrix $A = U\Sigma V^T$. Then, the cosines in the reduced-rank approximation that compare terms i and j are

$$\begin{aligned} \cos\theta_{ij} &= \frac{(e_i^T A_k)\,(A_k^T e_j)}{\|A_k^T e_i\|\,\|A_k^T e_j\|} \\ &= \frac{(e_i^T U_k \Sigma_k V_k^T)(V_k \Sigma_k^T U_k^T e_j)}{\|V_k \Sigma_k^T U_k^T e_i\|\,\|V_k \Sigma_k^T U_k^T e_j\|} = \frac{(e_i^T U_k \Sigma_k)(\Sigma_k U_k^T e_j)}{\|\Sigma_k U_k^T e_i\|\,\|\Sigma_k U_k^T e_j\|}, \end{aligned} \tag{3.64}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, n$. Here we have used the facts that $\Sigma_k^T = \Sigma_k$, and $\|QA\|_2 = \|A\|_2$, for any orthogonal matrix Q.

If we let $b_j = \Sigma_k U_k^T e_j$, we get

$$\cos\theta_{ij} = \frac{b_i^T b_j}{\|b_i\|\,\|b_j\|}, \qquad i, j = 1, \ldots, m. \tag{3.65}$$

And again, each vector in this computation is only k-dimensional.

■ **EXAMPLE 3.48**

Let us use a rank-4 approximation of the matrix A above to perform a term–term comparison following the formula (3.65), where

$$b_j = \Sigma_4 U_4^T e_j, \qquad \text{for} \quad j = 1, \ldots, 7.$$

The new matrix of cosines, which we denote by C_4, is given by

$$C_4 = \left[\begin{array}{c|ccc|cc|c} 1 & 0.4153 & 0.6655 & 0.6655 & 0.3137 & 0.3137 & 0.6493 \\ \hline & 1 & 0.6692 & 0.6692 & 0.0025 & 0.0025 & 0.0185 \\ & & 1 & 1 & -0.0039 & -0.0039 & -0.0291 \\ & & & 1 & -0.0039 & -0.0039 & -0.0291 \\ \hline & & & & 1 & 1 & -0.0086 \\ & & & & & 1 & -0.0086 \\ \hline & & & & & & 1 \end{array}\right]. \tag{3.66}$$

Thus, the clustering coincides with the one given by (3.63), giving the correct groups of term according to their corresponding meanings, but this time the vectors b_j used to compute the cosines in (3.65) are only 4-dimensional.

3.6 PARTITION OF SIMPLE SUBSTITUTION CRYPTOGRAMS

We have seen in Section 3.5.3 how the SVD can be used to cluster documents and terms into concepts. Not surprisingly, such matrix factorization has other similar applications, that

is, to cluster, group, or partition elements according to certain properties or characteristics. The SVD of a given matrix A will be used here to partition letters of the alphabet in a given text into vowels and consonants, and to find out about their corresponding frequencies in the text. This is a central part in the process of code breaking or obtaining the meaning of encrypted information. On one side there is a first person trying to send some encrypted information to a second party. The art of creating such a coded message is known as **cryptography**. In between the first person and the second party is the code breaker who studies the encrypted data and tries to unencrypt it without use of the key known only by the first and the second persons. **Cryptanalysis** is the art of breaking codes mainly by finding weaknesses within it. The study of both, cryptography and cryptanalysis, is called **cryptology**. Finally, we say a **cipher** is a method for encrypting a message.

Although each letter of the alphabet can be replaced by an arbitrary symbol when creating the encrypted message, here we consider only simple substitution ciphers, whereby each letter of the original text (called plain text) is just replaced by a different letter to create the encrypted text or cryptogram. For example, suppose we want to encipher:

P: There is no applied mathematics if there is no mathematics to apply.

And assume we decide to replace A with Z, B with Y, C with X, and so on; then the cryptogram would be

C: Gsviv rh ml zkkorw nzgsvnzgrxh ru gsviv rh ml nzgsvnzgrxh gl zkkob.

More generally, a cryptogram is obtained by making substitutions for the letters in the plain text with corresponding letters obtained from a *cipher alphabet*, which could be nothing else than a permutation of the letters in the original alphabet.

Faced with a cryptogram C, the cryptanalyst first has to find out which of the letters may represent a vowel and which a consonant. The partition of the letters into vowels and consonants is based on the typical frequency that each letter has on any given (not too short) text. In addition, this partition is plausible under the assumption that the proportion of vowel–vowel pairs is less than the proportion of the consonant–vowel pairs, that is

$$\frac{\text{number of vowel} - \text{vowel pairs}}{\text{number of vowels}} < \frac{\text{number of consonant} - \text{vowel pairs}}{\text{number of consonants}} \tag{3.67}$$

Texts satisfying condition (3.67) are called **"vfc"** texts. English is a language that produces such texts. In fact, it is not very common in English for a vowel to follow another vowel; however it is very common that a vowel follows a consonant (vfc).

Let n be the number of letters in the alphabet (for English, $n = 26$). We define the entries of the *digram frequency matrix* A as

$$a_{ij} = \text{number of times the } i\text{th letter is followed by the } j\text{th letter.} \tag{3.68}$$

Blanks and punctuation are ignored, and the text is thought of as circular; that is, the first letter of the text is assumed to follow the last letter.

The row sum and the column sum of A are the same, and they represent the number of times the corresponding letter occurs in the text. That is,

$$f_i = \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} a_{ji} \qquad (i = 1, \ldots, n) \tag{3.69}$$

is the number of occurrences of the ith letter.

Since the purpose is to partition the letters into vowels and consonants, we can define two vectors w and c by

$$w_i = \begin{cases} 1, & \text{if the } i\text{th letter is a vowel} \\ 0, & \text{otherwise} \end{cases} \tag{3.70}$$

$$c_i = \begin{cases} 1, & \text{if the } i\text{th letter is a consonant} \\ 0, & \text{otherwise.} \end{cases} \tag{3.71}$$

With these definitions, we have the following consequences:

1. The vectors w and c are orthogonal.

2. The quadratic form $w^T A\, w$ is the number of vowel-vowel pairs.

3. The quadratic form $c^T A\, w$ is the number of consonant-vowel pairs.

4. The vfc rule (3.67) can be written as

$$\frac{w^T A w}{w^T A(w + c)} < \frac{c^T A w}{c^T A(w + c)}. \tag{3.72}$$

In Theorem 3.21, we introduced rank-k approximations of a matrix $A_{m \times n}$ via its SVD $A = U \Sigma V^T$. Namely, a rank-k approximation of A is written as

$$A_k = U_k \Sigma_k V_k^T, \tag{3.73}$$

where the columns of U_k are the first k columns of U, the columns of V_k are the first k columns of V, and Σ_k is the $k \times k$ diagonal matrix whose diagonal elements are the k largest singular values of A. For the application in mind, the matrix A is square, order $n = 26$.

The central idea here is that given a text (in practice, a cryptogram C), and its corresponding digram frequency matrix A, we can use low–rank SVD approximations of A to obtain relevant information as a first step in deciphering such text.

e	12.70	d	4.25	p	1.93
t	9.06	l	4.03	b	1.50
a	8.17	c	2.78	v	0.98
o	7.51	u	2.76	k	0.77
i	6.97	m	2.41	j	0.16
n	6.75	w	2.36	x	0.15
s	6.33	f	2.23	q	0.10
h	6.09	g	2.02	z	0.07
r	5.99	y	1.97		

Table 3.1 Relative frequencies of letters in English alphabet.

3.6.1 Rank-1 approximation

Although with $k = 1$ in (3.73), we expect a poor approximation of A, it is still possible to obtain information on the frequency of each letter of the text, thus allowing the cryptanalyst to guess in a reliable way which is the true letter of the alphabet that the letter on the text (or cryptogram) is supposed to represent. Of course, there is not a closed formula to represent the relative frequencies of the letters in English alphabet that appear on a typical text, because that varies according to different factors such as the subject or topic, the writer, or whether it is American or British English. One possible list of typical relative frequencies is shown in Table 3.1.

Let us consider a rank-1 approximation of the matrix A:

$$A \approx A_1 = \sigma_1 u_1 v_1^T. \tag{3.74}$$

If we define e to be a vector of ones ($e = [1 \cdots 1]^T$) of length n, and f the vector with each entry f_i representing the number of occurrences of the ith letter in the text (see also (3.69)), then it is clear that $Ae = A^T e = f$ and therefore

$$\begin{aligned} f &= Ae \approx A_1 e = \sigma_1 u_1 (v_1^T e) = [\sigma_1 (v_1^T e)]\, u_1, \\ f &= A^T e \approx A_1^T e = \sigma_1 v_1 (u_1^T e) = [\sigma_1 (u_1^T e)]\, v_1. \end{aligned} \tag{3.75}$$

This indicates that both, the first left and the first right singular vectors u_1 and v_1, respectively, are approximately proportional to the frequencies of the letters of the text.

Consider for instance the following text:

"The Best Teacher"

There are no secrets about the best strategies for the most effective teaching.

Student engagement will be on every list of things to seriously consider when teaching. If you manage to engage your students in every single class, if you make them be a very active part of the learning experience, and if they continuously and enthusiastically participate by asking questions, giving answers, working in groups to solve sample problems or projects, presenting their results to the rest of the class, helping others individually or in groups, then your teaching experience will be

filled with energy and with students eager to learn, to participate and to say at the end of the class: that was cool! Make sure everyone participates. Knowing all their names is an excellent starting point.

Immediate feedback is what students always expect from instructors, and they do appreciate it more than what you think. The classical immediate feedback happens during office hours when students come for help and instructors do their best to contribute with the students' learning. Now there is communication through email. If a student sends you an email asking a question, you may be out of office hours, or you may be at home, but it is extremely important that you reply that email as soon as possible. The student may be at the library or at home unable to solve a problem or unable to move forward on a topic because there is something that he or she simply cannot understand. By promptly replying such email, you are allowing the student to keep on studying with no interruption, and make true progress through the material. Otherwise, the student will feel frustrated, may abandon the effort, and dedicate his or her time to something else, probably never returning again to that part of the material he or she did not understand. With smart phones you can now answer those emails from anywhere you are and at any time.

Targeting different learning styles will broaden your impact as a teacher. Some students learn best by seeing things. A teacher has to make efforts to provide some relevant graphs, charts, videos or displays that help summarize or highlight the main ideas behind an abstract concept, or that help compare two or more different methods and their corresponding efficiencies or performances. Other students learn best by communicating directly through questions, answers, and group interaction with peers. A teacher should always be asking questions to students as to what the next step in the process just being explained is, a teacher should ask them what they think it would be the best way to go from that point; a teacher should very strongly encourage students to ask questions at anytime: a question does not interrupt the class, it makes it richer. No matter which subject is being taught, group work in and outside class is always an excellent way for students to learn at their own pace and using their own style, besides contributing to improving their social skills. An important component of group work is to ask students to present their findings to the whole class, and to encourage the whole class to actively participate in the studentsŠ presentations. Finally, some other students learn best by getting involved in hands-on experiences. Group assignment already is part of this approach, but wherever applicable or feasible, one should also make efforts to schedule lab hours, computer simulations experiments, field trips, etc. They all represent the most unforgettable learning experiences for the students, something that will stay in their minds for much longer time.

None of the three points just commented above (student engagement, immediate feedback, and targeting different learning styles) can be addressed or accomplished if the teacher does not have the personal motivation or commitment to do it, or does not have the conviction that it can be done. If you love your job as a teacher, you will do your best every day to get student engagement, give them immediate feedback and make enormous efforts to reach different learning styles. The positive feedback from your students will be the best reward you can get.

Last but not least: technology can be at times of great help to address some of the issues within the three main points mentioned above, and we should always be eager to learn new ways of implementing technology into our teaching. Technology can help the teacher getting more student engagement, give more immediate feedback and be more effective and even more efficient in reaching students with different learning styles. How much technology and what kind of technology can or should be used in class varies greatly, depending on the level of the class, and the subject being taught, and here is where the discussion can start.

In Table 3.2, we show the corresponding digram frequency matrix, and in Table 3.3 we show the first left and right singular vectors u_1 and v_1 and the frequencies of the 26 letters of the

alphabet in the text "The Best Teacher". It is clear to see that in fact, the frequency of each letter of the alphabet within the text is approximately proportional to the corresponding entry in the first left and right singular vectors u_1 and v_1.

This information provided by the rank-1 SVD approximation can be used by the cryptanalyst to estimate the alphabetic permutations used to create the cryptogram. Obviously, this would not be enough to come with a tentative answer and more information would be needed (such as the one provided by a rank-2 approximation), but it provides us with a first step into analyzing the cryptogram.

This first analysis of single letter frequencies could be complemented with the known relative frequencies of double letters (digrams) or even triple letters; see, e.g. , Table 3.4.

3.6.2 Rank-2 approximation

If we are interested in finding a correlation between pairs of letters in the text, then we need to use a rank-2 approximation of the digram frequency matrix A. This is exactly what will allow us to establish an approximate partition of the alphabet into vowels and consonants. Thus, let us consider a rank-2 SVD approximation of $A_{n\times n}$; that is

$$A \approx A_2 = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T. \tag{3.76}$$

If we observe at Table 3.3, first we realize that the first left and right singular vectors u_1 and v_1 have the same sign and they are close to each other, and therefore they cannot be used to propose the desired partition. However, the second left and right singular vectors u_2 and v_2 are not only different but also they have a special pattern of signs. In particular, we can observe that for all vowels, the entry in u_2 is positive while the one in v_2 is negative. Similarly, for most consonants, u_2 is negative and v_2 is positive. This is a pattern that will persist for any text we consider. Thus, by considering the vectors defined in (3.70) and (3.71), a simple partition of the alphabet into vowels and consonants can be establish as

$$w_i = \begin{cases} 1, & \text{if } u_{i2} > 0,\ v_{i2} < 0 \\ 0, & \text{otherwise,} \end{cases} \qquad c_i = \begin{cases} 1, & \text{if } u_{i2} < 0,\ v_{i2} > 0 \\ 0, & \text{otherwise,} \end{cases} \tag{3.77}$$

for $i = 1, \ldots n$. But then, we are left with the letters that cannot be classified as either vowels or consonants under partition (3.77); such letters are called "neuter". Fortunately enough, there are in general very few letters that fall into this category (among them, usually H, and Y). Such letters are those for which the u_2 and v_2 have the same sign.

Thanks to this partition of the alphabet using the signs of the second singular vectors components, we can think of the letters as distributed in the plane (u_2, v_2) according

Table 3.2 Digram frequency matrix of text "The Best Teacher"

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	0	11	28	4	0	0	11	0	11	0	7	17	2	52	0	6	2	35	30	54	3	3	0	0	15	0
B	9	0	0	0	29	0	0	1	0	2	0	9	0	0	3	0	0	2	1	0	5	0	0	0	4	0
C	16	1	1	0	12	0	0	27	9	0	6	10	0	0	19	0	0	1	0	16	1	0	0	0	0	0
D	9	11	0	3	37	0	2	1	25	0	0	0	3	1	14	0	0	2	4	7	4	1	5	0	3	0
E	52	6	25	26	22	17	1	4	21	0	0	20	24	74	11	16	0	62	63	38	2	9	6	11	7	0
F	1	0	0	0	15	15	2	0	9	0	0	0	0	0	13	0	0	5	0	11	0	0	0	0	3	0
G	10	0	0	3	24	0	0	7	7	0	0	2	1	2	4	1	2	11	9	13	0	0	1	0	6	0
H	22	1	0	2	94	0	0	0	18	0	0	2	0	6	17	1	1	5	4	6	1	0	0	0	0	0
I	10	5	18	9	10	10	3	0	0	0	0	14	18	83	16	5	0	10	20	18	0	10	0	0	0	1
J	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	2	0	0	0	0	0					
K	3	0	0	0	8	1	0	1	10	0	0	0	0	1	0	0	1	0	1	2	0	0	0	0	0	0
L	17	4	0	9	38	2	0	1	7	0	0	17	2	0	10	6	0	2	6	2	0	3	4	0	17	0
M	26	1	0	0	33	0	0	0	4	0	0	0	10	0	12	12	0	0	1	1	5	0	1	0	2	0
N	12	6	11	38	18	1	58	2	12	0	2	0	2	3	21	2	0	1	17	73	1	2	2	0	6	0
O	7	4	3	2	5	17	8	0	7	1	1	14	27	38	4	4	0	46	14	15	47	6	8	0	1	0
P	16	0	1	0	11	0	0	2	4	0	0	9	0	0	11	4	0	15	4	6	1	0	2	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	11	2	2	5	51	3	5	4	17	1	3	0	5	16	27	2	0	5	26	27	7	0	4	0	9	0
S	23	4	9	2	20	6	3	11	21	2	8	6	2	3	22	6	0	0	22	83	7	2	11	0	2	0
T	20	5	5	0	47	1	4	99	48	0	1	10	6	1	45	3	0	13	32	23	29	0	9	0	7	0
U	4	2	7	25	8	0	5	0	0	0	0	11	5	7	1	8	0	16	11	10	0	0	1	0	0	0
V	3	0	0	0	26	0	0	0	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
W	11	0	0	0	4	0	0	13	18	0	0	0	1	2	5	0	0	0	0	1	0	0	1	0	0	0
X	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	7	0	0	0	2	0	0	0	0	0	0
Y	8	3	7	3	1	1	1	0	7	0	0	6	0	1	22	3	0	1	10	7	0	0	1	0	0	0
Z	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

	Freq	u_1	v_1	u_2	v_2
A	290	−0.2870	−0.2680	0.3490	−0.0564
B	65	−0.0695	−0.0636	−0.1056	0.0200
C	119	−0.1184	−0.1373	−0.1436	0.1569
D	131	−0.1345	−0.1336	−0.1739	0.0928
E	517	−0.4944	−0.4014	0.3972	−0.5294
F	74	−0.0747	−0.0756	−0.0611	0.0780
G	103	−0.1161	−0.1052	−0.0752	0.0307
H	180	−0.2333	−0.2080	−0.3518	−0.4015
I	270	−0.2577	−0.2387	0.3394	−0.2142
J	6	−0.0065	−0.0047	−0.0127	−0.0016
K	28	−0.0319	−0.0274	−0.0400	0.0026
L	147	−0.1305	−0.1330	−0.1393	0.0757
M	108	−0.1078	−0.1176	−0.1394	0.1134
N	290	−0.2788	−0.3502	0.0082	0.5335
O	279	−0.2230	−0.2437	0.2402	−0.2284
P	86	−0.0870	−0.0765	−0.0313	0.0432
Q	6	−0.0027	−0.0042	−0.0015	0.0010
R	232	−0.2575	−0.2591	−0.1415	0.2726
S	275	−0.2995	−0.3270	−0.0634	0.1399
T	415	−0.3919	−0.4365	−0.5136	0.1068
U	121	−0.1037	−0.1154	0.0759	−0.0379
V	36	−0.0501	−0.0423	−0.1033	0.0585
W	56	−0.0546	−0.0581	−0.0782	−0.0220
X	11	−0.0065	−0.0211	0.0055	0.0292
Y	82	−0.0749	−0.0750	−0.0163	−0.0061
Z	1	−0.0016	−0.0010	−0.0035	0.0023

Table 3.3 Left and right singular vectors.

th	3.88	ou	1.29
he	3.68	ed	1.28
in	2.28	ha	1.27
er	2.18	to	1.17
an	2.14	or	1.15
re	1.75	it	1.13
nd	1.57	is	1.11
on	1.42	hi	1.09
en	1.38	es	1.09
at	1.34	ng	1.05

Table 3.4 Relative frequencies of digrams in English alphabet.

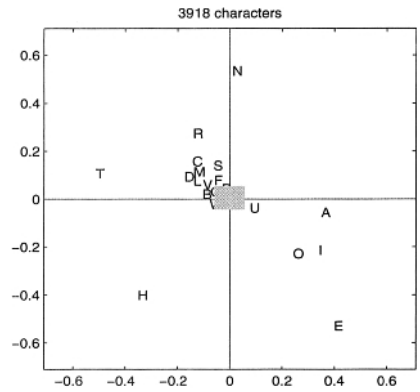


Figure 3.2 High frequency letters in the text.

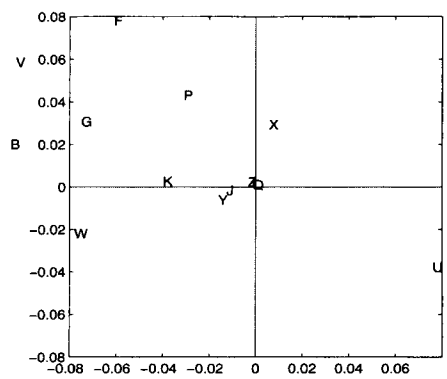


Figure 3.3 Zooming low frequency letters in the text.

to those signs. For instance, A would correspond to the point $(0.3490, -0.0564)$, B to $(-0.1056, 0.0200)$, and so on. This will place most vowels in the fourth quadrant and most consonants in the second quadrant. Neuter words will be placed into the first and third quadrants. In Figure 3.2 we plot the most frequent letters in the text "The Best Teacher". The little square in the middle contains the lower frequency letters, which are shown in detail in Figure 3.3.

Now, let us suppose that the text "The Best Teacher" was presented as a cryptogram that reads:

"Gsv Yvhg Gvzxsvi"

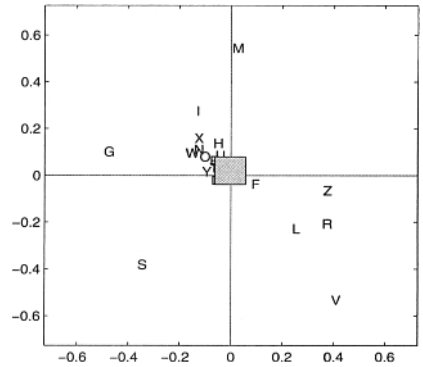


Figure 3.4 High frequency letters in cryptogram.

Gsviv ziv ml hvxivgh zylfg gsv yvhg hgizgvtrvh uli gsv nlhg vuuvxgrev gvzxsrmt.

Hgfwvmg vmtztvnvmg droo yv lm vevib orhg ...

A digram frequency matrix A is constructed, and its second left and right singular vectors u_2, v_2 are computed and used to establish an approximate partition of the letters in the text into vowels and consonants. By using again the relationship between the signs of their entries, we obtain such a partition using (3.77). In Table 3.5, we show each letter in the text along with its frequency as well as the first and second singular vectors.

In Figure 3.4, we have plotted the most frequent letters and in Figure 3.5 the least frequent. It is clear that such a partition reveals that the letters L, R, V, Z, and probably F represent in fact vowels and that the most frequent one, V (farthest from the origin), represents E. We can also deduce that S most likely represents H.

3.7 FINAL REMARKS AND FURTHER READING

We have presented four matrix factorizations (LU, QR, SVD and Schur), their main properties, and some of their applications. The LU factorization corresponds to the (naive) Gaussian elimination process, and for full-rank matrices, QR factorization corresponds to Gram–Schmidt. QR, SVD, and Schur factorizations exploit orthogonality and provide us with orthonormal bases of special subspaces. Schur factorization is the only one that gives an orthogonal similarity transformation.

Besides the classic books by Golub and Van Loan [25], and Horn and Johnson [38], the two volumes written by Stewart [62], [63] are excellent references on matrix factorizations. Meyer [49] and Watkins [66] also provide a detailed discussion on these topics. Matrix factorizations have been extended to smooth factorizations of matrix functions, which are

	Freq	u_1	v_1	u_2	v_2
A	1	−0.0016	−0.0010	−0.0035	0.0023
B	82	−0.0749	−0.0750	−0.0163	−0.0061
C	11	−0.0065	−0.0211	0.0055	0.0292
D	56	−0.0546	−0.0581	−0.0782	−0.0220
E	36	−0.0501	−0.0423	−0.1033	0.0585
F	121	−0.1037	−0.1154	0.0759	−0.0379
G	415	−0.3919	−0.4365	−0.5136	0.1068
H	275	−0.2995	−0.3270	−0.0634	0.1399
I	232	−0.2575	−0.2591	−0.1415	0.2726
J	6	−0.0027	−0.0042	−0.0015	0.0010
K	86	−0.0870	−0.0765	−0.0313	0.0432
L	279	−0.2230	−0.2437	0.2402	−0.2284
M	290	−0.2788	−0.3502	0.0082	0.5335
N	108	−0.1078	−0.1176	−0.1394	0.1134
O	147	−0.1305	−0.1330	−0.1393	0.0757
P	28	−0.0319	−0.0274	−0.0400	0.0026
Q	6	−0.0065	−0.0047	−0.0127	−0.0016
R	270	−0.2577	−0.2387	0.3394	−0.2142
S	180	−0.2333	−0.2080	−0.3518	−0.4015
T	103	−0.1161	−0.1052	−0.0752	0.0307
U	74	−0.0747	−0.0756	−0.0611	0.0780
V	517	−0.4944	−0.4014	0.3972	−0.5294
W	131	−0.1345	−0.1336	−0.1739	0.0928
X	119	−0.1184	−0.1373	−0.1436	0.1569
Y	65	−0.0695	−0.0636	−0.1056	0.0200
Z	290	−0.2870	−0.2680	0.3490	−0.0564

Table 3.5 Singular vectors of digram matrix for cryptogram.

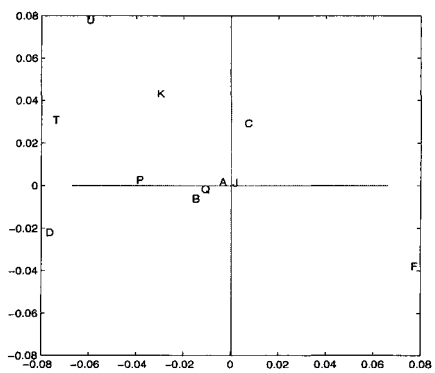


Figure 3.5 Low frequency letters in cryptogram.

crucial in several applications, such as in the smooth continuation of solutions in dynamical systems. See, e.g., the articles by Dieci and Eirola [17] and Rebaza [56].

The application of SVD on information retrieval is based on the article by M. Berry et al. [5], and the application on simple substitution cryptograms is based on an article by Moler and Morrison [52].

EXERCISES

3.1 Show that the matrix

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

does not have an LU factorization.

3.2 Find an LU factorization of $A = \begin{bmatrix} 1 & 3 & 2 \\ 3 & 5 & 9 \\ 2 & 2 & 5 \end{bmatrix}$ with L being unit lower triangular.

3.3 Let $A = \begin{bmatrix} 2 & 3 & 2 \\ 4 & -5 & 7 \\ 8 & 23 & -1 \end{bmatrix}$. Find a matrix X such that

$$XA = \begin{bmatrix} 2 & 3 & 2 \\ 0 & -11 & 3 \\ 0 & 0 & -6 \end{bmatrix}.$$

3.4 Given the matrix $A = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 3 & 0 & 0 & 1 \\ 1 & 2 & 2 & 0 \end{bmatrix}$, find the LU factorization of the corresponding permuted matrix. That is, $PA = LU$.

3.5 Use Gaussian elimination and backward substitution to solve the system

$$\begin{aligned} 3x - 2y + 3z &= 1 \\ 4x - y + 6z &= 0 \\ x - y + 4z &= -1 \end{aligned}$$

and identify the matrices L and U in the corresponding LU factorization of the coefficient matrix.

3.6 Let $A = LU$ be an LU factorization of $A_{n\times n}$, where L has a diagonal of 1's. Show that the determinant of A equals the product of the diagonal entries of U.

3.7 Build an arbitrary matrix $A_{n\times n} = LU$ so that $\det(A) = \frac{1}{n^n}$.

3.8 Show that if a zero pivot a_{kk} is encountered in the Gaussian elimination process for a nonsingular matrix $A_{n\times n}$, then it is always possible to exchange it with a nonzero entry a_{ik} of A, where $k + 1 \leq i \leq n$.

3.9 Show that the matrix L in (3.9) is unit lower triangular.

Hint: Let $\mathcal{L}_1 = L_1^{-1}$, $\mathcal{L}_k = P_k \mathcal{L}_{k-1} P_k^T L_k^{-1}$. Use induction.

3.10 Let $L_{n\times n}$ be a nonsingular lower triangular matrix with the property that $l_{i,i-1} \neq 0$ for $2 \leq i \leq n$, and let y be the solution of $Ly = b$, where $b_i = 0$ for $1 \leq i < k$ and $b_k \neq 0$. Show that

$$y_i = 0 \text{ for } 1 \leq i < k, \quad \text{and} \quad y_i \neq 0 \text{ for } k \leq i \leq n.$$

3.11 Consider a matrix $A_{m\times n}$. Show that there exist permutation matrices P_1 and P_2 such that

$$P_1^T A P_2 = LU,$$

where $L_{m\times k}$ is unit lower trapezoidal and $U_{k\times n}$ is upper trapezoidal with nonzero diagonal entries, and $k \leq \min\{m, n\}$.

3.12 Consider a matrix $A_{m\times n}$. Show that there exist two full-rank matrices $X_{m\times k}$ and $Y_{n\times k}$ such that

$$A = XY^T.$$

Hint: Use Exercise 3.11, and set $X = P_1 L$.

3.13 Let $A = LU$ be an LU factorization of a symmetric matrix $A_{n\times n}$. Show that

$$A = LDL^T,$$

where D is a diagonal matrix.

3.14 Let $A_{n\times n}$ be a symmetric positive definite matrix. Show that there exists a lower triangular matrix $\mathcal{L}$ such that

$$A = \mathcal{L}\mathcal{L}^T.$$

This is called the **Choleski factorization** of A.

3.15 Does the Choleski factorization exist for any of the following matrices? Why?

$$A = \begin{bmatrix} 15 & -4 & 1 \\ -4 & 8 & 0 \\ 1 & 0 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & 6 \\ 6 & 1 \end{bmatrix}.$$

3.16 Find the Choleski factorization of $A = \begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix}$.

3.17 Apply the Gram–Schmidt process to the following set of linearly independent vectors: $v_1 = [2 \quad 1 \quad -5]^T$, $v_2 = [5 \quad 1 \quad 0]^T$, $v_3 = [1 \quad 0 \quad 3]^T$, and verify that the resulting orthonormal vectors are the columns of Q, where $A = QR$ is the QR factorization of $A = [v_1 \quad v_2 \quad v_3]$.

3.18 Let $\{v_1, ..., v_n\}$ be a set of linearly independent vectors in R^n, and let $\{q_1, ..., q_n\}$ be its Gram–Schmidt orthonormalization. Prove that for each $j = 1, \ldots, n$,

$$v_j = \sum_{i=1}^{j} q_i r_{ij}, \tag{3.78}$$

where

$$r_{ij} = q_i^T v_j = v_j^T q_i, \qquad j = 1, \ldots, n. \tag{3.79}$$

3.19 From the expression given in (3.79) for the entries of R in the QR factorization of a matrix A, show that $r_{ij} = 0$ if $i > j$ and that $r_{ii} > 0$.

3.20 Find the QR factorization of $A = \begin{bmatrix} 2 & 1 & -5 \\ 5 & 1 & 0 \\ 1 & 0 & 3 \\ 1 & 3 & 4 \end{bmatrix}$. Do the columns of Q form an orthonormal basis of $\mathrm{col}(A)$?

3.21 Find the QR factorization of the matrix $A = \begin{bmatrix} 2 & 1 & 2 \\ 5 & 1 & 2 \\ 1 & 0 & 0 \\ 1 & 3 & 6 \end{bmatrix}$. Do the columns of Q form an orthonormal basis of $\mathrm{col}(A)$?

3.22 True or False? If $A_{m\times n}$ is a matrix with $m \geq n$, and $A = QR$ is its QR factorization, then $\mathrm{rank}(A) = \mathrm{rank}(R)$.

3.23 Let $A_{m\times n} = QR$ be the QR factorization of A. Show that $\|A\| = \|R\|$, where $\|\cdot\|$ is either the Frobenius or the 2-norm.

3.24 Show that a QR factorization of a general matrix $A_{m\times n}$ is not necessarily unique. Show however that the QR factorization of a full-rank matrix $A_{m\times n}$ is unique, where R has positive diagonal entries.

3.25 Solve the linear system

$$\begin{aligned} x + 2y + 3z &= 12 \\ -3x - y + 2z &= -3 \\ 4x - 3y - z &= 7, \end{aligned}$$

by first reducing the coefficient matrix to a triangular form. Do it in two ways, using LU and QR factorization of the coefficient matrix.

3.26 Show that the matrices $A_2, A_3, \ldots$ in the QR algorithm of (3.20) are all similar to the original matrix A.

3.27 Consider the matrices

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \text{and} \qquad Q = \begin{bmatrix} c & -s \\ s & c \end{bmatrix}.$$

Find the values of c and s so that Q is orthogonal and QA is upper triangular.

3.28 Define a Householder matrix as $H_1 = I - \frac{2uu^T}{u^Tu}$, where u is an n-dimensional vector defined by

$$u = \begin{bmatrix} a_{11} + \mathrm{sign}(a_{11})\|a_1\| \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}, \tag{3.80}$$

and a_1 denotes the first column of A. Show that if A is an $n \times n$ matrix, then

$$H_1 A = \begin{bmatrix} \alpha & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \cdots & * \end{bmatrix},$$

where $|\alpha| = \|a_1\|$. Then, following the same idea, find a Householder matrix H_2 such that

$$H_2 H_1 A = \begin{bmatrix} \alpha & * & \cdots & * \\ 0 & \beta & \cdots & * \\ 0 & 0 & \cdots & * \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & * \end{bmatrix},$$

for some $\beta \in \mathbb{R}$.

Hint: See Example 3.19.

3.29 Show that if $A_{n\times n} = xy^T$, for some nonzero vectors x and y, then $A^2 = A \cdot \mathrm{tr}(A)$.

3.30 Let $A_{n\times n}$ be a rank-1 matrix. What is col(A)? What is $N(A)$? What are their dimensions?

3.31 Find the singular value decomposition of

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

3.32 Show that if $A = U\Sigma V^T$ is the SVD of a matrix $A_{m\times n}$, then the columns of U are eigenvectors of the matrix AA^T with eigenvalues σ_i^2.

3.33 Let $A = U\Sigma V^T$ be the SVD of $A_{m\times n}$. Show that the columns $u_{r+1}, \ldots, u_m$ of U form an orthonormal basis of $N(A^T)$, and that the columns $v_{r+1}, \ldots, v_n$ of V form an orthonormal basis of $N(A)$.

3.34 Let $A = \begin{bmatrix} 2 & 0 & 1 & -1 \\ 2 & -3 & 1 & 2 \\ -1 & -1 & 0 & 2 \\ 1 & -4 & 1 & 4 \end{bmatrix}$. Find orthonormal bases for col(A) and row(A).

3.35 Let σ_1 be the largest singular value of a matrix $A_{n\times n}$. Show that if $\sigma < 1$, then $A^k \to 0$ as $k \to \infty$.

3.36 Let $A_{m\times n}$ be a matrix, and as usual let $p = \min\{m, n\}$. Show that

$$\|A\|_F = \left(\sum_{i=1}^{p} \sigma_i^2\right)^{1/2}.$$

3.37 Let $A_{m\times n} = U\Sigma V^T$ be the SVD of A, and let $j \le \min\{m, n\}$. If u_j and v_j denote the j–th column of U and V respectively, show that the j–th singular value of A

satisfies

$$\sigma_j = u_j^T A v_j.$$

3.38 Let $A_{m\times n} = U\Sigma V^T$ be the SVD of A. Show that

$$Ax = b \quad \Longleftrightarrow \quad \begin{cases} \Sigma\tilde{x} = \tilde{b} \\ x = V\tilde{x}, \end{cases}$$

where $\tilde{b} = U^T b$.

3.39 True or False? Given two square matrices A and B, the singular values of AB and BA are the same.

3.40 Consider the pseudoinverse matrix defined in (1.52). Show that if $A = U\Sigma U^T$ is the SVD of $A_{m\times n}$, then,

$$A^\dagger = V\Sigma^\dagger U^T, \tag{3.81}$$

where $\Sigma^\dagger$ denotes the transpose of Σ, with the positive singular values replaced by their reciprocals.

Note: The definition (3.81) is true for matrices of any rank.

3.41 Show the following properties of the pseudoinverse of a general matrix $A_{m\times n}$:

(a) $AA^\dagger A = A$,

(b) $A^\dagger AA^\dagger = A^\dagger$,

(c) $(A^\dagger A)^T = A^\dagger A$,

(d) $(AA^\dagger)^T = AA^\dagger$.

3.42 Let $A_{m\times n}$ be a matrix with pseudoinverse $A^\dagger$. Show that if $y \in \mathrm{col}(A^T)$, then $A^\dagger Ay = y$.

3.43 Let $A_{m\times n}$ have singular values $\sigma_1 \geq \sigma_2 \geq \cdots$, and let A_k be a rank-k approximation of A; that is,

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.$$

Show that the singular values of $A - A_k$ are $\sigma_{k+1} \geq \sigma_{k+2} \geq \cdots \geq \sigma_r$, where $r = \mathrm{rank}(A)$.

3.44 Let $A_{n\times n}$ be nonsingular and have singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$. Show that in the 2-norm,

$$\mathrm{cond}(A) = \frac{\sigma_1}{\sigma_n}.$$

3.45 Consider the SVD of a matrix $A_{m\times n}$, with $\mathrm{rank}(A) = r$, and let A_k be the rank-k approximation of A, with $k < r$, obtained with such factorization as usual. Show that

$$\sigma_{k+1} = \|A - A_k\|_2 = \min_{rank(B)=k} \|A - B\|_2.$$

3.46 Let $A_{n\times n}$ be a symmetric matrix with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$. Show that

$$\lambda_1 = \min_{x\neq 0} \frac{x^T Ax}{x^T x} \quad \text{and} \quad \lambda_n = \max_{x\neq 0} \frac{x^T Ax}{x^T x}.$$

3.47 Let $\sigma_n > 0$ be the smallest singular value of a matrix $A_{n\times n}$. Show that the singular matrix B closest to A satisfies

$$\|A - B\|_2 = \sigma_n.$$

Hint: Start by using Exercise 3.46 for $A^T A$.

3.48 Let $A_{n\times n}$ be an spd matrix, and let $A = \mathcal{L}\mathcal{L}^T$ be its Choleski factorization. If $\mathcal{L} = U\Sigma V^T$ is the SVD of $\mathcal{L}$ and we define $B = U\Sigma U^T$, show that B is spd and

$$B^2 = A.$$

B is called a **square root** of A.

3.49 Find the square root of the matrix

$$A = \begin{bmatrix} 15 & -4 & 1 \\ -4 & 8 & 0 \\ 1 & 0 & 5 \end{bmatrix}.$$

3.50 True or False? Given an spd matrix $A_{n\times n}$, there is a unique matrix B such that $B^2 = A$.

3.51 Let as usual A_k represent the rank-k approximation of a matrix $A_{m\times n}$ with SVD $A = U\Sigma V^T$. Show that

$$\sigma_1(A - A_{k-1}) = \sigma_k(A).$$

3.52 Let $A_{m\times n}$, $B_{m\times n}$ be any two matrices with $m \geq n$, and assume that $\text{rank}(B) = r$. Show that

$$\sigma_1(A - B) \geq \sigma_{r+1}(A).$$

3.53 Let $A_{m\times n}, B_{m\times n}$ be two matrices. Show that

$$\sigma_{i+j-1}(A + B) \leq \sigma_i(A) + \sigma_j(B).$$

Hint: Apply Exercise 3.51 to each matrix A and B. Then use Exercise 3.52.

3.54 Let $A_{m\times n}$ be a matrix with $m \geq n$, that is perturbed by $E_{m\times n}$. Show that

$$|\sigma_i(A + E) - \sigma_i(A)| \leq \|E\|_2, \qquad i = 1, \ldots, n.$$

Hint: Use Exercise 3.53 twice.

3.55 Show that for a matrix $A_{n\times n}$, the set

$$\text{col}(A - \lambda I)^k, \qquad k = 1, 2, \ldots$$

is A-invariant.

3.56 Find the Schur factorization of $A = \begin{bmatrix} 5 & 3 & 0 \\ -3 & 5 & 0 \\ 1 & -2 & 7 \end{bmatrix}$.

3.57 Given a symmetric matrix $A_{n\times n}$, use its Schur factorization to show that it always has a complete set of orthonormal eigenvectors.

3.58 Let $A_{n\times n}$ be a symmetric matrix. Show that there exists an orthogonal matrix Q such that

$$Q^T A\, Q = \text{diag}(\lambda_1, \ldots, \lambda_n),$$

where $\lambda_1, \ldots \lambda_n$ are the eigenvalues of A.

3.59 **(Gershgorin)** Let $A_{n\times n}$ be symmetric and $Q_{n\times n}$ orthogonal. If $Q^T A\, Q = D+E$, where $D = \text{diag}(d_1, \ldots, d_n)$ and E has zero diagonal entries, show that

$$\sigma(A) \subseteq \cup_{i=1}^{n} [d_i - r_i,\ d_i + r_i],$$

where $r_i = \sum_{j=1}^{n} |e_{ij}|, \quad i = 1, \ldots, n.$

Hint: $(D - \lambda I) + E$ is singular. Use Exercise 1.23.

3.60 Show that for any given eigenvalue λ of a matrix $A_{n\times n}$,

$$\text{geometric multiplicity}(\lambda) \ \leq\ \text{algebraic multiplicity}.$$

3.61 For which type of matrices are the columns of Q in the Schur factorization $Q^T A\, Q$ eigenvectors of A?

3.62 Consider an arbitrary matrix $A_{n\times n}$. Show that there is an orthogonal matrix $Q_{n\times n}$ such that $Q^T A\, Q$ is upper Hessenberg.

3.63 For a matrix $A_{n\times n}$ and a nonzero vector $x \in \mathbb{R}^n$, we define the j–th Krylov subspace associated with A and x as

$$\mathcal{K}_j(A, x) = \text{span}\,\{x, Ax, A^2x, \ldots, A^{j-1}x\}.$$

Assume now that H is an upper Hessenberg matrix with the property that $h_{ij} \neq 0$ for all $i = j + 1$ (we say H is properly upper Hessenberg). Show that

$$\mathcal{K}_j(H, e_1) = \text{span}\,\{e_1, \ldots, e_j\}.$$

3.64 Let $A_{n\times n}$ be similar to a properly upper Hessenberg matrix B; that is, $B = Q^{-1} A\, Q$. If $Q = [q_1\ \cdots\ q_n]$, show that

$$\mathcal{K}_j(A, q_1) = \text{span}\,\{q_1, \ldots, q_j\}, \qquad j = 1, \ldots, n.$$

3.65 Let $U\Sigma V^T$ be the SVD of a matrix $A_{m\times n}$, with $m \geq n$. Use the orthogonal matrices U and V to find Schur factorizations of the matrices $A^T A$ and AA^T, respectively.

3.66 Let $U\Sigma V^T$ be the SVD of a matrix $A_{m\times n}$, with $m \geq n$ and $U = [U_1\quad U_2]$, where U_1 has n columns. Find an orthogonal matrix Q such that

$$Q^T \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} Q \ = \ \text{diag}(\sigma_1, \ldots, \sigma_n, -\sigma_1, \ldots, -\sigma_n, 0, \ldots, 0).$$

Hint: Note that $U_1^T AV = V^T A^T U_1$ and $U_2^T AV = 0$. Define Q as a 2×3 block square matrix of order $m + n$.

3.67 Consider two matrices $A_{m\times n}$ and $B_{m\times n}$. Show that the problem

$$\text{minimize } \|A - BQ\|_F, \qquad Q_{n\times n} \text{ orthogonal},$$

is equivalent to the problem of maximizing the trace of the matrix $Q^T B^T A$.

3.68 Consider two matrices $A_{m\times n}$ and $B_{m\times n}$, and let $B^T A = U\Sigma V^T$ be the singular value decomposition of the matrix $B^T A$. Show that the trace of the matrix $Q^T B^T A$ is maximized by $Q = UV^T$.

3.69 Consider two matrices $A_{m\times n}$ and $B_{r\times n}$. Let the columns of $Q = [q_1 \cdots q_k]$ form an orthonormal basis of $N(A)$ and the columns of $U = [u_1 \cdots u_q]$ form an orthonormal basis of $N(BQ)$. Show that the columns of QU form an orthonormal basis of $N(A) \cap N(B)$.

3.70 Let A and Q be two square matrices of order n, with Q orthogonal and partitioned as $Q = [Q_1 \quad Q_2]$, where Q_1 has k columns. Show that if col (Q_1) is an A-invariant subspace, then

$$Q^T A Q = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix},$$

where B is square of order k. What if A is symmetric?

3.71 A matrix $A_{n\times n}$ is **nilpotent** if $A^k = 0$, for some $k \geq 1$. Show that a matrix A is nilpotent if and only if all its eigenvalues are zero.

3.72 In the proof of Theorem 3.27, the deflation process takes place via an orthogonal matrix U whose first column is a unit eigenvector u_1 with eigenvalue $\lambda = \lambda_1$. How would this deflation work if instead we define an orthogonal matrix $U = [X \quad q]$, where q is a left eigenvector of A with eigenvalue λ?

3.73 Show that an eigenpair (R_{22}, Z) of $R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$ can be defined as $Z = \begin{bmatrix} C \\ I \end{bmatrix}$, for some matrix C, and that C is uniquely determined if the eigenvalues of R_{11} and R_{22} are disjoint. The dimensions of the matrices C and the identity I are according to the dimensions of R_{11} and R_{22}.

3.74 Give the details on the proof of Theorem 3.42.

3.75 Let λ be a simple eigenvalue of a matrix $A_{n\times n}$ with corresponding eigenvector v. Define the matrix $U = [v \quad X]$ to be nonsingular, and write its inverse as $U^{-1} = \begin{bmatrix} w^T \\ Y^T \end{bmatrix}$, for some w, Y with dimensions corresponding to those of v and X, respectively. Show that w is a left eigenvector of A.

3.76 Consider the matrix

$$R = \left[\begin{array}{cc|cc} 1 & -0.7639 & 4 & 5 \\ 5.2361 & 1 & 2 & 7 \\ \hline 0 & 0 & 4 & -1 \\ 0 & 0 & 1 & 4 \end{array}\right].$$

Interchange the eigenblocks via orthogonal transformations.

Hint: See Theorem 3.37 and Example 3.40.

3.77 Let $[X \quad Y]$ be an orthogonal matrix, and let $R = X^T A - LX^T$. Show that the matrix

$$L = X^T A X$$

minimizes the norm $\|R\|_2$.

3.78 Consider the following terms:

T1:	fire(d)	T4:	reelection
T2:	health(ier)	T5:	run(ning)
T3:	president	T6:	time

and consider the following documents

D1: Running is good for your health

D2: Bank president fired for running dirty business

D3: Time running out: fire destroys farms

D4: President Smith runs for reelection one more time

D5: University president offers healthier food at dining halls

Assume that you enter the query running for president. Following Example 3.45, set a cutoff value equal to 0.55 and find the relevant documents.

3.79 Following Exercise 3.78, compute approximations of rank 2 and rank 3. Which of the three, the original, the rank-2, or the rank-3 seem to pull the relevant documents accurately?

3.80 Create a term-document matrix $A_{7\times 5}$ with rank$(A) \geq 4$ and apply two low-rank approximations for query matching. Your query should have two words.

3.81 Create a term-document matrix $A_{7\times 5}$ with rank$(A) \geq 4$ and apply a low-rank approximation for term–term comparison.

3.82 Apply the rank-reduction approach to the matrix of Example 3.47, with the same query, using the SVD rank-2 approximation A_2. Use a cutoff of $\cos\theta \geq 0.6$. Does this approximation retrieve the right relevant documents?

3.83 Show that the digram frequency matrix of M_c is PAP^T, where P is the permutation matrix used to generate the cryptogram by permuting the letters of the alphabet.

3.84 What is the relationship between the singular vectors of the digram matrix A of the plain text and those of the digram matrix of the cryptogram?

3.85 Obtain paragraphs written in German and Spanish. Using the singular vectors of their digram frequency matrices, find out which letters of their corresponding alphabets are classified as vowels, consonants, and neuter. Show plots similar to Figures 3.2 and 3.3.

3.86 Following the notation in (3.70) and (3.71), show that the product $w^T Aw$ gives the number of vowel-vowel pairs, and, similarly, that $c^T Aw$ gives the number of consonant–vowel pairs.

CHAPTER 4

LEAST SQUARES

Beginning with the problem of a linear system $Ax = b$ that does not have a solution, in this short chapter, we investigate the possibility of finding a vector x^* that even though it does not solve the given system, it minimizes the error $\|Ax - b\|$. The matrix A under consideration is an arbitrary rectangular $m \times n$ matrix, but since the minimizing solution is related to the matrix $A^T A$, we start the exposition by studying quadratic functions where the matrix involved is symmetric positive definite. The minimization problem of least squares is directly related to orthogonal projection matrices; thus, we study such projections onto the column space of A and then we find explicit expressions for these orthogonal projections onto an arbitrary subspace S. The least squares problem can also be studied, especially in the computational framework, through matrix factorizations introduced in Chapter 3, in particular, QR and Choleski factorization. Finally, we also consider minimization problems under certain constraints by using Lagrange multipliers. We show that this can boil down to solving a linear system of equations.

4.1 PROJECTIONS AND NORMAL EQUATIONS

Recall that according to Theorem 1.49, the system $Ax = b$ has a solution if and only if b is in the column space of A, col(A). However, in several systems of this type coming from some applications, b is not necessarily in col(A). In such a case, the system $Ax = b$ has

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

no solution, and instead, what we do is to project the vector b onto col(A) and solve the modified system. This will give us some sort of the "closest" solution to that of the original problem. This is a minimization problem where we compute a solution to the modified system so that the distance or error $Ax - b$ is minimum in a certain sense.

Here we consider the case of a linear system of m equations in n unknowns, with $m > n$; i.e., the coefficient matrix $A_{m\times n}$ has more rows than columns. In this case we say that the problem is *overdetermined*, and it is not likely to have a solution because there are more equations than variables. We can also see this by first denoting the columns of A with a_i, $i = 1, \ldots, n$. Since Ax is a combination of the columns of A (see Definition 1.43), we have

$$Ax = c_1 a_1 + \cdots + c_n a_n,$$

for some scalars $c_1, \ldots, c_n$. If rank(A) $= r$, where $r \leq n$, then Ax lies in the r-dimensional column space of A. However, the m-dimensional vector b must also be in col(A) for the system $Ax = b$ to have a solution, which is very unlikely, since $m > n \geq r$. See Figure 4.1.

■ **EXAMPLE 4.1**

Consider solving the system $Ax = b$, where

$$A = \begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad b = \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix}.$$

The columns of A are linearly independent, and therefore the column space of A, col(A), is 2-dimensional. However, b is a three-dimensional vector, and it does not lie in col(A). See Figure 4.1. Therefore, the system $Ax = b$ does not have a solution. Observe however that for the vector $\tilde{b} = [5 \quad 5 \quad 4]^T$, the system $Ax = \tilde{b}$ does have a solution, because $\tilde{b}$ is precisely a linear combination of the columns of A. The solution is $x = [2 \quad 1]^T$.

Remark 4.1 *For overdetermined systems, the larger m is with respect to n, the more unlikely it is for b to "land" on col(A), whose dimension is at most n, and therefore the more unlikely it is for $Ax = b$ to have a solution.*

Before we show how to find a solution with minimum error in the case where the original system $Ax = b$ does not have a solution, we first study the following general **quadratic form**

$$Q(x) = \frac{1}{2} x^T A x - x^T b, \tag{4.1}$$

where A is an $n \times n$ symmetric positive definite (spd) matrix and $b \in \mathbb{R}^n$. This quadratic form $Q(x)$ is nothing more than a generalization of the usual quadratic function we are familiar with in one dimension:

$$f(x) = \frac{1}{2} a x^2 - bx.$$

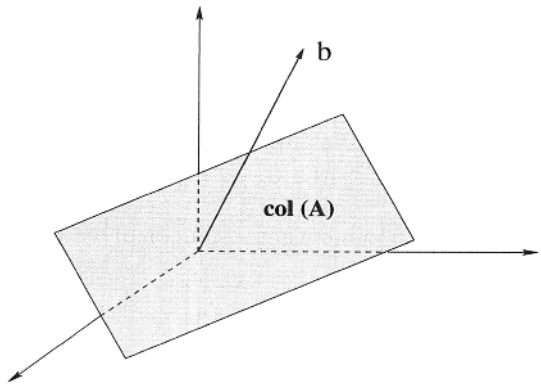


Figure 4.1 The vector b does not lie in $\mathrm{col}(A)$.

The coefficient $1/2$ is just for technical convenience. We know from calculus that the graph of f is a parabola and that $f(x)$ attains its minimum at the point $x = b/a$, if $a > 0$. In other words, the point x at which $f(x)$ reaches its minimum is the one for which

$$ax = b.$$

It should not be surprising then, but very interesting to learn, that the quadratic form $Q(x)$ is minimized exactly at the point x for which $Ax = b$ as the theorem below shows. This means that the minimization problem reduces to the problem of solving a linear system.

Recall that for a symmetric matrix $A_{n\times n}$ and any vectors $x, y \in \mathbb{R}^n$,

$$y^T A x = x^T A y. \tag{4.2}$$

Theorem 4.2 *Let A be an spd matrix. Then, the solution to the system $Ax = b$ is the vector x that minimizes the quadratic form (4.1).*

Proof. Let x be the solution of $Ax = b$, and let $y \in \mathbb{R}^n$ be arbitrary. Then,

$$\begin{aligned}
Q(y) - Q(x) &= \tfrac{1}{2}y^T A y - y^T b - \tfrac{1}{2}x^T A x + x^T b \\
&= \tfrac{1}{2}y^T A y - y^T A x - \tfrac{1}{2}x^T A x + x^T A x \\
&= \tfrac{1}{2}y^T A y - y^T A x + \tfrac{1}{2}x^T A x \\
&= \tfrac{1}{2}(y^T A y - y^T A x - x^T A y + x^T A x) \\
&= \tfrac{1}{2}(y - x)^T A (y - x) \geq 0
\end{aligned}$$

because A is positive definite. Therefore, $Q(y) \geq Q(x)$, for any $y \in \mathbb{R}^n$. This implies the minimum of Q happens at x, which is the solution of $Ax = b$.

□

■ **EXAMPLE 4.2**

Consider the quadratic form

$$Q(x_1, x_2) = \frac{1}{2}[x_1 \quad x_2]\begin{bmatrix} 4 & 10 \\ 10 & 6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - [x_1 \quad x_2]\begin{bmatrix} 5 \\ 3 \end{bmatrix}.$$

This quadratic form attains its minimum at the solution of the system

$$\begin{bmatrix} 4 & 10 \\ 10 & 6 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix},$$

that is, at $[x_1 \quad x_2]^T = [0 \quad 1/2]^T$.

Now, by going back to the problem of finding the least squares solution to a system $Ax = b$, where $A_{m\times n}$ is a general rectangular matrix, the illustration in Figure 4.3 indicates that when projecting b onto col(A), we would like to minimize the norm of the vector $b - Ax$. This is equivalent to finding the vector x that minimizes $\|Ax - b\|^2$. (This is where the name *least squares* comes from). However,

$$\|Ax - b\|^2 = (Ax - b)^T (Ax - b) = x^T A^T Ax - x^T A^T b - b^T Ax + b^T b.$$

We are not interested in the minimum value of $\|Ax - b\|^2$, but only on the value x for which that minimum is attained. Therefore, we can drop the constant term $b^T b$ since this does not affect the solution of the minimization problem. In that case, and since $b^T Ax = x^T A^T b$, we are in fact interested in the vector x that minimizes

$$x^T A^T Ax - 2x^T A^T b.$$

Finally, we divide by two (again, this does not change the solution of the minimization problem), and the problem reduces to minimizing

$$\frac{1}{2}x^T A^T Ax - x^T A^T b. \tag{4.3}$$

By comparing now (4.3) with the general quadratic form (4.1) and by using Theorem 4.2, we conclude that $\|Ax - b\|^2$ is minimized at the point x where

$$A^T A\, x = A^T b, \tag{4.4}$$

which is simply a linear systems of equations. The expression in (4.4) is known as the **normal equations**, whose solution is called the least squares solution of the system $Ax = b$. Thus, we have proved the following important result:

Theorem 4.3 *The least squares solution of the overdetermined system of equations $Ax = b$, where A is an $m \times n$ matrix is given by the solution of the normal equations*

$$A^T A\, x = A^T b.$$

Note: To make the notation more clear, when the overdetermined system $Ax = b$ does not have a solution, we find the least squares solutions x^* to the problem, which is the solution of the normal equations (4.4)

$$A^T A x^* = A^T b. \tag{4.5}$$

(i.e., x^* does not solve $Ax = b$).

■ **EXAMPLE 4.3**

Find the least squares solution of the system

$$\begin{array}{rcrcl} x_1 & + & 2x_2 & = & 2 \\ & & x_2 & = & 5 \\ x_1 & + & x_2 & = & 8. \end{array}$$

In this case, $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$, $b = [2 \quad 5 \quad 8]^T$. The vector b does not lie in the 2-dimensional col(A), and therefore, the system $Ax = b$ does not have a solution. To compute the least squares solution, first we find $A^T A = \begin{bmatrix} 2 & 3 \\ 3 & 6 \end{bmatrix}$ and $A^T b = [10 \quad 17]^T$. Therefore, the normal equations are

$$\begin{array}{rcl} 2x + 3y & = & 10, \\ 3x + 6y & = & 17, \end{array}$$

whose solution is $[x^* \quad y^*] = [3 \quad 4/3]$.

One classical application of least squares is the problem of fitting a line to a set of points. It is obvious that given a set of points, there is in general no line that will pass through all the points: The corresponding system $Ax = b$ is overdetermined and has no solution.

■ **EXAMPLE 4.4**

(Linear Regression) Consider the set of points.

t_i	0.78	0.80	0.82	0.84	0.86	0.88
y_i	0.70	0.72	0.73	0.74	0.77	0.77

We would like to find a line $y = ct + d$ that hopefully passes through all these points, that is, we would like to have

$$y_i = ct_i + d \quad \text{for} \quad i = 1, ..., 6.$$

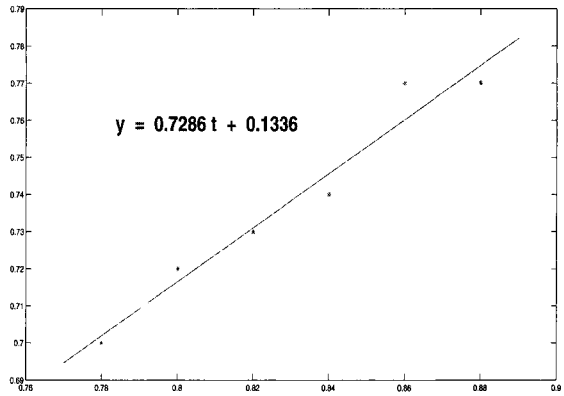


Figure 4.2 Least squares solution of (4.6).

In matrix form this can be written as

$$\begin{bmatrix} 0.78 & 1 \\ 0.80 & 1 \\ 0.82 & 1 \\ 0.84 & 1 \\ 0.86 & 1 \\ 0.88 & 1 \end{bmatrix} \begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} 0.70 \\ 0.72 \\ 0.73 \\ 0.74 \\ 0.77 \\ 0.77 \end{bmatrix}. \tag{4.6}$$

This overdetermined system $Ax = b$, with $x = [c \quad d]^T$ and $b = [y_1 \cdots y_6]^T$ has no solution, so instead we find the least squares solution by solving the normal equations $A^T Ax^* = A^T b$, where

$$A^T A = \begin{bmatrix} 4.1404 & 4.980 \\ 4.98 & 6 \end{bmatrix} \quad \text{and} \quad A^T b = \begin{bmatrix} 3.682 \\ 4.430 \end{bmatrix}.$$

The solution is $x^* = [\,c^* \quad d^*]^T = [\,0.7286 \quad 0.1336\,]^T$.

This means that the line that best fits the data in the sense of least squares is $y = 0.7286t + 0.1336$. See Figure 4.2.

Algorithm. A widely used method to solve the least squares problem in the case where the matrix $A_{m\times n}$ has full rank is using a Choleski factorization to solve the normal equations $A^T Ax = A^T b$. Namely,

- Factor $A^T A = LL^T$.
- Solve $Ly = A^T b$.
- Solve $L^T x = y$.

As mentioned before, the least squares solutions obtained in the examples above can also be obtained through projection matrices. (Orthogonal projection matrices were discussed in Section 1.9). The vector b is not in the col(A), and we can work instead with a vector $\tilde{b}$ in col(A) whose distance to b is minimum in the sense of least squares. Orthogonal projection matrices will provide the desired property of minimization.

First, we give a useful way of writing a general orthogonal projection matrix onto a vector subspace S.

Theorem 4.4 *Let P be the orthogonal projection onto a subspace S of $\mathbb{R}^n$, and let $\mathcal{B} = \{q_1, \ldots, q_k\}$ be an orthonormal basis of S. Then, for any $x \in \mathbb{R}^n$, we have*

$$Px = \sum_{i=1}^{k} (q_i^T x) q_i. \tag{4.7}$$

Proof. Since $\dim(S) = k$, then $\dim(S^\perp) = n-k$. If $\mathcal{C} = \{q_{k+1}, \ldots, q_n\}$ is an orthonormal basis of $S^\perp$, we can use the elements of $\mathcal{C}$ to extend the basis $\mathcal{B}$ into an orthonormal basis $\{q_1, \ldots, q_n\}$ of $\mathbb{R}^n$. Thus, any $x \in \mathbb{R}^n$ can be uniquely expressed as the linear combination

$$x = c_1 q_1 + \cdots + c_n q_n,$$

for some scalars $c_1, \ldots, c_n$. By using orthogonality, we note from the last equality that for $i = 1, \ldots, n$, we have $q_i^T x = c_i$. Then,

$$x = \sum_{i=1}^{n} (q_i^T x) q_i,$$

which we can split into two parts, one belonging to S and the other to $S^\perp$. More precisely, we can write

$$x = \sum_{i=1}^{k} (q_i^T x) q_i + \sum_{i=k+1}^{n} (q_i^T x) q_i.$$

Thus, by applying P, we get

$$Px = \sum_{i=1}^{k} (q_i^T x) q_i.$$

□

Remark 4.5 *In Section 1.9, we explicitly wrote the orthogonal projection matrix onto a subspace S as $P = QQ^T$, where the columns q_i of Q form an orthonormal basis of S. Then, with this P, Px gives exactly (4.7).*

The next theorem gives explicit expressions for the orthogonal projection matrix onto a vector subspace S, in particular, one that uses an orthogonal matrix Q formed with the vectors of an orthonormal basis of S. It then states the minimization property.

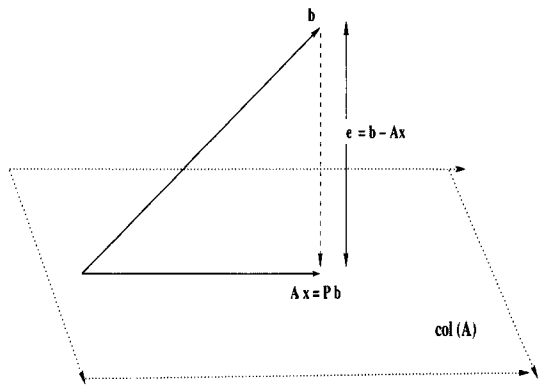


Figure 4.3 Projection of b onto col(A).

Theorem 4.6

1. *Assume the columns of A are linearly independent. If $S = \mathrm{col}\,(A)$, then, the orthogonal projection onto S is defined as*

$$P = A(A^T A)^{-1} A^T. \tag{4.8}$$

2. *If a vector subspace S is spanned by an orthonormal set $\{q_1, ..., q_k\}$, then the projection matrix onto S is $P = Q\,Q^T$, where $Q = [q_1 \cdots q_k]$.*

3. *Let S be a subspace of $\mathbb{R}^n$, and let P be the orthogonal projection onto S. Then $c := Pb$ is in S and for any $b \in \mathbb{R}^n$:*

$$\|b - c\| \leq \|b - w\|, \quad \forall\, w \in S.$$

Proof.

1. We proved in Section 1.9 that the matrix P in (4.8) satisfies $P^2 = P$ and $P^T = P$ and therefore is an orthogonal projection. Given an arbitrary $x \in \mathbb{R}^n$, it is clear that the vector $Px = A[(A^T A)^{-1} A^T x]$ is in col(A).

2. Since $Q = [q_1 \ \ldots \ q_k]$, for an arbitrary $x \in \mathbb{R}^n$, we have

$$Q^T x = \begin{bmatrix} q_1^T \\ \vdots \\ q_k^T \end{bmatrix} x = \begin{bmatrix} q_1^T x \\ \vdots \\ q_k^T x \end{bmatrix}. \quad \text{Then,}$$

$$QQ^T x = [q_1 \ \ldots \ q_k] \begin{bmatrix} q_1^T x \\ \vdots \\ q_k^T x \end{bmatrix} = \sum_{i=1}^{k} (q_i^T x) q_i = Px.$$

The last equality is true because of Theorem 4.4.

3. Let w be an arbitrary vector in S. Recall first that $P^{\perp} = I - P$ so that we can write $b = Pb + P^{\perp}b$. On the other hand, since $(Pb - w) \in S$ and $P^{\perp}b \in S^{\perp}$, they are orthogonal. Then, we have

$$\begin{aligned} \|b - w|^2 &= \|(Pb + P^{\perp}b) - w\|^2 = \|(Pb - w) + P^{\perp}b\|^2 \\ &= \|Pb - w\|^2 + \|P^{\perp}b\|^2 \\ &\geq \|P^{\perp}b\|^2 = \|b - Pb\|^2 = \|b - c\|^2. \end{aligned}$$

□

Part 1 of Theorem 4.6 gives an explicit representation of the projection matrix P onto the column space of A. As already noted in Section 1.9 however, this representation is not very useful in practice due to the computation of inverse matrices.

Part 2 gives the most efficient way to compute such a projection matrix by using orthogonal matrices whose columns are vectors of an orthonormal basis of the vector subspace S under consideration, which in particular could be col(A).

Finally, Part 3 states that the projected vector $c = Pb$ onto S is the closest to b among all other vectors in S. This is exactly what we want for the least squares solutions of $Ax = b$, where in this case S =col(A).

As expected, finding the least squares solution through the normal equations is equivalent to using the orthogonal projection matrix P to project b onto col(A) and then solving the new system on that space. More precisely, suppose that for a given matrix $A_{m\times n}$, the set $\{q_1, \ldots, q_k\}$ is an orthonormal basis for the col(A), where $k \leq n$. Then we define the matrix $Q = [q_1 \ldots q_k]$, and the projection matrix onto col(A) is given by

$$P = QQ^T. \tag{4.9}$$

Recall that the original system $Ax = b$ does not have a solution because b does not lie in col(A). The vector Pb however does lie in col(A) by definition, and therefore the system

$$Ax = Pb \tag{4.10}$$

has a solution. The solution $x = x^*$ of (4.10) is precisely the least squares solution of the original system $Ax = b$. That is, x^* coincides with the solution to the normal equations (4.5). This is illustrated in the following example.

■ **EXAMPLE 4.5**

Consider the linear regresion problem of Example 4.4; that is, we want to find the least squares solution of the system (4.6). The two columns of A are linearly independent, and we can find an orthonormal basis $\{q_1, q_2\}$ for col(A), e.g., applying Gram–Schmidt to the columns of A to obtain

$$Q = [q_1 \quad q_2] = \begin{bmatrix} -0.3982 & -0.6044 \\ -0.4022 & -0.3654 \\ -0.4062 & -0.1264 \\ -0.4102 & 0.1127 \\ -0.4142 & 0.3517 \\ -0.4182 & 0.5907 \end{bmatrix},$$

so that now the orthogonal projection matrix onto col(A) is

$$P = QQ^T = \begin{bmatrix} 0.5238 & 0.3810 & 0.2381 & 0.0952 & -0.0476 & -0.1905 \\ 0.3810 & 0.2952 & 0.2095 & 0.1238 & 0.0381 & -0.0476 \\ 0.2381 & 0.2095 & 0.1810 & 0.1524 & 0.1238 & 0.0952 \\ 0.0952 & 0.1238 & 0.1524 & 0.1810 & 0.2095 & 0.2381 \\ -0.0476 & 0.0381 & 0.1238 & 0.2095 & 0.2952 & 0.3810 \\ -0.1905 & -0.0476 & 0.0952 & 0.2381 & 0.3810 & 0.5238 \end{bmatrix}$$

and the projected vector is

$$Pb = [0.7019 \;\; 0.7165 \;\; 0.7310 \;\; 0.7456 \;\; 0.7602 \;\; 0.7748]^T.$$

Since Pb is in col(A), the system $Ax = Pb$:

$$\begin{bmatrix} 0.78 & 1 \\ 0.80 & 1 \\ 0.82 & 1 \\ 0.84 & 1 \\ 0.86 & 1 \\ 0.88 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.7019 \\ 0.7165 \\ 0.7310 \\ 0.7456 \\ 0.7602 \\ 0.7748 \end{bmatrix}$$

has the solution $x = x^* = [0.7286 \quad 0.1336]^T$, which is exactly what we obtained through normal equations in Example 4.4.

We have seen in Example 4.5 what is a general fact: The least squares solutions of $Ax = b$ obtained through (4.4) or (4.10) are equivalent. Indeed, if we assume A is full-rank, then the matrix A^TA is nonsingular. Thus, let us start with equation (4.4) and solve for x

$$A^TAx = A^Tb \quad \Longrightarrow \quad x = (A^TA)^{-1}A^Tb.$$

Now multiply by A on both sides and use (4.8) to get

$$Ax = A(A^TA)^{-1}A^Tb = Pb.$$

Similarly,

$$Ax = Pb \quad \Longrightarrow \quad A^TAx = A^TPb = A^TA(A^TA)^{-1}A^Tb = A^Tb.$$

4.2 LEAST SQUARES AND QR FACTORIZATION

In Theorem 4.6, we saw how an orthonormal set of vectors can be used to obtain a simple way of defining a projection matrix. Orthonormal sets are the ideal set of vectors to be used in numerical computations. We have already mentioned in Section 3.2 some applications of QR factorization. Recall also that the QR factorization of a full–rank matrix A is nothing

else but a matrix form of describing the Gram–Schmidt orthonormalization process; the columns of Q form a set of orthonormal vectors obtained from the (linearly independent) columns of A.

Given a general matrix $A_{m\times n}$, consider now solving the linear system

$$A\,x = b. \tag{4.11}$$

Suppose that A has the QR factorization $A = QR$; then by substituting this into (4.11), we get $QRx = b$. Multiplying by Q^T on both sides of this equation, we obtain

$$R\,x = Q^T b. \tag{4.12}$$

Two important observations are in order. First, we have transformed the original system (4.11) into a much simpler triangular system (4.12). Second, this reduction to a triangular system works for general $m \geq n$. The case $m = n$ has been already discussed in Section 3.2, but in general, the matrix A is not necessarily square, in which case (4.12) gives the least squares solution of (4.11); that is, it gives the vector x for which $\|Ax - b\|$ is minimum. As a matter of fact, it can be shown that equation (4.12) is equivalent to the normal equations (4.4). See Exercise 4.17.

■ **EXAMPLE 4.6**

Consider again the linear regression problem of Example 4.4. For the given 6×2 matrix A, we have $A = QR$ with

$$Q = \begin{bmatrix} -0.3833 & -0.6139 \\ -0.3932 & -0.3751 \\ -0.4030 & -0.1362 \\ -0.4128 & 0.1026 \\ -0.4226 & 0.3415 \\ -0.4325 & 0.5803 \end{bmatrix}, \qquad R = \begin{bmatrix} -2.0348 & -2.4474 \\ 0 & -0.1007 \end{bmatrix}.$$

Then we solve the triangular system $Rx = Q^T b$

$$\begin{bmatrix} -2.0348 & -2.4474 \\ 0 & -0.1007x \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1.8095 \\ -0.0135 \end{bmatrix}$$

to obtain the least squares solution $[x_1 \quad x_2]^T = [0.7286 \quad 0.1336]^T$ of the original system $Ax = b$.

■ **EXAMPLE 4.7**

Consider fitting the best line to the three points:

$$(1,2),\ (2,5),\ (3,7).$$

From the linear regression point of view, this means that we are trying to solve $Ax = b$, or

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 7 \end{bmatrix}.$$

QR factorization of A gives

$$Q = \begin{bmatrix} 1/\sqrt{14} & 4/\sqrt{21} \\ 2/\sqrt{14} & 1/\sqrt{21} \\ 3/\sqrt{14} & -2/\sqrt{21} \end{bmatrix}, \quad R = \begin{bmatrix} \sqrt{14} & 6/\sqrt{14} \\ 0 & \sqrt{21}/7 \end{bmatrix}.$$

Then, $Q^T b = \begin{bmatrix} -33/\sqrt{4} \\ -1/\sqrt{21} \end{bmatrix}$ and the solution to $Rx = Q^T b$ is $x = \begin{bmatrix} 5/2 \\ -1/3 \end{bmatrix}$, which means that the best fit line is $y = \frac{5}{2}x - \frac{1}{3}$.

The rank–deficient case. If the matrix $A_{m\times n}$ has rank $r < n$, then we can still solve the least squares problem but the solution is not unique. In fact, in such a case, we have dim $N(A) = n - r$, and the null space contains more than just the zero vector.

Thus, suppose x minimizes the norm $\|Ax - b\|$ and let $v \in N(A)$ be arbitrary. Then, we have $\|A(x+v) - b\| = \|Ax + Av - b\| = \|Ax - b\|$, which is minimum. This means that now we have an infinite number of solutions. Then, we can define the set of all minimizers

$$M = \{x \in \mathbb{R}^n : \|Ax - b\|_2 = \min\}. \tag{4.13}$$

One can prove that this set is convex (Exercise 4.14).

Out of the many least squares solutions, we are interested in computing the one with the smallest norm, denoted with $\hat{x}$. This can be done using orthogonal matrices.

Theorem 4.7 *Let $A_{m\times n}$ be a matrix with rank$(A) = r$ and $b \in \mathbb{R}^m$ arbitrary. If Q and U are square orthogonal matrices such that*

$$Q^T A U = \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix}_{m\times n},$$

where B is a square matrix of order r, then the least squares solution of $Ax = b$ with the smallest norm is given by

$$\hat{x} = U \begin{bmatrix} B^{-1}u \\ 0 \end{bmatrix},$$

for some vector u.

Proof. *Let us denote*

$$U^T x = \begin{bmatrix} v \\ w \end{bmatrix}, \quad Q^T b = \begin{bmatrix} u \\ z \end{bmatrix}, \quad \text{and} \quad C = Q^T A U,$$

where $u, v \in \mathbb{R}^r$, $w \in \mathbb{R}^{n-r}$, *and* $z \in \mathbb{R}^{m-r}$. *Then, we have*

$$\begin{aligned} \|Ax - b\|^2 &= \|Q^T A U U^T x - Q^T b\|^2 = \|C U^T x - Q^T b\|^2 \\ &= x^T U C^T C U^T x - 2x^T U C^T Q^T b + b^T Q Q^T b \\ &= [v^T \quad w^T] \begin{bmatrix} B^T B & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} - 2[v^T \quad w^T] \begin{bmatrix} B^T u \\ 0 \end{bmatrix} + \|b\|^2 \\ &= v^T B^T B v - 2v^T B^T u + \|u\|^2 + \|z\|^2 \\ &= \|Bv - u\|^2 + \|z\|^2, \end{aligned}$$

where we have used the fact that $\|b\|^2 = \|Q^T b\|^2 = \|u\|^2 + \|z\|^2$. *Thus, for* x *to minimize* $\|Ax - b\|^2$, *we must have* $v = B^{-1}u$, *and if we want* x *to have minimal norm, then we must have* $w = 0$. *Hence, the least squares solution with the smallest norm is*

$$x = U \begin{bmatrix} v \\ w \end{bmatrix} = U \begin{bmatrix} B^{-1}u \\ 0 \end{bmatrix}.$$

□

What if b is in col(A)? The problem of least squares deals with the case of solving systems $Ax = b$ when the vector b is not in col(A). Now let us assume that $b \in \text{col}(A)$, where again A is an $m \times n$ matrix, that is, the overdetermined system has at least one solution. The new question is, out of all the solutions to the system $Ax = b$, which one has the smallest norm? The answer is given in the following theorem.

Theorem 4.8 *Let* A *be an* $m \times n$ *matrix, and let* P_r *denote the orthogonal projection onto* row(A). *If* x *solves the system* $Ax = b$, *then* $x^* = P_r x$ *is also a solution of the system and*

$$\|x^*\|^2 \leq \|x\|^2.$$

Proof. Let x be a solution of $Ax = b$. First note that the vector $x - P_r x = (I - P_r)x = P_r^{\perp} x \in \text{row}(A)^{\perp}$, and by Theorem 1.46 we know that $\text{row}(A)^{\perp} = N(A)$. Then, we must have $A(x - P_r x) = 0$. Thus,

$$b = Ax = A(P_r x + x - P_r x) = AP_r x + A(x - P_r x) = AP_r x = Ax^*.$$

Then, $x^* = P_r x$ is indeed a solution of $Ax = b$. Finally, by using Theorem 4.6, we have that $\|x^*\|^2 = \|P_r x\|^2 = \|x - P_r^{\perp} x\|^2 \leq \|x - 0\|^2 = \|x\|^2$.

□

■ **EXAMPLE 4.8**

Let $A = \begin{bmatrix} 1 & 3 & 2 \\ 2 & 1 & 4 \\ 1 & 2 & 2 \\ 4 & 1 & 8 \end{bmatrix}$, and let $b = \begin{bmatrix} 0 \\ 5 \\ 1 \\ 11 \end{bmatrix}$. It is simple to verify that $b \in \text{col}(A)$. A solution to the system $Ax = b$ is $x = [1 \ -1 \ \ 1]^T$, and an orthonormal basis of row(A) is given by $\mathcal{B} = \{q_1, q_2\}$, where

$$q_1 = [0.2673 \quad 0.8018 \quad 0.5345]^T, \quad q_2 = [-0.3586 \quad 0.5976 \ -0.7171]^T.$$

Then, with $Q = [q_1 \ \ q_2]$, the matrix $P_r = QQ^T = \begin{bmatrix} 0.2 & 0 & 0.4 \\ 0 & 1 & 0 \\ 0.4 & 0 & 0.8 \end{bmatrix}$ is the orthogonal projection onto row(A). If we define

$$x^* = P_r x = [0.6 \ \ -1 \ \ \ 1.2]^T,$$

then one can verify that indeed x^* is also a solution of $Ax = b$, and

$$1.6733 = \|x^*\| \leq \|x\| = 1.7321.$$

4.3 LAGRANGE MULTIPLIERS

In Section 4.1 we have dealt with the problem of minimizing the quadratic function $Q(x) = \frac{1}{2}x^T A x - x^T b$, and we have seen its relationship with the least squares solution of an overdetermined system of linear equations. Now we want to extend this minimization problem to the case where a certain constraint has to be satisfied.

Recall from calculus, that given two functions $f, g : R^m \to R$, in order to

$$\text{minimize } f \qquad \text{subject to} \quad g(x) = 0,$$

we solve the equation

$$\nabla(f + \lambda g) = 0, \quad \text{or} \qquad \nabla f = \lambda \nabla g, \tag{4.14}$$

where ∇F denotes the gradient of F, and $\lambda \in R$. The scalar λ is called the Lagrange multiplier.

■ **EXAMPLE 4.9**

To minimize $f(x, y) = x^2 + (y - 2)^2$, subject to $x^2 - y^2 = 1$, we let $g(x, y) = x^2 - y^2 - 1$. Then $\nabla f(x) = \lambda \nabla g(x)$ gives $[\,2x \quad 2(y-2)\,]^T = \lambda[\,2x \quad -2y\,]^T$, or

$$\begin{aligned} 2x &= 2\lambda x, \\ 2(y-2) &= -2\lambda y. \end{aligned} \qquad \text{Then,} \qquad \lambda = 1, \quad x = \pm\sqrt{2}, \quad y = 1,$$

where we have used the equation $x^2 - y^2 - 1 = 0$. The corresponding minimum value is $f(\pm\sqrt{2},\ 1) = 3$.

In general, observe that if we have two functions $f, g : \mathbb{R}^n \to \mathbb{R}$, and if we define

$$L(x, y, \lambda) = f(x, y) + \lambda g(x, y),$$

then

$$\nabla L(x, y, \lambda) = \begin{bmatrix} \frac{\partial f}{\partial x} + \lambda \frac{\partial g}{\partial x} \\ \frac{\partial f}{\partial y} + \lambda \frac{\partial g}{\partial y} \\ g \end{bmatrix} = \begin{bmatrix} \nabla f + \lambda \nabla g \\ g \end{bmatrix}.$$

The solution to the minimization problem is found by solving $\nabla L = 0$.

Example 4.9 was a one-dimensional minimization problem. We now cover the same subject, from the point of view of matrices and linear algebra, with n-dimensional Lagrange multipliers, and for a special case of functions. More precisely, we consider the problem

$$\text{Minimize} \quad Q = \frac{1}{2} x^T A x - x^T b, \quad \text{subject to} \quad B^T x = h, \tag{4.15}$$

where A is an spd matrix of order n, B is an $n \times m$ matrix, $b \in \mathbb{R}^n$, and $h \in \mathbb{R}^m$.

Observe that the quadratic form Q in (4.15) is the same as the one introduced in (4.1), only that now we want to perform a constrained minimization of such a quadratic function.

To establish the solution of the minimization problem, we introduce a vector of Lagrange multipliers $y = [y_1 \ \cdots \ y_n]^T$ and following the idea from the one-dimensional case of forming the function $f + \lambda g$, here we form a new function:

$$L = Q + y^T G, \quad \text{with} \quad G = B^T x - h.$$

That is,

$$L(x, y) = \frac{1}{2} x^T A x - x^T b + (By)^T x - y^T h. \tag{4.16}$$

Then we need to solve $\nabla L = 0$: $\frac{\partial}{\partial x} L(x, y) = 0$, and $\frac{\partial}{\partial y} L(x, y) = 0$, which gives

$$\begin{aligned} Ax - b + By &= 0 \\ B^T x - h &= 0, \end{aligned} \tag{4.17}$$

which can also be written as

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ h \end{bmatrix}. \tag{4.18}$$

The systems (4.17) or (4.18) are known as **equilibrium equations**, which as we see are nothing else but the critical equation $\nabla L = 0$ written in matrix form.

Remark 4.9 *What we learn from above is that the constrained minimization problem (4.15) is equivalent (as it was in the one-dimensional case) to solving a linear system of equations.*

■ **EXAMPLE 4.10**

Minimize $Q = \frac{1}{2}(x_1^2 + x_2^2)$, subject to $x_1 + x_2 = 2$.
This can be written as:

Minimize $Q = \frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix}$,

subject to $\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2.$

Then, the matrices and vectors in (4.15) are $A = I$, $b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $h = 2$, and the equilibrium equilibrium equations (4.18) are

$$\left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 1 \\ \hline 1 & 1 & 0 \end{array}\right]\begin{bmatrix} x_1 \\ x_2 \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}.$$

Solving this system we get $x_1 = 1$, $x_2 = 1$ and the Lagrange multiplier $y = -1$. Then, the vector $\begin{bmatrix} 1 & 1 \end{bmatrix}^T$ minimizes Q subject to $x_1 + x_2 = 2$.

In Example 4.10, we have only the one-dimensional case (only one Lagrange multiplier) so that it can also be solved directly by applying the calculus technique introduced above. The next example is two-dimensional.

■ **EXAMPLE 4.11**

Minimize $Q = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2)$, subject to $x_1 - x_2 = 1$, $x_2 - x_3 = 2$.
This can be written as:

Minimize $Q = \frac{1}{2}\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$,

subject to $\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$

Then, in this case, the matrices and vectors in (4.15) are $A = I$, $b = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$,

$h = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $B = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$, and the equilibrium equations (4.18) are

$$\left[\begin{array}{ccc|cc} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & 1 \\ 0 & 0 & 1 & 0 & -1 \\ \hline 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 2 \end{bmatrix},$$

where y_1, y_2 are the Lagrange multipliers. The solution to the system is $x_1 = 4/3$, $x_2 = 1/3$, $x_3 = -5/3$, $y_1 = -4/3$, $y_2 = -5/3$. Then, $[\frac{4}{3} \ \ \frac{1}{3} \ \ -\frac{5}{3}]^T$ minimizes Q, subject to the conditions $x_1 - x_2 = 1$, $x_2 - x_3 = 2$.

4.4 FINAL REMARKS AND FURTHER READING

Because of the importance and usefulness of least squares, we have devoted one short chapter to highlighting the main concepts involved in this problem. It is an application of several topics in linear algebra and matrix computations, including: linear independence, rank, column space, positive definite matrices, projection matrices, linear systems, and QR and Choleski factorizations.

The discussion started with the problem that a vector b is not in the column space of A, and therefore the system $Ax = b$ does not have a solution. However, we have also discussed the case when b is in col(A) and we are instead interested in finding the solution with smallest magnitude. Our study has also considered the rank-deficient case, that is, when the matrix $A_{m\times n}$ does not have full rank. We have finished the chapter with a short exposition of Lagrange multipliers as tools for solving a minimization problem under constraints.

The books by Strang [64], Meyer [49], and Golub and Van Loan[25] are excellent references for studying least squares and Lagrange multipliers, both, theoretically and with an eye to some simple applications.

EXERCISES

4.1 Let $A = \begin{bmatrix} 2 & 3 \\ 1 & 1 \\ 7 & -5 \end{bmatrix}$, and let $b_1 = [2 \ \ 1 \ \ 0]^T$, $b_2 = [-5 \ \ -1 \ \ 29]^T$. Are b_1, b_2 in col(A)?

4.2 For which vector $[x \ \ y]^T$ is the quadratic form $f(x, y) = 2x^2+10xy+13y^2-2x-4y$ minimum?

4.3 Find the least squares solution of the system

$$\begin{array}{rcrcl} x_1 & + & 3x_2 & = & 1 \\ x_1 & & 4x_2 & = & 5 \\ x_1 & + & x_2 & = & 0. \end{array}$$

4.4 Consider the function

$$f(x, y) = ax^2 + 2bxy + cx^2.$$

Under what conditions on a and b is $f(x, y) > 0$?

4.5 Let $A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$. Row reduce A to an upper triangular matrix using the usual Gaussian elimination. What are the pivot entries? How is this matrix A related to the function f in Exercise 4.4?

4.6 First find the least squares solution for

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-5} \\ 0 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}.$$

Then, modify the entry $(3, 2)$ of A from 0 to 10^{-8}, and solve the new least squares problem. How does this small perturbation affect the solution to the least squares problem?

4.7 Solve the least squares problem of Example 4.4 by using the Choleski factorization of the matrix $A^T A$.

4.8 Consider the data

t_i	1.0	1.2	1.6	1.8	2.0	2.4
y_i	3	5	2	4	6	5

Solve the linear regression problem following the method used in Example 4.5.

4.9 The linear regression problem of finding a line $y = ct + d$ that fits a set of points (refer to Example 4.4) reduces to solving the normal equations $A^T Ax = A^T b$, where

$$A = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}, \qquad \text{and} \quad x = \begin{bmatrix} c \\ d \end{bmatrix}.$$

Show that these normal equations amount to solving the equations

$$\begin{aligned} md + \left(\sum_{i=1}^{m} t_i\right) c &= \sum_{i=1}^{m} b_i, \\ \left(\sum_{i=1}^{m} t_i\right) d + \left(\sum_{i=1}^{m} t_i^2\right) c &= \sum_{i=1}^{m} t_i b_i. \end{aligned}$$

4.10 Suppose you want to fit a cubic polynomial $P(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$ to a set of m points $(t_1, b_1), \ldots, (t_m, b_m)$ so as to minimize the error in the sense of least squares.

Show that the problem is equivalent to minimize $\|Ax - b\|^2$, where A is an $m \times 4$ matrix, $b \in \mathbb{R}^m$, and $x \in \mathbb{R}^4$.

4.11 Determine the cubic polynomial $p(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$ that best fits the following data in the sense of least squares.

t_i	-0.8	-0.4	0	0.2	0.6	0.8
y_i	-2	2	3	2.8	4	5

4.12 Let $Q_{m\times m}$ be an orthogonal matrix such that

$$QA = \begin{bmatrix} R \\ 0 \end{bmatrix} \quad \text{and} \quad Qb = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix},$$

where A is an $m \times n$ matrix $(m > n)$, $c_1 \in \mathbb{R}^n$, $c_2 \in \mathbb{R}^{m-n}$, and R is a square matrix of order n. Show that the least squares solutions of $Ax = b$ and $Rx = b_1$ are the same.

4.13 Let $A_{n\times n}$ be an spd matrix. Show that the function $F(x) = \frac{1}{2}x^T Ax - x^T b$ can be written as

$$F(x) = \frac{1}{2}(x - A^{-1}b)^T A\,(x - A^{-1}b) - \frac{1}{2}\,b^T A^{-1} b.$$

(Note that F attains its minimum at $x = A^{-1}b$).

4.14 Show that the set M in (4.13) is convex.

4.15 Assume that a matrix $A_{m\times n}$ has full rank. Show that the matrix $A^T A$ is spd.

4.16 Suppose that you are finding the least squares solution of $Ax = b$ and

$$A^T A = \begin{bmatrix} 2 & 4 & 2 \\ 4 & 5 & 4 \\ 2 & 4 & 2 \end{bmatrix}.$$

Is $A^T A$ symmetric positive definite? What can you say about the least squares solution?

4.17 Show that equation (4.12) is equivalent to the normal equations (4.4).

4.18 Let $S = \mathrm{col}(A)$, and let $W = S^\perp$, where A is an $m \times n$ matrix. If for an arbitrary vector $b \in \mathbb{R}^m$ we denote with d_s its distance to S and with d_w its distance to W, show that

$$\|b\|^2 = d_s^2 + d_w^2 = \min_x \|Ax - b\|^2 + \min_{A^T y = 0} \|b - y\|^2.$$

4.19 What is the distance from the point $(0, 0, 0)$ to the plane $x_1 + 4x_2 + 4x_3 = 12$?

4.20 Let $V = U \oplus W$, with $\dim(V) = n$ and $\dim(U) = r$, and let $\mathcal{B}_\mathcal{U} = \{x_1, \ldots, x_r\}$ and $\mathcal{B}_\mathcal{W} = \{y_1, \ldots, y_{n-r}\}$ be bases of U and W, respectively. Define the matrix

$$M = [\,x_1 \;\cdots\; x_r \,|\, y_1 \;\cdots\; y_{n-r}\,].$$

Let I denote the identity matrix of order r. Show that the projection matrix $P_{n\times n}$ onto U along W can be written as

$$P = M \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} M^{-1}.$$

4.21 Show that if P is a projection matrix ($P^2 = P$), then $\text{col}(P)$ and $N(P)$ are complementary spaces.

4.22 Let U and W be subspaces of V with $\dim(V) = n$, $\dim(U) = r$, and $W = V^{\perp}$. Show that there exists an orthogonal matrix Q such that the orthogonal projection matrix $P_{n\times n}$ onto U along W can be written as

$$P = Q \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} Q^T.$$

Hint: See Exercise 4.20

4.23 Let U and W be subspaces of V with $\dim(V) = n$, $\dim(U) = r$, and $W = V^{\perp}$. If B is a matrix whose columns form a basis of U, show that the orthogonal projection matrix onto U is given by

$$P = B(B^T B)^{-1} B^T.$$

4.24 Let U be a subspace of a vector space V, and let $P_{n\times n}$ be the orthogonal projection matrix onto U. For $x \in V$, define the subspace $S = x + U$. If $b \in V$, find the orthogonal projection of b onto S.

4.25 Use Lagrange multipliers to find the points (x, y) at which the function $f(x, y) = xy$ attains its minimum value when subject to the constraint $\frac{x^2}{8} + \frac{y^2}{2} = 1$.

4.26 Use Lagrange multipliers to find the dimensions of a box of maximum volume such that its surface area is fixed at 52 in^2.

4.27 Find the line that best fits the following points:

$$(1, 1.9), (2, 3.7), (3, 6.2), (4, 7.7), (5, 10.5),$$

in the sense of least squares, by using QR factorization.

CHAPTER 5

IMAGE COMPRESSION

One of the central issues in information technology is the representation of data by arrays of bits in the most efficient way possible, a never-ending quest for improvement in the representation of bits that are smaller, faster and cheaper. This is exactly the role of data compression: to convert strings of bits into shorter ones for more economical transmission, storage and processing. Abundant applications require such compression process: medical imaging, publishing, graphic arts, digital photography, wire photo transmission, and so on.

For the past few years, the Joint Photographic Experts Group (**JPEG**) has been working to keep an international compression standard for both, grayscale and color images. It is no surprise that strong mathematical research in this direction has been going on since then, and it is important to remark that when JPEG conducted a first selection process in 1988, they reported that a proposal based on the discrete cosine transform had produced the best picture quality. As a matter of fact, JPEG is a format for image compression based on the discrete cosine transform, which is used to reduce the file size of an image as much as possible without affecting the quality of the image as experienced by the human sensory system.

In this chapter, we present an elegant application of mathematical tools and concepts (in particular from linear algebra and numerical analysis) to the problem of image compression, and we illustrate how certain theoretical mathematical results can be effectively used

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

in applications that include the response of our vision system to changes in the image representation.

5.1 COMPRESSING WITH DISCRETE COSINE TRANSFORM

We will study two main techniques for compressing images. First we introduce a technique that is currently used for compressing most images available on the Internet and that uses a square orthogonal matrix of order eight to perform the corresponding transformation of coordinates (from space to frequency). Later on, and for completion, we study image compression as an application of the SVD factorization of a matrix A. Thus, orthogonality is present in both approaches.

We are interested in the two-dimensional discrete cosine transform, but we start with its one-dimensional version.

5.1.1 1-D discrete cosine transform

Through the discrete cosine transform we can combine and apply concepts such as orthogonality, interpolation, least squares, as well as linear combinations of basis functions in vector spaces. This transform is a very special mathematical tool that will allow us to separate and order an image into parts of differing importance, with respect to the image visual quality. We start with the one-dimensional case.

Definition 5.1 *Define the following $n \times n$ orthogonal matrix:*

$$C = \sqrt{\frac{2}{n}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \cdots & \frac{1}{\sqrt{2}} \\ \cos\frac{\pi}{2n} & \cos\frac{3\pi}{2n} & \cdots & \cos\frac{(2n-1)\pi}{2n} \\ \cos\frac{2\pi}{2n} & \cos\frac{6\pi}{2n} & \cdots & \cos\frac{2(2n-1)\pi}{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \cos\frac{(n-1)\pi}{2n} & \cos\frac{(n-1)3\pi}{2n} & \cdots & \cos\frac{(n-1)(2n-1)\pi}{2n} \end{bmatrix}. \tag{5.1}$$

Given a vector $x = [x_0 \cdots x_{n-1}]^T$, the discrete cosine transform (DCT) of x is the vector $y = [y_0 \cdots y_{n-1}]^T$ given by

$$y = Cx. \tag{5.2}$$

■ **EXAMPLE 5.1**

For $n = 8$, the matrix C in (5.1) is

$$\begin{bmatrix} 0.3536 & 0.3536 & 0.3536 & 0.3536 & 0.3536 & 0.3536 & 0.3536 & 0.3536 \\ 0.4904 & 0.4157 & 0.2778 & 0.0975 & -0.0975 & -0.2778 & -0.4157 & -0.4904 \\ 0.4619 & 0.1913 & -0.1913 & -0.4619 & -0.4619 & -0.1913 & 0.1913 & 0.4619 \\ 0.4157 & -0.0975 & -0.4904 & -0.2778 & 0.2778 & 0.4904 & 0.0975 & -0.4157 \\ 0.3536 & -0.3536 & -0.3536 & 0.3536 & 0.3536 & -0.3536 & -0.3536 & 0.3536 \\ 0.2778 & -0.4904 & 0.0975 & 0.4157 & -0.4157 & -0.0975 & 0.4904 & -0.2778 \\ 0.1913 & -0.4619 & 0.4619 & -0.1913 & -0.1913 & 0.4619 & -0.4619 & 0.1913 \\ 0.0975 & -0.2778 & 0.4157 & -0.4904 & 0.4904 & -0.4157 & 0.2778 & -0.0975 \end{bmatrix}.$$

We can readily verify that this matrix (up to rounding) is in fact orthogonal; that is, $C^T C = I$. Now define the vector

$$x = [1 \quad 2 \quad -2 \quad 0 \quad 1 \quad 4 \quad 0 \quad -1]^T$$

Then, the DCT of x is $y = Cx$, where

$$y = [1.7678 \quad 0.0480 \quad -0.4619 \quad 3.8565 \quad -1.0607 \quad -1.4262 \quad -0.1913 \quad -2.3645]^T.$$

Note: Observe the sign pattern in the rows or columns of the matrix C in Example 5.1.

To appreciate how the DCT will allow us to compress data, we introduce a theorem that through interpolation of an input vector x, it explicitly arranges the elements of its DCT $y = Cx$ in order of importance, as coefficients of a linear combination of (basis) cosine functions.

Theorem 5.2 (DCT Interpolation Theorem) *Let C be the matrix in (5.1), and let $x = [x_0 \cdots x_{n-1}]^T$. If $y = [y_0 \cdots y_{n-1}]^T$ is the DCT of x ($y = Cx$), then the function*

$$P_n(t) = \frac{1}{\sqrt{n}} y_0 + \sqrt{\frac{2}{n}} \sum_{k=1}^{n-1} y_k \cos \frac{k(2t+1)\pi}{2n} \tag{5.3}$$

satisfies

$$P_n(i) = x_i, \qquad \text{for} \quad i = 0, \ldots, n-1.$$

That is, $P_n(t)$ interpolates the data $(0, x_0), (1, x_1), \ldots, (n-1, x_{n-1})$; i.e., $P_n(t)$ passes through the n points (i, x_i).

Proof. From (5.3), we have

$$\begin{aligned}
P_n(0) &= \tfrac{1}{\sqrt{n}} y_0 + \sqrt{\tfrac{2}{n}} \sum_{k=1}^{n-1} y_k \cos \tfrac{k\pi}{2n} \\
P_n(1) &= \tfrac{1}{\sqrt{n}} y_0 + \sqrt{\tfrac{2}{n}} \sum_{k=1}^{n-1} y_k \cos \tfrac{3k\pi}{2n} \\
\vdots \quad & \quad \vdots \\
P_n(n-1) &= \tfrac{1}{\sqrt{n}} y_0 + \sqrt{\tfrac{2}{n}} \sum_{k=1}^{n-1} y_k \cos \tfrac{k(2n-1)\pi}{2n}.
\end{aligned}$$

Using orthogonality, $y = Cx$ implies $x = C^T y$. Then, the equations above can be written as

$$\begin{bmatrix} P_n(0) \\ \vdots \\ P_n(n-1) \end{bmatrix} = C^T \begin{bmatrix} y_0 \\ \vdots \\ y_{n-1} \end{bmatrix} = C^T y = x = \begin{bmatrix} x_0 \\ \vdots \\ x_{n-1} \end{bmatrix}.$$

□

Remark 5.3 *In terms of linear algebra, the DCT interpolation statement in (5.3) is nothing else but expressing $P_n(t)$ as a unique linear combination of n cosine basis functions of increasing frequencies (the first term $\frac{1}{\sqrt{n}} y_0$ corresponds to the cosine of zero frequency), weighted by appropriate coefficients. For $n = 8$, in Figure 5.1 we plot the cosine basis functions in (5.3) and the corresponding eight-point basis (denoted with '$\times$') from the rows of the matrix C in Example 5.1.*

Before we present an example of a one-dimensional DCT interpolation, let us stress the fact that in the proof of Theorem 5.2, we have exploited the fact that the matrix C is orthogonal and therefore $C^T C = I$. Thus, from (5.2), we can take $C^T y = C^T C x = x$. That is, we can recover x as

$$x = C^T y, \tag{5.4}$$

which is known as the (one-dimensional) *inverse discrete cosine transform* of y.

Now for a moment let us take $n = 3$. Then, (5.4) is

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \sqrt{\frac{2}{3}} \begin{bmatrix} \frac{1}{\sqrt{2}} & \cos\frac{\pi}{6} & \cos\frac{2\pi}{6} \\ \frac{1}{\sqrt{2}} & \cos\frac{3\pi}{6} & \cos\frac{6\pi}{6} \\ \frac{1}{\sqrt{2}} & \cos\frac{5\pi}{6} & \cos\frac{10\pi}{6} \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix},$$

thus, componentwise we have

$$\begin{aligned}
x_0 &= \tfrac{1}{\sqrt{3}} y_0 + \sqrt{\tfrac{2}{3}} \left[y_1 \cos\tfrac{\pi}{6} + y_2 \cos\tfrac{2\pi}{6} \right], \\
x_1 &= \tfrac{1}{\sqrt{3}} y_0 + \sqrt{\tfrac{2}{3}} \left[y_1 \cos\tfrac{3\pi}{6} + y_2 \cos\tfrac{6\pi}{6} \right], \\
x_2 &= \tfrac{1}{\sqrt{3}} y_0 + \sqrt{\tfrac{2}{3}} \left[y_1 \cos\tfrac{5\pi}{6} + y_2 \cos\tfrac{10\pi}{6} \right].
\end{aligned} \tag{5.5}$$

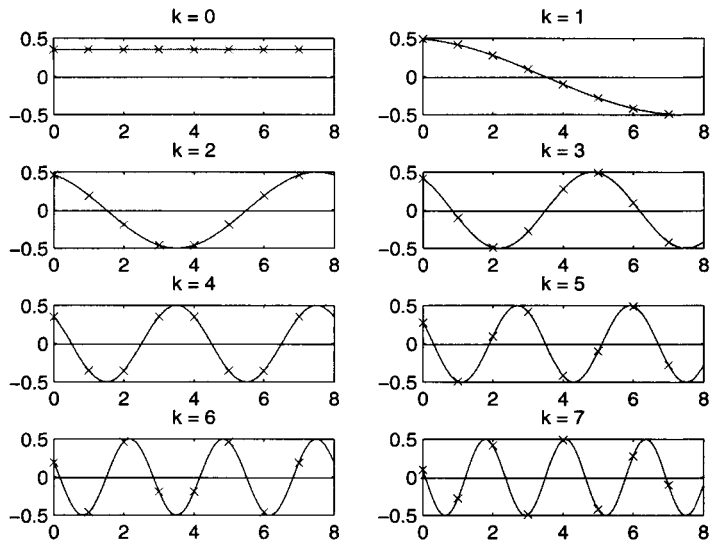


Figure 5.1 Cosine basis functions for $n = 8$.

These equations (5.5) are nothing else but (5.3) with the interpolation property $P_n(j) = x_j$, for $j = 0, \ldots, n-1$. This illustrates a general fact about the connection between the DCT interpolation (5.3) and the inverse DCT given by (5.4).

■ EXAMPLE 5.2

Interpolate the points

$$(0,2),\ (1,0),\ (2,-1),\ (3,0),\ (4,0.25),\ (5,-1.5),\ (6,-2)$$

using the DCT.

In this case we use the DCT matrix (5.1) with $n = 7$, and then for the vector $x = [2 \quad 0 \quad -1 \quad 0 \quad 0.25 \quad -1.5 \quad -2]^T$, we compute (after rounding)

$$y = Cx = [-0.8504 \quad 2.4214 \quad 0.0715 \quad 1.9751 \quad 0.8116 \quad -0.3764 \quad 0.1387]^T.$$

Then, from Theorem 5.2, the function interpolating the seven data points is

$$\begin{aligned} P_7(t) &= \tfrac{1}{\sqrt{7}}(-0.8504) + \sqrt{\tfrac{2}{7}}\Big[2.4214\cos\tfrac{(2t+1)\pi}{14} + 0.0715\cos\tfrac{2(2t+1)\pi}{14} \\ &\quad + 1.9751\cos\tfrac{3(2t+1)\pi}{14} + 0.8116\cos\tfrac{4(2t+1)\pi}{14} - 0.3764\cos\tfrac{5(2t+1)\pi}{14} \\ &\quad + 0.1387\cos\tfrac{6(2t+1)\pi}{14}\Big]. \end{aligned}$$

The interpolant $P_7(t)$, which is a combination of the seven cosine basis functions, is shown in Figure 5.2 as a solid curve, where the data points are represented by stars.

There are a couple of remarks to point out about the interpolation via DCT. First, the frequencies of the cosine functions in (5.3) are in increasing order, and the coefficients y_k act as weights of these cosine functions. As it will turn out, the terms with the highest frequencies (the last terms in the expansion) will be the least important in terms of accuracy of interpolation, so that they can be safely dropped without substantially altering the final interpolation, resulting in a saving of terms (and storage).

Second, when using the interpolating polynomial $P_n(t)$ in (5.3), the coefficients y_k of the interpolation are easily computed through a matrix-vector multiplication $y = Cx$. Finding such coefficients is precisely the difficult part when finding other interpolating functions (such as Lagrange polynomials, splines, etc.). In addition, the basis functions are just cosines of increasing frequency. This makes DCT interpolation very simple and inexpensive to compute.

Finally, a more remarkable fact about DCT interpolation is that we can drop some of the last terms in the polynomial $P_n(t)$, and the error involved will be minimum in the sense of least squares. This is exactly our first step into compression.

Theorem 5.4 DCT Least Squares Approximation. *Let C be the matrix in (5.1). For $x = [x_0 \cdots x_{n-1}]^T$, let y be its DCT, that is, $y = [y_0 \cdots y_{n-1}]^T$ with $y = Cx$, and let m be an integer with $1 \le m < n$. Then, choosing the first m coefficients $y_0, \ldots, y_{m-1}$ to form*

$$P_m(t) = \frac{1}{\sqrt{n}} y_0 + \sqrt{\frac{2}{n}} \sum_{k=1}^{m-1} y_k \cos \frac{k(2t+1)\pi}{2n} \tag{5.6}$$

minimizes the error $\sum_{i=0}^{n-1} (P_m(i) - x_i)^2$, when approximating the n data points.

Proof. We are trying to find coefficients $y_0, \ldots, y_{m-1}$ so that the error in matching the equations

$$P_m(i) = \frac{1}{\sqrt{n}} y_0 + \sqrt{\frac{2}{n}} \sum_{k=1}^{m-1} y_k \cos \frac{k(2i+1)\pi}{2n} = x_i$$

is minimum. Following the notation in the proof of Theorem 5.2, the last equality above can be written as

$$C_m^T y = x,$$

where C_m is the matrix formed with the first m rows of C. This means that the columns of C_m^T are orthonormal and therefore $I = (C_m^T)^T C_m^T = C_m C_m^T$. The equation $C_m^T y = x$ is an overdetermined linear system and therefore we can find its least squares solution by using the corresponding normal equations. This gives

$$C_m C_m^T y = C_m x, \quad \text{or} \quad y = C_m x.$$

Thus, the minimum least squares error is obtained by choosing the first m coefficients $y_0, \ldots, y_{m-1}$.

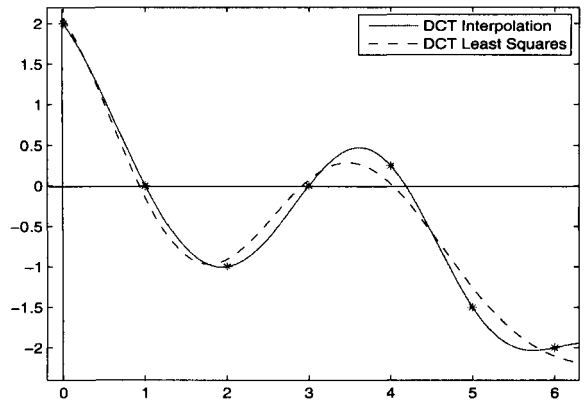


Figure 5.2 Interpolation and least squares using DCT.

□

■ **EXAMPLE 5.3**

Consider the data vector x from Example 5.2. We perform DCT least squares approximation by dropping the last two terms of $P_7(t)$ to obtain

$$P_5(t) = \frac{1}{\sqrt{7}}(-0.8504) + \sqrt{\frac{2}{7}}\left[2.4214\cos\frac{(2t+1)\pi}{14} + 0.0715\cos\frac{2(2t+1)\pi}{14}\right.$$

$$\left. +1.9751\cos\frac{3(2t+1)\pi}{14} + 0.8116\cos\frac{4(2t+1)\pi}{14}\right].$$

According to Theorem 5.4, this new function $P_5(t)$ (although not an interpolant anymore) approximates the data points with a minimum error in the sense of least squares. Figure 5.2 shows $P_5(t)$ as a dashed curve.

Thus, the DCT gives not just a more general application of least squares to problems such as the one studied in linear regression, but more importantly, it provides an approximation with the terms arranged in a very special fashion.

Remark 5.5 *Dropping terms in such a way for another class of interpolating functions such as Lagrange polynomials or splines would completely alter the interpolant, resulting in a function that is very far from being an approximation of the data points. However, we are able to do this with a DCT interpolating function because the terms are already arranged in order of importance.*

Since we are interested in image compression, we need to move to a 2-dimensional framework. But the above introduction to the one-dimensional DCT has given us a clear overall

idea involved in the process of compressing. Now we just need to extend everything to two dimensions.

5.1.2 2-D discrete cosine transform

We start with the definition of the 2-dimensional version of the DCT, which is simply speaking the 1-D DCT applied twice. Given an input matrix X, the DCT is applied to the rows of X, and then the DCT is applied again to the rows of the resulting matrix. That is, we perform the matrix multiplications $C(CX^T)^T = C\,X\,C^T$. More formally, we have the following definition.

Definition 5.6 *Let C be the $n \times n$ matrix defined in (5.1), and let X be an arbitrary $n \times n$ real matrix. Then, the 2-D DCT of X is defined as*

$$Y = C\,X\,C^T. \tag{5.7}$$

Remark 5.7 *Recall that the DCT matrix C is orthogonal and square, which implies that $C^{-1} = C^T$. Thus, the expression in (5.7) is a statement of similarity of matrices (see (1.63)); or, more properly, a change of coordinates.*

One of the goals in image processing is to be able to recover the original image stored in an input matrix X. Here is where again the concept of orthogonality plays a crucial role. Observe that we can first multiply (5.7) by C^T from the left and then multiply by C from the right to obtain $C^T Y\,C = C^T C\,X\,C^T C = I\,X\,I = X$. That is, we have

$$X = C^T Y\,C. \tag{5.8}$$

The matrix X is then what is known as the **2-D Inverse Discrete Cosine Transform (IDCT)** of the $n \times n$ matrix Y.

In a similar way as we did in Section 5.1.1, here we illustrate how the mathematical concept of interpolation is related to the IDCT in (5.8), and both in turn are related to the technique of compressing images.

In a general 1-dimensional interpolation problem, a function is found so that its graph is a curve that passes through a given set of points (t_i, x_i), $i = 0, \ldots, n-1$ in $\mathbb{R}^2$. For the case of 1-dimensional interpolation with DCT studied above, those points were (i, x_i), $i = 0, \ldots, n-1$.

The 2-dimensional case is similar, but now given a set of points (t_i, t_j, x_{ij}) in $\mathbb{R}^3$, with $i = 1, \ldots, n$, we want to find a function whose graph is a surface that passes through the given points. See Figure 5.3. For the particular case of 2-dimensional interpolation with DCT, those points are (i, j, x_{ij}), with $i, j = 0, \ldots, n-1$.

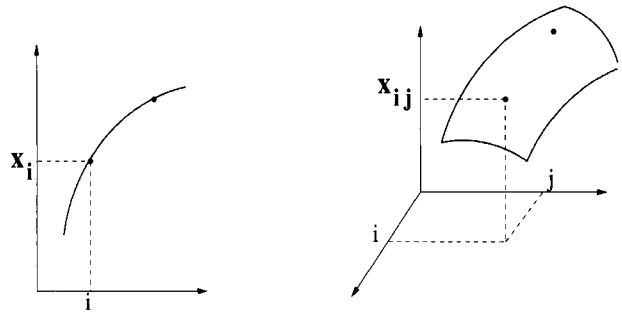


Figure 5.3 1-D and 2-D interpolation

Theorem 5.8 (2-d DCT Interpolation) *Let C be the matrix in (5.1), and let X be any real $n \times n$ real matrix. If Y is the 2-D DCT of X, then the function*

$$P_n(s,t) = \frac{2}{\sqrt{n}} \sum_{k=0}^{n-1} \sum_{l=1}^{n-1} y_{kl} a_k a_l \cos \frac{k(2s+1)\pi}{2n} \cos \frac{l(2t+1)\pi}{2n} \tag{5.9}$$

satisfies $P_n(i,j) = x_{ij}$, for $i,j = 0, \ldots, n-1$, where

$$a_k = \begin{cases} 1/\sqrt{2}, & k = 0 \\ 1, & k > 0. \end{cases}$$

In other words, the function $P_n(s,t)$ interpolates the input data (i, j, x_{ij}), for $i,j = 0, 1, \ldots, n-1$.

■ **EXAMPLE 5.4**

Consider the input data matrix

$$X = \begin{bmatrix} 1.0 & 0.8 & 1.0 & 1.0 & 0.8 & 1.0 \\ 1.0 & 0.5 & 0.3 & 0.0 & 0.5 & 1.0 \\ 1.0 & 0.3 & 0.2 & 0.0 & 0.3 & 1.0 \\ 1.0 & 0.2 & 0.0 & 0.0 & 0.2 & 1.0 \\ 1.0 & 0.3 & 0.2 & 0.0 & 0.3 & 1.0 \\ 1.0 & 0.8 & 1.0 & 1.0 & 0.8 & 1.0 \end{bmatrix}.$$

We want to perform 2-dimensional interpolation of this data by using the DCT through Theorem 5.8. We can consider each entry x_{ij} of the matrix X as an assigned value at each grid point (i,j), like in Figure 5.3. First, we compute the DCT of

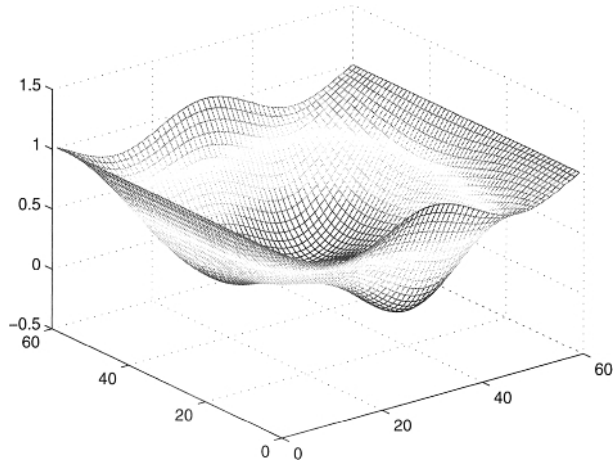


Figure 5.4 2-D DCT interpolation of Example 5.4.

$X,\ Y = C\,X\,C^T$:

$$Y = \begin{bmatrix} 3.7500 & 0.0427 & 1.4901 & -0.1167 & 0.6010 & 0.1594 \\ 0.1077 & 0.0106 & -0.0354 & -0.0289 & -0.0911 & 0.0394 \\ 1.2247 & -0.0149 & -0.9500 & 0.0408 & -0.0866 & -0.0558 \\ -0.1500 & -0.0183 & 0.0612 & 0.0500 & 0.1061 & -0.0683 \\ 0.4950 & -0.0345 & -0.4619 & 0.0943 & 0.1000 & -0.1288 \\ 0.0077 & 0.0106 & -0.0354 & -0.0289 & 0.0503 & 0.0394 \end{bmatrix}.$$

Then, we compute the function $P(s,t)$ as in (5.9):

$$P_6 = \frac{2}{6}\left[\frac{1}{2}(3.75) + \frac{1}{\sqrt{2}}(0.0427)\cos\frac{(2t+1)\pi}{12} + \frac{1}{\sqrt{2}}(1.4901)\cos\frac{2(2t+1)\pi}{12} + \right.$$

$$\cdots + \frac{1}{\sqrt{2}}(0.0503)\cos\frac{5(2s+1)\pi}{16}\cos\frac{4(2t+1)\pi}{16}$$

$$\left. + \frac{1}{\sqrt{2}}(0.0394)\cos\frac{5(2s+1)\pi}{16}\cos\frac{5(2t+1)\pi}{16}\right].$$

This function, which passes through all the points (i, j, x_{ij}), is plotted in Figure 5.4.

There are several important facts that need to be explained from Theorem 5.8; and, in particular, from (5.9). We start by realizing that what we observed in Section 5.1.1 for the one–dimensional case also applies here. Namely, by recalling the definition of the DCT matrix C in (5.1) and by performing the matrix multiplication $X = C^T Y\, C$ in (5.7) componentwise, we can easily deduce that this gives exactly (5.9) with the property

$P_n(i,j) = x_{ij}$, establishing the connection between the IDCT and the Interpolation Theorem 5.8. Once again, we remark on the fact that the coefficients y_{kl} of the interpolation in (5.9) are easily obtained through matrix multiplication $Y = C\,X\,C^T$, that is, by computing the DCT of the input matrix X.

By Remark 5.7, applying the DCT to an input matrix X amounts to a similarity transformation, or in other words, the DCT can be understood as a change of coordinates. In fact, in the applications language (say, image processing), the DCT is understood as a technique to convert a spatial domain waveform into its constituent frequency components (represented by a set of coefficients). Thus, the DCT is a change from spatial to frequency coordinates. It is exactly in the frequency framework where compression can take place, as we will see in the next section.

2-D DCT Least Squares Approximation. With the obvious modifications, Theorem 5.4 still applies here. Namely, we can zero some coefficients corresponding to large frequencies (some of the last few terms in (5.9)) and the error involved will be minimum in the sense of least squares. As in the 1-D case, the function obtained is not an interpolant anymore but it approximates the data points in an optimal way.

Since we are walking our way toward image compression, we are interested in dropping terms with high frequency. Now, given two distinct terms like $\cos(4t)\cos(5t)$ and $\cos(t)\cos(6t)$, which one has higher frequency, and therefore should be dropped? We can use the convention that the frequency of the term is given by the sum of the individual frequencies. Thus, e.g. ,$\cos(4t)\cos(5t)$ has a "total" frequency that is higher than that of $\cos(t)\cos(6t)$. Then, by following the index notation in (5.9); that is, by considering the matrix Y as having elements y_{kl}, with $k, l = 0, 1, \ldots n-1$, we want to zero, e.g.. those elements for which $k + l > m$, for a given $m < n$.

■ **EXAMPLE 5.5**

Consider the input data x_{ij} from Example 5.4, where $n = 6$. Out of a total of 36 terms in the interpolation function $P_6(s,t)$, first, we zero a total of 21 by requiring that we keep only those terms for which $k+l \leq 4$, and then we are less demanding and impose $k + l \leq 6$, which still eliminates 10 terms. Both least squares approximations are shown in Figure 5.5. Compare those approximations with the original interpolation of Figure 5.4.

5.1.3 Image compression and the human visual system

To illustrate the idea of the great need for compressing images, consider a color picture measuring three by five inches that you shot using your digital camera, and at a resolution (which defines the quality of the image) of 400 dots per inch (dpi). The image is 15 in^2, and since each square inch has $400 \times 400 = 160,000$ dots (or pixels), the image will contain a total of 2,400,000 pixels. Now, each pixel requires 3 bytes of data to store the

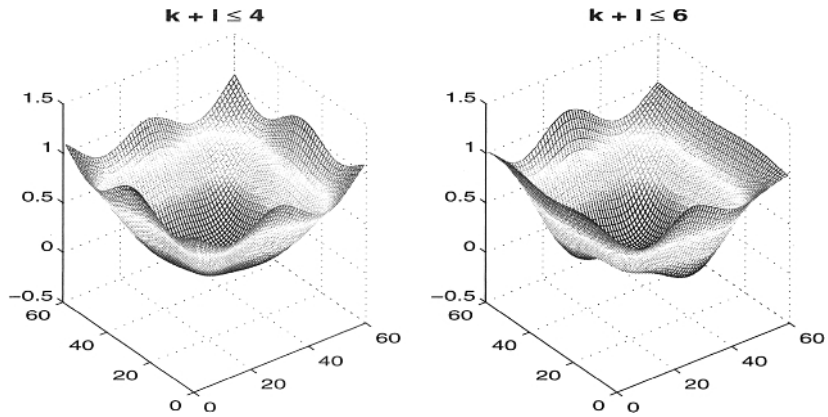


Figure 5.5 2-d DCT least squares approximation

different colors in the picture. Therefore, the image would require 7,200,000 bytes, which is about 7 MB of storage. Thus, storing such digital images without compression would mean using huge amounts of storage. In addition, these images would require large transfer times when sent electronically, especially for those with slow connection speeds. Several compression techniques have been developed to deal with this problem, all of them with the goal of compressing images to several times less than their original size, allowing for easier storage and transmission.

There are two main types of data compression: lossless and lossy. In **lossless compression** (such in zip and tar files), one is able to regain the original data after compression so that the quality of the image is not sacrificed; in fact, only redundant data are removed from the original file. In **lossy compression** some of the data are lost during the compression process, resulting in a final image that is of a lower quality than the original image, but such a loss of quality is in general not easily perceived by the human eye. Obviously, compression rates in this case are much higher than those achieved with lossles compression.

Here we mostly discuss lossy compression, in particular, the most common form of image compression, known as JPEG (about 80% of all web images today are JPEG encoded). A newer and more sophisticated version, JPEG 2000, has been developed, but it is not yet widely used.

A word on the human visual system. The central idea in image processing is to exploit the unique characteristics of the human visual system in order to deliver images of optimal quality in color and detail, at a minimum cost. The human vision is sensitive to the visible portion of the electromagnetic spectrum we know as light. The incident light is focused on the retina, which contains photoreceptors called rods and cones. Rods give us the ability to see at very low light levels, whereas at relatively higher light levels, cones take over. However, we have fewer cones than rods. This may explain why we can discern fewer colors than we can discern a larger number of shades of gray.

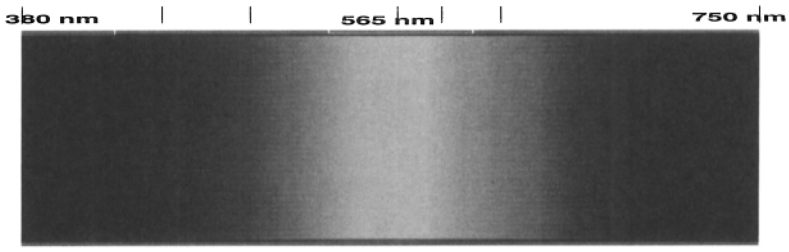


Figure 5.6 Visible spectrum.

We will see later that black and white pixels can be represented by a single number denoting the so–called luminance. However, colors have three attributes: brightness, hue, and saturation, and therefore, they cannot be specified by a single number. We also know that the human vision has the highest sensitivity to yellow-green light, the lowest sensitivity to blue light, and red somewhere in between. In fact, evidence shows that the cones of the human retina can be classified into three types, with overlapping spectral sensitivities centered at about 700nm (red), 535 nm (green), and 445 nm (blue). See Figure 5.6. According to the tri-stimulus theory, the color of light entering the human visual system may be specified by only three numbers associated with three independent color sources. In optical systems, these colors are Red, Green and Blue (RGB).

5.1.4 Basis functions and images

If we consider an input matrix X as an image block of grayscale values $(0-255)$, the statements of interpolation (5.9) and that of the IDCT (5.8) tell us that such an image can be written as the unique linear combination of basis functions given in terms of cosines (and it is possible to visualize such basis functions for any value of n). We want to start by illustrating the case $n = 4$.

In the vector space of 4×4 real matrices, we have the canonical basis

$$\mathcal{B} = \left\{ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \cdots \quad , \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right\}. \tag{5.10}$$

Thus, any 4×4 real matrix A can be expressed as a unique linear combination of the matrices of the basis $\mathcal{B}$. Following the change of coordinates reasoning, a simple way to obtain and visualize the standard DCT 4×4 basis functions is to compute the 2-D DCT of each matrix X in $\mathcal{B}$ as $Y = C\,X\,C^T$, where C is the matrix in (5.1), with $n = 4$, and then to display Y as an image, through the MATLAB commands

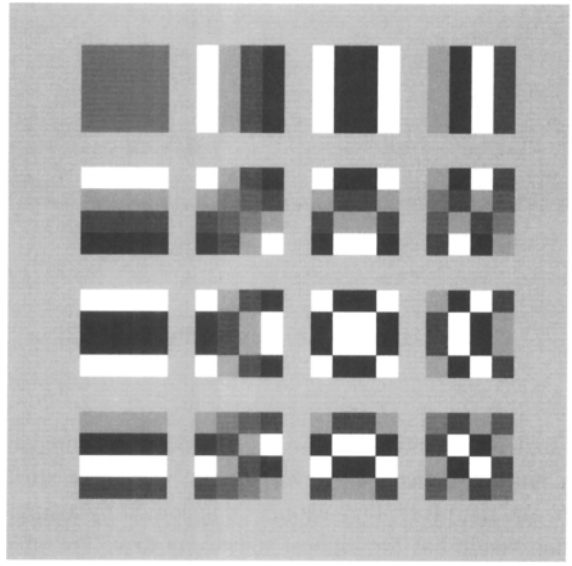


Figure 5.7 DCT 4×4 basis images.

$$
\begin{aligned}
&Y = C * X * C';\\
&\mathsf{colormap(gray)};\\
&\mathsf{imagesc}(Y).
\end{aligned}
$$

See Figure 5.7, where we show the DCT 4×4 basis functions resulting from these calculations. For illustration, here are the Y matrices in the new basis, corresponding to the first two X matrices in the basis $\mathcal{B}$ above.

$$
\begin{bmatrix} 0.2500 & 0.2500 & 0.2500 & 0.2500 \\ 0.2500 & 0.2500 & 0.2500 & 0.2500 \\ 0.2500 & 0.2500 & 0.2500 & 0.2500 \\ 0.2500 & 0.2500 & 0.2500 & 0.2500 \end{bmatrix}, \begin{bmatrix} 0.3266 & 0.1353 & -0.1353 & -0.3266 \\ 0.3266 & 0.1353 & -0.1353 & -0.3266 \\ 0.3266 & 0.1353 & -0.1353 & -0.3266 \\ 0.3266 & 0.1353 & -0.1353 & -0.3266 \end{bmatrix}.
$$

This means that each of the 16 images in Figure 5.7 is the image display of the 2-D DCT of each of the corresponding matrices in the basis $\mathcal{B}$. But more importantly, this means that any 4×4 grayscale image block can be obtained or expressed as a unique linear combination of the 16 basis functions shown in Figure 5.7.

Remark 5.9 *The* MATLAB *command* imagesc(Y) *rescales the entries of* Y *to values in the interval* $[0, 255]$ *and then prints grayscales corresponding to each entry, where* 0 *corresponds to black and* 255 *corresponds to white.*

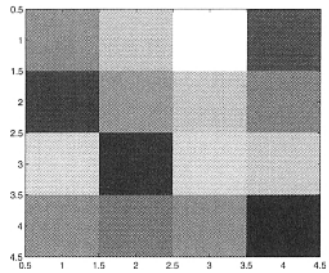


Figure 5.8 4×4 image Y of Example 5.6.

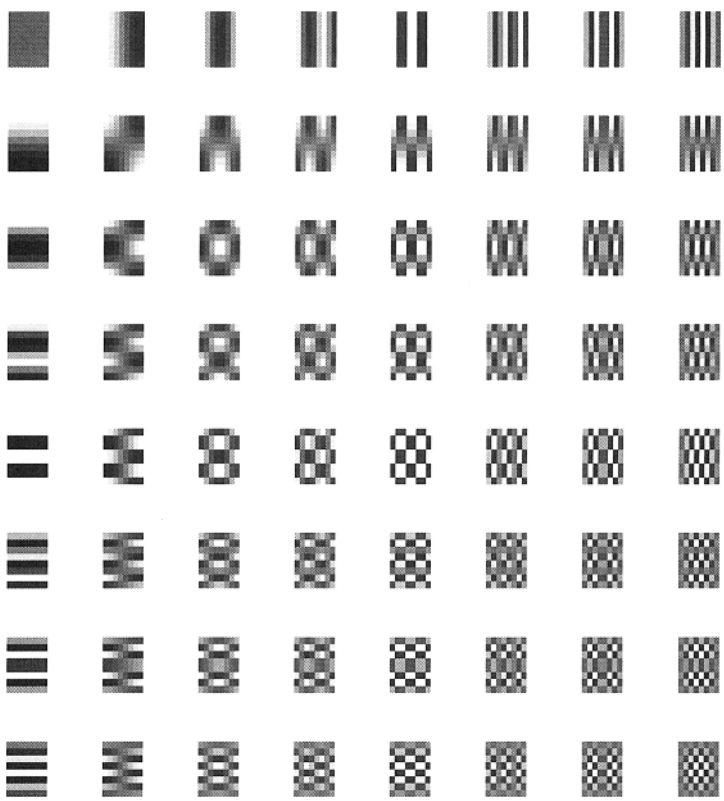


Figure 5.9 DCT 8×8 basis images.

■ **EXAMPLE 5.6**

Consider the 4×4 image of Figure 5.8; call it Y. Then, Y can be uniquely written as the linear combination of the DCT 4×4 basis elements shown in Figure 5.7. More precisely, if we name such basis as $\{Y_0, Y_1, \ldots, Y_{15}\}$, then the image can be decomposed as

$$Y = Y_0 + 0.5Y_1 - 3Y_2 + 4Y_3 + 2Y_4 - 1.5Y_5 - 2.5Y_6 - Y_7 - Y_8 + 4Y_9$$
$$-3Y_{10} - 0.5Y_{11} + 2.5Y_{12} + 3.5Y_{13} + 3Y_{14} - Y_{15}.$$

Similarly, any other 4×4 grayscale image can also be written as a unique combination of the basis functions $Y_0, \ldots, Y_{15}$.

But beyond this illustration, we are mostly interested in 8×8 images, because for image compression, an arbitrary figure will be decomposed into hundreds or thousands of 8×8 pixel values. To obtain the DCT 8×8 basis shown in Figure 5.9, we can proceed in exactly the same way we did to obtain the basis in the 4×4 case, and any 8×8 grayscale image will be the unique linear combination of such 64 basis elements.

Remark 5.10 *There are other transforms that can be used in a similar fashion for the purpose of image compression, like the Haar transform (Exercise 5.17) or the Hadamard transform (Exercise 5.18), but it has been shown that with the DCT, the mean square error between the original image and the reconstructed image decreases fastest as the number of basis images used increases.*

5.1.5 Low-pass filtering

We want to start by considering grayscale images, and later on we will generalize this discussion to color images. Any digital image is composed of pixels, which can be thought of as small dots on the screen. Consider for instance the image of Figure 5.10(a), which is a 592×752 array of pixels. Mathematically speaking, this grayscale image is a 592×752 matrix X (the input matrix), where each entry has a value between 0 and 255, corresponding to how dark or bright the pixel at the corresponding position should be; the value 0 corresponds to black and 255 to white. Thus, if we zoom-in the picture enough, we will see only small boxes of different grayscales, such as the one in Figure 5.10(b), which corresponds to an area around the right eye of the girl in the picture.

Assume the picture is stored as a JPEG image face.jpg. Then, we can import it into MATLAB through the command

```
A=imread('face.jpg');
```

Although the DCT can be applied to the whole matrix A at once, we will consider this matrix A as composed of several 8×8 image blocks, like the one showed in Figure 5.10 (b), and we will successively apply the DCT to each block. At the same time, this will allow

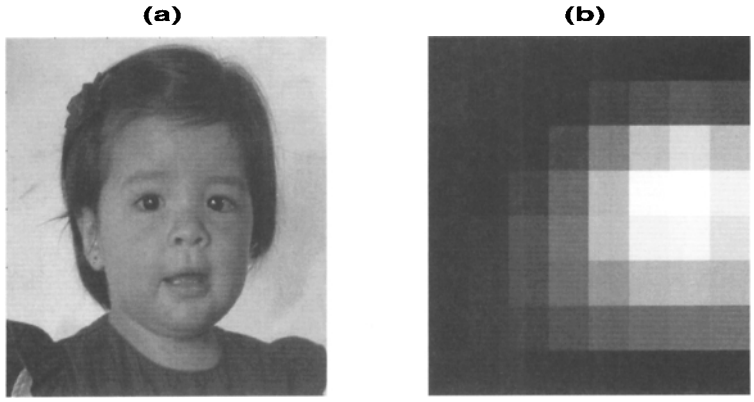


Figure 5.10 Original image and one 8×8 block.

us to illustrate how the DCT works on such matrices. The grayscales of Figure 5.10(b) are the entries in the matrix

$$X = \begin{bmatrix} 33 & 29 & 27 & 21 & 21 & 20 & 11 & 8 \\ 41 & 31 & 23 & 28 & 58 & 81 & 93 & 86 \\ 25 & 13 & 23 & 65 & 142 & 198 & 220 & 196 \\ 24 & 22 & 51 & 107 & 192 & 251 & 255 & 232 \\ 37 & 43 & 80 & 127 & 194 & 239 & 243 & 223 \\ 36 & 50 & 74 & 118 & 165 & 193 & 197 & 191 \\ 38 & 47 & 60 & 83 & 107 & 122 & 129 & 136 \\ 42 & 45 & 46 & 46 & 45 & 45 & 49 & 55 \end{bmatrix}, \tag{5.11}$$

which can be thought of as an input matrix. Before applying the DCT to X, there is an optional and technical step called **level shifting**, which changes the values in the interval $[0, 255]$ to $[-128, 127]$, thus converting them into signed bytes, centered around zero. This can be achieved by subtracting $128 = 2^7$ (in general, we subtract 2^{n-1}, where 2^n is the maximum number of gray levels). The shifted matrix, which we still call X, is

$$X = \begin{bmatrix} -95 & -99 & -101 & -107 & -107 & -108 & -117 & -120 \\ -87 & -97 & -105 & -100 & -70 & -47 & -35 & -42 \\ -103 & -115 & -105 & -63 & 14 & 70 & 92 & 68 \\ -104 & -106 & -77 & -21 & 64 & 123 & 127 & 104 \\ -91 & -85 & -48 & -1 & 66 & 111 & 115 & 95 \\ -92 & -78 & -54 & -10 & 37 & 65 & 69 & 63 \\ -90 & -81 & -68 & -45 & -21 & -6 & 1 & 8 \\ -86 & -83 & -82 & -82 & -83 & -83 & -79 & -73 \end{bmatrix}. \tag{5.12}$$

Our first step consists of applying the DCT to the matrix X. This gives (after rounding):

$$Y = CXC^T = \begin{bmatrix} -283 & -369 & -23 & 71 & -11 & 1 & 1 & 5 \\ -92 & 11 & 22 & 25 & -6 & 3 & 1 & 4 \\ -340 & 276 & 35 & -54 & 9 & -4 & -1 & -4 \\ 10 & 44 & -27 & -20 & 4 & -3 & 2 & -3 \\ -26 & 34 & -18 & -4 & -1 & 1 & 6 & 0 \\ 16 & -30 & -9 & 5 & 2 & -1 & 1 & -2 \\ 1 & -11 & -5 & 4 & 0 & 1 & -2 & 0 \\ 9 & -10 & 2 & 1 & -1 & 3 & 0 & 2 \end{bmatrix}. \tag{5.13}$$

It is now evident that using X, the DCT has produced a matrix Y such that its entries around the upper left corner have the largest magnitude, whereas the ones around the lower right corner have the lowest one (and this will be true for any given input matrix X). The entries of the matrix Y are known as the **DCT coefficients**. By recalling the discussion about equations (5.8) and (5.9), that is, the image X can be represented as a combination of cosine basis functions with the DCT coefficients acting as weights, we observe that the largest weights are associated with the basis elements with lower frequency (the upper left corner), and the smallest weights are associated with the basis elements with higher frequency (lower right corner).

We are at the heart of the DCT action: It has produced a change of coordinates from the image input signal to the frequency coordinates, and it has arranged it in increasing order of frequency. In terms of image compression, it is ideal, since the human visual system is more sensitive to lower frequencies than to higher ones. Thus, in terms of human vision, the first terms in (5.9) are far more important than the last terms. Accordingly, the DCT has therefore given (through the entries in the upper left corner of Y) more weight to those functions with lower frequency.

Now we can try our first compression strategy, in a similar way as we did for the one-dimensional case: Drop some terms in the lower right corner of Y (that is, a few of the last terms in (5.9)) to obtain a new matrix $\overline{Y}$ and then apply the IDCT to this new matrix. Of course, we will not obtain the original image but a compressed one, technically of lower quality, but still for the most part the difference is not very much perceived by the human eye. This simple technique is called low-pass filtering.

Thus, suppose we decide to zero the diagonal and the lower triangular part of Y. Then, this filtering gives

$$\overline{Y} = \begin{bmatrix} -283 & -369 & -23 & 71 & -11 & 1 & 1 & 0 \\ -92 & 11 & 22 & 25 & -6 & 3 & 0 & 0 \\ -340 & 276 & 35 & -54 & 9 & 0 & 0 & 0 \\ 10 & 44 & -27 & -20 & 0 & 0 & 0 & 0 \\ -26 & 34 & -18 & 0 & 0 & 0 & 0 & 0 \\ 16 & -30 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5.14}$$

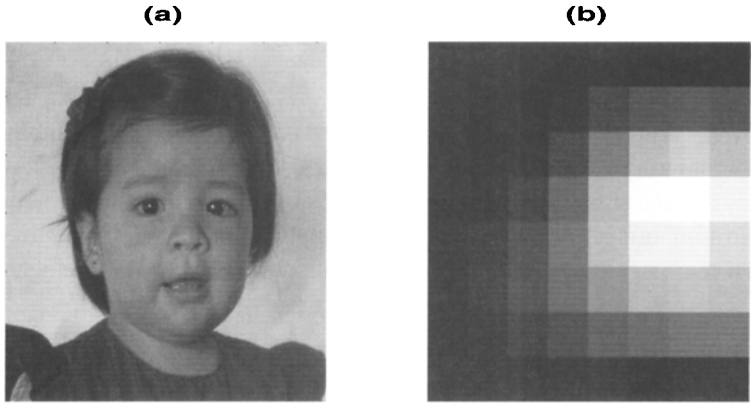


Figure 5.11 Compressed image with DCT low-pass filtering

To reconstruct the (compressed) image, we apply the IDCT to this matrix, that is we calculate $C^T \overline{Y} C$ and then we add back 128 to each entry to obtain (after rounding)

$$\overline{X} = \begin{bmatrix} 35 & 31 & 27 & 21 & 17 & 17 & 13 & 7 \\ 37 & 26 & 21 & 32 & 61 & 89 & 95 & 87 \\ 27 & 18 & 24 & 66 & 136 & 195 & 209 & 196 \\ 22 & 24 & 47 & 107 & 192 & 254 & 260 & 238 \\ 34 & 45 & 75 & 129 & 196 & 240 & 238 & 217 \\ 37 & 51 & 79 & 118 & 163 & 194 & 199 & 193 \\ 36 & 44 & 60 & 81 & 104 & 121 & 131 & 137 \\ 44 & 44 & 45 & 47 & 47 & 45 & 48 & 54 \end{bmatrix}.$$

Obviously, the compressed image is not exactly the same as the original image ($\overline{X} \neq X$), but the difference between them is not easily perceived by the human eye. Compare Figure 5.10(b) with Figure 5.11(b). Furthermore, these 8×8 image blocks are just very small pieces of a given actual image. Thus, we can expect not to notice the small changes in the compressed image even though, as in this case, we have reduced storage requirements by about 50%. To apply this method to an entire image, we apply the above technique to each 8×8 block and then build up the compressed image from the compressed blocks. Figure 5.11 (a) shows the compressed image, which should be compared with the original Figure 5.10(a).

Remark 5.11 *This low-pass filtering technique is related to the 2-D DCT interpolation and least squares approximation discussed in Section 5.1.2; that is, the error involved when dropping some of the last DCT coefficients is minimum in the sense of least squares.*

5.1.6 Quantization

The low-pass filtering compression technique presented above is effective, but it sure can be improved. While still trying to zero the DCT coefficients associated with the largest frequencies, now at the same time we want to rescale the remaining nonzero coefficients in such a way that fewer bits are necessary for their storage. Since the lower frequency terms are the most important ones, we would like to apply a moderate rescaling to them, while applying a more aggressive rescaling (if possible down to zero) to the higher frequency terms.

There are several possible methods to perform this nonuniform rescaling, which in the language of compression is known as quantization. Most of these methods define a so–called quantization matrix Q so that Y is entrywise divided by Q and then rounded to obtain a new quantized matrix

$$Y_Q = \text{round}\left(\frac{y_{kl}}{q_{kl}}\right). \tag{5.15}$$

Clearly, here an error is introduced due to rounding; this is why this technique falls into the category of lossy compression. One such a quantization matrix Q can be defined as

$$q_{kl} = 8s\,(k+l+1), \quad 0 \le k, l \le 7.$$

That is,

$$Q = s\begin{bmatrix} 8 & 16 & 24 & 32 & 40 & 48 & 56 & 64 \\ 16 & 24 & 32 & 40 & 48 & 56 & 64 & 72 \\ 24 & 32 & 40 & 48 & 56 & 64 & 72 & 80 \\ 32 & 40 & 48 & 56 & 64 & 72 & 80 & 88 \\ 40 & 48 & 56 & 64 & 72 & 80 & 88 & 96 \\ 48 & 56 & 64 & 72 & 80 & 88 & 96 & 104 \\ 56 & 64 & 72 & 80 & 88 & 96 & 104 & 112 \\ 64 & 72 & 80 & 88 & 96 & 104 & 112 & 120 \end{bmatrix}. \tag{5.16}$$

Thus, for larger values of s, more compression will be applied. Observe that the entries of Q at the upper left corner are small, because we expect to have large values in the matrix Y at those positions, and entrywise division by Q and rounding will merely rescale to numbers of smaller magnitude, requiring therefore smaller bits for storage. At the same time, the elements at the lower right corner of Q are large, and since we expect to have small values in the matrix Y at those positions, entrywise division by Q and rounding will set most of them to zero and the rest will be rescaled to a smaller magnitude and require fewer bits for storage. This is clearly more efficient than low-pass filtering.

The JPEG standard in its Appendix K ("Examples and Guidelines") recommends quantization matrices that are based on psychovisual thresholding and, that are derived empirically in experiments with the human visual system, and therefore, from the practical point of view, they are more reliable. For the case of grayscales, the so–called **luminance quantization matrix** they recommend is

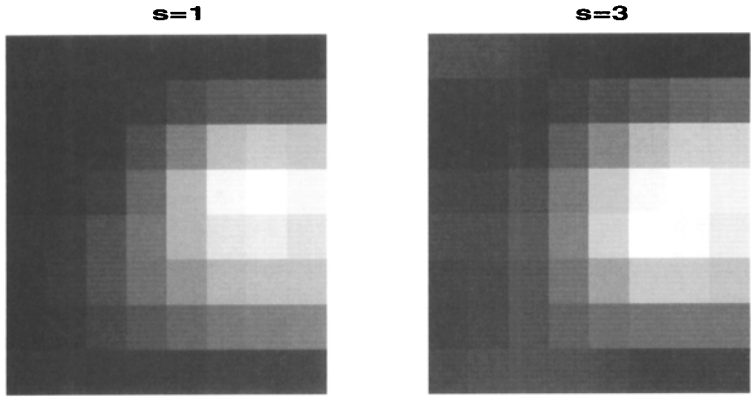


Figure 5.12 DCT luminance quantization using (5.17).

$$Q = s \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}. \tag{5.17}$$

Now let us apply this luminance quantization to (5.13), with the parameter $s = 1$. This gives

$$Y_Q = \begin{bmatrix} -18 & -34 & -2 & 4 & 0 & 0 & 0 & 0 \\ -8 & 1 & 2 & 1 & 0 & 0 & 0 & 0 \\ -24 & 21 & 2 & -2 & 0 & 0 & 0 & 0 \\ 1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{5.18}$$

Compare this matrix Y_Q with the one in (5.14) from low-pass filtering. More zeros were obtained, and the nonzero entries are smaller in magnitude than those in (5.14), thus requiring fewer bits to be stored.

To recover back the (compressed) image, we apply the reverse process; that is, we first multiply entrywise Y_Q by Q (this is where an error is introduced) to obtain a modified $Y = QY_Q$. Then we apply the IDCT to Y : $X = C^T Y C$, and finally we add back 128

Figure 5.13 DCT luminance quantization using (5.17).

to X. In Figure 5.12, we show the images obtained when using $s = 1$ and $s = 3$. For $s = 1$, the compressed image is quite similar to the original one in Figure 5.10 (b), while for $s = 3$, some differences may already be noticeable.

We can now compress the image of Figure 5.10(a) by applying the above process to each 8×8 image block and then reconstructing the image by putting together the compressed blocks. Figure 5.13 shows the results for $s = 1$ and $s = 3$.

As an illustration to estimate how much storage in terms of bits we save by applying luminance quantization, consider an arbitrary 8×8 image block, and as a worst–case scenario, assume each entry in $Y = CXC^T$ is the number 255, the largest possible. If we apply quantization through (5.15) and (5.17) to Y, we obtain the matrix Y_Q below.

$$Y_Q = \begin{bmatrix} 16 & 23 & 26 & 16 & 11 & 6 & 5 & 4 \\ 21 & 21 & 18 & 13 & 10 & 4 & 4 & 5 \\ 18 & 20 & 16 & 11 & 6 & 4 & 4 & 5 \\ 18 & 15 & 12 & 9 & 5 & 3 & 3 & 4 \\ 14 & 12 & 7 & 5 & 4 & 2 & 2 & 3 \\ 11 & 7 & 5 & 4 & 3 & 2 & 2 & 3 \\ 5 & 4 & 3 & 3 & 2 & 2 & 2 & 3 \\ 4 & 3 & 3 & 3 & 2 & 3 & 2 & 3 \end{bmatrix}, \text{Bits}: \begin{bmatrix} 6 & 6 & 6 & 6 & 5 & 4 & 4 & 4 \\ 6 & 6 & 6 & 5 & 5 & 4 & 4 & 4 \\ 6 & 6 & 6 & 5 & 4 & 4 & 4 & 4 \\ 6 & 5 & 5 & 5 & 4 & 3 & 3 & 4 \\ 5 & 5 & 4 & 4 & 4 & 3 & 3 & 3 \\ 5 & 4 & 4 & 4 & 3 & 3 & 3 & 3 \\ 4 & 4 & 3 & 3 & 3 & 3 & 3 & 3 \\ 4 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix}.$$

Since the number r of bits necessary to store a given number n can be estimated as

$$r = \lfloor \log_2(n) \rfloor + 2,$$

where the function $\lfloor x \rfloor$ is the largest integer less than or equal to x, we have calculated the bits necessary to represent the entries in Y_Q, and we have shown them on the matrix next to it. Thus, adding up the 64 numbers in the matrix of bits, we get 266, which is about half the bits necessary to store the original 8×8 image without compression.

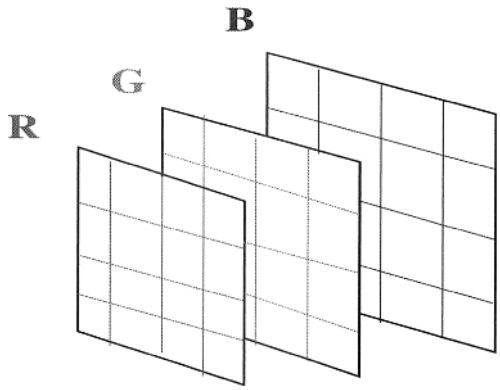


Figure 5.14 Three–dimensional array of a color image.

Thus, for general images, by applying quantization with $s = 1$, we can save about 50% of storage and still obtain a compressed image that to the human eye has been perfectly reconstructed. For larger values of the parameter s, more compression can be applied to the image, resulting in a smaller file size and therefore in more storage saving, but at the same time it also means losing more quality. Thus, it all depends on the application at hand, or on how much quality we want to trade off for storage. In any case, the parameter s in (5.17) allows flexibility and an easy way for testing different compression rates (for $s = 4$, the number of total number of bits is 166, which is about 32% of the original size).

5.1.7 Compression of color images

When we think of matrices, we automatically think of 2-dimensional arrays of rows and columns, just like the ones we have been working with so far. However, if we import a color picture into MATLAB, say through the command A=imread('face.jpg'), and then check the size of A, we observe that such a matrix is 3-dimensional, e.g., $512 \times 512 \times 3$. This matrix can be understood as three layers of two-dimensional matrices. In particular, for color images, the three layers correspond to Red, Green and Blue **(RGB)** intensities (see Figure 5.14). For black–and–white images, each pixel corresponds to a number between 0 and 255 according to its grayscale. For color images, each pixel is given three numbers representing the three color intensities.

Several approaches can be taken to compress color images. The simplest one would be to treat each color (or layer) independently, that is, compression can be applied to each color as if we were dealing with grayscale intensities, and then we reconstruct the (compressed) image from the superposition of the colors. This works, but it is not efficient. A second, and very popular approach is the one outlined by the so–called Baseline JPEG. The central idea again comes from the practical point of view: The human eye is more sensitive to luminance (changes in brightness) than to chroma (changes in color). This real fact gives a hint: We should be able to perform higher rates of compression in the chrominance

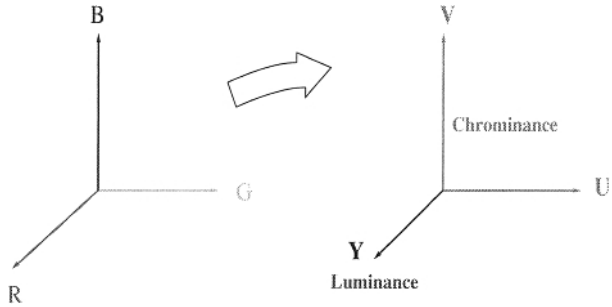


Figure 5.15 Change of coordinates from RGB to Y-UV.

coordinates and still make it unnoticeable to the human eye. Recall that for grayscale images, we only had luminance.

Therefore, instead of working just with plain colors (RGB), we perform a change of coordinates to color differences, or chroma **(YUV)**:

$$Y = 0.299R + 0.587G + 0.114B, \qquad U = B - Y, \quad V = R - Y. \tag{5.19}$$

Remark 5.12 *The coefficients of R, G, and B in the Y coordinate agree with the fact that out of the three colors, the human eye is most sensitive to green and least sensitive to blue, with red somewhere in the middle.*

Through this change of coordinates, the color image can be represented in (y, u, v) form. We perform compression in the Y coordinate as if we were working with grayscale images; that is, we can quantize the data using the luminance matrix in (5.17). Then, independently we can perform a more aggressive compression in the UV coordinates, by using a less conservative quantization matrix.

JPEG recommends the following **Chrominance matrix:**

$$Q_c = s \begin{bmatrix} 17 & 18 & 24 & 47 & 99 & 99 & 99 & 99 \\ 18 & 21 & 26 & 66 & 99 & 99 & 99 & 99 \\ 24 & 26 & 56 & 99 & 99 & 99 & 99 & 99 \\ 47 & 66 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \\ 99 & 99 & 99 & 99 & 99 & 99 & 99 & 99 \end{bmatrix}. \tag{5.20}$$

The difference between the luminance and chrominance matrices (5.17) and (5.20), respectively, is obvious.

For compression of a color image, we proceed in the following way:

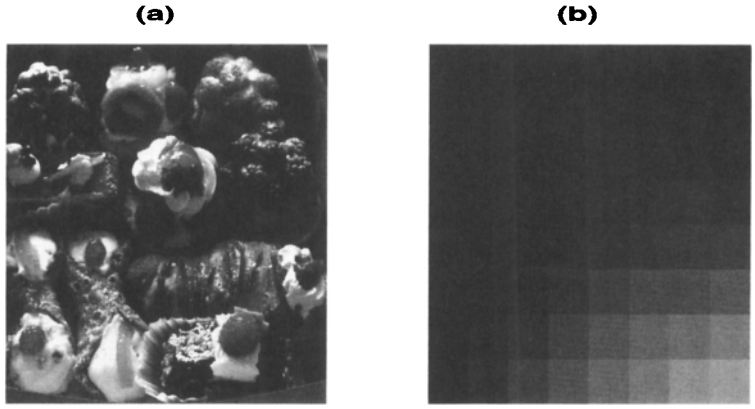


Figure 5.16 Original image and 8×8 block.

- Group pixel values for each of the three components into 8×8 blocks.
- Transform each block X by applying the DCT to it to obtain the DCT coefficients.
- Apply luminance quantization (5.17) to the Y coordinates, and chrominance quantization (5.20) to the U and V coordinates (5.19).

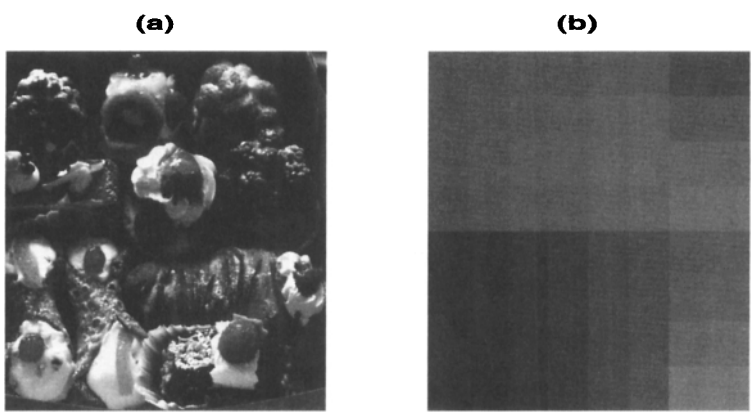


Figure 5.17 Compressed images with $s = 3$.

The DCT coefficients either will be zeroed or reduced in size for storage. The decompression process is just the reverse algorithm, as explained before, except that now we also need

to go back to the RGB coordinates through the equations

$$\begin{aligned} B &= U + Y, \\ R &= V + Y, \\ G &= (Y - 0.299R - 0.114B)/0.587. \end{aligned}$$

We are going to apply these ideas to the color image of Figure 5.16(a); the 8×8 block was taken from a part of the big bread in the center. As usual, we load the image into MATLAB with the command imread, which represents the image in this case as a $1336 \times 1336 \times 3$ matrix: 1336 rows, 1336 columns, and (combinations of) three colors: red, green, and blue. Figures 5.17 (a) and (b) show the results of applying the recommended JPEG luminance and chrominance quantization matrices for $s = 3$. Compare this with $s < 1$ for black and white pictures.

Note. To obtain a similar quality of compressed image for the grayscale case, we had to take just $s = 1$ and no higher.

5.2 HUFFMAN CODING

A central step in the lossy compression technique described above consists of applying a quantization matrix of the type (5.17) or (5.20) to the DCT of the original image matrix X to obtain a matrix Y of quantized coefficients of the form (5.18). The entries of this matrix Y with a large number of zeros are the numbers we need to store. But this has to be done in an efficient way. This is the point where an additional modest amount of compression can be achieved, but this time it is lossless compression.

Those entries of Y will be stored as binary digits, or bits, that is, as sequences of 0's and 1's. We want to explain how this is done.

The first question to answer is the following: Given a general string of symbols, what is the minimum number of bits needed to code such string? It turns out that the answer is related to the probability with which each symbol occurs in the string. Suppose there are n different symbols available, and let p_i be the probability of the occurrence of symbol i in any given string. Then we define the **entropy** of the string as

$$H = -\sum_{i=1}^{n} p_i \log_2(p_i). \tag{5.21}$$

This entropy H tries to quantify the average minimum number of bits per symbol needed to code the string.

■ EXAMPLE 5.7

Let us find the entropy of the string BDABBCDB. The probabilities of the symbols A,B,C,D are, respectively:

$$p_1 = 1/8, \; p_2 = 4/8, \; p_3 = 1/8, \; p_4 = 2/8,$$

or expressed as powers of two: $p_1 = 2^{-3}$, $p_2 = 2^{-1}$, $p_3 = 2^{-3}$, $p_4 = 2^{-2}$. Then, the entropy of the string is

$$H = -\sum_{i=1}^{4} p_i \log_2(p_i) = \frac{1}{8}(3) + \frac{4}{8}(1) + \frac{1}{8}(3) + \frac{2}{8}(2) = \frac{14}{8} = 1.75.$$

Thus, the entropy formula indicates that the minimum number of bits per symbol needed to code the string BDABBCDB is 1.75.

By taking this as a starting point, several coding techniques have been developed to code strings of symbols, but it is the Huffman coding the one closest to achieving this minimum. This process is better explained through a detailed example.

Suppose we have the following symbols, and their corresponding probabilities of occurrence in a string:

Symbol	Probability
A	0.35
B	0.25
C	0.14
D	0.11
E	0.11
F	0.04

Then, to code these symbols, we proceed as follows:

Building the tree. (See Figure 5.18).

1. We combine two symbols with the lowest probabilities, say E and F, to obtain the symbol EF with probability 0.15.

2. Now we have five symbols left, A, B, C, D, and EF. We combine the two with lowest probabilities, C and D, to obtain the symbol CD with probability 0.25.

3. From the four symbols left A, B, CD, and EF, we combine two with the lowest probabilities, say CD and EF to obtain the symbol CDEF with probability 0.40.

4. Now we have three symbols left, A, B, and CDEF. We combine the two with the lowest probabilities, A and B, to obtain the symbol AB with probability 0.60.

5. Finally, we combine the remaining two symbols AB and CDEF to obtain the symbol ABCDEF with probability 1.0.

Assigning the codes.

At this step, we translate the string of symbols into a bit stream, by first obtaining the Huffman code for each symbol. This is done by arbitrarily assigning a bit of 0 to a left branch and a bit of 1 to a right branch. Once this is done, we start at the top of the tree and we read the symbols as:

A=00	C=100	E=110
B=01	D=101	F=111

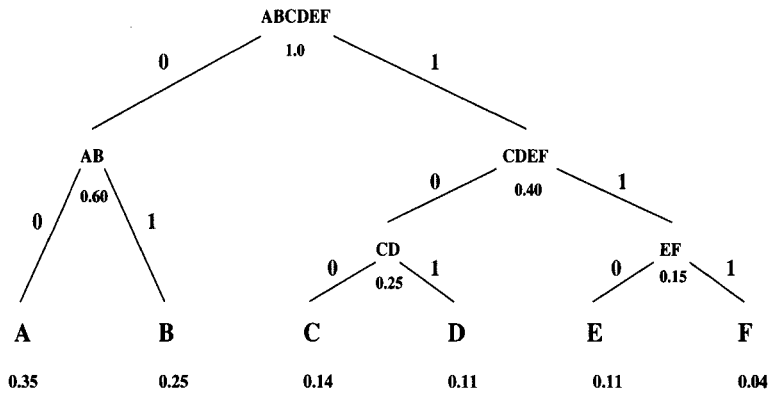


Figure 5.18 Huffman coding tree.

Now we can translate a string of those symbols into bits. For instance, the string

ACDAACBBEB

is translated as

(00)(100)(101)(00)(00)(100)(01)(01)(110)(01).

This bit stream has length 24, and therefore it uses $24/10 = 2.4$ bits per symbol.

Uniqueness. In step 1 of building the tree, we could have also combined the symbols D and F first. Similarly, in step 3, we could have chosen to combine the symbols B and EF instead of combining CD and EF. The idea is to combine arbitrary symbols with the lowest probabilities. By picking different choices, we obtain in general different codes for the symbols, which implies that a Huffman code is not unique. However, the average size will remain the same. For the example above, it will always be 2.4 bits per symbol.

5.2.1 Huffman coding and JPEG

With the basic background introduced above, we can now explain how to encode the DCT coefficients, (the entries of the quantized matrix Y). Recall that we partition the matrices in 8×8 blocks so that in fact we are dealing with 64 DCT coefficients at a time. We also know that the first of these coefficients, which is known as the DC coefficient, is the most important as it has the largest weight or magnitude, and that all other 63 coefficients are smaller and decrease in magnitude as we read the matrix Y toward the lower right corner (in fact the majority are zeros). These 63 coefficients are known as AC coefficients. Because of this main difference, they are coded separately.

5.2.1.1 Coding the DC coefficients Since we expect some correlation between neighboring 8×8 blocks, instead of coding individual DC coefficients for each block, the

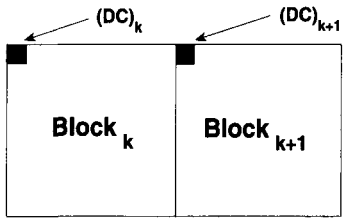


Figure 5.19 DC coefficients of neighboring blocks.

Bit Size	Code	Bit Size	Code
0	00	6	1110
1	010	7	11110
2	011	8	111110
3	100	9	1111110
4	101	10	11111110
5	110	11	111111110

Table 5.1 Codes for DC symbol 1.

strategy is to code their differences (see Figure 5.19). The larger the correlation, the smaller the difference. That is, we will code the difference D between the DC coefficients of two neighboring blocks k and $k+1$:

$$D = (DC)_{k+1} - (DC)_k, \quad k = 1, 2, \ldots, \tag{5.22}$$

where $(DC)_k$ is initially set to zero.

The DC coefficient difference D will be represented as two symbols, the first one for its bit size and the second one for its value D in (5.22). That is,

$$\begin{pmatrix}\text{Symbol 1 for}\\ \text{Bit Size}\end{pmatrix}\begin{pmatrix}\text{Symbol 2 for}\\ \text{Diff Value } D\end{pmatrix}. \tag{5.23}$$

Here we define the **bit size** of an integer z as

$$S = \begin{cases} \lfloor \log_2 |z| \rfloor + 1, & z \neq 0 \\ 0 & z = 0. \end{cases} \tag{5.24}$$

Given a particular DC coefficient difference, we first find the bit size S of that difference through (5.24). Next, to get symbol 1 for S we use Table 5.1, where the codes shown were obtained by building a tree similar to that in Figure 5.18 (see Exercise 5.27).

S	Difference Value D	Code
0	0	
1	-1, 1	0, 1
2	-3, -2, 2, 3	00, 01, 10, 11
3	-7, -6, -5, -4, 4, 5, 6, 7	000, 001, 010, 011, 100, 101, 110, 111
4	-15, -14, . . . , -8, 8, . . . ,14, 15	0000, 0001,. . . ,0111, 1000,. . . ,1110, 1111
5	-31, -30,. . . ,-16, 16,. . . ,30, 31	00000, 00001,. . . ,01111, 10000,. . . ,11110, 11111
6	-63, -62,. . . ,-32, 32,. . . ,62, 63	000000, 0000001,. . . ,011111, 100000,. . . ,111110, 111111
7	-127, -126,. . . ,-64, 64,. . . ,126, 127	0000000, 0000001,. . . ,0111111, 1000000,. . . ,1111110, 1111111
⋮	⋮	⋮

Table 5.2 Codes for DC/AC symbol 2

■ **EXAMPLE 5.8**

Suppose the DC coefficient difference between two neighboring blocks is $D = 9$. From (5.24), its bit size is $S = 4$, and according to Table 5.1, symbol 1 should be **101**.

To obtain symbol 2 in (5.23) we use n bits if the bit size S of the difference is n. But since there are several integer coefficients (positive and negative) that have the same size S, they are grouped together by bit size. Then each one in the group is assigned a unique combination of 0's and 1's according to Table 5.2.

■ **EXAMPLE 5.9**

In Example 5.8, we had $D = 9$, with $S = 4$. Then, by looking at Table 5.2 we conclude that symbol 2 is **1001**. Thus, from (5.23), the complete representation of the DC coefficient difference $D = 9$ is

$$(101)(1001),$$

where the parenthesis is only for notational convenience and clarity.

5.2.1.2 Coding the AC coefficients We know that a great majority of the 63 AC coefficients will likely be zero, as a result of the quantization process, and it is very likely that a high frequency coefficient will be zero given that its predecessors are zero. This implies that there will be runs of zeros in the AC coefficients. We exploit the presence of these runs of zeros by using a zigzag scanning as illustrated in Figure 5.20 when reading the coefficients, because this scanning tends to group longer runs of zeros.

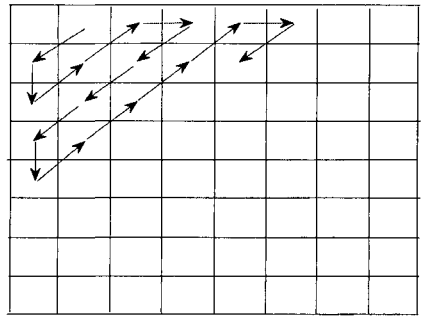


Figure 5.20 Zigzag pattern for AC coefficients.

The AC coefficient will be coded as two symbols. We use the first symbol to represent the pair

$$(r, S),$$

where r is the length of a run of zeros, that is, the number of consecutive zero AC coefficients, and S is the bit size of the next nonzero entry. The corresponding code for each pair is obtained from Table 5.3. The second symbol represents the value of the AC coefficient; the corresponding code for this value comes as before, from Table 5.2.

Thus, the representation has the form

$$\begin{pmatrix}\text{Symbol 1 for}\\ (r, S)\end{pmatrix}\begin{pmatrix}\text{Symbol 2 for}\\ \text{AC Value}\end{pmatrix}, \tag{5.25}$$

where as usual, S comes from (5.24).

■ **EXAMPLE 5.10**

Suppose we have the following AC coefficients:

$$9, 6, 0, 0, 0, 0, -3.$$

For the first coefficient 9, we have $(r, S) = (0, 4)$ because it contains no zeros and because from (5.24) the size of 9 is $S = 4$. Thus, from Table 5.3, its symbol 1 is **1011**. For symbol 2, we conclude from Table 5.2 that the code for 9 is **1001**.

Similarly, for the coefficient 6, we have $(r, S) = (0, 3)$, because $r = 0$ and from (5.24) the size of 6 is $S = 3$. Thus, from Table 5.3, symbol 1 is **100**. To obtain symbol 2, we observe from Table 5.2 that the code for 6 is **110**.

Next we have four consecutive zeros followed by -3. Then, we have $(r, S) = (4, 2)$, because the run of zeros has length 4, and the size of -3 is $S = 2$. Thus, from Table

(r, S)	Code	(r, S)	Code
(0,1)	00	(5,1)	1111010
(0,2)	01	(5,2)	11111110111
(0,3)	100	(5,3)	1111111110011110
(0,4)	1011	(5,4)	1111111110011111
(0,5)	11010	(5,5)	1111111110100000
⋮	⋮	⋮	⋮
(1,1)	1100	(6,1)	1111011
(1,2)	11011	(6,2)	111111110110
(1,3)	1111001	(6,3)	1111111110100110
(1,4)	111110110	(6,4)	1111111110100111
(1,5)	11111110110	(6,5)	1111111110101000
⋮	⋮	⋮	⋮
(2,1)	11100	(7,1)	11111010
(2,2)	11111001	(7,2)	111111110111
(2,3)	1111110111	(7,3)	1111111110101110
(2,4)	111111110100	(7,4)	1111111110101111
(2,5)	1111111110001001	(7,5)	1111111110110000
⋮	⋮	⋮	⋮
(3,1)	111010	(8,1)	111111000
(3,2)	111110111	(8,2)	111111111000000
(3,3)	111111110101	(8,3)	1111111110110110
(3,4)	1111111110001111	(8,4)	1111111110110111
(3,5)	1111111110010000	(8,5)	1111111110111000
⋮	⋮	⋮	⋮
(4,1)	111011	(9,1)	111111001
(4,2)	1111111000	(9,2)	1111111110111110
(4,3)	1111111110010110	(9,3)	1111111110111111
(4,4)	1111111110010111	(9,4)	1111111111000000
(4,5)	1111111110011000	(9,5)	1111111111000001
⋮	⋮	⋮	⋮
		EOB	1010

Table 5.3 AC table, symbol 1.

5.3, symbol 1 is **1111111000**. Finally, we observe from Table 5.2 that the code for -3 is **00**.

Thus, the given seven AC coefficients are coded as

$$(1011)(1001)\ (100)(110)\ (1111111000)(00),$$

where again the parentheses are just for notational convenience and clarity.

Note: In the example above, if -3 was the very last coefficient from the quantized matrix, then the code above must be finished with EOB (end of block), that is, with (1010). See Table 5.3.

5.3 COMPRESSION WITH SVD

In Section 3.3, we introduced a matrix factorization of a general matrix $A_{m\times n}$ as the product of two orthogonal matrices and a diagonal one. This factorization is expressed as

$$A = U\Sigma V^T, \tag{5.26}$$

where $U_{m\times m}$ and $V_{n\times n}$ are orthogonal matrices and $\Sigma_{m\times n}$ is a diagonal matrix whose diagonal entries σ_i are known as the *singular values* of the matrix A.

We learned that this factorization provides us with plenty of information about the matrix A: It gives orthonormal bases for col(A) and row(A), it reveals the rank of A, it provides the spectral norm of A, and so on. One very important result was that the factorization (5.26) can be written as

$$A = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T, \tag{5.27}$$

where r is the rank of A and u_i, v_i represent the i–th columns of the matrices U and V respectively.

Writing the SVD of a matrix A as the expansion (5.27) allowed us to introduce low-rank approximations of the matrix A. A rank-k matrix A_k that approximates A with minimum error in the sense of least squares is given by a truncation of the expansion (5.27) to k terms. That is,

$$A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T, \qquad k \le r. \tag{5.28}$$

We have already studied two direct applications of these SVD low-rank approximations, namely, in information retrieval (Section 3.5) and simple substitution cryptograms (Section 3.6). Now we discuss one more application of SVD low-rank approximations, this time to image compression, of both, gray scale, and color images.

5.3.1 Compressing grayscale images

As remarked before, given a grayscale image, this can be understood as an $m \times n$ matrix X whose entries are values between 0 and 255 indicating different gray intensities between black (0) and white (255). Let us assume that such a matrix X has rank r. Then, by using the notation in (5.27), its SVD factorization $X = U\Sigma V^T$ can be written as

$$X = \sigma_1 u_1 v_1^T + \cdots + \sigma_r u_r v_r^T. \tag{5.29}$$

One very important fact to remember about this factorization is that the singular values satisfy the inequalities

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0.$$

This means that the importance of the terms in the expansion (5.29) decreases as more terms are considered, or equivalently, the first terms of the expansion must contain the most important information about the matrix A. This remarkable fact about the SVD of X is exactly what we can exploit to achieve compression: Instead of storing the whole expansion (5.29), we can try to store just a truncation of such expansion to k terms, with $k < r$, dropping all terms with coefficients $\sigma_{k+1}, \ldots, \sigma_r$.

For a chosen value of $k < r$, we know from Theorem 3.21 that

$$X_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T$$

is an optimal rank-k approximation to X, and therefore, we expect that the compressed image X_k will look very similar to the original one X.

As expected, we have a trade-off between quality and storage savings. The lower the value of k, the more we save and compress, but at the same time we may be losing some quality of the compressed image. We want to illustrate this with an example.

■ **EXAMPLE 5.11**

Consider again the image in Figure 5.10. We show in Figure 5.21 this original image along with three different compression rates, corresponding to the rank-k approximations. The original matrix has rank $r = 593$. Observe that with $k = 95$ we already obtain a very good approximation to the original image. This means that from the 593 terms in (5.29), we can drop $593 - 95 = 498$ terms and still obtain a good quality image.

5.3.2 Compressing color images

A very simple approach to compress color images via low-rank SVD approximations is to treat each color coordinate in (R, G, B) independently. Recall that a color image is understood as a three-dimensional array (see Figure 5.14). Since each layer is a usual

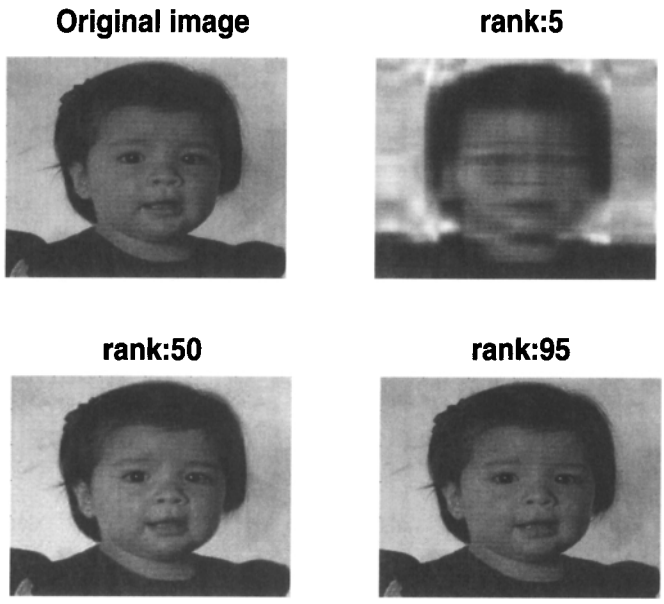


Figure 5.21 Compression with low-rank SVD approximations.

two-dimensional matrix, we compute the SVD factorization of each one of them, and then apply low-rank approximation just the way we did to grayscale images. The final step is to reconstruct the whole (compressed) image by putting the three layers back together again.

■ **EXAMPLE 5.12**

Consider the color image in Figure 5.22 and one of its 8×8 blocks. We want to apply SVD compression to both images by truncating the the SVD expansion (5.29) on each coordinate of (R, G, B). Each of these layers has rank $r = 1773$. With only $k = 55$, we are able to get a good quality compressed image. See Figure 5.23.

5.4 FINAL REMARKS AND FURTHER READING

In this section, we have studied two different approaches to grayscale and color image compression. The first and most important one is done via the discrete cosine transform, which is used by JPEG. The second approach is presented for completion, as an application of the singular value decomposition.

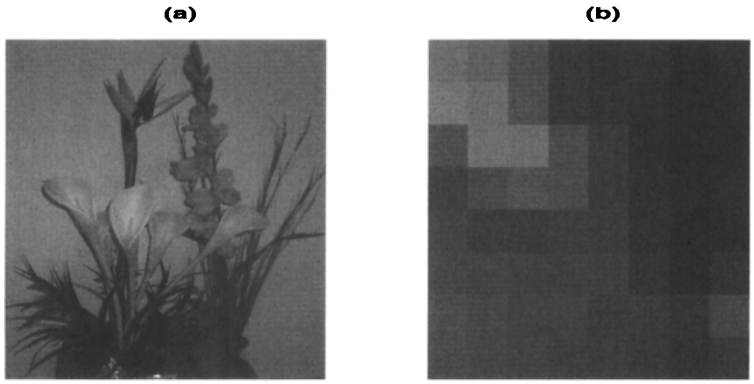


Figure 5.22 Original image and 8×8 block.

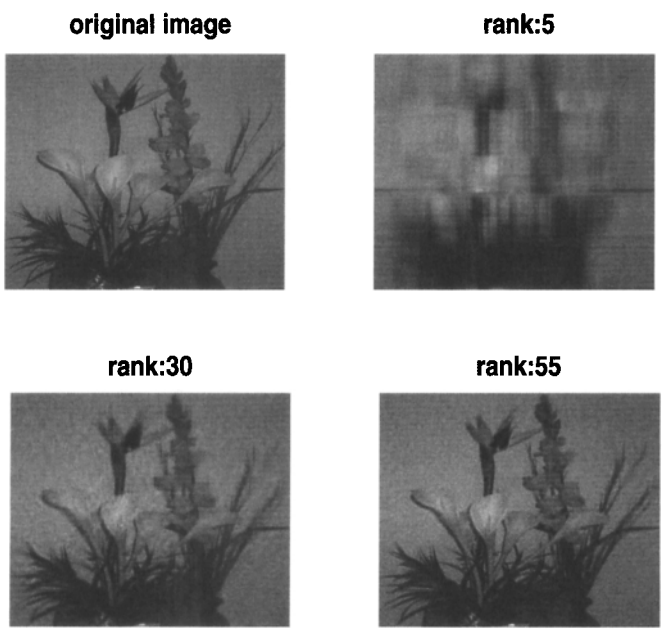


Figure 5.23 Original image and low-rank approximations.

The topic of image compression is discussed by a large list of books and articles. A great reference on data compression in general is the book by D. Salomon [59]. The book by K. S. Thyagarajan [65] offers a detailed discussion on image processing, including applications to digital cinema. A brief and clear exposition of image and sound compression can also be found in the book by T. Sauer [60]. The reader can always have full access to documents online with detailed and complete tables for Huffman coding and other information. See, for example, http://www.w3.org/Graphics/JPEG/itu-t81.pdf

Although not widely used yet, JPEG2000, based on wavelets, is the latest effort in achieving even more efficiency when compressing images, and it may soon become the standard choice for images in web browsers.

Ultimately, both, the JPEG and SVD approaches are excellent and current real-world applications of linear algebra and numerical analysis, and they represent a very interesting topic to convey to students.

EXERCISES

5.1 Let C be the orthogonal matrix in (5.1), and define $A_{n\times n}$ as

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

Show that the rows of C are unit eigenvectors of A.

5.2 Observe the DCT matrix in Example 5.1 row by row. What pattern do you see in the signs of the entries, and how is this related to low-high frequencies?

5.3 Following Exercise 5.2, make up a matrix $C_{4\times 4}$ of positive and negative 1's that would follow a pattern similar to that in Example 5.1. Then normalize it to make it orthogonal. Apply this transform to the matrix corresponding to a 4×4 grayscale image (CXC^T).

5.4 Prove Theorem 5.8.

5.5 Interpolate the following data:

$$(0,3),\ (1,1),\ (2,-1),\ (3,3),\ (4,1.5),\ (5,-0.5),\ (6,-2)$$

using the DCT. Plot the data and the interpolating function together.

5.6 Consider the data of Exercise 5.5. Apply DCT least squares approximation by dropping the last two terms of its interpolating polynomial. Plot both the least squares approximation and the interpolating polynomials as well as the data points.

5.7 Consider the input data matrix

$$X = \begin{bmatrix} -3.50 & -1.50 & -0.75 & -0.70 & -0.75 & -1.50 & -3.50 \\ -3.50 & -1.25 & -0.65 & -0.60 & -0.65 & -1.25 & -3.50 \\ -3.50 & -1.50 & -1.00 & -0.50 & -1.00 & -1.50 & -3.50 \\ -3.50 & -1.00 & -0.40 & 0.60 & -0.40 & -1.00 & -3.50 \\ -3.50 & -1.25 & -0.25 & 0.10 & -0.25 & -1.25 & -3.50 \\ -3.50 & -2.00 & -0.25 & 0.00 & -0.25 & -2.00 & -3.50 \\ -3.50 & -3.00 & -2.50 & -2.00 & -2.50 & -3.00 & -3.50 \end{bmatrix}.$$

Find the DCT of X and plot the graph of the interpolating function.

5.8 Consider again the data X of Exercise 5.7. By following Example 5.5, compute two least squares approximations by requiring that $k + l \leq 4$ and $k + l \leq 6$.

5.9 Try the following compression technique: Given a grayscale image, crop it so that the number of rows and columns is a multiple of 8. Then, replace each entry of each 8×8 block with its corresponding average pixel value in that block. Plot both, the original and the compressed image.

5.10 True or False? Any file can be compressed.

5.11 Find a scaling function that transforms an arbitrary interval $[a,\ b]$ into the interval $[0,\ 255]$.

5.12 Obtain and plot 4-bit (16 variations) and 8-bit (256 variations) black to white gradients.

5.13 Consider the 4×4 block image Y of Example 5.6. Plot the original image together with three approximations to it, according to the number of basis images Y_i used: (a) $i = 0$, (b) $i = 0, 1, 2, 3$, (c) $i = 0, 1, \ldots, 8$.

5.14 Import a grayscale image into MATLAB.

(a) Extract an 8×8 block from the image and compress it by using the quantization matrices (5.16) and (5.17), with $s = 3$. Compare your results.

(b) Apply the same process to the whole image.

5.15 Repeat Exercise 5.14 but now using the following quantization matrix, for $s = 5$:

$$K = s \begin{bmatrix} 5 & 5 & 5 & 5 & 5 & 6 & 6 & 8 \\ 5 & 5 & 5 & 5 & 5 & 6 & 7 & 8 \\ 5 & 5 & 5 & 5 & 6 & 7 & 8 & 9 \\ 5 & 5 & 5 & 6 & 7 & 8 & 9 & 10 \\ 5 & 5 & 6 & 7 & 8 & 9 & 11 & 12 \\ 6 & 6 & 7 & 8 & 9 & 11 & 13 & 14 \\ 6 & 7 & 8 & 9 & 11 & 13 & 15 & 16 \\ 8 & 8 & 9 & 10 & 12 & 14 & 16 & 19 \end{bmatrix}.$$

5.16 Denote with X the original image and with Z the compressed one. If there was no loss in the compression, then Z is identical to X and the image $A = X - Z$ is a matrix of zeros and therefore black. For the 8×8 block of Exercise 5.14, obtain the corresponding

matrices A corresponding to both quantization matrices (5.16) and (5.17), and display their images. Which one is farther from a black image?

5.17 Consider the following orthogonal (Haar) matrix:

$$H = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{2} & \sqrt{2} & -\sqrt{2} & -\sqrt{2} \\ 2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & -2 \end{bmatrix}.$$

Starting with the canonical basis of the vector space of 8×8 matrices, obtain the basis images associated with H, to obtain an image similar to Figure 5.9.

5.18 Repeat Exercise 5.17 for the (Hadamard) matrix

$$H = \frac{1}{\sqrt{8}} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}.$$

5.19 Consider the following 8×8 block of pixels values:

$$X = \begin{bmatrix} 154 & 161 & 188 & 197 & 200 & 181 & 134 & 111 \\ 153 & 155 & 185 & 199 & 199 & 191 & 145 & 108 \\ 154 & 149 & 176 & 196 & 198 & 194 & 161 & 112 \\ 160 & 150 & 168 & 190 & 200 & 196 & 173 & 122 \\ 164 & 157 & 167 & 188 & 201 & 201 & 181 & 130 \\ 165 & 162 & 168 & 186 & 197 & 195 & 188 & 142 \\ 173 & 166 & 162 & 181 & 194 & 191 & 191 & 160 \\ 184 & 171 & 155 & 176 & 197 & 198 & 191 & 176 \end{bmatrix}.$$

Compute its DCT transform $Y = CXC^T$, and verify that the sums of squares of the entries in both matrices X and Y are equal. Why is this true?

5.20 We know that we can compress an image by filtering out high frequency terms, retaining only the low frequency ones, which are the most important to the human eye. Experiment compressing a grayscale image but this time filtering out the low frequency terms and retaining the high frequency ones.

5.21 Suppose you have a color RGB image. Change it to grayscale by expressing the grayscale intensities as a linear combination of the three coordinates R, G and B. Then, compare your result to the one you obtain with the MATLAB command rgb2gray.

5.22 By setting $x = [R \quad G \quad B]^T$ and $z = [Y \quad U \quad V]^T$, write the change of coordinates in (5.19) as the transformation

$$z = Tx + b,$$

for some matrix T and some vector b.

5.23 Import a grayscale image into MATLAB.

(a) Extract an 8×8 block from the image and compress it by individually applying the luminance quantization matrix (5.17) to each color R, G, B.

(b) Apply the same process to the whole image.

5.24 Import a grayscale image into MATLAB.

(a) Extract an 8×8 block from the image and compress it by first changing from RGB to YUV coordinates as in (5.19) and then using the luminance quantization matrix (5.17) for Y and the chrominance matrix (5.20) for UV.

(b) Apply the same process to the whole image.

5.25 Assume we have the set of symbols: $\{A, B, C, D, E\}$, with probabilities: $A = 0.25$, $B = 0.10$, $C = 0.15$, $D = 0.15$, $E = 0.35$. Find the entropy (5.21). According to this entropy, find out the optimal number of bits needed to code the string DECEEEAA.

5.26 Refer to the Huffman tree of Figure 5.18. Build a different tree for the same symbols, but this time taking different choices; e.g., at step 1, combine D and F instead of E and F. What codes do you get for the symbols A, B, C, D, E? Next, translate the string ACDAACBBEB into a bit string. How many bits per symbol are needed?

5.27 Construct a Huffman tree that generates the Table 5.1.

5.28 Suppose two neighboring 8×8 blocks have the DC coefficients $DC_4 = 35$, $DC_5 = 42$. Find the coding of the difference coefficient $D = DC_5 - DC_4$ as given in (5.23).

5.29 From Table 5.2, find the codes for (a)12, and (b)-60.

5.30 Suppose we have the following AC coefficients:

$$8, -5, 0, 0, 0, 0, 0, 4.$$

Translate this into a bit stream following to (5.25).

5.31 Consider the following quantized matrix:

$$Y_Q = \begin{bmatrix} -35 & 14 & -2 & -1 & 0 & 0 & 0 & 0 \\ -27 & -11 & 3 & 0 & 0 & 0 & 0 & 0 \\ 12 & 6 & -1 & 0 & 0 & 0 & 0 & 0 \\ 3 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Following the zigzag pattern of Figure 5.20, find the Huffman code for all the quantized coefficients.

5.32 Import a grayscale image into MATLAB.

(a) Extract an 8×8 block from the image and apply rank-k SVD approximations to the corresponding 8×8 matrix for three different values of k. Print the rank-k images together with the original image in a 4×4 figure.

(b) Apply the same process to the whole image.

5.33 Use SVD compression on a color image by first individually compressing each color (RGB) as if you were dealing with grayscale intensities, and then reconstruct the compressed image from the superposition of the three colors.

CHAPTER 6

ORDINARY DIFFERENTIAL EQUATIONS

A great variety of natural phenomena can be modeled as continuous problems, those whose solutions are functions that depend on or vary with time. In general, a relationship between a function and its derivatives is known, and we have to find the function itself out of that relation. A relation of this type is what is called a differential equation: an equation that depends on a function and its derivatives.

It is here where the most fundamental theories and methods from calculus, analysis and linear algebra are crucial for establishing the solutions of such differential equations. When we are dealing with one-dimensional differential equations, calculus suffices, but when we move to solve n-dimensional differential equations, we will need some of the most elegant and useful tools of linear algebra for computing a solution. At the same time, some theorems from analysis (or advanced calculus) become essential for establishing existence, uniqueness and other properties of the solutions of differential equations. This is probably why differential equations are considered one area of applied mathematics: Calculus, analysis, and linear algebra are applied to the theory and methods of differential equations.

Once we leave the textbook or classroom examples, most differential equations are still theoretically solvable by analytical methods, but, impossible to solve in practice due to their complexity. At that point, we need to make use of numerical approximations to the solutions of differential equations. For initial value problems, there exists a number of very reliable software applications, (e.g., MATLAB) that can readily compute the numerical

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

solution with such a degree of accuracy that it can be considered to be the exact solution. We can also study the solution of boundary value problems by analytical methods, and numerically approximate some solutions using some scientific software. Solving boundary value problems is by far (analytically and numerically) more difficult than solving initial value problems. Some boundary value problems will be studied in Chapter 7.

As we move through this chapter, several simple examples will be provided to illustrate the theory and methods introduced. Some of these examples will be application problems, which in some cases will be developed in full detail, starting with the modeling stage. We also take care to provide with some theorems on existence and uniqueness of solutions to differential equations.

In this chapter we will focus on one-dimensional differential equations, as well as on linear systems of differential equations. We will deal with general nonlinear systems of differential equations and their applications in Chapter 7.

6.1 ONE-DIMENSIONAL DIFFERENTIAL EQUATIONS

We start our study of models in differential equations with the scalar or one-dimensional case, and we consider equations of first and second order. Later we consider systems of linear differential equations, leaving the case of systems of nonlinear equations for a subsequent chapter. Before doing anything, we need to establish a result on the existence (and uniqueness) of solution to this type of differential equations.

6.1.1 Existence and uniqueness

In Chapter 7, we will study nonlinear systems of differential equations and will consider some theorems on existence and uniqueness of solutions in a general framework. In this chapter, before presenting a list of classical models of one-dimensional differential equations and the qualitative behavior of their solutions, we need to establish a result on the existence and uniqueness of such solutions.

We follow the typical approach of rewriting an initial value problem as an integral equation and then of finding a fixed point of the corresponding operator.

Let us consider the initial value problem (IVP)

$$\begin{cases} \dot{x} &= f(t,x), \\ x(t_0) &= x_0, \end{cases} \tag{6.1}$$

where we use the notation $\dot{x} = \frac{dx}{dt}$.

The IVP (6.1) can be solved analytically for a few particular cases. For instance, when f is the linear function $f(t,x) = ax$, we can write the solution explicitly. Namely,

$$\dot{x} = ax, \quad x(t_0) = x_0 \quad \Longrightarrow \quad x(t) = x_0\, e^{a(t-t_0)}. \tag{6.2}$$

■ **EXAMPLE 6.1**

The solution to the IVP

$$\dot{x} = 5x, \; x(0) = 3$$

is $x(t) = 3e^{5t}, \quad t \geq 0.$

One can also prove that if p and q are continuous functions of t, then the general solution of the differential equation

$$\dot{x} = p(t)x + q(t) \tag{6.3}$$

can be written as

$$x(t) = \left[\int q(t)e^{-\int p(t)dt}\, dt + c\right] e^{\int p(t)dt}, \tag{6.4}$$

where c is a constant to be determined from the initial condition.

■ **EXAMPLE 6.2**

Consider the differential equation

$$\dot{x} = -2tx + t.$$

Then, following (6.4), we have

$$x(t) = \left[\int te^{t^2} dt + c\right] e^{-t^2} = \left[\frac{1}{2}\, e^{t^2} + c\right] e^{-t^2} = \frac{1}{2} + ce^{-t^2}.$$

Similarly, we can get some explicit solutions for a few more differential equations, so that the problem of existence is taken care of, but we are mostly interested in conditions under which we can guarantee not only the existence but also the uniqueness of solutions of the general IVP (6.1), for a given function f.

We need to introduce a few definitions and theorems as preparation for our main theorem on existence and uniqueness.

Definition 6.1 *A function $f : I \subseteq \mathbb{R} \to \mathbb{R}$ is called* **Lipschitz** *on the interval I if there exists $L \geq 0$ such that for any $x, y \in I$,*

$$|f(x) - f(y)| \leq L\,|x - y|. \tag{6.5}$$

The smallest constant L that satisfies (6.5), is called Lipschitz constant.

■ **EXAMPLE 6.3**

Let $f(x) = x^2$ be defined on an arbitrary interval $[a, b]$. Then, for any $x, y \in [a, b]$ we have

$$|f(x) - f(y)| = |x^2 - y^2| = |x + y|\,|x - y| \leq L\,|x - y|,$$

where $L = 2\max\{|a|, |b|\}$. Therefore, f is Lipschitz on $[a, b]$.

Remark 6.2 *With the obvious modifications, we can generalize (6.5) for a function $f : X \to Y$ between any two spaces where the corresponding distance functions (or norms) are well–defined.*

There are two important facts that relate Lipschitz functions with continuous and differentiable ones. The proofs are left as exercises to the reader.

(a) If f is Lipschitz in an interval I, then it is continuous in I.

(b) If f is differentiable in an interval I and $|f'(x)| \leq L$ for all $x \in I$, then f is Lipschitz on that interval I.

Remark 6.3 *When the constant L in (6.5) satisfies $L < 1$, we say that the function f is a* **contraction.**

Definition 6.4 *A function $f : I \subseteq \mathbb{R} \to \mathbb{R}$ is said to have a* **fixed point** *if there exists $x \in I$ such that*

$$f(x) = x.$$

■ **EXAMPLE 6.4**

Let $f(x) = x^3$. By solving the equation $f(x) = x$, we observe that f has three fixed points in $\mathbb{R}$, namely $-1, 0, 1$, but only one fixed point in $(0, 2)$.

Note: The definition of a fixed point of a function f can be generalized to an operator T that maps a function to another function. In this case, we say that a function $f(x)$ is a fixed point of the operator T if

$$T[f(x)] = f(x).$$

■ **EXAMPLE 6.5**

For any continuous function f on the interval $[0, 2]$, define

$$T[f(x)] = 1 + \int_0^x f(t)dt.$$

Then, the function $f(x) = e^x$ is a fixed point of T. Indeed,

$$T[f(x)] = 1 + \int_0^x e^t dt = 1 + e^x - 1 = e^x = f(x).$$

Before we introduce a theorem that gives sufficient conditions for the existence and uniqueness of a fixed point of a real function, let us recall that a sequence of real numbers $\{x_k\}$ is said to **converge** to some $x \in \mathbb{R}$ if

$$\lim_{k\to\infty} x_k = x. \tag{6.6}$$

One very useful means of determining whether a sequence $\{x_k\}$ is converging without actually computing the limit in (6.6) is to look at two arbitrary iterates x_m and x_n and see how close they are to each other, as m and n get larger and larger.

Definition 6.5 *We say a sequence of real numbers $\{x_k\}$ is a* **Cauchy sequence** *if*

$$|x_m - x_n| \to 0, \quad \text{as} \quad m, n \to \infty. \tag{6.7}$$

Cauchy and convergent sequences are related by the following statement:

Every Cauchy sequence in $\mathbb{R}^n$ is a convergent sequence.

Remark 6.6 *Later in Chapter 7 we will come back to this and other results in a more formal and general framework.*

Theorem 6.7 *Let I be a closed interval in $\mathbb{R}$, and let $f : I \to I$ be a contraction in I. Then, f has a unique fixed point.*

Proof. Let $x_0 \in I$ be arbitrary, and define the following sequence (known as *fixed point iteration*):

$$x_{k+1} = f(x_k), \quad k = 0, 1, 2, \ldots$$

Then,

$$\begin{aligned} |x_2 - x_1| &= |f(x_1) - f(x_0)| \le L|x_1 - x_0|, \\ |x_3 - x_2| &= |f(x_2) - f(x_1)| \le L|x_2 - x_1| \le L^2|x_1 - x_0|, \end{aligned}$$

and by induction

$$|x_{k+1} - x_k| \leq L^k |x_1 - x_0|.$$

Thus, for any $j \geq 1$,

$$\begin{aligned} |x_{k+j} - x_k| &\leq |x_{k+1} - x_k| + |x_{k+2} - x_{k+1}| + \cdots + |x_{k+j} - x_{k+j-1}| \\ &\leq \left(L^k + L^{k+1} + \cdots + L^{k+j-1}\right) |x_1 - x_0| \\ &\leq L^k(1 + L + \cdots + L^{j-1}) |x_1 - x_0| \\ &= \tfrac{L^k(1-L^j)}{1-L} |x_1 - x_0| \\ &\leq \tfrac{L^k}{1-L} |x_1 - x_0|. \end{aligned}$$

Since $L < 1$, we have $|x_{k+j} - x_k| \to 0$ as $k \to \infty$, so that $\{x_k\}$ is a Cauchy sequence and therefore it converges to some $x^* \in I$. Then, we have

$$|f(x^*) - x^*| \leq |f(x^*) - x_k| + |x_k - x^*| \leq L|x^* - x_{k-1}| + |x_k - x^*|,$$

which converges to zero as $k \to \infty$. Hence, $f(x^*) = x^*$, and f has a fixed point.
As for uniqueness, suppose y^* is another fixed point of f. Then,

$$|x^* - y^*| = |f(x^*) - f(y^*)| \leq L|x^* - y^*| < |x^* - y^*|,$$

which implies that $x^* = y^*$.

□

Remark 6.8 *The statement of Theorem 6.7 remains valid if instead of real functions we use operators that map continuous functions on some interval $[a, b]$ into continuous functions on the same interval.*

The key observation for proving existence and uniqueness is that the initial value problem

$$\dot{x} = f(t, x), \quad x(t_0) = x_0 \tag{6.8}$$

is equivalent to the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds. \tag{6.9}$$

Also observe that equation (6.9) can be written as $T[x(t)] = x(t)$, where

$$T[x(t)] = x_0 + \int_{t_0}^{t} f(s, x(s))\, ds. \tag{6.10}$$

Thus, the solution $x(t)$ of the initial value problem (6.8) is a fixed point of the operator T, and the converse is also true.

Now we are ready to state and prove the theorem on existence and uniqueness.

Theorem 6.9 Existence and Uniqueness. *Assume f is continuous on the set $D = \{(t,x) : a \le t \le b,\ x \in \mathbb{R}\}$ and Lipschitz in x. If (t_0, x_0) is an arbitrary point in D, then the IVP*

$$\dot{x} = f(t,x), \quad x(t_0) = x_0 \tag{6.11}$$

has a unique solution in $[a,\ b]$.

Proof. Let us define the distance between two continuous functions in $[a,\ b]$ as

$$d(x_1, x_2) = \max_{t \in [a,\, b]} |x_1(t) - x_2(t)|.$$

Then, using the fact that f is Lipschitz with respect to x, we have

$$|f(t, x_1(t)) - f(t, x_2(t))| \le L|x_1(t) - x_2(t)| \le L\, d(x_1, x_2).$$

Thus, for the operator in (6.10), we have

$$\begin{aligned} |T[x_1(t)] - T[x_2(t)]| &= \left|\int_{t_0}^{t} [f(s, x_1(s)) - f(s, x_2(s))]\, ds\right| \\ &\le Ld(x_1, x_2) \left|\int_{t_0}^{t} ds\right| = L\, d(x_1, x_2)|t - t_0|. \end{aligned}$$

Hence,

$$\begin{aligned} \left|T^2[x_1(t)] - T^2[x_2(t)]\right| &= \left|\int_{t_0}^{t} \{f\,(s,\, T[x_1(s)]) - f\,(s,\, T[x_2(s)])\}\, ds\right| \\ &\le L\left|\int_{t_0}^{t} |\,T[x_1(s)] - T[x_2(s)]\,|\, ds\right| \\ &\le L^2 d(x_1, x_2) \left|\int_{t_0}^{t} |s - t_0|\right| \\ &\le L^2 d(x_1, x_2) \tfrac{|t-t_0|^2}{2}. \end{aligned}$$

By following the same reasoning we get for $k \ge 1$:

$$\begin{aligned} \left|T^k[x_1(t)] - T^k[x_2(t)]\right| &\le L^k d(x_1, x_2) \tfrac{|t-t_0|^k}{k!} \\ &\le L^k \tfrac{(b-a)^k}{k!} d(x_1, x_2). \end{aligned}$$

However, we know that the series $\sum_{k=1}^{\infty} L^k \frac{(b-a)^k}{k!}$ is convergent, so we must have that $L^k \frac{(b-a)^k}{k!} < 1$ and therefore, T^k is a contraction, which implies that it has a unique fixed point $x(t)$, but (see Exercise 6.8) $x(t)$ is the same unique fixed point of T, or equivalently, it is the unique solution of the IVP (6.11).

□

We have established that the IVP (6.11) has a unique solution if the right–hand side f is continuous in (x,t) and Lipschitz in x. Next, we study some simple problems that can be modeled as initial value problems of ordinary differential equations in the one-dimensional case.

6.1.2 A simple population model

Consider the initial value problem

$$\begin{aligned} \dot{x} &= ax, \quad a \in R \\ x(0) &= x_0. \end{aligned} \tag{6.12}$$

As we already know, the solution of (6.12) is given by

$$x(t) = e^{at} x_0,$$

which means the solution is always exponentially increasing or decreasing for all time t, depending on whether a is positive or negative, respectively.

The most classic and illustrative problem that can be modeled as (6.12) is that of population. Let $P(t)$ be the size of a certain population at time t. If we denote with $bP(t)$ (respectively $dP(t)$) the rate at which individuals are added (respectively removed), then the number of individuals added (respectively removed) from time t to $t + \Delta t$ is $bP(t)\Delta t$ (respectively, $dP(t)\Delta t$). Then, by a simple law of conservation, we get

$$P(t + \Delta t) - P(t) = bP(t)\Delta t - dP(t)\Delta t. \tag{6.13}$$

By dividing (6.13) by Δt and taking the limit as $\Delta t \longrightarrow 0$, we get

$$\frac{dP}{dt} = k\,P, \qquad \text{where} \quad k = b - d,$$

which, provided with an initial population $P(0) = P_0$, is of the form (6.12).

■ **EXAMPLE 6.6**

Assume the net growth of the USA population is $k = 0.01035$ and that in 1979 the population size was 225.1 million. Assume also that the rate of change of the population is proportional to the current population size. Find out what the population would be in 2009.

We can set the year 1979 to correspond to $t = 0$; that is, $P(0) = 225.1$. The differential equation modeling the population growth is then

$$\frac{dP}{dt} = 0.01035\,P, \qquad P(0) = 225.1,$$

whose solution is $P(t) = e^{0.01035\,t}(225.1)$, for any time $t \geq 0$. In particular for the year 2009 ($t = 30$), we have $P(30) = 307.06$ million, a number quite close to the actual population size.

6.1.3 Emigration

A simple modification of (6.12) corresponds to a constant emigration E of individuals per unit time. In this case, the rate of change of population gets negatively affected by E and the initial value problem becomes

$$\begin{aligned} \dot{x} &= ax - E, \\ x(0) &= x_0. \end{aligned} \tag{6.14}$$

Here, the solution of (6.14) is given by (see Exercise 6.13)

$$x(t) = e^{at}x_0 - \frac{E}{a}(e^{at} - 1). \tag{6.15}$$

Compare this solution with that of (6.12) to see the effect that emigration has in the population for any time t. Also, observe from the solution that, at least theoretically, if a is negative, then $x(t) \to \frac{E}{a}$, as $t \to \infty$.

■ **EXAMPLE 6.7**

Assume the net growth rate in a country is $a = 0.014$, with an initial population of 25 million, and that half a million people per year leaves the country. What is the population after 45 years?

This problem can be modeled as (6.14):

$$\begin{aligned} \dot{P} &= 0.014P - 0.5, \\ P(0) &= 25. \end{aligned}$$

The solution is

$$P(t) = e^{0.014t}(25) - \frac{0.5}{0.014}(e^{0.014t} - 1),$$

for any time t. In particular, $P(45) = 15.597$ million.

6.1.4 Time-varying emigration

A generalization of model (6.14) is to consider a nonconstant emigration term $E(t)$. This gives the following IVP:

$$\begin{aligned} \dot{x} &= ax - E(t), \\ x(0) &= x_0. \end{aligned} \tag{6.16}$$

The solution of this IVP is given by (Exercise 6.15)

$$x(t) = e^{at}x_0 + e^{at}\int E(t)e^{-at}\,dt. \tag{6.17}$$

Compare this general solution with the one in (6.15).

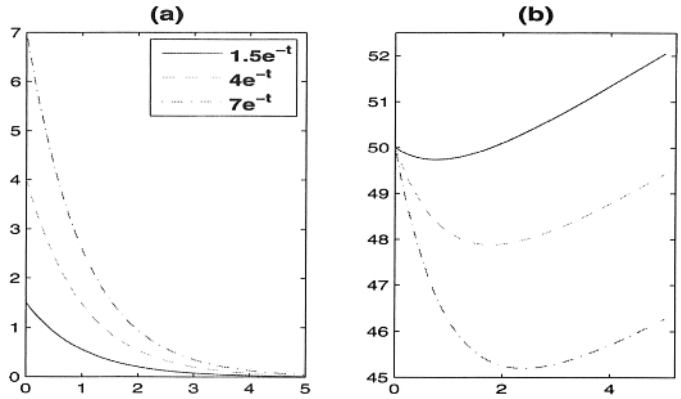


Figure 6.1 Exponential functions and solutions of Example 6.8.

■ **EXAMPLE 6.8**

Consider a population model with exponentially decaying emigration, i.e., $E(t) = re^{-t}$, and with $a = 0.014$. In this case, (6.16) corresponds to a population with fewer individuals emigrating as time passes. It is clear that the larger the coefficient r of e^{-t} , the longer it takes to see the effect of the exponential $E(t)$. See Figure 6.1, where in part(a) we show the plots of the exponential functions themselves and in part (b) we show the plots of the corresponding solutions, for an initial population $x(0) = 50$. Observe that initially, the population decreases, but as the emigration decays, the population starts to grow again.

Note. In Chapter 7, we will study two-dimensional population (predator–prey) models, where harvesting will play the role of emigration.

6.1.5 Competition

A very well–known and studied initial value problem is the so–called **logistic equation**:

$$\begin{aligned} \dot{x} &= ax - bx^2, \\ x(0) &= x_0. \end{aligned} \tag{6.18}$$

This problem is different from the previous ones in that it is nonlinear. The solution is given by (see Exercise 6.16)

$$x(t) = \frac{a}{b + e^{-at}(a - bx_0)/x_0}. \tag{6.19}$$

It is clear from this solution that $x(t) \to \frac{a}{b}$ as $t \to \infty$. This is in contrast to (6.12) where the population is supposed to always increase or decrease exponentially for all time t. This

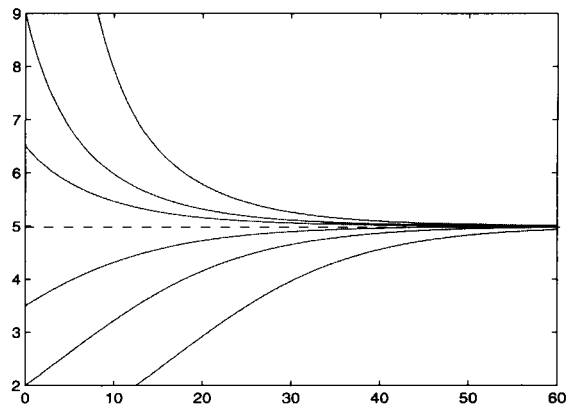


Figure 6.2 Solutions of Example 6.9.

is a little more realistic model because we take into account the fact that as population increases, competition for sources also will increase, which in turn indicates the death rate must increase with the size of population. This is reflected in the term $-bx^2$ in (6.18).

■ **EXAMPLE 6.9**

Suppose in a certain species population the net rate of growth is $a = 0.1$ and the rate of competition is $b = 0.02$. Assume also that initially, the population size is 9 million. Then, according to the general solution (6.19), competition should prevent unbounded growth and all solutions must converge to a steady state. In fact, the solution approaches $u = \frac{a}{b} = 5$ as $t \to \infty$, and this is independent of the initial population. In Figure 6.2, we show several solution curves corresponding to different initial populations.

Remark 6.10 *Later in Chapter 7 all these ideas of emigration, competition, and so on. will be generalized to systems of differential equations to model population interactions and harvesting of some species.*

6.1.6 Spring systems

We now move to differential equations of second order. Later we will see that such equations can be represented as first order differential equations in two dimensions, after a change of variables, and a theorem on existence and uniqueness of solutions can then be established.

We will denote $\ddot{x} = \frac{d^2x}{dt^2}$.

We start considering the following homogeneous differential equation:

$$m\ddot{x} + c\dot{x} + kx = 0, \tag{6.20}$$

for some positive real numbers m, c, and k. Typically, the solutions of (6.20) are classified according to the roots of its characteristic equation

$$m\lambda^2 + c\lambda + k = 0.$$

There are three cases: (i) the roots are real and distinct, (ii) the roots are real and equal, (iii) the roots are complex conjugate. Accordingly, for some constants c_1, c_2, we have the following solutions:

$$\begin{array}{llll} \text{(i)} & \lambda_1 \neq \lambda_2 & \longrightarrow & x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}. \\ \text{(ii)} & \lambda_1 = \lambda_2 & \longrightarrow & x(t) = c_1 e^{\lambda_1 t} + c_2 t\, e^{\lambda_2 t}. \\ \text{(iii)} & \lambda = a \pm bi & \longrightarrow & x(t) = e^{at}(c_1 \cos t + c_2 \sin bt). \end{array} \tag{6.21}$$

Consider now the classic problem of a fixed spring, to which an object of mass m is attached. The weight of this object stretches down the spring a distance x, and the spring itself pushes back up with a force that is proportional to the distance it has been stretched. This is **Hooke's law**, which can be expressed as

$$F_k = -kx,$$

where k is the spring constant representing the spring stiffness. The negative sign indicates that it is acting in the opposite direction. If we also consider there is friction or damping involved, it is assumed that the force from friction is proportional to the velocity with which the object is moving, that is,

$$F_c = -c\dot{x},$$

where c is called the damping constant. Now, Newton's law

$$F = ma,$$

with the total force $F = F_c + F_k$, gives $-c\dot{x} - kx = m\ddot{x}$, or

$$m\ddot{x} + c\dot{x} + kx = 0,$$

which is (6.20). Again, the solutions to this equation will be classified according to the roots of the characteristic equation. The roots are

$$\lambda_{1,2} = -\frac{c}{2m} \pm \frac{1}{2m}\sqrt{c^2 - 4mk}. \tag{6.22}$$

By following (6.22), we say the spring system represents:

- Overdamping, if $c^2 > 4mk$ ($\lambda_1 \neq \lambda_2$, both negative).
- Critical damping, if $c^2 = 4mk$ ($\lambda_1 = \lambda_2$, negative).

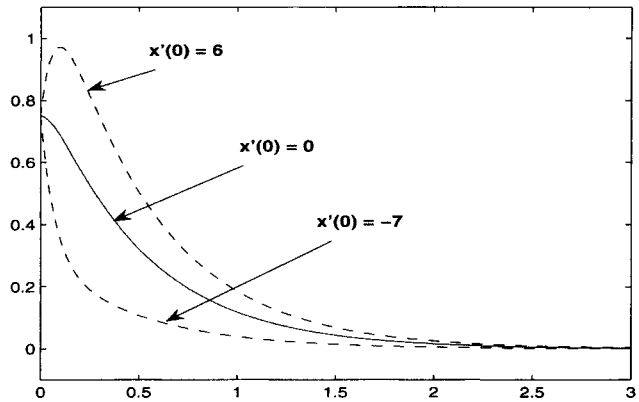


Figure 6.3 Overdamped solutions in Example 6.10.

- Underdamping, if $c^2 < 4mk$ ($\lambda_{1,2} = a \pm bi$, a negative).

In any of these three cases, the solutions will eventually decay to zero, due to the presence of damping (or of the negative sign of λ_1, λ_2, or a). With overdamping the solution will decay to zero exponentially; something similar happens with critical damping, but the solution will decay to zero a little more slowly, as a result of a linear term t in part (ii) of (6.21). With underdamping, the solution will oscillate to zero. These three cases of damping are clearly expressed in the solutions given in (6.21).

■ EXAMPLE 6.10

Consider a spring system with $m = 1/4$, $c = 4$, and $k = 7$. Assume that $x(0) = 0.75$ and $x'(0) = 0$. This means the spring is pulled 0.75 units below its equilibrium position and then it is simply released.

In this case, the general solution of (6.20) is

$$x(t) = c_1 e^{-14t} + c_2 e^{-2t}.$$

In particular, by substituting the given initial conditions, we get $c_1 = -1/8$ and $c_2 = 7/8$. It is clear that the solution goes to zero exponentially fast (overdamping). In Figure 6.3, we show this particular solution (in solid curve) as well as other solutions with different initial velocities (dashed curves).

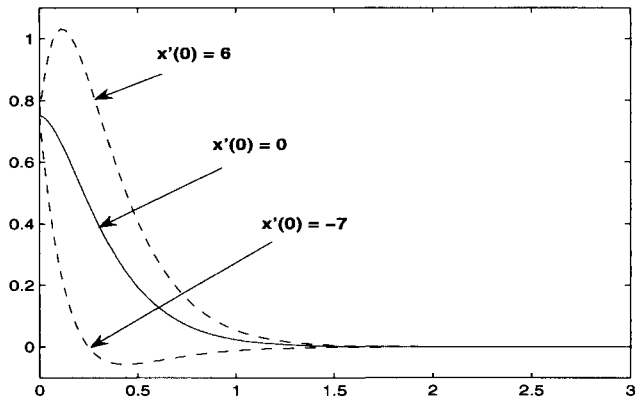


Figure 6.4 Critically damped solutions in Example 6.11

■ EXAMPLE 6.11

Let us change the value of c in Example 6.10 to $c = \sqrt{7}$, so that now we have $c^2 = 4mk$, and therefore the system is critically damped. The roots of the characteristic equation are $\lambda_{1,2} = -2\sqrt{7}$; then according to (6.21), the general solution is of the form

$$x(t) = c_1 e^{-2\sqrt{7}t} + c_2\, t\, e^{-2\sqrt{7}t}.$$

Some solutions are shown in Figure 6.4, all of them with $x(0) = 0.75$, but with a different initial velocity $x'(0)$.

The reader should observe the clear differences between the solutions of overdamped and critically damped systems, especially by looking at and comparing Figures 6.3 and 6.4. In particular, the critically damped solution with $x'(0) = -7$ first wanders (it "almost" oscillates) on negative values before settling down to a zero equilibrium solution, as opposed to the corresponding overdamped solution, which monotonically decays to zero.

■ EXAMPLE 6.12

We now consider the case $c^2 < 4km$. We modify the value of the damping constant c in Example 6.12 to $c = 1/2$, so that the inequality is satisfied. This causes the system to be underdamped and it will oscillate before settling to zero. The roots of the characteristic equation are $\lambda_{1,2} = -0.5 \pm 3.7081i$, and following (6.21), the general solution is

$$x(t) = e^{-0.5t}\,(c_1 \cos 3.7081\, t + c_2 \sin 3.7081t).$$

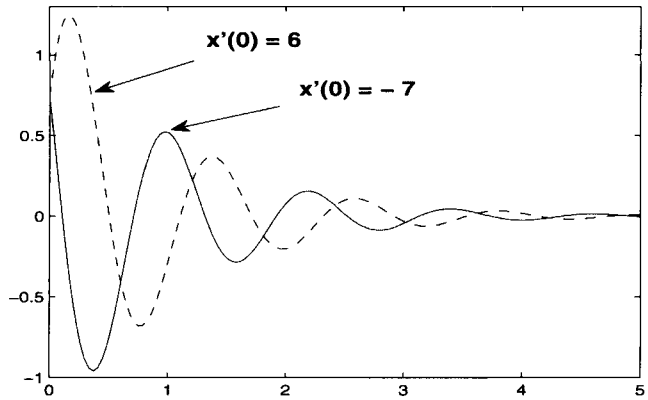


Figure 6.5 Underdamped solutions in Example 6.12

In Figure 6.5, we show two solutions, both with $x(0) = 0.75$ but with different initial velocities.

A very different situation happens when there is no damping or friction, that is, when $c = 0$. In this case, solutions will oscillate for all time t, since the roots of the characteristic equations will be purely imaginary and the solution to (6.20) will be combinations of sines and cosines. The characteristic equation corresponding to the differential equation

$$m\ddot{x} + kx = 0 \tag{6.23}$$

is $m\,\lambda^2 + k = 0$, and the roots are $\lambda_{1,2} = \pm\sqrt{\frac{k}{m}}\,i$; therefore, the solution of (6.23) is given by

$$x(t) = A\,\cos\sqrt{\frac{k}{m}}\,t + B\,\sin\sqrt{\frac{k}{m}}\,t, \tag{6.24}$$

where A and B are arbitrary constants.

■ **EXAMPLE 6.13**

Consider a spring system with no damping, where an object with mass $m = 0.5$ kg is attached to it. Suppose that a force of 100 Newtons stretches the spring 2 meters, and that the system is set in motion with $x(0) = 0.6$ and $\dot{x}(0) = -5$.

First we need to find the spring constant k, by using Hooke's law $F_s = -kx$. In magnitude, this means $100 = k(2)$, which gives $k = 50$. Thus, according to (6.24), the general solution is given by

$$x(t) = A\cos 10t + B\sin 10t.$$

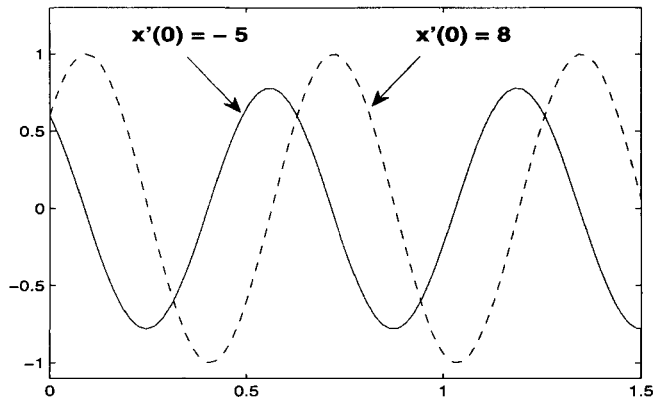


Figure 6.6 Undamped solutions in Example 6.13.

By using the initial conditions, we find $A = 0.6$, $B = -0.5$. Thus, no damping results in oscillations for all time t. This is illustrated in Figure 6.6 where we also show a second solution for a different initial velocity. Observe that a larger initial velocity causes a larger amplitude of the oscillations.

Nonhomogeneous Equations. We now generalize the differential equations (6.20) and (6.23) to the nonhomogeneous case, which models systems subject to some external force $f(t)$. If we combine this force with the damping and spring forces F_c and F_k introduced before, then applying again Newton's law $F = ma$, with

$$F = F_c + F_k + f(t),$$

we get $-c\dot{x} - kx + f(t) = m\ddot{x}$, or

$$m\ddot{x} + c\dot{x} + kx = f(t). \tag{6.25}$$

Although the function $f(t)$ describing the external force can be very general, here we restrict ourselves to the case of periodic external forces, those that can be described in terms of sines or cosines. That is, we consider

$$f(t) = a\cos\omega t \quad \text{or} \quad f(t) = a\sin\omega t, \tag{6.26}$$

where a is a constant real number. Later we will also study the case where $f(t) = e^{at}\cos bt$.

The models considered here come mostly from spring and other mechanical systems, although some other applications are also possible. We start studying the case where the system has no damping or friction.

6.1.7 Undamped equations

Suppose a vehicle of mass m is moving along a bumpy road with a velocity of 30 miles per hour. Suppose also that the road can be simply represented as

$$y = 0.04 \cos 3\pi t. \tag{6.27}$$

Let us oversimplify the vehicle to a wheel with one spring and no shock absorber, and let us represent its upward displacement as $x(t)$. Then, as the vehicle moves, the spring is stretched by the amount $x - y$. Thus, by Newton's law, we have $m\ddot{x} = -k(x - y)$, where k is the spring constant. Rewriting this equation and using (6.27), we get

$$m\ddot{x} + kx = 0.04k \cos 3\pi t. \tag{6.28}$$

Equation (6.28) represents a particular example of the general undamped nonhomogeneous differential equations of the form:

$$m\ddot{x} + kx = a \cos \omega t. \tag{6.29}$$

The general solution $x(t)$ of the nonhomogeneous equation (6.29) is the sum of the general solution $x_h(t)$ of the corresponding homogeneous differential equation $\ddot{x} + kx = 0$, and a **particular solution** $x_p(t)$ of the nonhomogeneous equation (6.29) itself. That is,

$$x(t) = x_h(t) + x_p(t). \tag{6.30}$$

From (6.24), we know $x_h(t)$ is a combination of sines and cosines:

$$x_h(t) = A \cos \omega_0 t + B \sin \omega_0 t, \tag{6.31}$$

where we have denoted

$$\omega_0 = \sqrt{\frac{k}{m}} \tag{6.32}$$

to be the so–called **natural frequency** of the system.

To find $x_p(t)$, we let $x_p = E \cos \omega t$, we compute its first and second derivatives, and then we substitute this into (6.29). Finally, by comparing coefficients (see Exercise 6.17), we find that $E = \frac{a/m}{\omega_0^2 - \omega^2}$. That is,

$$x_p(t) = \frac{a/m}{\omega_0^2 - \omega^2} \cos \omega t. \tag{6.33}$$

Then, the general solution to the undamped differential equation (6.29) is

$$x(t) = A \cos \omega_0 t + B \sin \omega_0 t + \frac{a/m}{\omega_0^2 - \omega^2} \cos \omega t, \tag{6.34}$$

where $\omega \neq \omega_0$.

Since sines and cosines are just the same curve up to a horizontal shift, it is possible and very convenient to write (6.34) in terms of sines or cosines only. In fact (see Exercise 6.18), we can write the general solution (6.34) as

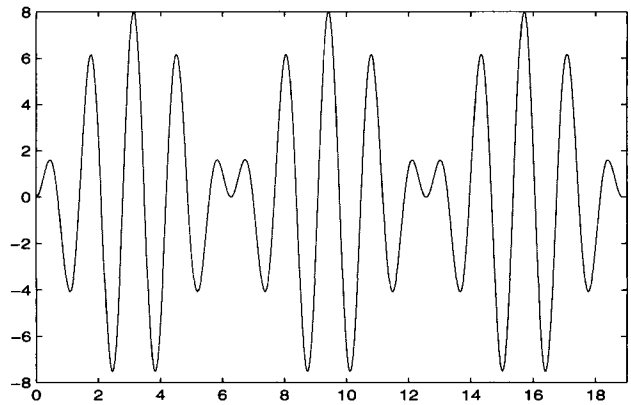


Figure 6.7 Solution of Example 6.14.

$$x(t) = C\cos(\omega_0 t - \alpha) + \frac{a/m}{\omega_0^2 - \omega^2}\cos\omega t, \tag{6.35}$$

where $C = \sqrt{A^2 + B^2}$, and $\alpha \in [0,\ 2\pi]$ is called the *phase angle*.

In both forms (6.34) or (6.35), we observe that $x(t)$ is the superposition of two oscillations, one with the natural frequency ω_0 and the other with the frequency ω of the external force. Again, since there is no damping or friction, the oscillations happen for all time t.

■ EXAMPLE 6.14

Suppose we have $m = 2, k = 32, a = 72, w = 5$, with the initial conditions $x(0) = \dot{x}(0) = 0$. Then $w_0 = \sqrt{k/m} = 4$ and the differential equation (6.29) is

$$\ddot{x} + 9x = 72\cos 5t.$$

The initial conditions give $A = 4, B = 0$, and the solution from (6.34) is

$$x(t) = 4\cos 4t - 4\cos 5t.$$

This is clearly the superposition of two oscillations with different frequencies, a phenomenon that is illustrated in Figure 6.7.

■ EXAMPLE 6.15

Consider the case of the vehicle on a bumpy road, represented through the equation (6.27), and assume that $m = 800$ kg and $k = 80,000$. The differential equation

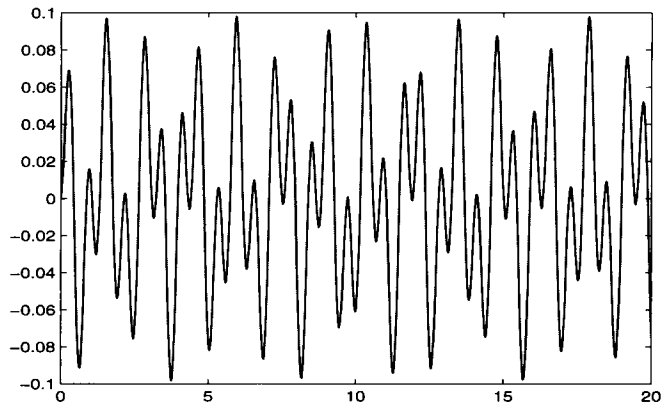


Figure 6.8 Solution of Example 6.15

representing the motion of the vehicle is given in (6.28). The solution corresponding to the initial conditions $x(0) = 0$ and $\dot{x}(0) = 0.1$ is shown in Figure 6.8.

Observe that the solutions (6.34) and (6.35) are well defined only when $\omega \neq \omega_0$, and we also expect something particular to happen when the distance between the two frequencies approaches zero, that is, when $\omega \approx \omega_0$. As a matter of fact, there are two important and well–known phenomena that happen when the two frequencies ω_0 and ω are close to each other, and when they are exactly equal.

Beats

We begin by considering the initial conditions

$$x(0) = 0, \qquad \dot{x}(0) = 0$$

for the differential equation (6.29). In this case, the general solution (6.34) can be written as (see Exercise 6.19)

$$x(t) = \frac{2a}{m(\omega_0^2 - \omega^2)} \sin \frac{1}{2}(\omega_0 - \omega)t \, \sin \frac{1}{2}(\omega_0 + \omega)\, t. \tag{6.36}$$

If $\omega \approx \omega_0$, then $(\omega_0 + \omega)$ is very large relative to $|\omega_0 - \omega|$, which means that the function $\sin \frac{1}{2}(\omega_0 + \omega)\, t$ oscillates rapidly, whereas $\sin \frac{1}{2}(\omega_0 - \omega)\, t$ is a function with slower oscillations. Furthermore, we can write (6.36) as

$$x(t) = A(t) \sin \frac{1}{2}(\omega_0 + \omega)\, t,$$

which indicates that we can understand the solution $x(t)$ as a rapid oscillation (with frequency $\frac{1}{2}(\omega_0 + \omega)$), but with a slowly varying amplitude

$$A(t) = \frac{2a}{m(\omega_0^2 - \omega^2)} \sin \frac{1}{2}(\omega_0 - \omega)\, t. \tag{6.37}$$

Tuning instruments I. For an illustration of the phenomenon of beats, think of two sound sources that are vibrating at two different but relatively close frequencies ω and ω_0. Then, what it is apparently heard is a sound with a frequency that is the average of the two given frequencies, that is, $\frac{1}{2}(\omega_0+\omega)$, and whose intensity (amplitude) is not constant but increases and decreases at a frequency proportional to what is known as the **beat frequency** $|\omega_0 - \omega|$. Piano tuners try to listen for beats when comparing the pitch of a tuning fork with that of a vibrating string; when no beats are heard, it means that the fork and string are at the same frequency.

■ **EXAMPLE 6.16**

Consider $m = 0.2$, $a = 65$, $k = 245$, and $\omega = 45$. Then, we have a natural frequency of $\omega_0 = \sqrt{k/m} = 35$, which is relatively close to the external frequency $\omega = 45$. With initial conditions $x(0) = 0$, $\dot{x}(0) = 0$, the slowly varying frequency in (6.37) is

$$A(t) = 2 \sin \frac{5}{2}\, t,$$

and the solution (6.36) takes the form

$$x(t) = 2 \sin \frac{5}{2}\, t \, \sin \frac{65}{2}\, t.$$

The oscillation with frequency $\frac{1}{2}(\omega_0 + \omega)\, t = \frac{65}{2} t$ is "modulated" by the amplitude $A(t) = 2 \sin \frac{5}{2} t$. This phenomenon is what we know as beats. See Figure 6.9.

Tuning instruments II. Going back to the relationship between beats and tuning instruments, assume that the first instrument has a frequency ω and the second a frequency of ω_0; that is, they have different number of cycles or oscillations per second. How many beats will one hear during one second?

Let b be the number of beats one hears in one second; then one hears one beat every $1/b$ seconds; that is, after every $1/b$ second the difference between the oscillations executed by the two tones, is 1. However, in $1/b$ sec, the first instrument tone executes ω/b oscillations, while the second instrument ω_0/b. Therefore,

$$\frac{1}{b}\,(\omega_0 - \omega) = 1,$$

from which $b = \omega_0 - \omega$.

In general, one defines $b = |\omega_0 - \omega|$, to account for the case $\omega > \omega_0$.

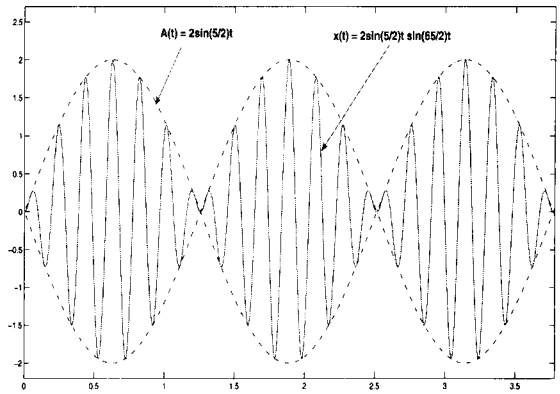


Figure 6.9 Beats solution of Example 6.16.

Remark 6.11 *There is another instance where beats are encountered. It is related to the generation of beats in the telephone receiver of telephony. The border of audibility of tones lies around* $20,000$ *to* $50,000$ *Hertz, while the Hertz number of the modulated oscillations coming from the antenna is in the order of several hundred thousands, say,* $350,000$. *In order to separate the tone arriving in the form of electric vibrations, and to give it an audible Hertz number, the idea is to send simultaneously through the receiver a second oscillation, produced by a special generator, at a relatively close frequency, for example, at* $351,000$ *Hertz. The two oscillations will then interfere and yield 1000 beats per second and thereby give an audible tone (beat reception).*

Resonance

As noted before, it is clear from (6.34) and (6.35) that such solutions are well defined only when the external frequency ω is well away from the natural frequency ω_0. We want to study briefly what occurs when this condition is not fulfilled. First, let us consider the differential equation (6.29) for the case when $(\omega - \omega_0) \to 0$.

As before, we let the particular solution of the differential equation be $x_p(t) = D\cos\omega t$. We compute for $x_p(t)$ its first and second derivatives and substitute into (6.29), and after comparing coefficients, we obtain

$$D = \frac{a}{k - m\omega^2} = \frac{a}{k[1 - (\omega/\omega_0)^2]}. \tag{6.38}$$

Thus, we see that as ω approaches ω_0, the amplitude D of the oscillations approaches infinity. For each value of ω, we have a fixed constant value (with respect to t) of the amplitude D, which will be bigger as ω gets closer to ω_0.

Now let us now consider the case when we exactly have $\omega = \omega_0$. Then, we can take

$$x_p(t) = t(E\cos\omega_0 t + F\sin\omega_0 t),$$

with E and F arbitrary constants, as a particular solution of the differential equation (6.29). By substituting x_p and its derivatives into the differential equation and comparing coefficients, we get

$$x_p(t) = \frac{a}{2m\omega_0}\, t \sin \omega_0 t. \tag{6.39}$$

This time we see a fundamental difference between the constant amplitude in (6.38) and the nonconstant amplitude in the solution (6.39): The presence of the linear term t in the amplitude of the sine oscillation means that as t increases, $x_p(t)$ will also increase, without bound, and the oscillations will linearly increase to infinity. This case $\omega = \omega_0$ is called pure resonance.

■ **EXAMPLE 6.17**

Consider a spring system consisting of an object of mass $m = 2$, a spring constant $k = 1800$, and of external frequency $\omega = \omega_0 = \sqrt{k/m} = 30$. Then, the differential equation (6.29), with $a = 80$, is

$$2\ddot{x} + 1800x = 80 \cos 30t.$$

The particular solution in (6.39) is

$$x_p(t) = D(t) \sin \omega_0 t = \frac{a}{2m\omega_0}\, t \sin \omega_0 t = \frac{2}{3}\, t \sin 30t,$$

which is shown in Figure 6.10.

Remark 6.12 *The general solution in Example 6.17 is of the form*

$$x(t) = A \cos 30t + B \sin 30t + \frac{2}{3}\, t \sin 30t,$$

with A and B arbitrary constants to be determined through initial conditions. If we set $x(0) = 0$, $\dot{x}(0) = 0$, then $x(t) = x_p(t)$. If we set instead, say, $x(0) = 5$, $\dot{x}(0) = 0$, then the resonance still occurs, although it will take longer to be clearly observed.

Tuning fork and hollow cylindrical tube. One of the classic and simplest illustrations of resonance is that of a hollow cylindrical tube partially filled with water that is forced into vibration by a tuning fork. When the tuning fork vibrates at a frequency ω, it creates sound waves that in turn force air inside of the tube to vibrate. If the frequency at which the tuning fork vibrates is not identical to one of the natural frequencies ω_0 of the air column inside the tube, resonance will not occur. However, the location of the water level can be altered by raising and lowering a reservoir of water, thus decreasing or increasing the length of the air column. In this way, the natural frequency of the air in the tube could be matched to the frequency at which the tuning fork vibrates. When the match is achieved, resonance occurs in the form of a big vibration; that is, a loud sound, or a sound with large intensity.

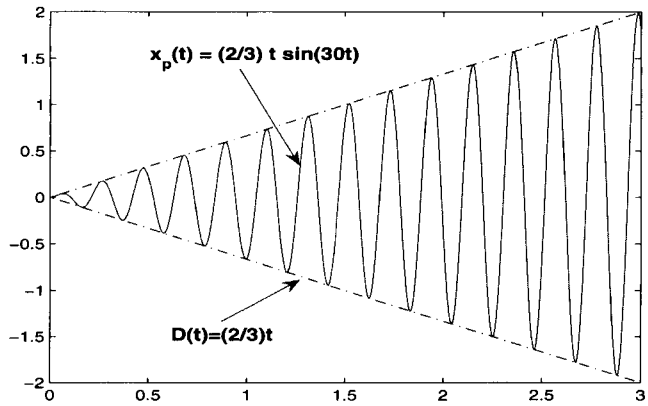


Figure 6.10 Resonance solution of Example 6.17.

6.1.8 Damped equations

In most physical systems, some friction or damping is present. Here we generalize (6.29) to include a damping or friction term. This gives the nonhomogeneous equation

$$m\ddot{x} + c\dot{x} + kx = a\cos\omega t, \tag{6.40}$$

where c is the damping (or friction) constant. Again, the solution of (6.40) is the sum of the solution of the general homogeneous equation and a particular solution of the nonhomogeneous one. That is,

$$x(t) = x_h(t) + x_p(t).$$

We also know that the solution of the homogeneous part can take one of these forms:

$$\begin{array}{rcll} x_h(t) & = & c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}, & \lambda_1,\, \lambda_2 < 0, \\ x_h(t) & = & c_1 e^{\lambda_1 t} + c_2 t\, e^{\lambda_1 t}, & \lambda_1 < 0, \\ x_h(t) & = & e^{at}(c_1 \cos bt + c_2 \sin bt), & a < 0, \end{array} \tag{6.41}$$

depending on whether $c^2 > 4mk$, $c^2 = 4mk$, or $c^2 < 4mk$, respectively. This means that in any case, the solution to the homogeneous part will decay to zero as t increases. This is why $x_h(t)$ is called the **transient solution**.

To find $x_p(t)$, we let

$$x_p(t) = E\cos\omega t + F\sin\omega t, \tag{6.42}$$

and substitute this into (6.40), to get

$$E = \frac{(k - m\omega^2)a}{(k - m\omega^2)^2 + (c\omega)^2}, \qquad F = \frac{c\omega a}{(k - m\omega^2)^2 + (c\omega)^2},$$

and just as we did before, we can combine sines and cosines into just cosine oscillations so that the particular solution (6.42) can be written as

$$x_p(t) = C\cos(\omega t - \alpha), \tag{6.43}$$

with amplitude

$$C = C(\omega) = \sqrt{E^2 + F^2} = \frac{a}{\sqrt{(k - m\omega^2)^2 + (c\omega)^2}}. \tag{6.44}$$

The main observation in (6.44) is that independent of whether $\omega = \omega_0$ or not, the amplitude remains finite in the presence of damping. Compare this with (6.38), where resonance happens when both frequencies are equal.

As remarked before, since $x_h(t) \to 0$ as $t \to \infty$, the particular solution $x_p(t)$ is the one that dominates the motion for large t, and it is called the **steady solution**.

■ **EXAMPLE 6.18**

Consider solving the initial value problem

$$\ddot{x} + 2\dot{x} + 17x = 100\cos 3t, \qquad x(0) = 5, \quad \dot{x}(0) = 0.$$

For the homogeneous part, the roots of the characteristic equation are $\lambda = -1 \pm 4i$, so that

$$x_h(t) = e^{-t}(c_1\cos 4t + c_2\sin 4t).$$

For the particular solution, let $x_p(t) = E\cos 3t + F\sin 3t$, and substitute this into the differential equation to get $E = 8$ and $F = 6$, so that the particular solution is

$$x_p(t) = 8\cos 3t + 6\sin 3t.$$

Hence, the general solution is

$$x(t) = e^{-t}(c_1\cos 4t + c_2\sin 4t) + 8\cos 3t + 6\sin 3t. \tag{6.45}$$

Finally, imposing the initial conditions $x(0) = 5, \dot{x}(0) = 0$ into (6.45), we get $c_1 = -3, c_2 = -7$. See Figure 6.11.

We can also write $x_p(t)$ only in terms of cosines, following (6.43) and (6.44), so that (6.45) can be expressed as (Exercise 6.21):

$$x(t) = e^{-t}(-3\cos 4t - 7\sin 4t) + 10\cos(3t - \alpha),$$

where $\tan\alpha = 6/8$.

The contributions of the transient and steady solutions to the final solution $x(t)$ of the initial value problem are clearly observed in Figure 6.11. As time passes, it is the steady solution

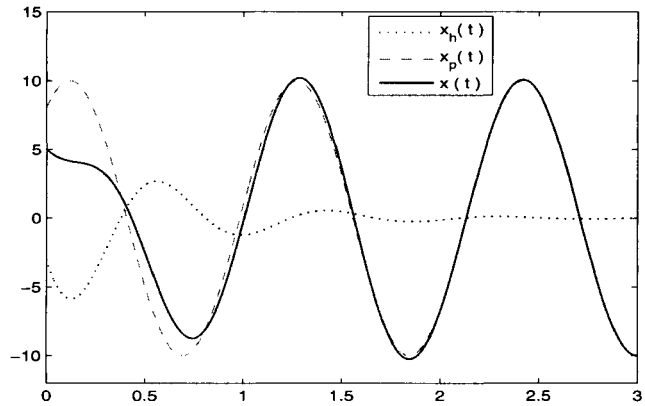


Figure 6.11 Solution $x(t) = x_h(t) + x_p(t)$ of Example 6.18.

$x_p(t)$ that takes over while the transient solution $x_h(t)$ decays to zero exponentially fast, so that

$$x(t) \to x_p(t) \quad \text{as} \quad t \to \infty.$$

Practical Resonance

We have already mentioned that the amplitude $C(\omega)$ in (6.44) is constant with respect to time and does not grow unbounded when $\omega = \omega_0$, so resonance is not possible. However, we can also view the behavior of $C(\omega)$ as a function of the external frequency ω. Since C is a continuous function, it attains maximum or minimum value. In fact, since

$$C'(\omega) = -\frac{aw[c^2 - 2m(k - m\omega^2)]}{[(k - m\omega^2)^2 + (c\omega)^2]^{3/2}},$$

$C'(\omega) = 0$ when $\omega^2 = \frac{k}{m} - \frac{c^2}{2m^2}$. That is,

$$\omega_c = \sqrt{\frac{k}{m} - \frac{c^2}{2m^2}} \tag{6.46}$$

is the critical point of C. This also indicates that $\omega_c < \omega_0$. In addition, it can be proved that if $c < \sqrt{2km}$, then C attains a maximum value at ω_c.

Going back to Example 6.18, let us look at the amplitude C in (6.44), that is, when $a = 100, m = 1, c = 2$, and $k = 17$. The amplitude as a function of ω is

$$C = C(\omega) = \frac{100}{\sqrt{\omega^4 - 30\omega^2 + 289}}. \tag{6.47}$$

The maximum of this function is attained at $\omega_c = \sqrt{15} \approx 3.8730$ (see Figure 6.12), a value very close to the natural frequency of the system $\omega_0 = \sqrt{17} \approx 4.1231$. This is what is known as "practical resonance".

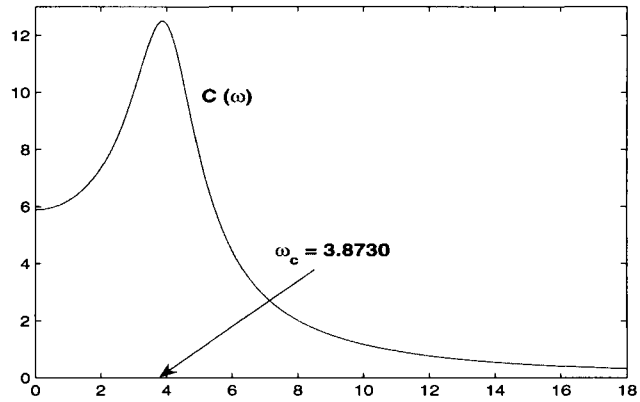


Figure 6.12 Practical resonance in Example 6.18.

A different external force. The external forces in the examples above are only cosine oscillations, and the solutions tend to keep the oscillations for all time t. We can also consider external forces of the form

$$f(t) = ae^{-rt}\cos\omega t,$$

where r is a positive real number.

■ **EXAMPLE 6.19**

Consider the IVP

$$16\ddot{x} + 8\dot{x} + 145x = 960e^{-t/4}\cos 3t, \qquad x(0) = 0, \quad \dot{x}(0) = 0.$$

The roots of the characteristic equation corresponding to the homogeneous differential equation are $\lambda_{1,2} = -\frac{1}{4} \pm 3i$, and therefore,

$$x_h(t) = e^{-t/4}\,(c_1\cos 3t + c_2\sin 3t).$$

This means that the system has the same frequency ω_0 as the external force. For a particular solution, we set then

$$x_p(t) = e^{-t/4}\,(E\,t\cos 3t + F\,t\sin 3t).$$

If we substitute $x_p(t)$ and its derivatives into the given differential equation, we obtain $E = 0$, $F = 10$. In addition, with the initial conditions $x(0) = 0$, $\dot{x}(0) = 0$, we obtain $c_1 = c_2 = 0$, and therefore the solution to the IVP is

$$x(t) = x_p(t) = 10\,e^{-t/4}t\sin 3t.$$

Observe that this is an oscillation with time–varying amplitude of the form $A(t) = 10e^{-t/4}t$, so that the oscillation decays to zero as time increases. See Figure 6.13.

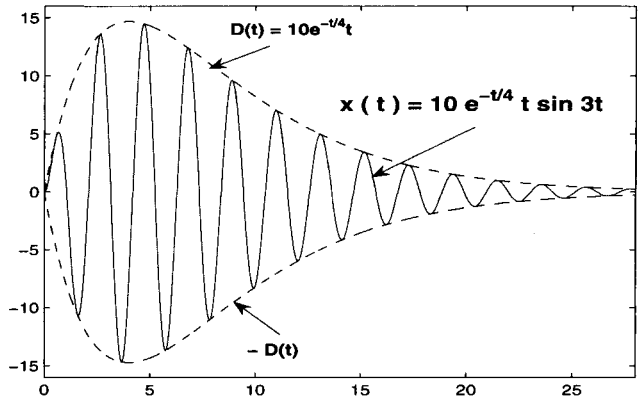


Figure 6.13 Solution of Example 6.19.

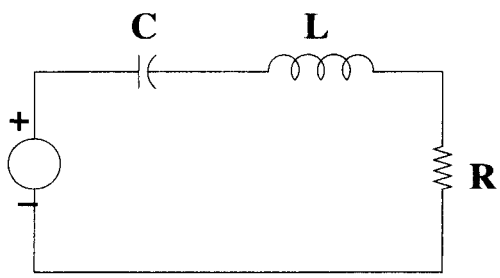


Figure 6.14 RLC circuit.

6.1.9 RLC circuits

It is not only mechanical systems where differential equations find applications; but also several other types of problems are equivalent in mathematical terms; one of them is electrical circuits. We consider here a particular kind of circuit. An RLC circuit is an electrical circuit composed of a resistor (with R ohms of resistance), an inductor (with L henries of inductance), and a capacitor (with C farads of capacitance), connected in series with a voltage source providing $E(t)$ volts at time t. See Figure 6.14. The circuit components are linear (later in Chapter 7 we will consider nonlinear elements in a circuit).

There is a current of $I(t)$ amperes flowing in the circuit and a charge of $Q(t)$ coulombs on the capacitor at time t. The rate of change of $Q(t)$ is equal to the current $I(t)$; that is,

$$\frac{dQ}{dt} = I(t). \tag{6.48}$$

Mechanical System		RLC Circuit	
position	x	charge	Q
mass	m	inductance	L
friction constant	c	resistance	R
spring constant	k	(reciprocal) capacitance	$1/C$
force	F	electromagnetic force	E

Table 6.1 Equivalence between mechanical and electrical systems.

An RLC circuit is a simple example of an electrical harmonic oscillator, and as we will see, it is equivalent to a mechanical system described by the differential equation (6.40). The similarities between the two systems are collected in Table 6.1.

To build up the mathematical model for a mechanical system, we used Newton's and Hooke's laws. Here we need the **Kirchhoff's voltage law:**

The sum of all the voltage drops in a closed circuit is equal to the sum of all voltage sources in that circuit.

The voltage drops across the inductor, resistor, and capacitor are given, respectively, by

$$L\frac{dI}{dt}, \quad RI, \quad \frac{1}{C}Q. \tag{6.49}$$

Then, Kirchhoff's law translates into

$$L\dot{I} + RI + \frac{1}{C}Q = E(t), \tag{6.50}$$

where as usual, $\dot{I}$ denotes dI/dt. By using the relation (6.48), and by taking the derivative with respect to t in (6.50), we can write

$$L\ddot{I} + R\dot{I} + \frac{1}{C}I = f(t), \tag{6.51}$$

where $f(t) = \dot{E}(t)$.

Equation (6.51) has exactly the same form as (6.25), which was used to described mechanical systems. This implies that the results derived for mechanical systems can be applied to electric circuits.

By following our simplification considered in (6.40), here we restrict our study to the case when the voltage is periodic. More precisely, we consider the following differential equation:

$$L\ddot{I} + R\dot{I} + \frac{1}{C}I = a\,\omega\cos\omega t, \tag{6.52}$$

where a is a constant real number.

The solution of (6.52) is again the sum of the general solution of the corresponding homogeneous differential equation and a particular solution to the nonhomogeneous one.

That is,

$$I(t) = I_h(t) + I_p(t).$$

By taking advantage of the similarities between the components of mechanical systems and electrical circuits, and since

$$\frac{\omega}{\sqrt{(\frac{1}{C} - L\omega^2)^2 + (R\omega)^2}} = \frac{1}{\sqrt{(\frac{1}{\omega C} - L\omega)^2 + R^2}},$$

we rewrite the particular solution (6.43), (6.44) of (6.40), with the appropriate changes for the particular solution of (6.52) as

$$x_p(t) = C\ \cos(\omega t - \alpha), \tag{6.53}$$

where $\tan\alpha = \frac{RC\omega}{1-LC\omega^2}$, and the amplitude is given by

$$C = C(\omega) = \frac{a}{\sqrt{(\frac{1}{\omega C} - L\omega)^2 + R^2}}. \tag{6.54}$$

■ **EXAMPLE 6.20**

Consider an RLC circuit with $L = 0.15$, $R = 70$, and $\frac{1}{C} = 1250$, which is connected to a 165V, 25Hz alternating current generator, and assume that $I(0) = \dot{I}(0) = 0$.

First observe that for the frequency, we have that 25 Hz is equal to $25(2\pi)$ radians per second, that is, $\omega = 50\pi$. This implies that $E(t) = 165\sin 50\pi t$ and $f(t) = \dot{E}(t) = 165(50\pi)\cos 50\pi t$. Thus, the differential equation modeling this circuit is

$$0.15\ddot{I} + 70\dot{I} + 1250I = 165(50\pi)\cos 50\pi t,$$

with $I(0) = \dot{I}(0) = 0$.

Since the roots of the corresponding characteristic equation are $\lambda_1 \approx -448.07$ and $\lambda_2 \approx -18.61$, the transient solution is

$$I_h(t) = c_1 e^{-448.07t} + c_2 e^{-18.61t}.$$

Thus, $I_h(t)$ decays to zero exponentially fast, which means that the final solution $I(t)$ is mainly dominated or determined by the particular steady solution $I_p(t)$, which according to (6.53) is given by

$$I_p(t) = C\ \cos(50\pi t - \alpha),$$

with corresponding amplitude (6.54)

$$C = \frac{a}{\sqrt{(\frac{1}{\omega C} - L\omega)^2 + R^2}} \approx 2.3.$$

Therefore, after a very short time t, the current

$$I(t) = I_h(t) + I_p(t)$$

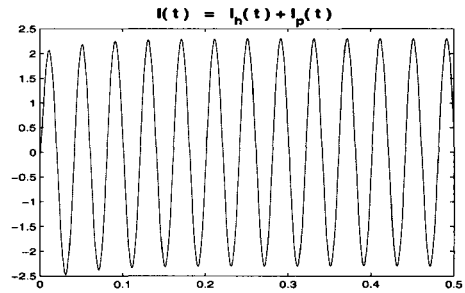


Figure 6.15 Solution of Example 6.20.

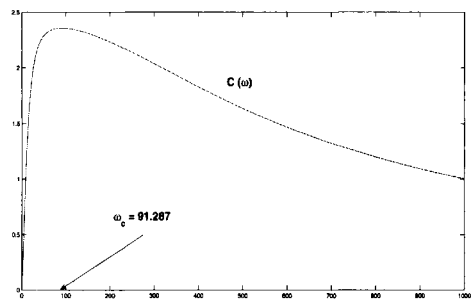


Figure 6.16 Electrical resonance in Example 6.20

will be just cosine oscillations of frequency $\omega = 50\pi$ and with an amplitude of about 2.3. See Figure 6.15.

Electrical Resonance

Following the ideas from mechanical systems, we can also study the behavior of the amplitude corresponding to the steady part of the solution. This amplitude can be written as

$$C = C(\omega) = \frac{a}{Z} = \frac{a}{\sqrt{(L\omega - \frac{1}{\omega C})^2 + R^2}}, \tag{6.55}$$

where Z is called the *impedance* of the circuit. The function $C(\omega)$ reaches its maximum at $\omega_c = \frac{1}{\sqrt{CL}}$. See Figure 6.16, where we have plotted the amplitude function corresponding to Example 6.20. The critical frequency is the resonance frequency of the circuit, and the phenomenon is known as electrical resonance, much as the practical resonance in mechanical systems.

Radio tuning. On a very simplified model, a radio tuning circuit is an RLC circuit, where R, L are fixed constants and C varies as one operates the tuning dial. If we want to pick up a particular radio station broadcasting at frequency ω (providing an input voltage $E(t) = E_0 \sin \omega t$), we want to choose C to maximize I_0 (as a function of C), and tune out stations broadcasting at other frequencies. Thus, we tune the dial to set the capacitance to this value.

6.2 LINEAR SYSTEMS OF DIFFERENTIAL EQUATIONS

One goal in this book is to study general nonlinear n-dimensional differential equations or dynamical systems and analyze in detail their qualitative behavior. In Chapter 7, we consider IVPs of the form

$$\dot{x} = f(t, x), \quad x(t_0) = x_0, \tag{6.56}$$

where $x : R \to R^n$ and $f : R \times R^n \to R^n$. Since several of the theoretical results, and analytical as well as numerical methods, concerning nonlinear systems of the form (6.56) are based on their linear counterpart, in this section we consider the case when f is defined by matrix function $f(t, x) = A(t)x$, and we start with the constant case $A(t) = A$, where A is real matrix.

That is, we consider the following IVP:

$$\dot{x} = A\,x, \quad x(0) = x_0, \tag{6.57}$$

where A is a square matrix of order n, and without loss of generality, we have chosen $t_0 = 0$, which is usually the initial value of time.

It is possible to give the solution of (6.57) explicitly. In fact, recall from Exercise 1.32 that $\frac{d}{dt}\, e^{At} = Ae^{At}$; therefore the solution to system (6.57) is

$$x(t) = e^{At}x_0, \quad t \in R. \tag{6.58}$$

Remark 6.13 *The expression in (6.58) is our first connection between linear algebra and differential equations. But there is more.*

6.3 SOLUTIONS VIA EIGENVALUES AND EIGENVECTORS

We will not always attempt to explicitly compute the matrix exponential to construct the solution (6.58). Instead, since the linear system (6.57) and its solution are completely determined by the constant matrix A, it is not surprising that such a solution can be

written in terms of the eigenvalues and (generalized) eigenvectors of A. We begin with the following theorem.

Theorem 6.14 *Let λ be an eigenvalue of $A_{n\times n}$ with corresponding eigenvector v. Then, the function $x = x(t) = ve^{\lambda t}$ is a solution of (6.57).*

Proof. Since $Av = \lambda v$, we have $Ax = A(ve^{\lambda t}) = \lambda ve^{\lambda t} = \lambda x$. On the other hand, $\dot{x} = \lambda ve^{\lambda t} = \lambda x = Ax$.

□

One can show (Exercise 6.28) that solutions of the type $x(t) = ve^{\lambda t}$, formed with distinct eigenvalues, are linearly independent. But of course, for an arbitrary matrix $A_{n\times n}$, one should not expect always to have distinct eigenvalues. Instead, the goal is to be able to construct n linearly independent solutions whose combination gives the general solution of (6.57), regardless of whether the eigenvalues are distinct or repeated.

In the case of complex eigenvalues, the conjugates also form a solution; that is (see Exercise 6.29), if λ and v are complex eigenvalue and eigenvector respectively of A, then

$$ve^{\lambda t} \quad \text{and} \quad \overline{v}e^{\bar{\lambda} t} \tag{6.59}$$

are solutions of the system $\dot{x} = Ax$. The solutions in (6.59) are complex. However, it is possible to find linearly independent solutions corresponding to complex eigenvalues and eigenvectors using only real expressions (Exercise 6.30).

Whatever kind of eigenvalues we encounter (distinct, complex, nondefective, defective), the goal is to construct n linearly independent solutions of the systems and form with them the general solution (see also Theorem 6.19 ahead).

Theorem 6.15 *If $\{x_1(t), \ldots, x_n(t)\}$ are linearly independent solutions of*

$$\dot{x} = Ax,$$

then the general solution of this system is

$$x(t) = c_1x_1(t) + \cdots + c_nx_n(t). \tag{6.60}$$

for arbitrary constants $c_1, \ldots, c_n$.

As we have already mentioned, the qualitative behavior of the solution to (6.57) is contained in the matrix A. In fact, the linear combination in (6.60) that gives the solution of the system $\dot{x} = Ax$ can be built using the eigenvalues λ_i and the (generalized) eigenvectors v_i of A. We briefly discuss and illustrate the possible cases we may encounter.

■ **EXAMPLE 6.21**

(Distinct Eigenvalues) Consider the IVP

$$\begin{array}{rcl} \dot{x}_1 & = & x_1 - 2x_2 \\ \dot{x}_2 & = & 3x_2. \end{array}, \quad x_0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The eigenvalues and eigenvectors of the coefficient matrix $A = \begin{bmatrix} 1 & -2 \\ 0 & 3 \end{bmatrix}$ have been computed in Example 1.68, namely $\lambda_1 = 3$, $\lambda_2 = 1$, $v_1 = [1 \quad -1]^T$, and $v_2 = [1 \quad 0]^T$. The eigenvectors v_1 and v_2 are linearly independent, and by Theorems 6.14 and 6.15, the general solution is

$$x(t) = c_1 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{3t} + c_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{t}.$$

Now, by using the initial conditon x_0, we get the constants $c_1 = 2, c_2 = -1$. Therefore, the (componentwise) solution to the IVP is

$$x_1(t) = 2e^{3t} - e^{t}, \qquad x_2(t) = -2e^{3t}.$$

- MATLAB commands: ode45, odeset

■ **EXAMPLE 6.22**

(Complex Eigenvalues) Consider the system

$$\begin{array}{rcl} \dot{x}_1 & = & 3x_1 + 5x_2, \\ \dot{x}_2 & = & -x_1 + 5x_2. \end{array}$$

The eigenvalues of the coefficient matrix A are $\lambda = 4 \pm 2i$. An eigenvector associated with $\lambda = 4 + 2i$ is $v = [5 \quad 1 + 2i]^T$. The corresponding complex-valued solution of the system is then

$$x(t) = ve^{\lambda t} = \begin{bmatrix} 5 \\ 1 + 2i \end{bmatrix} e^{(4+2i)t}.$$

A second solution is formed using the complex conjugate eigenvalue and eigenvector, as indicated in (6.59). We can however obtain linearly independent real-valued solutions of the system (refer to Exercise 6.30) as

$$x_1(t) = e^{4t} \left(\begin{bmatrix} 5 \\ 1 \end{bmatrix} \cos 2t + \begin{bmatrix} 0 \\ -2 \end{bmatrix} \sin 2t \right),$$

$$x_2(t) = e^{4t} \left(\begin{bmatrix} 0 \\ -2 \end{bmatrix} \cos 2t - \begin{bmatrix} 5 \\ 1 \end{bmatrix} \sin 2t \right).$$

■ **EXAMPLE 6.23**

(Repeated Eigenvalues) Consider the system

$$\begin{aligned} \dot{x}_1 &= 9x_1 + 4x_2, \\ \dot{x}_2 &= -6x_1 - x_2, \\ \dot{x}_3 &= 6x_1 + 4x_2 + 3x_3. \end{aligned}$$

The eigenvalues of the coefficient matrix are $\lambda_1 = 5$, $\lambda_2 = 3$, $\lambda_3 = 3$. The eigenvector corresponding to λ_1 is $v_1 = [1 \quad -1 \quad 1]^T$. The eigenvectors corresponding to λ_2 are of the general form (see Exercise 6.24) $v = [-\frac{2}{3}x_2 \quad x_2 \quad x_3]^T$. In particular, we take $v_2 = [0 \quad 0 \quad 1]^T$, $v_3 = [2 \quad -3 \quad 0]^T$. The eigenvectors v_1, v_2, and v_3 are linearly independent. Then, by Theorem 6.15, the general solution is

$$x(t) = c_1 \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} e^{5t} + c_2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} e^{3t} + c_3 \begin{bmatrix} 2 \\ -3 \\ 0 \end{bmatrix} e^{3t}.$$

■ **EXAMPLE 6.24**

(Defective Eigenvalue) Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1 - 2x_2 - 2x_3, \\ \dot{x}_2 &= -2x_1 + 2x_2 + 3x_3, \\ \dot{x}_3 &= 2x_1 - 3x_2 - 4x_3. \end{aligned}$$

The eigenvalues of the coefficient matrix are $\lambda_1 = 1, \lambda_2 = -1, \lambda_3 = -1$. Any eigenvector associated with λ_1 has the form $v = [x_1 \quad -x_1 \quad x_1]^T$. In particular, we can take $v_1 = [1 \quad -1 \quad 1]^T$. Any eigenvector associated with $\lambda_2 = \lambda_3$ has the form $v = [0 \quad x_2 \quad -x_2]^T$, so we can choose only one linearly independent eigenvector (there's only one free variable, x_2), say, $v = [0 \quad 1 \quad -1]^T$. We need to compute generalized eigenvector(s). By following the definition in (1.61) with $\lambda = -1$, we solve the system

$$(A - \lambda I)^2 v = 0,$$

which reduces to the equation $x_1 - x_2 - x_3 = 0$. Therefore, any generalized eigenvector is of the form $[x_1 \quad x_2 \quad x_1 - x_2]^T$. In particular, we can take $v_2 = [1 \quad 1 \quad 0]^T$ and $v_3 = [0 \quad 1 \quad -1]^T$. Observe that v_3 can also be obtained using the second expression in (1.61). In fact,

$$v_3 = (A - \lambda I)v_2 = [0 \; 1 \quad -1]^T.$$

The general solution to the system of differential equations is

$$x(t) = c_1 x_1(t) + c_2 x_2(t),$$

where $x_1(t) = v_1 e^t$ and $x_2(t) = (v_2 + v_3 t)e^{-t}$ are two linearly independent solutions. The solution can be rewritten as

$$x(t) = c_1 \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} e^{t} + c_2 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} e^{-t} + c_2 \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} t\, e^{-t}. \tag{6.61}$$

6.3.1 Chains of generalized eigenvectors

Given a coefficient matrix $A_{n\times n}$, one can expect to have some defective eigenvalues, especially for large n, and therefore we must construct some of the linearly independent solutions of the system $\dot{x} = Ax$ by using generalized eigenvectors, as we did in Example 6.24.

Here we briefly introduce the steps to take for computing such linearly independent solutions in the general case. Suppose that the given eigenvalue λ has algebraic multiplicity k, so that we expect to obtain k linearly independent solutions associated with it. Assume also that λ only has r linearly independent ordinary eigenvectors associated with it, with $r < k$. Let then

$$d = k - r.$$

d is called the **defect** of λ.

We start by finding a nonzero vector v that solves the equation

$$(A - \lambda I)^{d+1} v = 0$$

and then successively multiply v by $(A - \lambda I)$ until the zero vector is obtained. More precisely, with $v_k = v$ (see also Exercise 6.31),

$$\begin{array}{rcl} (A-\lambda I)v_k & = & v_{k-1} \neq 0 \\ (A-\lambda I)v_{k-1} & = & v_{k-2} \neq 0 \\ & \vdots & \\ (A-\lambda I)v_2 & = & v_1 \neq 0, \end{array}$$

but

$$(A - \lambda I)v_1 = 0.$$

Thus, the set $\{v_1, \ldots, v_k\}$ forms a chain of generalized eigenvectors, based on the ordinary eigenvector v_1.

The key point is that this chain of length k determines a set of k linearly independent solutions of $\dot{x} = Ax$ corresponding to the defective eigenvalue λ. They are given by

$$\begin{array}{rcl} x_1(t) & = & v_1 e^{\lambda t} \\ x_2(t) & = & (v_1 t + v_2)e^{\lambda t} \\ x_3(t) & = & (\frac{1}{2}v_1 t^2 + v_2 t + v_3)e^{\lambda t} \\ & \vdots & \\ x_k(t) & = & \left(\frac{1}{(k-1)!}v_1 t^{k-1} + \cdots + \frac{1}{2!}v_{k-2}t^2 + v_{k-1}t + v_k\right)e^{\lambda t}. \end{array} \tag{6.62}$$

6.4 FUNDAMENTAL MATRIX SOLUTION

From the previous section, we realize that the main idea in solving the linear system of differential equations $\dot{x} = Ax$ is to be able to write the general solution as a combination of linearly independent solutions. In this context, it is natural to introduce the following definition.

Definition 6.16 *A* **fundamental matrix solution** *of the system $\dot{x} = Ax$ is a nonsingular matrix function $\Phi(t)$ that satisfies the matrix differential equation*

$$\dot{X} = AX. \tag{6.63}$$

This means that the columns of Φ are linearly independent solutions of (6.57), and the solution to the IVP $\dot{x} = Ax$, $x(0) = x_0$ can be expressed as

$$x(t) = \Phi(t)\Phi^{-1}(0)x_0. \tag{6.64}$$

In fact, $\dot{x}(t) = \dot{\Phi}(t)\Phi^{-1}(0)x_0 = A\Phi(t)\Phi^{-1}(0)x_0 = Ax(t)$.

Looking back at (6.58), we observe that $\Phi(t) = e^{At}$ is a fundamental matrix solution, for which $\Phi(0) = I$. Furthermore, any fundamental matrix solution $\Phi(t)$ of $\dot{x} = Ax$ is given by

$$\Phi(t) = Ce^{At},$$

where C is some nonsingular constant matrix.

Remark 6.17 *The Definition 6.16 is similar if $A = A(t)$.*

■ **EXAMPLE 6.25**

Consider the differential equation from Example 6.23. We calculated three independent solution, namely, $\phi_1 = [1\ \ -1\ \ 1]^T e^{5t}$, $\phi_2 = [0\ \ 0\ \ 1]^T e^{3t}$, and $\phi_3 = [2\ \ -3\ \ 0]^T e^{3t}$. Therefore, a fundamental matrix solution is

$$\Phi(t) = \begin{bmatrix} e^{5t} & 0 & 2e^{3t} \\ -e^{5t} & 0 & -3e^{3t} \\ e^{5t} & e^{3t} & 0 \end{bmatrix}.$$

Assuming that $x(0) = [1\ \ 2\ \ -3]^T$ and using (6.64), we can express the solution as

$$x(t) = \begin{bmatrix} e^{5t} & 0 & 2e^{3t} \\ -e^{5t} & 0 & -3e^{3t} \\ e^{5t} & e^{3t} & 0 \end{bmatrix} \begin{bmatrix} 3 & 2 & 0 \\ -3 & -2 & 1 \\ -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix}.$$

By performing the multiplications, we can write the solution in componentwise form as

$$x_1(t) = 7e^{5t} - 6e^{3t}, \quad x_2(t) = -7e^{5t} + 9e^{3t}, \quad x_3(t) = 7e^{5t} - 4e^{3t}.$$

Remark 6.18 *The matrix $\Phi(t)$ in Example 6.25 is not the matrix exponential e^{At}; observe that $\Phi(0) \neq I$. For such a system, the matrix exponential is given by*

$$\Phi(t) = e^{At} = \begin{bmatrix} -2e^{3t} + 3e^{5t} & -2e^{3t} + 2e^{5t} & 0 \\ -3e^{3t} - 3e^{5t} & 3e^{3t} - 2e^{5t} & 0 \\ -3e^{3t} + 3e^{5t} & -2e^{3t} + 2e^{5t} & e^{3t} \end{bmatrix}.$$

However, the solution $x(t)$ of the IVP is of course the same.

Now we give a theorem on the solutions of a system of differential equations, where the matrix A is not necessarily a constant anymore but a function of t.

Theorem 6.19 *Consider the system*

$$\dot{x} = A(t)x. \tag{6.65}$$

Then,

(a) (Superposition) *If $x_1(t), \ldots, x_n(t)$ are solutions of the system (6.65), then the linear combination*

$$x(t) = c_1 x_1(t) + \cdots + c_n x_n(t)$$

is also a solution, for any constants $c_1, \ldots, c_n$.

(b) If $x_1(t), \ldots, x_n(t)$ are linearly independent solutions of the system (6.65) on an open interval I, then any solution of the system can be uniquely written as

$$x(t) = c_1 x_1(t) + \cdots + c_n x_n(t).$$

Proof.

(a)

$$\begin{aligned} \dot{x}(t) &= c_1\dot{x}_1(t) + \cdots + c_n\dot{x}_n(t) = c_1 A(t)x_1(t) + \cdots + c_n A(t)x_n(t) \\ &= A(t)\,(\,c_1 x_1(t) + \cdots + c_n x_n(t)\,) = A(t)x(t). \end{aligned}$$

Therefore, $x(t)$ is a solution of (6.65), for any constants $c_1, \ldots, c_n$.

(b) Let $x(t)$ be any solution of (6.65), and let $t_0 \in I$ be arbitrary, with $x(t_0) = b$, for some $b = [b_1 \quad \cdots \quad b_n]^T$. We want to find unique scalars $c_1, \ldots, c_n$ such that $x(t_0) = c_1 x_1(t_0) + \cdots + c_n x_n(t_0)$. If for any $t \in I$ we denote the vectors

$$x_i(t) = \begin{bmatrix} x_{1i}(t) \\ x_{2i}(t) \\ \cdots \\ x_{ni}(t) \end{bmatrix}, \qquad i = 1, \ldots, n,$$

then the equation $c_1x_1(t_0) + \cdots + c_nx_n(t_0) = x(t_0)$ can be written as

$$\begin{aligned} x_{11}(t_0)c_1 + \cdots + x_{1n}(t_0)c_n &= b_1 \\ x_{21}(t_0)c_1 + \cdots + x_{2n}(t_0)c_n &= b_2 \\ &\vdots \\ x_{n1}(t_0)c_1 + \cdots + x_{nn}(t_0)c_n &= b_n. \end{aligned}$$

This is an algebraic system of linear equations that has a unique solution $[c_1 \quad \cdots \quad c_n]^T$, because the coefficient matrix has linearly independent columns.

□

■ **EXAMPLE 6.26**

Consider the system

$$\begin{aligned} \dot{x}_1 &= -8x_1 - 3x_2, \\ \dot{x}_2 &= 10x_1 + 3x_2. \end{aligned}$$

The functions

$$x(t) = [3e^{-3t} \quad -5e^{-3t}]^T \quad \text{and} \quad y(t) = [e^{2t} \quad -2e^{-2t}]^T$$

are independent solutions of the system. The function

$$z(t) = [9e^{-3t} - \frac{1}{2}e^{-2t} \quad -15e^{-3t} + e^{-2t}]^T$$

is another solution of the system, which can be written as $z(t) = 3x(t) - \frac{1}{2}y(t)$.

6.4.1 Nonhomogeneous systems

We now explain how a fundamental matrix solution of the homogeneous system $\dot{x} = Ax$ can be used to find an explicit expression for the solution of the associated nonhomogeneous system. We have the following theorem.

Theorem 6.20 *If $\Phi(t)$ is a fundamental matrix solution of $\dot{x} = Ax$, and $g(t)$ is a continuous function, then the solution of the IVP*

$$\dot{x} = Ax + f(t), \quad x(0) = x_0 \tag{6.66}$$

is given by

$$x(t) = \Phi(t)\Phi^{-1}(0)x_0 + \Phi(t)\int_0^t \Phi^{-1}(s)f(s)ds. \tag{6.67}$$

Proof. By taking the derivative of (6.67), we have

$$\begin{aligned}
\dot{x} &= \dot{\Phi}(t)\Phi^{-1}(0)x_0 + \Phi(t)\Phi^{-1}(t)f(t) + \dot{\Phi}(t)\textstyle\int_0^t \Phi^{-1}(s)f(s)ds \\
&= A\Phi(t)\Phi^{-1}(0)x_0 + f(t) + A\Phi(t)\textstyle\int_0^t \Phi^{-1}(s)f(s)ds \\
&= A\left(\Phi(t)\Phi^{-1}(0)x_0 + \Phi(t)\textstyle\int_0^t \Phi^{-1}(s)f(s)ds\right) + f(t) \\
&= Ax(t) + f(t).
\end{aligned}$$

□

Remark 6.21 *Recall that if we choose $\Phi(t) = e^{At}$ as the fundamental matrix solution, then $\Phi(0) = I$, and therefore the solution (6.67) takes the form*

$$x(t) = e^{At}x_0 + e^{At}\int_0^t e^{-As}f(s)ds, \tag{6.68}$$

which generalizes (6.58).

■ **EXAMPLE 6.27**

Consider the IVP

$$\begin{aligned}
\dot{x}_1 &= x_1 - 2x_2 + e^{5t} \\
\dot{x}_2 &= 3x_2
\end{aligned}, \qquad x_0 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}.$$

In this case, we have

$$A = \begin{bmatrix} 1 & -2 \\ 0 & 3 \end{bmatrix}, \quad e^{At} = \begin{bmatrix} e^t & e^t - e^{3t} \\ 0 & e^{3t} \end{bmatrix}, \quad f(t) = \begin{bmatrix} e^{5t} \\ 0 \end{bmatrix}.$$

Thus, the solution is

$$\begin{aligned}
x(t) &= e^{At}x_0 + e^{At}\textstyle\int_0^t e^{-As}f(s)ds \\
&= \begin{bmatrix} 4e^t \\ 0 \end{bmatrix} + e^{At}\textstyle\int_0^t \begin{bmatrix} e^{-s} & e^{-s} - e^{-3s} \\ 0 & e^{-3s} \end{bmatrix}\begin{bmatrix} e^{5s} \\ 0 \end{bmatrix} ds \\
&= \begin{bmatrix} 4e^t \\ 0 \end{bmatrix} + e^{At}\textstyle\int_0^t \begin{bmatrix} e^{4s} \\ 0 \end{bmatrix} ds \\
&= \begin{bmatrix} 4e^t + \frac{1}{4}(e^{5t} - e^t) \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{4}(15e^t + e^{5t}) \\ 0 \end{bmatrix}.
\end{aligned}$$

6.5 FINAL REMARKS AND FURTHER READING

We have studied one-dimensional as well as linear systems of ordinary differential equations. For the scalar case, we have established theorems on existence and uniqueness of solutions of initial value problems, and we have introduced some very simple models of first and second order differential equations. The focus was not on techniques of solutions for given classes of differential equations, but on the study of the behavior of solutions for homogeneous and non homogeneous cases. Several results from analysis have been explicitly stated and immediately applied to establish some of the theoretical results on one-dimensional differential equations, keeping the level as simple as possible but without losing rigor.

For the higher dimensional case, we have study only linear systems of differential equations, leaving the nonlinear case for Chapter 7. Again, some results from analysis, this time coupled with some from linear algebra, have proved to be essential to establishing some theoretical results. The solutions have been explicitly given as combinations of linearly independent solutions, which in turn have been formed using eigenvalues and (generalized) eigenvectors. The section of exercises provides an opportunity to discuss further other important theorems on linear systems of differential equations.

We have also taken care of introducing the concept of fundamental matrix solutions, in particular, the one constructed with the matrix exponential function. All this material and discussion provides us with a simple but clear preparation and introduction to the theory and applications of nonlinear systems of differential equations and dynamical systems covered in the next chapter.

There are several textbooks on differential equations at both, the undergraduate and graduate level, that offer excellent presentations of these topics. Worth mentioning are the books by Boyce and Di Prima [9]; Brauer and Nohel [10]; Edwards and Penney [23]; Hale [28]; and Kreider et al. [42].

EXERCISES

6.1 Show that the solution of the differential equation

$$\dot{x} = p(t)x + q(t)$$

is given by

$$x(t) = \left[\int q(t)e^{-\int p(t)dt}\,dt + c\right] e^{\int p(t)dt},$$

where p and q are continuous functions of t, and c is a constant.

6.2 Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ that is Lipschitz and of another function that is not.

6.3 Show that if a function f is differentiable on an interval I and $|f'(x)| \leq L$, for some constant L and for every $x \in I$, then f is Lipschitz in I.

6.4 Show that if a function f is Lipschitz on an interval I, then it is continuous there.

6.5 Let D be any set containing the point $(0,\ 0)$, and consider the function $f(t,x) = t|x|$. Is f Lipschitz in D? Does it have continuous partial derivatives in D?

6.6 Let f be a continuous function on a set $D = \{(t,x) : t \in I, x \in \mathbb{R}\}$, where I is some interval, and let (t_0, x_0) be an arbitrary point in D. Show that $\phi(t)$ is a solution of the initial value problem

$$\dot{x} = f(t,x), \quad x(t_0) = x_0$$

if and only if $\phi(t)$ is a solution of the integral equation

$$x(t) = x_0 + \int_{t_0}^{t} f(s, x(s))ds.$$

6.7 Let f be continuous on $D = \{(t,x) : |t-t_0| \leq a,\ |x-x_0| \leq b\}$, with $|f(t,x)| \leq M$, and define the sequence

$$\phi_0(t) = x_0, \quad \phi_{k+1} = x_0 + \int_{t_0}^{t} f(s, \phi_k(s))ds, \quad k = 0, 1, 2, \ldots.$$

Show that the functions ϕ_k are defined on the interval $|t-t_0| \leq r$, where $r = \min(a,\ b/M)$ and that

$$|\phi_k(t) - x_0| \leq M|t - t_0| \leq b.$$

6.8 Suppose that T is a function or operator so that T^k has a unique fixed point x^* for some $k > 0$. Show that x^* is also the unique fixed point of T.

6.9 In the proof of Theorem 6.9, we used the distance between two continuous function on $[a,\ b]$: $d(f,g) = \max_{t\in[a,b]} |f(t) - g(t)|$. Show that for any three continuous functions f, g, h, this distance satisfies the triangle inequality

$$d(f,h) \leq d(f,g) + d(g,h).$$

6.10 Let $f(x) = \frac{x}{2} + 1$, and define the sequence $x_{k+1} = f(x_k)$ for $k = 0, 1, 2, \ldots$ with $x_0 = 0$. Observe that $x_1 = 1, x_2 = 3/2, x_3 = 7/4, x_4 = 15/8$. Find a general expression for x_k in terms of k, and show that the sequence converges to the fixed point $x^* = 2$ of f.

6.11 Consider the IVP $\dot{x} = \sqrt{x},\ x(0) = 0$, where $x \geq 0$. Show that two solutions of this IVP are $x(t) = 0$, for $t \in (-\infty,\ \infty)$ and also

$$x(t) = \begin{cases} \frac{(t-1)^2}{4}, & t \geq 1 \\ 0, & t \leq 1. \end{cases}$$

Does this contradict Theorem 6.9?

6.12 Consider the IVP $\dot{x} = 2x,\ x(0) = 1$. Establish the corresponding integral equation as in (6.9), and (6.10), and find the first three iterates of the sequence $x_{k+1} = T(x_k)$. Show that $x_k \to e^{2x}$.

6.13 Show that the solution of the IVP $\dot{x} = ax - E,\ x(0) = x_0$ is given by

$$x(t) = e^{at}x_0 - \frac{E}{a}(e^{at} - 1).$$

6.14 Show that a function $x(t) = A\cos w_0 t + B\sin w_0 t$ can be written as

$$x(t) = C\cos(w_0 t - \alpha),$$

where $C = \sqrt{A^2 + B^2}$, $\cos\alpha = A/C$, and $\sin\alpha = B/C$ (C is called the amplitude, w_0 the natural frequency, and α the phase angle).

6.15 Show that the solution of the IVP $\dot{x} = ax - E(t),\quad x(0) = x_0$ is given by

$$x(t) = e^{at}x_0 + e^{at}\int E(t)e^{-at}\,dt.$$

6.16 Show that the solution of the IVP $\dot{x} = ax - bx^2,\quad x(0) = x_0$ is given by

$$x(t) = \frac{a}{b + e^{-at}(a - bx_0)/x_0}.$$

6.17 Show that

$$x_p(t) = \frac{a/m}{\omega_0^2 - \omega^2}\cos\omega t,$$

where $\omega_0 = \sqrt{k/m}$ is a particular solution of $m\ddot{x} + kx = a\cos\omega t$.

6.18 Show that the function

$$x(t) = A\cos\omega_0 t + B\sin\omega_0 t + \frac{a/m}{\omega_0^2 - \omega^2}\cos\omega t,$$

where $\omega \neq \omega_0$ can be written as

$$x(t) = C\cos(\omega_0 t - \alpha) + \frac{a/m}{\omega_0^2 - \omega^2}\cos\omega t,$$

where $C = \sqrt{A^2 + B^2}$ and $\alpha \in [0,\ 2\pi]$.

6.19 Show that the general solution of $m\ddot{x} + kx = a\cos\omega t$ with initial conditions $x(0) = 0$, $\dot{x}(0) = 0$, is given by

$$x(t) = \frac{2a}{m(\omega_0^2 - \omega^2)}\sin\frac{1}{2}(\omega_0 - \omega)t\,\sin\frac{1}{2}(\omega_0 + \omega)\,t.$$

6.20 Solve the IVP

$$2\ddot{x} + 1800x = 80\cos 30t, \qquad x(0) = 5,\ \dot{x}(0) = 0,$$

and plot the solution. Comment on the resonance effect.

6.21 Show that the solution of

$$\ddot{x} + 2\dot{x} + 17x = 100\cos 3t, \qquad x(0) = 5, \quad \dot{x}(0) = 0$$

is given by

$$x(t) = e^{-t}(-3\cos 4t - 7\sin 4t) + 10\cos(3t - \alpha),$$

where $\tan\alpha = 6/8$.

6.22 Let $x_1(t), \ldots, x_n(t)$ be solutions of the differential equation

$$a_n(t)\frac{d^n x}{dt^n} + \cdots + a_1(t)\frac{dx}{dt} + a_0(t)x = 0,$$

where the functions $a_i(t)$, $i = 1, \ldots, n$ are continuous on some interval I and $a_n(t)$ does not vanish in I. Assume that for some $t_0 \in I$, these solutions satisfy

$$\begin{bmatrix} x_1 & \cdots & x_n \\ x_1' & \cdots & x_n' \\ \vdots & & \vdots \\ x_1^{(n-1)} & \cdots & x_n^{(n-1)} \end{bmatrix}(t_0) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

Show that $x_1(t), \ldots, x_n(t)$ are linearly independent on I.

6.23 Define the Wronskian of a set of functions $x_1(t), \ldots, x_n(t)$ as the determinant

$$W(x_1, \ldots, x_n] = \det \begin{bmatrix} x_1(t) & x_2(t) & \cdots & x_n(t) \\ x_1'(t) & x_2'(t) & \cdots & x_n'(t) \\ \vdots & \vdots & & \vdots \\ x_1^{(n-1)}(t) & x_2^{(n-1)}(t) & \cdots & x_n^{(n-1)}(t) \end{bmatrix}.$$

Show that if $W[x_1, \ldots, x_n] \neq 0$ for all t on some interval I, then $x_1(t), \ldots, x_n(t)$ are linearly independent on I.

6.24 Let A be the coefficient matrix of the system

$$\begin{aligned} \dot{x}_1 &= 9x_1 + 4x_2, \\ \dot{x}_2 &= -6x_1 - x_2, \\ \dot{x}_3 &= 6x_1 + 4x_2 + 3x_3. \end{aligned}$$

Show that any eigenvector corresponding to the eigenvalue $\lambda = 3$ of A is of the form $[-\frac{2}{3}a \quad a \quad b]^T$, where $a, b \in \mathbb{R}$.

6.25 Find the matrix exponential e^{At} for $A = \begin{bmatrix} 1 & 9 \\ 1 & 1 \end{bmatrix}$.

6.26 Let $A = \begin{bmatrix} 3 & -2 \\ 2 & 3 \end{bmatrix}$. Show that

$$e^{At} = e^{3t}\begin{bmatrix} \cos 2t & -\sin 2t \\ \sin 2t & \cos 2t \end{bmatrix}.$$

6.27 Are the functions $x(t) = e^{2t}$ and $y(t) = e^{-2t}$ linearly independent on $\mathbb{R}$?

6.28 Let $\lambda_1, \ldots, \lambda_k$ be distinct eigenvalues of $A_{n\times n}$ with corresponding eigenvectors $v_1, \ldots, v_k$. Show that the functions $v_1 e^{\lambda_1 t}, \ldots, v_k e^{\lambda_k t}$ are linearly independent solutions of the system $\dot{x} = Ax$.

6.29 Let λ be a complex eigenvalue of $A_{n\times n}$ with corresponding eigenvector v. Show that $ve^{\lambda t}$ and $\bar{v}e^{\bar{\lambda} t}$ are solutions of the system $\dot{x} = Ax$.

6.30 Let $\lambda = a + bi$ be an eigenvalue of $A_{n\times n}$ with corresponding eigenvector u. Show that the functions

$$x_1(t) = e^{at}(v\cos bt + w\sin bt), \qquad x_2(t) = e^{at}(w\cos bt - v\sin bt),$$

where $v = \frac{1}{2}(u + \overline{u})$ and $w = \frac{i}{2}(u - \overline{u})$, are linearly independent solutions of $\dot{x} = Ax$.

6.31 Let $\{v_1, v_2, v_3\}$ be a "chain" of generalized eigenvectors of a coefficient matrix A, associated with an eigenvalue λ (this defective eigenvalue has algebraic multiplicity 3). That is,

$$(A - \lambda I)v_3 = v_2, \quad (A - \lambda I)v_2 = v_1, \quad (A - \lambda I)v_1 = 0.$$

Show that the following functions:

$$\begin{aligned} x_1(t) &= v_1 e^{\lambda t}, \\ x_2(t) &= (v_1 t + v_2)e^{\lambda t}, \\ x_3(t) &= (\tfrac{1}{2}v_1 t^2 + v_2 t + v_3)e^{\lambda t} \end{aligned}$$

are linearly independent solutions of the system $\dot{x} = Ax$.

6.32 (Distinct eigenvalues) Find the general solution of

$$\begin{aligned} \dot{x}_1 &= 4x_1 + x_3, \\ \dot{x}_2 &= -2x_1 + x_2, \\ \dot{x}_3 &= -2x_1 + x_3. \end{aligned}$$

6.33 (Repeated eigenvalues) Find the general solution of

$$\begin{aligned} \dot{x}_1 &= -6x_3, \\ \dot{x}_2 &= 3x_1 + 6x_2 + 3x_3, \\ \dot{x}_3 &= 3x_1 + 9x_3. \end{aligned}$$

6.34 (Complex eigenvalues) Find the general solution of

$$\begin{aligned} \dot{x}_1 &= x_1 - 5x_2, \\ \dot{x}_2 &= 2x_1 - 5x_2. \end{aligned}$$

6.35 (Defective eigenvalue) Find the general solution of

$$\begin{aligned} \dot{x}_1 &= -6x_1 + 9x_2 - 4x_3, \\ \dot{x}_2 &= -12x_1 + 20x_2 - 9x_3, \\ \dot{x}_3 &= -19x_1 + 33x_2 - 15x_3. \end{aligned}$$

6.36 Consider the system of Example 6.23, with $x(0) = [1 \quad 2 \quad -3]^T$. Show that both fundamental matrices, $\Phi(t) = e^{At}$ and

$$\Phi(t) = \begin{bmatrix} e^{5t} & 0 & 2e^{3t} \\ -e^{5t} & 0 & -3e^{3t} \\ e^{5t} & e^{3t} & 0 \end{bmatrix}$$

give exactly the same solution to the IVP. See Remark 6.18.

6.37 Consider the system

$$\begin{aligned} \dot{x}_1 &= 5x_1 + 3x_2, \\ \dot{x}_2 &= -6x_1 - 4x_2. \end{aligned}$$

Find a fundamental matrix solution $\Phi(t)$ using eigenvalues and eigenvectors, and then find the fundamental matrix solution e^{At} by taking

$$e^{At} = \Phi(t)\Phi^{-1}(0).$$

6.38 Find a 3×3 nonconstant matrix function $\Phi(t)$ such that its determinant is identically zero for all $t \in \mathbb{R}$, and yet its columns are linearly independent.

6.39 Let $X(t)$ be a matrix whose columns satisfy the system $\dot{x} = A(t)x$. Show that either $\det X(t) = 0$ for all t, or $\det X(t) \neq 0$ for all t.

6.40 Let $\Phi(t)$ be a fundamental matrix solution of $\dot{x} = A(t)x$. Show that the rows of $\Phi^{-1}(t)$ satisfy

$$(\dot{y}^T) = -y^T A(t).$$

6.41 Consider the system

$$\begin{aligned} \dot{x}_1 &= 4x_1 + 3x_2, \\ \dot{x}_2 &= 10x_1 + 3x_2. \end{aligned}$$

Show that $x(t) = [2(e^{9t} + \frac{1}{3}e^{-2t}) \quad \frac{1}{3}(10e^{9t} - 4e^{-2t})]^T$ is a solution of the system, and then write it as a linear combination of any two linearly independent solutions.

6.42 Let A be a square matrix of order n. Show that the function $x(t) = e^{At}x_0$ is the unique solution of the IVP $\dot{x} = Ax$, $x(0) = x_0$.

6.43 Using (6.68), find the solution to the IVP

$$\begin{aligned} \dot{x}_1 &= 2x_1 + 3x_2 + e^{5t} \\ \dot{x}_2 &= 2x_1 + x_2 \end{aligned}, \qquad x_0 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}.$$

Hint: The corresponding matrix exponential is

$$\frac{1}{5}\begin{bmatrix} 2e^{-t} + 3e^{4t} & -3e^{-t} + 3e^{4t} \\ -2e^{-t} + 2e^{4t} & 3e^{-t} + 2e^{4t} \end{bmatrix}.$$

6.44 Let $x_1(t), \ldots, x_k(t)$ be solutions of the system $\dot{x} = A(t)x$ on some interval I, and let $t_0 \in I$ be arbitrary. Show that $x_1(t), \ldots, x_k(t)$ are linearly independent if and only if $x_1(t_0), \ldots, x_k(t_0)$ are linearly independent.

6.45 Let $A(t)$ be an $n \times n$ matrix function. Show that the dimension of the solution space of $\dot{x} = A(t)x$ is n.

6.46 Let $A_{n\times n}$ have n linearly independent eigenvectors $v_1, \ldots, v_n$ with corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. Denote each eigenvector v_i as

$$v_i = [v_{1i} \; v_{2i} \quad \cdots \quad v_{ni}]^T, \quad i = 1, \ldots, n.$$

Show that a fundamental matrix solution of $\dot{x} = Ax$ is given by

$$\Phi(t) = \begin{bmatrix} v_{11}e^{\lambda_1 t} & \cdots & v_{1n}e^{\lambda_n t} \\ v_{21}e^{\lambda_1 t} & \cdots & v_{2n}e^{\lambda_n t} \\ \vdots & & \vdots \\ v_{n1}e^{\lambda_1 t} & \cdots & v_{nn}e^{\lambda_n t} \end{bmatrix}.$$

6.47 Let $\Phi(t)$ be a fundamental matrix solution of $\dot{x} = A(t)x$, and let C be any nonsingular constant matrix of the same size as $\Phi(t)$. Show that $\Phi(t)C$ is also a fundamental matrix solution of $\dot{x} = A(t)x$.

6.48 Let v be a rank-r generalized eigenvector of a matrix $A_{n\times n}$, with corresponding eigenvalue λ. Show that the vectors

$$v,\ (A - \lambda I)v,\ \ldots,\ (A - \lambda I)^{r-1}v$$

are linearly independent.

6.49 Let λ be an eigenvalue of a matrix $A_{n\times n}$, and let X be the space of all generalized eigenvectors of A corresponding to λ, together with the zero vector. Let r be the largest integer so that $(A - \lambda I)^r x = 0$ for all $x \in X$ and $(A - \lambda I)^{r-1}v \neq 0$, for some $v \in X$. If Y is the column space of $(A - \lambda I)^r$, show that $\mathbb{R}^n = X \oplus Y$, and that X and Y are invariant under A.

6.50 Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the distinct eigenvalues of a matrix $A_{n\times n}$, with multiplicities $m_1, m_2, \ldots, m_k$, respectively (here $n = m_1 + \cdots + m_k$). Let $X_1, \ldots, X_k$ denote the spaces of generalized eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$, respectively. Show that for $j = 1, \ldots, k$, the spaces X_j have dimension m_j, are invariant under A,

$$\mathbb{R}^n = X_1 \oplus \cdots \oplus X_k,$$

and $(A - \lambda I)^{m_j} x = 0$, for all $x \in X_j$.

6.51 Let A be an $n \times n$ matrix. As in Exercise 6.50, let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of A with multiplicities $m_1, \ldots, m_k$, and $v_1, \ldots, v_k$ corresponding vectors in X_j. Show that for each $j = 1, \ldots, k$, we have

$$e^{At}v_j = e^{\lambda_j t}\left[I + (A - \lambda_j I) + \frac{t^2}{2!}(A - \lambda_j I)^2 + \cdots + \frac{t^{m_j - 1}}{(m_j - 1)!}(A - \lambda_j I)^{m_j - 1}\right]$$

and that $e^{At}v_j \in X_j$.

6.52 With the notation of Exercise 6.51, let A be an $n \times n$ matrix and $x_0 \in \mathbb{R}^n$ be arbitrary. Show that the solution $x(t) = e^{At}x_0$ of the IVP $\dot{x} = Ax,\ x(0) = x_0$ is given by

$$x(t) = \sum_{j=1}^{k} e^{\lambda_j t}\left[\sum_{i=0}^{m_j - 1} \frac{t^i}{i!}(A - \lambda I)^i\right] v_j,$$

where x_0 has been uniquely written as $x_0 = v_1 + \cdots + v_k$ for some vectors in $X_1, \ldots, X_k$.

6.53 Find a fundamental matrix solution for the system

$$\begin{aligned} \dot{x}_1 &= 3x_1 - x_2 + x_3 \\ \dot{x}_2 &= 2x_1 + x_3 \\ \dot{x}_3 &= x_1 - x_2 + 2x_3. \end{aligned}$$

6.54 Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of a matrix $A_{n\times n}$ with multiplicities $m_1, \ldots, m_k$ and $m_1 + \cdots + m_k = n$. If r is a real number such that

$$r > \max_{1\le j\le k} \mathrm{Re}(\lambda_j),$$

show that there exists a constant $C > 0$ such that

$$\|e^{At}\| \le Ce^{rt}, \qquad t \in [0, \infty].$$

6.55 (Variation of Constants Formula) Extend Theorem 6.20 to the following: Let $A(t)$ and $f(t)$ be continuous on some real interval I, and consider the IVP

$$\dot{x} = A(t)x + f(t), \quad x(t_0) = x_0.$$

If $\Phi(t)$ is a fundamental matrix solution of $\dot{x} = A(t)x$, show that the solution to the IVP is

$$x(t) = \Phi(t)\Phi^{-1}(t_0)x_0 + \Phi(t)\int_{t_0}^{t} \Phi^{-1}(s)f(s)ds.$$

6.56 Solve the IVP $\dot{x} = Ax + f(t)$, $x(0) = [3 \quad 2]^T$, where

$$A = \begin{bmatrix} 2 & 1 \\ -1 & 4 \end{bmatrix} \quad \text{and} \quad f(t) = \begin{bmatrix} e^{4t} \\ 1 \end{bmatrix}.$$

Hint: The matrix exponential is $e^{At} = e^{3t}\begin{bmatrix} 1-t & t \\ -t & 1+t \end{bmatrix}$.

6.57 Consider the system $\dot{x} = Ax + f(t)$, where f satisfies

$$\|f(t)\| \le Ce^{at}, \quad t \ge \tau,$$

for some $C > 0$ and $\tau \ge 0$. Show that every solution of the given system satisfies

$$\|x(t)| \le Ke^{bt}, \quad t \ge \tau,$$

for some constants b and $K > 0$.

6.58 Let $A(t)$ and $f(t)$ be continuous on some closed interval I, and consider the IVP

$$\dot{x} = A(t)x + f(t), \quad x(t_0) = x_0,$$

where $t_0 \in I$. Show that the unique solution $x(t)$ of such a system is bounded, in the sense that

$$\|x(t)\| \le C_1 + C_2\int_{t_0}^{t} \|x(s)\|ds,$$

where $C_1 = \|x_0\| + \max_{t \in I} \|f(t)\|(b-a)$, and $C_2 = \max_{t \in I} \|A(t)\|$.

6.59 Let $x_1(t)$ and $x_2(t)$ be any two solutions of the system $\dot{x} = A(t)x + b(t)$. Show that $x(t) = x_1(t) - x_2(t)$ is a solution of $\dot{x} = A(t)x$.

6.60 Let $A(t)$ and $b(t)$ be continuous on a closed interval I, and consider the linear system $\dot{x} = f(x,t) = A(t)x + b(t)$. Show that f is Lipschitz in x, with Lipschitz constant

$$L = \sum_{i,j=1}^{n} \max_{t \in I} |a_{ij}(t)|.$$

CHAPTER 7

DYNAMICAL SYSTEMS

In this chapter, we give an introduction to the very important subject of dynamical systems, the study of qualitative behavior and the computation of solutions of nonlinear systems of the form

$$\dot{x} = f(x), \qquad f : \mathbb{R}^n \to \mathbb{R}^n. \tag{7.1}$$

We present some of the most relevant theoretical results in this area, as well as several examples to illustrate better the idea being introduced. We also take care to introduce some numerical techniques necessary for the computation of special solutions. Some particular mathematical models of dynamical systems are studied in detail, and they give us a chance to highlight and apply the theory and techniques introduced.

In most mathematical theory, dealing with linear problems is by far simpler than studying nonlinear ones. Results on linear dynamical systems are well established, and are essential to understanding and developing the theory for their nonlinear counterpart. Thus, our starting point will be focusing on the simpler theory of linear dynamical systems and the qualitative behavior of their solutions.

The study of nonlinear dynamical systems is usually divided into local and global theory. The main ingredient for local theory will be the linearization of the system around equilibrium points or periodic solutions, and then we will study the resulting linear systems to understand the local behavior of the corresponding nonlinear system around those special

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

solutions. We will also consider some aspects of global theory, by studying some special solutions such as homoclinic and heteroclinic orbits, and we will give an elementary introduction to bifurcations, which represent radical changes in qualitative behavior of solutions for small changes in the problem parameters. Finally, we will give a brief introduction to chaos, characterized by unpredictable behavior of solutions.

7.1 LINEAR DYNAMICAL SYSTEMS

Most problems in real–world applications are nonlinear, and our main focus in this chapter is on nonlinear dynamical systems. But as explained above, the local theory of nonlinear systems can be studied through linearization; therefore, it is not only very illustrative but especially essential to first study and understand linear systems.

We have already seen in Section 6.2 (see (6.57) and (6.58)) that for an $n \times n$ matrix A, the solution to the initial value problem (IVP)

$$\dot{x} = A\,x, \quad x(0) = x_0 \tag{7.2}$$

is given by

$$x(t) = e^{At}x_0, \quad t \in \mathbb{R}, \tag{7.3}$$

where the matrix exponential is

$$e^{At} = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!}\,. \tag{7.4}$$

First, from (7.3), we can deduce that there is no unpredictability in the solution of a linear system, since such a solution is well determined and known for all $t \in \mathbb{R}$, and therefore we do not expect to have chaotic solutions. Second, we observe that the dynamics of the solution is contained in the matrix exponential e^{At}; in particular, the eigenvalues of the matrix A will tell us about the stability properties of the solutions.

In this context, the concept of similarity of matrices becomes particularly important in the study of linear dynamical systems. The following theorem says that if two matrices are similar, so are their corresponding matrix exponentials.

Theorem 7.1 *Let A and B be two $n \times n$ matrices. Then*

$$A = PBP^{-1} \qquad \text{implies} \qquad e^{At} = Pe^{Bt}P^{-1}. \tag{7.5}$$

In particular, if A is similar to a diagonal matrix D, we have that

$$A = PDP^{-1} \qquad \text{implies} \qquad e^{At} = P\begin{bmatrix} e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_n} \end{bmatrix} P^{-1}, \tag{7.6}$$

where the diagonal entries of D : $\lambda_1, \ldots, \lambda_n$, are the eigenvalues of A.

Proof. First observe that if $A = PBP^{-1}$, then $A^k = PB^kP^{-1}$, for any nonnegative integer k. Then

$$\begin{aligned} e^{At} &= \sum_{k=0}^{\infty} \frac{A^k t^k}{k!} = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{A^k t^k}{k!} = \lim_{n\to\infty} \sum_{k=0}^{n} \frac{PB^kP^{-1}t^k}{k!} \\ &= P\left(\lim_{n\to\infty} \sum_{k=0}^{n} \frac{B^k t^k}{k!}\right) P^{-1} = P\left(\sum_{k=0}^{\infty} \frac{B^k t^k}{k!}\right) P^{-1} \\ &= Pe^{Bt}P^{-1}. \end{aligned}$$

If $B = D$ is diagonal, so is e^{Bt}, as in (7.6).

□

Some terminology. Before we start studying the qualitative behavior of solutions, we need to introduce some notation and terminology, with linear systems in mind. Some of this notation and terminology will later be extended to the nonlinear case.

Let E be an open set of $\mathbb{R}^n$. We say that the function $\phi : \mathbb{R} \times E \to E$ defined by

$$\phi(t, x) = e^{At}x \tag{7.7}$$

defines a dynamical system on $E \subseteq \mathbb{R}^n$.

If we relate this definition with the system $\dot{x} = Ax$, and its solution $x(t) = e^{At}x_0$, we observe that the dynamical system defined by $\phi(t, x)$ is a description of how a given point or state $x \in E$ moves with respect to time; that is, how a solution $x = x(t)$ of $\dot{x} = Ax$ evolves with time.

The solution of $\dot{x} = Ax$ defines a motion that can be described geometrically by drawing the solution curves in the $\mathbb{R}^n$ space, which in this context is called the phase space (phase plane, in two dimensions).

The phase portrait of a dynamical system is the set of all solution curves on the phase space.

Clearly, the origin $x = 0$ satisfies $Ax = 0$; that is, $\dot{x} = Ax = 0$, for all $t \in \mathbb{R}$. This means the origin does not change with time and we say that it is an equilibrium point of $\dot{x} = Ax$.

7.1.1 Dynamics in two dimensions

On our way to understanding n-dimensional linear systems, we first study the very illustrative and helpful two-dimensional case. The focus is on classifying the linear dynamical systems according to the eigenvalues of the corresponding matrix A. We know that in two

dimensions the eigenvalues of a matrix will be either real and different, real and repeated, or complex conjugate, and therefore these cases will determine the various qualitative behaviors of the solutions.

Thus, we start by considering the matrices B below and their corresponding matrix exponentials (see also Exercises 7.10 and 7.11):

$$
\begin{aligned}
B &= \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \implies e^{Bt} = \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix}, \\
B &= \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix} \implies e^{Bt} = \begin{bmatrix} e^{\lambda t} & te^{\lambda t} \\ 0 & e^{\lambda t} \end{bmatrix} = e^{\lambda t} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \\
B &= \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \implies e^{Bt} = e^{at} \begin{bmatrix} \cos bt & -\sin bt \\ \sin bt & \cos bt \end{bmatrix}.
\end{aligned}
\tag{7.8}
$$

Remark 7.2 *The important fact is that given an arbitrary 2×2 matrix A, it can be shown that it is similar to one of the matrices B in (7.8). This means that since the solution of (7.2) is given in terms of a matrix exponential, and by recalling Theorem 7.1, we can restrict ourselves to the cases in (7.8) to cover all the possible solutions of (7.2) for $n = 2$.*

In the two-dimensional case, we can take real advantage of graphical illustrations to facilitate the study of a dynamical system. This will hep us understand the behavior of solutions, especially as t increases to infinity, and to classify the origin (equilibrium point of the system) according to the asymptotic behavior of the solutions.

As discussed before, the eigenvalues of the matrix A determine the qualitative behavior of solutions around the origin. In general, the presence of at least one eigenvalue with positive real part will cause the solutions to get farther and farther from the origin as $t \to \infty$. In this case, the origin is called **unstable**. If the eigenvalues are purely imaginary, then the solutions will stay around the origin, but without approaching it as $t \to \infty$. The origin is called **stable**. If all eigenvalues have negative real part, the solutions will approach the origin asymptotically (as $t \to \infty$), and the origin is called **asymptotically stable**.

We want to look at all the possible phase portraits of a two-dimensional dynamical system $\dot{x} = Bx$. That is, by following Remark 7.2, we consider the effect of the matrices B of (7.8) in the solution $x(t) = e^{Bt}x_0$. In the cases below, we are assuming none of the eigenvalues is equal to zero.

Case 1: Eigenvalues with opposite sign.

$B = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, with $\lambda_1 < 0 < \lambda_2$. The origin is unstable, and it is called a **saddle point**.

Case 2: Eigenvalues with equal sign.

$B = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, with $\lambda_1 \leq \lambda_2 < 0$, or $B = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$, with $\lambda < 0$. The origin is called an asymptotically stable **node**.

Case 3: Complex conjugate eigenvalues.

For $\lambda = a \pm b\,\mathrm{i}$, we have $B = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$, $a < 0$. The origin is called an asymptotically stable **focus**.

Case 4: Pure imaginary eigenvalues.

For $\lambda = \pm b\,\mathrm{i}$, we have $B = \begin{bmatrix} 0 & -b \\ b & 0 \end{bmatrix}$. The origin is called a stable **center**.

Remark 7.3

1. *For cases 2 and 3, by considering positive $\lambda_{1,2}$ and $a > 0$, we get the phase portraits for the corresponding* unstable *origin.*
2. *For a general 2×2 matrix A, the phase portrait will be equivalent to one of the four cases above, obtained by a linear transformation of coordinates (similarity transformation).*

We illustrate all these cases in the examples below.

■ **EXAMPLE 7.1**

(Saddle) Consider the system

$$\begin{aligned} \dot{x}_1 &= -x_1 - 3x_2, \\ \dot{x}_2 &= 2x_2, \end{aligned} \qquad x_0 = [c_1 \quad c_2]^T. \tag{7.9}$$

The eigenvalues of $A = \begin{bmatrix} -1 & -3 \\ 0 & 2 \end{bmatrix}$ are $\lambda_1 = -1$, $\lambda_2 = 2$. We inmediately know that this corresponds to case 1, and thus, the origin is an unstable saddle point. The corresponding eigenvectors $v_1 = [1 \quad 0]^T$, $v_2 = [-1 \quad 1]^T$ will help determine the shape and direction of the solution curves in the phase portrait. We keep in mind that v_1 is associated with the stable eigenvalue λ_1 and v_2 is associated with the unstable eigenvalue λ_2. Now let us use a similarity transformation to find the solution $x(t)$. Let $P = [v_1 \quad v_2] = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$. Then, $P^{-1}AP = B = \begin{bmatrix} -1 & 0 \\ 0 & 2 \end{bmatrix}$. Thus, the solution to (7.9) is

$$x(t) = e^{At}x_0 = P\,e^{Bt}P^{-1}x_0 = P \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{2t} \end{bmatrix} P^{-1}x_0, \quad \text{or,}$$

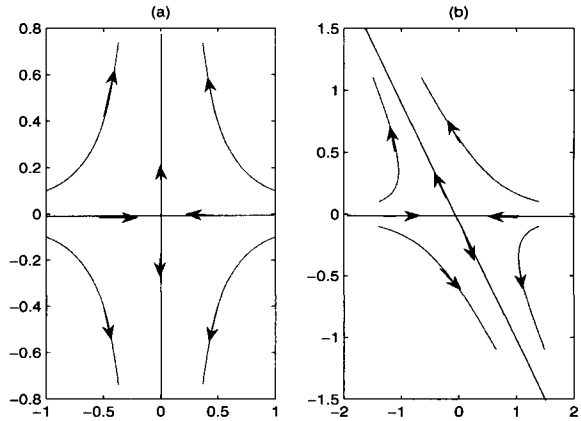


Figure 7.1 **Saddle:** (a) Solutions of (7.10). (b) Solutions of (7.9).

$$\begin{aligned} x_1(t) &= c_1 e^{-t} + c_2(e^{-t} - e^{2t}), \\ x_2(t) &= c_2 e^{2t}, \end{aligned}$$

We also observe that under the transformation $u = P^{-1}x$, the uncoupled system is

$$\begin{aligned} \dot{u}_1 &= -u_1, \\ \dot{u}_2 &= 2u_2, \end{aligned} \qquad u_0 = [k_1 \quad k_2]^T, \tag{7.10}$$

with solution $u_1(t) = k_1 e^{-t}$, $u_2(t) = k_2 e^{2t}$. By the similarity transformation $A = PBP^{-1}$, systems (7.9) and (7.10), and their corresponding solutions, are equivalent. Observe in Figure 7.1 that the phase portraits are equivalent, with the main difference that in (a) the directions are determined by the usual x– and y–axis (or more properly, by the canonical vectors e_1, e_2), while in (b) the directions are determined by the eigenvectors v_1 (which coincides with the x–axis) and v_2.

We should remark that the above example shows that the similarity transformation is in general a change of coordinates: The solutions in (a) are in the usual xy coordinates, while the solutions in (b) are in the coordinates of the eigenvectors v_1 and v_2.

Note: We could have solved system (7.9) by simply solving the second differential equation for x_2, and then substituting this into the first equation to solve for x_1, but here we want to illustrate the role of similarity transformations.

■ EXAMPLE 7.2

(Node) Consider the system

$$\begin{aligned} \dot{x}_1 &= -5x_1 - x_2, \\ \dot{x}_2 &= -x_1 - 5x_2, \end{aligned} \qquad x_0 = [c_1 \quad c_2]^T. \tag{7.11}$$

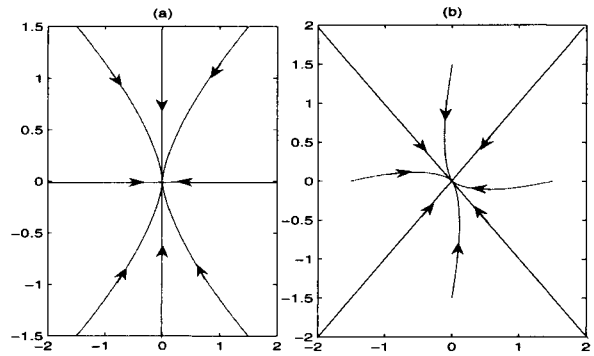


Figure 7.2 **Node:** (a) Solutions of (7.12). (b) Solutions of (7.11).

The eigenvalues of $A = \begin{bmatrix} -5 & -1 \\ -1 & -5 \end{bmatrix}$ are $\lambda_1 = -6$, $\lambda_2 = -4$. This corresponds to case 2, and hence, the origin is an (asymptotically) stable node. The corresponding eigenvectors are $v_1 = [1 \quad 1]^T$, $v_2 = [1 \ -1]^T$. Let us use a similarity transformation to find the solution $x(t)$. Let $P = [v_1 \quad v_2] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Then, $P^{-1}AP = B = \begin{bmatrix} -6 & 0 \\ 0 & -4 \end{bmatrix}$. Thus, the solution to (7.11) is

$$x(t) = e^{At}x_0 = Pe^{Bt}P^{-1}x_0 = P\begin{bmatrix} e^{-6t} & 0 \\ 0 & e^{-4t} \end{bmatrix} P^{-1}x_0, \quad \text{or,}$$

$$\begin{aligned} x_1(t) &= \tfrac{1}{2}(c_1 + c_2)e^{-6t} + \tfrac{1}{2}(c_1 - c_2)e^{-4t}, \\ x_2(t) &= \tfrac{1}{2}(c_1 + c_2)e^{-6t} + \tfrac{1}{2}(c_2 - c_1)e^{-4t}. \end{aligned}$$

As before, under the transformation $u = P^{-1}x$, the uncoupled system is

$$\begin{aligned} \dot{u}_1 &= -6u_1 \\ \dot{u}_2 &= -4u_2, \end{aligned} \qquad u_0 = [k_1 \quad k_2]^T, \tag{7.12}$$

with solution $u_1(t) = k_1e^{-6t}$, $u_2(t) = k_2e^{-4t}$. By similarity, systems (7.11) and (7.12), and their corresponding solutions, are equivalent. Also, as in the previous example, observe in Figure 7.2 that the phase portraits are equivalent.

■ EXAMPLE 7.3

(Focus) Consider the system

$$\begin{aligned} \dot{x}_1 &= -6x_1 - 5x_2 \\ \dot{x}_2 &= 10x_1 + 4x_2, \end{aligned} \qquad x_0 = [c_1 \quad c_2]^T. \tag{7.13}$$

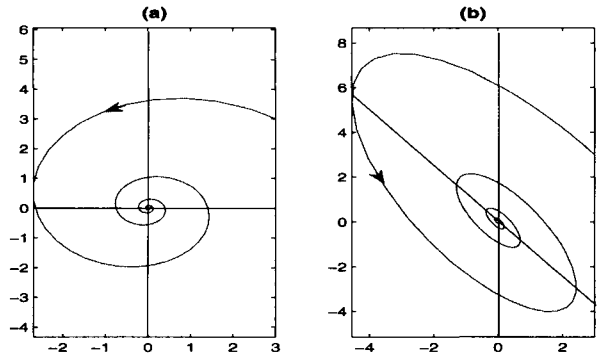


Figure 7.3 **Focus:** (a) Solutions of (7.14). (b) Solutions of (7.13).

The eigenvalues of $A = \begin{bmatrix} -6 & -5 \\ 10 & 4 \end{bmatrix}$ are $\lambda_{1,2} = -1 \pm 5i$. This corresponds to case 3, that is, the origin is an asymptotically stable focus. In fact, the solutions will *spiral* into the origin. The corresponding eigenvectors are $v_{1,2} = [1 \quad -1]^T \pm i\,[0 \quad -1]^T$. To find the solution $x(t)$, let $P = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}$. Then, $P^{-1}AP = B = \begin{bmatrix} -1 & -5 \\ 5 & -1 \end{bmatrix}$. Thus, the solution to (7.13) is

$$x(t) = e^{At}x_0 = P\,e^{Bt}P^{-1}x_0 = P\,e^{-t}\begin{bmatrix} \cos 5t & -\sin 5t \\ \sin 5t & \cos 5t \end{bmatrix} P^{-1}x_0, \quad \text{or,}$$

$$\begin{aligned} x_1(t) &= e^{-t}\,[c_1 \cos 5t - (c_1 + c_2)\sin 5t\,], \\ x_2(t) &= e^{-t}\,[c_2 \cos 5t + (2c_1 + c_2)\sin 5t\,]. \end{aligned}$$

As before, under the transformation $u = P^{-1}x$, the system is

$$\begin{aligned} \dot{u}_1 &= -u_1 - 5u_2, \\ \dot{u}_2 &= 5u_1 - u_2, \end{aligned} \qquad u_0 = [k_1 \quad k_2]^T, \tag{7.14}$$

with solution $u_1(t) = e^{-t}[k_1 \cos 5t - k_2 \sin 5t]$, $u_2(t) = e^{-t}[k_1 \sin 5t + k_2 \cos 5t]$. By similarity, systems (7.13) and (7.14), and their corresponding solutions, are equivalent. Also, as in the previous example, observe in Figure 7.3 that the phase portraits are equivalent. Also observe that although the solutions are given in terms of sines and cosines, the exponential part e^{-t} causes the solutions to spiral toward the origin.

■ EXAMPLE 7.4

(Center) Consider the system

$$\begin{aligned} \dot{x}_1 &= 2x_1 - 2x_2, \\ \dot{x}_2 &= 4x_1 - 2x_2, \end{aligned} \qquad x_0 = [c_1 \quad c_2]^T. \tag{7.15}$$

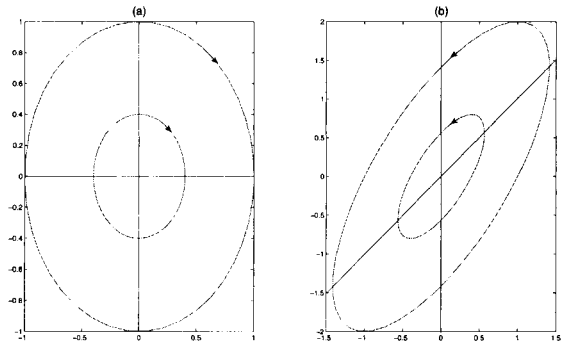


Figure 7.4 **Center:** (a) Solutions of (7.16). (b) Solutions of (7.15).

The eigenvalues of $A = \begin{bmatrix} 2 & -2 \\ 4 & -2 \end{bmatrix}$ are $\lambda_{1,2} = \pm 2\mathrm{i}$. This corresponds to case 4; that is, the origin is a stable center. The corresponding eigenvectors are $v_{1,2} = [1 \quad 1]^T \pm \mathrm{i}\,[0 \quad -1]^T$. Now let $P = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}$. Then, $P^{-1}AP = B = \begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix}$. The solution to (7.15) is

$$x(t) = e^{At}x_0 = P\,e^{Bt}P^{-1}x_0 = P \begin{bmatrix} \cos 2t & \sin 2t \\ -\sin 2t & \cos 2t \end{bmatrix} P^{-1}x_0, \quad \text{or,}$$

$$\begin{aligned} x_1(t) &= c_1 \cos 2t + (c_1 - c_2)\sin 2t, \\ x_2(t) &= c_2 \cos 2t + (2c_1 - c_2)\sin 2t. \end{aligned}$$

Under the transformation $u = P^{-1}x$, the uncoupled system is

$$\begin{aligned} \dot{u}_1 &= 2u_2 \\ \dot{u}_2 &= -2u_1 \end{aligned} \qquad u_0 = [k_1 \quad k_2]^T, \tag{7.16}$$

with solution $u_1(t) = k_1 \cos 2t + k_2 \sin 2t$, $u_2(t) = -k_1 \sin 2t + k_2 \cos 2t$. Systems (7.15) and (7.16), and their corresponding solutions, are equivalent. Also, as in the previous example, observe in Figure 7.4 that the phase portraits are equivalent.

Remark 7.4 *Observe in Figure 7.4 that the orientation of the trajectories has been reversed. This is because det(P) < 0. If we had chosen instead $P = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$, then det(P) > 0, and the orientation would have been preserved. In general if the complex eigenvector is $w = u + iv$, one chooses $P = [v \quad u]$ so that orientation is preserved. See Example 7.3.*

Through the examples above, we have covered the possible dynamical systems in the 2-dimensional case (as long as no eigenvalue is equal to zero). Although the very shapes of

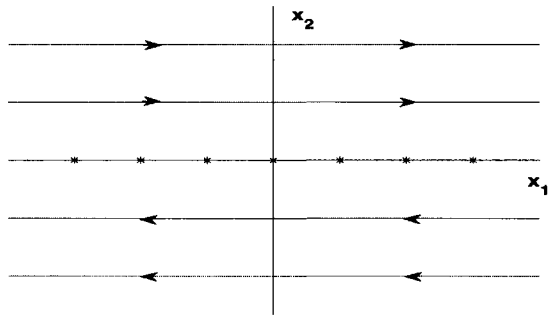


Figure 7.5 Degenerate equilibrium point.

the solution curves may vary a little for similar systems, the qualitative behavior around the origin remains the same. The central point is that the coefficient matrix A, through its eigenvalues and eigenvectors, fully determines the solution. The sign of the real part of the eigenvalues indicates whether the solution is stable or unstable. In the examples discussed, we have considered only stable cases. The unstable cases are identical to the stable cases, but with solutions moving away from the origin, that is, the direction of the arrows would be reversed.

The associated eigenvectors form a new coordinate system (see, e.g., the lines on part (b) of Figures 7.1–7.4) through the similarity transformation $A = PBP^{-1}$, and hence, they help determine the general shape of the solutions. For illustration, in all figures above we have plotted first the solutions of the corresponding canonical form and then the solutions of the general case in the given example.

Degenerate equilibrium point. In the case where $\det A = 0$, that is, when one or both of the eigenvalues of A are zero, we will call the origin a degenerate equilibrium point. In such a case, the origin is not the only equilibrium of the system. Consider for example the system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= 0, \end{aligned} \qquad x(0) = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

The corresponding matrix $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ has eigenvalues $\lambda_1 = \lambda_2 = 0$, and all the points of the form $(x_1,\ 0)$, for any real number x_1, are equilibrium points. We say they are not isolated equilibrium points. The solution to the system is

$$x_1(t) = c_2 t + c_1, \qquad x_2(t) = c_2.$$

Thus, for $c_2 > 0$, we have $x_2 > 0$, and x_1 increases as t increases; similarly, for $c_2 < 0$, we have $x_2 < 0$, and x_1 decreases as t increases. The phase portrait is shown in Figure 7.5. See also Exercise 7.14.

7.1.2 Trace-determinant analysis

Here we introduce a very useful result concerning the stability of the origin of a system $\dot{x} = Ax$, for the case when $\det(A) \neq 0$. This technique for determining the qualitative behavior of solutions around the origin becomes especially useful when we later consider a local study of two-dimensional nonlinear systems around equilibrium points and the system itself depends on some parameters. In most such systems, it becomes difficult to find an explicit expression for the eigenvalues in terms of the problem parameters, so that a good alternative is to find instead explicit expressions for the trace and the determinant of the associated matrix and then to arrive at conclusions about the stability of the origin.

From Theorem 1.59, we know that if $\det(A) \neq 0$, then the only solution of $Ax = 0$ is $x = 0$, that is, the origin is the only equilibrium of the system $\dot{x} = Ax$.

Theorem 7.5 *Let A be a 2×2 matrix, denote $D = \det A$, $T = \mathrm{trace} A$, and consider the system*

$$\dot{x} = Ax. \tag{7.17}$$

Then,

(a) *If $D < 0$, then the origin is a saddle point of* (7.17).

(b) *If $D > 0$ and $T^2 - 4D \geq 0$, then the origin is a stable node of* (7.17) *if $T < 0$ and an unstable node if $T > 0$.*

(c) *If $D > 0$ and $T^2 - 4D < 0$, then the origin is a stable focus of* (7.17) *if $T < 0$ and an unstable focus if $T > 0$.*

(d) *If $D > 0$ and $T = 0$, then the origin is a center of* (7.17).

Proof. Let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$. Then,

$$\det(A - \lambda I) = \lambda^2 - (a + d)\lambda + (ad - bc).$$

Then, the eigenvalues of A are

$$\lambda = \frac{T \pm \sqrt{T^2 - 4D}}{2}.$$

(a) If $D < 0$, then $\sqrt{T^2 - 4D} > T$, and therefore A has two real eigenvalues of opposite sign. This implies the origin is a saddle of (7.17).

(b) If $D > 0$ and $T^2 - 4D \geq 0$, then $\sqrt{T^2 - 4D} < T$, and therefore A has two real eigenvalues of the same sign as T. This implies the origin is a node; stable if $T < 0$ and unstable if $T > 0$.

(c) If $D > 0$ and $T^2 - 4D < 0$, then A has two complex conjugate eigenvalues (as long as $T \neq 0$). This implies the origin is a focus; stable if $T < 0$ and unstable if $T > 0$.

(d) If $D > 0$ and $T = 0$, then A has two pure imaginary complex conjugate eigenvalues. This implies the origin is a center.

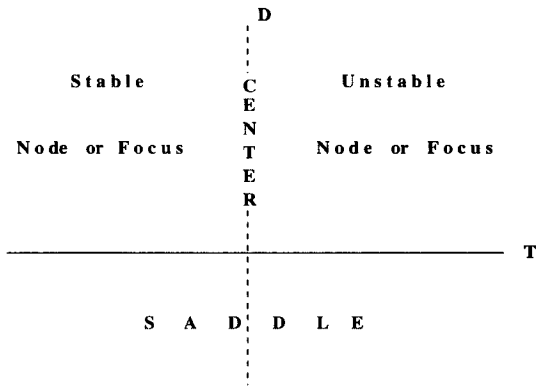


Figure 7.6 (T, D) plane. $T^2 - 4D \geq 0$: Node. $T^2 - 4D < 0$: Focus.

□

We can summarize the results of Theorem 7.5 in a diagram representing regions of different qualitative structure of solutions according to the cases (a)–(d) above. What we obtain is a (T, D) plane where each region is associated with the type of equilibrium point and its stability properties. See Figure 7.6.

■ **EXAMPLE 7.5**

Consider the system

$$\begin{aligned} \dot{x}_1 &= -3x_1 + 2x_2, \\ \dot{x}_2 &= -3x_2. \end{aligned}$$

The determinant of the coefficient matrix $A = \begin{bmatrix} -3 & 2 \\ 0 & -3 \end{bmatrix}$ is $D = 9 > 0$. The trace is $T = -6 < 0$, and therefore $T^2 - 4D = 0$. Thus, we are in case (b) of Theorem 7.5. That is, the origin is a stable node.

Indeed, the eigenvalues of A are $\lambda_1 = \lambda_2 = -3$, which indicates that the origin is a stable node. See Figure 7.7

The real power of Theorem 7.5 is appreciated when the given two-dimensional system depends on one or more parameters so that explicit expressions for the eigenvalues are too complicated to determine their sign. This often happens when the linear system results from a process of linearization applied to a given nonlinear system, as we will see in detail later on. As mentioned before, it may be then more convenient to look at the determinant D and the trace T of the corresponding coefficient matrix A, and decide accordingly about the properties of the origin. See also the equilibrium point analysis in Section 7.3.2.

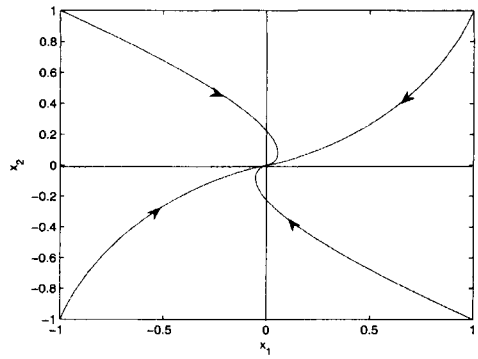


Figure 7.7 Phase portrait of Example 7.5.

■ **EXAMPLE 7.6**

Consider the following system:

$$\begin{aligned} \dot{x} &= (d-h)x + (h-1)y, \\ \dot{y} &= d^2x + (b-d)y, \end{aligned} \tag{7.18}$$

where b, d, and h are certain nonzero parameters. In this case, it is not possible to find a simple explicit expression for the eigenvalues of the coefficient matrix in terms of the given parameters. The trace and the determinant of the coefficient matrix are given by

$$T = b - h, \qquad D = b(d-h) - dh(d-1).$$

Without computing the eigenvalues explicitly we know that the origin will be a saddle point when the determinant is negative, that is, whenever

$$b(d-h) < dh(d-1).$$

Similarly, we can conclude that the origin will be a center when the determinant is positive and the trace is zero, that is, whenever $b(d-h) > dh(d-1)$ and $b = h$. Combining these two conditions, and using the fact that $h \neq 0$, the origin will be a center whenever $d - h > d(d-1)$, or alternatively whenever

$$h < d(d-2).$$

7.1.3 Stable, unstable, and center subspaces

We have already seen that the eigenvectors of the coefficient matrix in a dynamical system $\dot{x} = Ax$ help determine the shape and direction of the solution curves. In fact, something

more fundamental is true: the eigenvectors associated with eigenvalues with negative real part span together an invariant subspace with the property that solutions starting on this subspace will stay on such a subspace for all time t. Furthermore, all solutions on that subspace will approach the origin as t increases. Something similar happens with eigenvectors associated with eigenvalues with positive real part, but in this case, solutions will go away from the origin. We want to formalize these ideas next.

Definition 7.6 *A set $S \subset R^n$ is called* invariant *with respect to a system $\dot{x} = Ax$ if $e^{At}S \subset S$. In other words, for any initial vector $x(t_0) = x_0 \in S$, the solution $x(t)$ of $\dot{x} = Ax$ remains in S for all time $t \geq 0$.*

We can now formally define the stable, unstable, and center subspaces of $\dot{x} = Ax$. These subspaces are invariant under such a system and are a central tool to study the stability of a given dynamical system.

Definition 7.7 *Let $\lambda_j = a_j + i\,b_j$, $(j = 1, \ldots, n)$ be the eigenvalues of $A_{n\times n}$ with (generalized) eigenvectors $u_j = v_j + iw_j$. Then we define*

$$\begin{aligned} E^s &= \text{span}\{\, v_j, w_j : \; a_j < 0 \,\} \quad \text{stable subspace,} \\ E^u &= \text{span}\{\, v_j, w_j : \; a_j > 0 \,\} \quad \text{unstable subspace,} \\ E^c &= \text{span}\{\, v_j, w_j : \; a_j = 0 \,\} \quad \text{center subspace.} \end{aligned} \tag{7.19}$$

The three subspaces are invariant with respect to the system $\dot{x} = Ax$. All solutions in E^s approach the equilibrium $x = 0$ as $t \to \infty$; all solutions in E^u approach the equilibrium $x = 0$ as $t \to -\infty$.

Given an arbitrary system $\dot{x} = Ax$, not all three subspaces in (7.19) are necessarily nonempty (to be more precise, they may contain only the origin), but in any case, it is always true that

$$E^s \oplus E^c \oplus E^u = R^n. \tag{7.20}$$

As usual (see Definition 1.34), the notation $\oplus$ means that any element $x \in R^n$ can uniquely be written as $x = u + v + w$, with $u \in E^s$, $v \in E^c$, and $w \in E^u$. The whole space $\mathbb{R}^n$ is split into these three different subspaces, and the only intersection between them is the origin.

■ **EXAMPLE 7.7**

Let $\dot{x} = Ax$, with $A = \begin{bmatrix} -1 & -3 \\ 0 & 2 \end{bmatrix}$. Then the eigenvalues are $\lambda_1 = -1$, $\lambda_2 = 2$, with corresponding eigenvectors $v_1 = [1 \quad 0]^T$, $v_2 = [-1 \quad 1]^T$. Thus, E^s is the

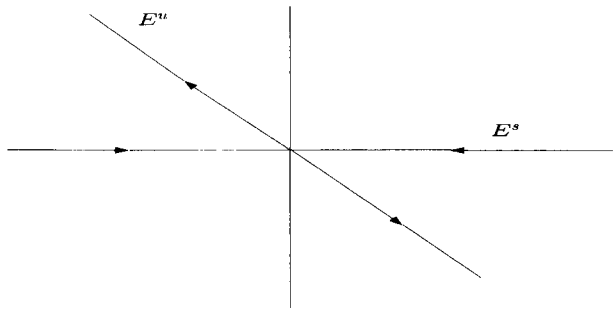


Figure 7.8 Invariant subspaces of Example 7.7.

one-dimensional subspace spanned by v_1; that is, E^s coincides with the x axis. E^u is the one-dimensional subspace spanned by v_2; that is, E^u is the line $y = -x$. There are no eigenvalues with zero real part, and therefore E^c is just the origin. See Figure 7.8.

■ **EXAMPLE 7.8**

Let $\dot{x} = Ax$, with $A = \begin{bmatrix} -2 & -1 & 0 \\ 1 & -2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$. Then the eigenvalues are $\lambda_{1,2} = -2 \pm i$, $\lambda_3 = 3$, with eigenvectors $u_{1,2} = v \pm iw = [0 \quad 1 \quad 0]^T \pm i\,[1 \quad 0 \quad 0]^T$ and $u_3 = [0 \quad 0 \quad 1]^T$. Thus, according to (7.19), E^s is the two-dimensional subspace spanned by v and w; that is, E^s coincides with the xy–plane. E^u is the one-dimensional subspace spanned by u_3; that is, E^u coincides with the z axis. E^c is just the origin. See Figure 7.9.

■ **EXAMPLE 7.9**

Let $\dot{x} = Ax$, with $A = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$. Then the eigenvalues are $\lambda_{1,2} = \pm i$, $\lambda_3 = 2$, with corresponding eigenvectors $u_{1,2} = v \pm iw = [0 \quad 1 \quad 0]^T \pm i\,[1 \quad 0 \quad 0]^T$, and $u_3 = [0 \quad 0 \quad 1]^T$. Thus, E^c is the two-dimensional subspace spanned by v and w; that is, E^c coincides with the xy–plane. E^u is the one-dimensional subspace spanned by u_3; that is, E^u coincides with the z–axis. E^s is just the origin. See Figure 7.9 with E^s replaced by E^c.

Remark 7.8 *The solutions of $\dot{x} = Ax$ on the center subspace E^c do not necessarily stay bounded. See Exercise 7.20.*

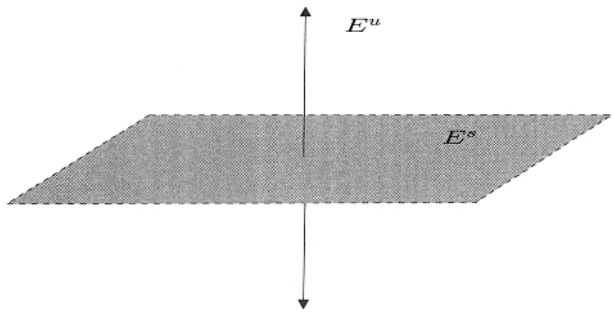


Figure 7.9 Invariant subspaces of Example 7.8.

7.2 NONLINEAR DYNAMICAL SYSTEMS

We now consider systems of the form

$$\dot{x} = f(x), \tag{7.21}$$

where $f : R^n \longrightarrow R^n$ is sufficiently smooth. The first and main difference between nonlinear and linear systems is that in general there is no closed formula for a solution of (7.21) as observed in (7.3) for the linear case. In most cases, it will be too difficult or simply impossible to find an exact or analytical solution of (7.21), and we will need to make use of numerical techniques to approximate a solution. However, without explicitly knowing the solution, it is still possible to study in detail the qualitative behavior of such a solution and decide on questions such as hyperbolicity, stability, sensitivity, asymptotic behavior, invariant sets, and so on.

As mentioned before, the study of nonlinear systems can be divided in two parts: local and global. We start with a local approach through linearization, and then study some aspects of global theory of nonlinear systems.

The main idea behind local theory is to study the nonlinear systems around certain special sets such as equilibrium points and periodic orbits, by means of linearization techniques, and then to arrive at conclusions about the original nonlinear system, but only in a certain vicinity of those special sets. Thus, the strategy is to locate the equilibrium points and periodic orbits of the nonlinear system, apply linearization about these sets, study the resulting linear systems, and finally make conclusions about the qualitative behavior of the nonlinear systems in some neighborhoods of those equilibria and periodic orbits.

One first classic result to introduce is the theorem on existence and uniqueness of solutions, as well as their dependence on the initial conditions. A clear proof of this statement can be found in [54]. The theorem says that if the function f is C^1, then so is the solution $u(t, y)$ in t and y, and for each fixed y, then $u(t, y)$ is C^2. We denote a neighborhood of a point x_0, with radius δ as $N_\delta(x_0)$.

Theorem 7.9 Existence, uniqueness, and dependence on initial conditions. *Let $E \subset \mathbb{R}^n$ be open, $x_0 \in E$, and $f \in C^1(E)$. Then,*

(a) There exist $a > 0$, $\delta > 0$ such that for all $y \in N_\delta(x_0)$, the IVP

$$\begin{cases} \dot{x} &= f(x), \\ x(0) &= x_0 \end{cases}$$

has a unique solution $u(t, y) \in C^1(G)$, where $G = [-a,\ a] \times N_\delta(x_0)$.

(b) For each $y \in N_\delta(x_0)$, $u(t, y) \in C^2([-a,\ a])$.

The proof of the theorem uses the very important Gronwall's inequality (Exercise 7.27) and the fact that if $f \in C^1(E)$, then f is locally Lipschitz on E (Exercise 7.28). The central idea is to show that the *successive approximations*

$$\begin{aligned} u_0(t, y) &= y, \\ u_{k+1}(t, y) &= y + \textstyle\int_0^t f(u_k(s, y))\, ds \end{aligned}$$

converge uniformly to a function $u(t, y)$ that satisfies

$$u(t, y) = y + \int_0^t f(u(s, y))\, ds,$$

so that

$$\dot{u}(t, y) = f(u(t, y)) \quad \text{and} \quad \ddot{u}(t, y) = Df(u(t, y))\, \dot{u}(t, y).$$

Note: We say f is **Lipschitz** on $E \subset \mathbb{R}^n$ if

$$\|f(x) - f(y)\| \leq K\|x - y\|, \tag{7.22}$$

for all $x, y \in E$ and some constant K. And we say f is locally Lipschitz on E, if for *each* point $x_0 \in E$, there is some neighborhood $N_\epsilon(x_0) \subset E$ such that

$$\|f(x) - f(y)\| \leq K\|x - y\|,$$

for all $x, y \in N_\epsilon(x_0)$ and some constant $K = K(x_0)$.

As usual, if $K < 1$ in (7.22), we say f is a *contraction*.

Before we introduce two of the most important theorems in dynamical systems, we give the following definitions.

Definition 7.10 *A point $x_0 \in R^n$ is called an* equilibrium point *of (7.21) if*

$$f(x_0) = 0.$$

This definition implies that at such a point, $\dot{x} = 0$; that is, the system remains static (in equilibrium). In other words, if a solution curve reaches an equilibrium point at time t_0, it will stay there for all time $t > t_0$.

Definition 7.11 *Let $E \subset \mathbb{R}^n$ be open, $x_0 \in E$, $f \in C^1(E)$, and let $\phi(t, x_0)$ denote the solution of the IVP $\dot{x} = f(x)$, $x(0) = x_0$ on some interval I. Then, for $t \in I$, the set of mappings $\phi_t : \mathbb{R}^n \to \mathbb{R}^n$ defined by*

$$\phi_t(x_0) = \phi(t, x_0)$$

is called the flow *of the system $\dot{x} = f(x)$.*

The function ϕ_t describes the motion of points $x_0 \in \mathbb{R}^n$ along trajectories of $\dot{x} = f(x)$.

Definition 7.12 *The linearization of the system (7.21) is defined as*

$$\dot{x} = Ax, \tag{7.23}$$

where $A = Df(x)$, the Jacobian of f.

Definition 7.12 provides us with a powerful tool to study nonlinear dynamical systems locally. First we consider linearization around equilibria.

7.2.1 Linearization around an equilibrium point

For the technique of linearization introduced above, we will need to use a special kind of equilibrium point x_0. Then, the Jacobian will be evaluated at this equilibrium point so that we obtain a linear system $\dot{x} = Ax$, where $A = Df(x_0)$.

Definition 7.13 *Let x_0 be an equilibrium point of (7.21). Then, x_0 is called* hyperbolic *if the Jacobian $A = Df(x_0)$ has no eigenvalues with zero real part.*

This definition implies that the eigenvalues associated with hyperbolic equilibria can be located anywhere in the plane $\mathbb{R}^2$ but the imaginary axis.

■ **EXAMPLE 7.10**

Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1(6 - 2x_1 - x_2) \\ \dot{x}_2 &= x_2(4 - x_1 - x_2). \end{aligned} \tag{7.24}$$

The equilibrium points of (7.24) are $(0,0), (3,0), (0,4), (2,2)$. The Jacobian of f is given by $A(x) = Df(x) = \begin{bmatrix} 6 - 4x_1 & -x_1 \\ -x_2 & 4 - x_1 - 2x_2 \end{bmatrix}$. The four points are hyperbolic; for instance, $A(3,0) = \begin{bmatrix} -6 & -3 \\ 0 & 1 \end{bmatrix}$, so that $\lambda_1 = -6$, $\lambda_2 = 1$.

Remark 7.14 *Observe that in general, given an arbitrary matrix $A_{n\times n}$, we expect only a very small subset of its eigenvalues to "land" on the imaginary axis. That is, by restricting ourselves to study hyperbolic equilibria we are still studying most cases possible.*

Recall that a homeomorphism is a function $h : A \to B$ that is a bijection, is continuous and whose inverse is also continuous.

We need to state two main theorems of dynamical systems that will allow us to use linearization. We start by defining a differentiable manifold of dimension n as a set that is locally homeomorphic to the usual Euclidean space $\mathbb{R}^n$. A differentiable manifold is in fact a topological space that generalizes the intuitive and geometric notion of a curve or a surface. Consider for example the one-dimensional space $\mathbb{R}$ (say, the usual x axis). Then a differentiable manifold homeomorphic to it, is the cubic parabola $y = x^3$: It is a continuous deformation of the x axis.

Recall that for linear systems we have defined the linear vector subspaces E^s, E^c and E^u. As observed in Examples 7.7, 7.8, and 7.9, solutions starting on any of these subspaces at time t_0 will stay there for all time $t > t_0$ (they are invariant). In particular, the solution will approach the equilibrium point if starting on E^s and will move away from it if starting on E^u. The equivalent objects for nonlinear dynamical systems are the so–called stable, center and unstable manifolds W^s, W^c, and W^u, respectively. They are not only invariant and have the same asymptotic behavior as the corresponding linear subspaces but under certain conditions, they are also tangent to them. Without loss of generality, we consider the equilibrium point to be the origin.

Theorem 7.15 (Stable Manifold Theorem) *Let $E \subset \mathbb{R}^n$ be open containing the origin, let $f \in C^1(E)$, and let ϕ_t be the flow of $\dot{x} = f(x)$. Suppose the origin is a hyperbolic equilibrium point and that $A = Df(0)$ has k eigenvalues with negative real part and the remaining $n-k$ eigenvalues have positive real part. Then,*

(a) There exists a k-dimensional differentiable manifold S tangent to E^s at the origin, such that $\phi_t(S) \subset S$, $\forall t \geq 0$ and $\lim_{t\to\infty} \phi_t(x) = 0$, $\forall x \in S$.

(b) There exists an $(n-k)$-dimensional differentiable manifold U tangent to E^u at the origin, such that $\phi_t(U) \subset U$, $\forall t \leq 0$ and $\lim_{t\to-\infty} \phi_t(x) = 0$, $\forall x \in U$.

Sketch of proof. First write the system $\dot{x} = f(x)$ as

$$\dot{x} = Ax + F(x), \tag{7.25}$$

where $A = Df(0)$ and $F(x) = f(x) - Ax$, and consider the similarity transformation

$$B = C^{-1}AC = \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix},$$

for some nonsingular C, where all the eigenvalues $\lambda_1, \ldots, \lambda_k$ of the matrix $P_{k\times k}$ have negative real part, and all the eigenvalues $\lambda_{k+1}, \ldots, \lambda_n$ of the matrix $Q_{(n-k)\times(n-k)}$ have positive real part. Also observe that we can always choose some $\alpha > 0$ sufficiently small so that $\mathrm{Re}(\lambda_j) < -\alpha < 0$, for $j = 1, \ldots, k$.

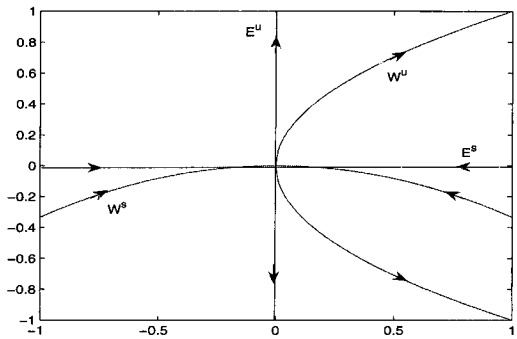


Figure 7.10 Schematic representation of the stable manifold theorem.

By letting $y = C^{-1}x$, the system (7.25) can now be written as

$$\dot{y} = By + G(y), \tag{7.26}$$

where $G(y) = C^{-1}F(Cy)$, so that this new system is split into a stable and an unstable part. Next define the matrix functions

$$U(t) = \begin{bmatrix} e^{Pt} & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad V(t) = \begin{bmatrix} 0 & 0 \\ 0 & e^{Qt} \end{bmatrix},$$

and consider the integral equation

$$u(t,a) = U(t)a + \int_0^t U(t-s)G(u(s,a))ds - \int_t^\infty V(t-s)G(u(s,a))ds. \tag{7.27}$$

The key here is that if $u(t,a)$ is a solution of this integral equation, then it is also a solution of the differential equation (7.26). Thus, we seek to solve the integral equation by successive approximations:

$$\begin{aligned} u^{(0)}(t,a) &= 0, \\ u^{(k+1)}(t,a) &= U(t)a + \textstyle\int_0^t U(t-s)G(u^{(k)}(s,a))ds - \\ &\quad \textstyle\int_t^\infty V(t-s)G(u^{(k)}(s,a))ds. \end{aligned} \tag{7.28}$$

The sequence $\{u^k(t,a)\}$ converges uniformly to a function $u(t,a)$ that satisfies the integral equation, and therefore also the differential equation (7.26), and

$$|u(t,a)| \leq K|a|e^{-\alpha t}, \quad \forall t \geq 0, \tag{7.29}$$

for some constant K. We are focusing our attention on the computation of the stable manifold. Observe from (7.27) that the last $n-k$ components of a do not affect the computation, and thus they are arbitrary. Therefore, we take $a = [a_1 \quad \cdots \quad a_k \; 0 \quad \cdots \quad 0]^T$. For $j = k+1, \ldots, n$, define the real functions

$$\psi_j(a_1, \ldots, a_k) = u_j(0,a). \tag{7.30}$$

Then, the initial values $y_j = u_j(0, a)$ satisfy

$$y_j = \psi_j(y_1, \ldots, y_k), \quad j = k+1, \ldots, n. \tag{7.31}$$

These are the equations that define a manifold $\tilde{S}$ of (7.26), with the property that if $y(0) = u(0, a) \in \tilde{S}$, then $y(t) = u(t, a) \in \tilde{S}$ for all $t \geq 0$, and that $y(t) \to 0$ as $t \to \infty$, according to (7.29). It can be shown that $\frac{\partial \psi_j}{\partial y_i}(0) = 0$, for $i = 1, \ldots, k$ and $j = k+1, \ldots, n$, which implies that $\tilde{S}$ is tangent to the stable subspace E^s of $\dot{y} = By$ at the origin. The stable manifold S is finally obtained from $\tilde{S}$ by changing back to x coordinates via $x = Cy$. The existence of the unstable manifold is established similarly by considering the system

$$\dot{y} = -By - G(y).$$

The stable manifold of this new system will be the unstable manifold of the system (7.26). It is necessary to read the vector y as $[y_{k+1} \cdots y_n \; y_1 \cdots y_k]^T$ in order to determine the $n-k$ manifold $\tilde{U}$. For the full details on this proof, see [54].

□

Remark 7.16 *The manifolds obtained in Theorem 7.15 are local. The global manifolds are obtained by extending the local ones in time. More precisely, the global stable and unstable manifolds, respectively, of $\dot{x} = f(x)$ are defined as*

$$W^s(0) = \bigcup_{t \leq 0} \phi_t(S) \quad \text{and} \quad W^u(0) = \bigcup_{t \geq 0} \phi_t(U).$$

■ **EXAMPLE 7.11**

Consider the system

$$\begin{aligned} \dot{x}_1 &= -x_1, \\ \dot{x}_2 &= x_1^2 - x_2, \\ \dot{x}_3 &= x_2^2 + x_3. \end{aligned}$$

Here, $A = Df(0) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $F(x) = f(x) - Ax = \begin{bmatrix} 0 \\ x_1^2 \\ x_2^2 \end{bmatrix}$. Since the eigenvalues in A are already ordered, then $B = A, C = I$, and $G(x) = F(x)$. We also have

$$U(t) = \begin{bmatrix} e^{-t} & 0 & 0 \\ 0 & e^{-t} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad V(t) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & e^t \end{bmatrix}, \quad a = \begin{bmatrix} a_1 \\ a_2 \\ 0 \end{bmatrix}.$$

The successive approximations (7.28) are
$u^{(0)}(t, a) = [0 \quad 0 \quad 0]^T$,

$$u^{(1)}(t,a) = [e^{-t}a_1 \quad e^{-t}a_2 \quad 0]^T,$$

$$\begin{aligned}
u^{(2)}(t,a) &= \begin{bmatrix} e^{-t}a_1 \\ e^{-t}a_2 \\ 0 \end{bmatrix} + \int_0^t \begin{bmatrix} e^{-(t-s)} & 0 & 0 \\ 0 & e^{-(t-s)} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ e^{-2s}a_1^2 \\ e^{-2s}a_2^2 \end{bmatrix} ds \\
&\quad - \int_t^\infty \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & e^{(t-s)} \end{bmatrix} \begin{bmatrix} 0 \\ e^{-2s}a_1^2 \\ e^{-2s}a_2^2 \end{bmatrix} ds \\
&= \begin{bmatrix} e^{-t}a_1 \\ e^{-t}a_2 \\ 0 \end{bmatrix} + \int_0^t \begin{bmatrix} 0 \\ e^{-(t+s)}a_1^2 \\ 0 \end{bmatrix} ds - \int_t^\infty \begin{bmatrix} 0 \\ 0 \\ e^{t-3s}a_2^2 \end{bmatrix} ds \\
&= \begin{bmatrix} e^{-t}a_1 \\ e^{-t}a_2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -e^{-2t}a_1^2 + e^{-t}a_1^2 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ -\frac{1}{3}e^{-2t}a_2^2 \end{bmatrix} \\
&= \begin{bmatrix} e^{-t}a_1 \\ (a_1^2 + a_2)e^{-t} - a_1^2 e^{-2t} \\ -\frac{1}{3}\, e^{-2t}a_2^2 \end{bmatrix}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
u^{(3)}(t,a) = [\, & e^{-t}a_1 \quad (a_1^2 + a_2)e^{-t} - a_1^2 e^{-2t} \quad -\tfrac{1}{5}\, a_1^4 e^{-4t} + \tfrac{1}{2}a_1^2(a_1^2 + a_2)e^{-3t} \\
& -\tfrac{1}{3}(a_1^2 + a_2)^2 e^{-2t}\,]^T, \\
u^{(4)}(t,a) = {} & u^{(3)}(t,a).
\end{aligned}$$

Since $Df(0)$ has eigenvalues $\lambda = -1, -1, 1$, the stable manifold $\tilde{S} = S$ is given according to (7.31) by one equation in two variables $x_3 = \psi_3(x_1, x_2)$, where (see (7.30))

$$\begin{aligned}
\psi(a_1, a_2) &= u_3(0, a_1, a_2, 0) = -\tfrac{1}{5}\, a_1^4 + \tfrac{1}{2}a_1^2(a_1^2 + a_2) - \tfrac{1}{3}(a_1^2 + a_2)^2 \\
&= -\tfrac{1}{30}a_1^4 - \tfrac{1}{6}a_1^2 a_2 - \tfrac{1}{3}a_2^2.
\end{aligned}$$

That is,

$$S = \{\, [x_1 \quad x_2 \quad x_3]^T \mid \quad x_3 = -\frac{1}{30}x_1^4 - \frac{1}{6}x_1^2 x_2 - \frac{1}{3}x_2^2 \,\}.$$

Remark 7.17 *It should be clear that if the solution of the system $\dot{x} = f(x)$ is explicitly available, then one can try to obtain the stable or unstable manifolds directly, without having to do the successive approximations, by analyzing the asymptotic behavior of the solution. See Exercise 7.30.*

The second main result is one of the most important theorems in the study of nonlinear systems. It is the one that gives the conditions under which a linear system obtained by a linearization around an equilibrium point is locally equivalent to the corresponding nonlinear system. The main condition imposed is hyperbolicity. The theorem states that we

can study a nonlinear system locally by considering the linearized system around hyperbolic equilibria.

Theorem 7.18 (Hartman–Grobman) *Let x_0 be a hyperbolic equilibrium of the nonlinear system (7.21). Then, in a neighborhood of x_0, the system (7.21) and its corresponding linearization*

$$\dot{x} = Ax, \tag{7.32}$$

where $A = Df(x_0)$, are equivalent; that is, there is a homeomorphism h that maps trajectories in (7.21) near x_0 onto trajectories in (7.32).

■ **EXAMPLE 7.12**

Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1, \\ \dot{x}_2 &= x_1^2 - x_2. \end{aligned} \tag{7.33}$$

The origin is the only critical point. The Jacobian is $Df(x) = \begin{bmatrix} 1 & 0 \\ 2x_1 & -1 \end{bmatrix}$, and the linearization of (7.33) is $\dot{x} = Ax$, where $A = Df(0,0) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$, with eigenvalues $\lambda_1 = 1$, $\lambda_2 = -1$, so that $(0,0)$ is hyperbolic. This means we can apply Theorem 7.18 and study the nonlinear system (7.33) around $(0,0)$ through its linearization. The eigenvector associated with the unstable eigenvalue λ_1 is $v_1 = [1 \quad 0]^T$, and the eigenvector associated with the stable eigenvalue $\lambda_2 = -1$ is $v_2 = [0 \quad 1]^T$. This means that the x–axis is the unstable direction, or E^u, and the y-axis is the stable direction, or E^s. It can be proved that in this case the (local) unstable manifold U is given by $y = \frac{1}{3}x^2$ and that the stable manifold S coincides with E^s. This gives a clear picture of how the solutions of (7.33) behave around the saddle point $(0,0)$. See Figure 7.11.

Observe in Example 7.12 that locally, the unstable manifold U is a continuous deformation of the corresponding unstable subspace E^u, and that solutions approach U (and move away from the origin) as $t \to \infty$, even if we start very close to the stable manifold S. In this case, the only way to approach the origin is to start exactly on S. The main point to observe is the local equivalence of both, the nonlinear and the linearized system around the equilibrium.

■ **EXAMPLE 7.13**

Let us study the local behavior of the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -\sin x_1 - cx_2. \end{aligned} \tag{7.34}$$

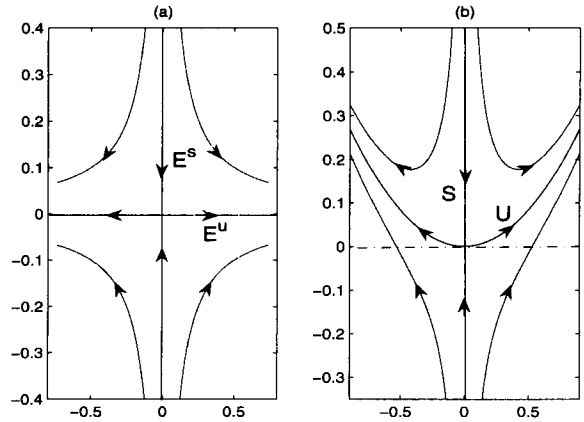


Figure 7.11 Linearized and nonlinear systems of Example 7.12.

There is an infinite number of equilibria: $(0,0), (\pm\pi, 0), (\pm 2\pi, 0)$, and so on. The Jacobian is $Df(x) = \begin{bmatrix} 0 & 1 \\ -\cos x_1 & -c \end{bmatrix}$ and we have

$$Df(0,0) = \begin{bmatrix} 0 & 1 \\ -1 & -c \end{bmatrix} \quad \text{and} \quad Df(\pm k\pi, 0) = \begin{bmatrix} 0 & 1 \\ -1 & -c \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & -c \end{bmatrix},$$

for k even or odd respectively. For $(0,0)$ or $(\pm k\pi, 0)$, with k even, the eigenvalues are $\lambda = (-c \pm \sqrt{c^2 - 4})/2$. For $c = 0$, there will be circles around those equilibria. For $(\pm k\pi, 0)$ with k odd, the eigenvalues are $\lambda = (-c \pm \sqrt{c^2 + 4})/2$. Thus, for arbitrary c, they are (hyperbolic) saddle equilibria. At these hyperbolic equilibria, we can apply Theorem 7.18 and study the nonlinear system (7.34) around $(\pm k\pi, 0)$ through its linearization: The unstable eigenvector (for $c = 0$) is $v_1 = [1 \quad 2]^T$ and the stable eigenvector is $v_2 = [1 \quad -2]^T$. Observe that for $(\pm k\pi, 0)$ with k even, the Jacobian coincides with the Jacobian at $(0,0)$, and there will also be circles around those equilibria. We have now a clear idea about the qualitative behavior solutions of (7.34) around the equilibrium points. See Figure 7.12.

7.2.2 Linearization around a periodic orbit

Just as we linearized a nonlinear system around equilibria, to understand at least locally the qualitative behavior of solutions of nonlinear systems, we can also linearize around periodic orbits. This will allow us to understand the behavior of solutions around such a periodic orbit. We start with the following definition.

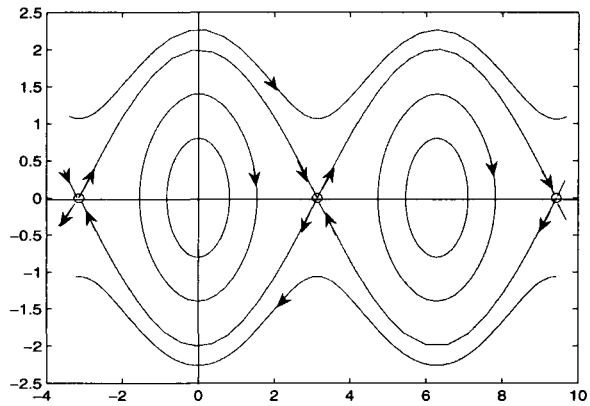


Figure 7.12 Solutions of (7.34)

Definition 7.19 *A solution of the system $\dot{x} = f(x)$ is called a periodic solution of period τ if*

$$x(t) = x(t + \tau), \quad \forall t \in R.$$

Note: Here, τ is in fact the minimum period, because $n\tau$ would also be a period for any $n = 1, 2, \ldots$.

We will call *periodic orbit* to the set of points in the phase space that corresponds to a periodic solution.

We mentioned before that finding equilibria of $\dot{x} = f(x)$ can be accomplished by solving the nonlinear system of algebraic equations $f(x) = 0$, say, by using Newton's method. A periodic orbit of $\dot{x} = f(x)$ of period τ can be computed as solution of the boundary value problem

$$\begin{aligned} \dot{x} &= f(x), \quad t \in [0, \tau], \\ x(0) &= x(\tau). \end{aligned} \tag{7.35}$$

In most cases, solutions to such boundary value problems must be approximated numerically, say, by multiple shooting.

Let $A = Df(x)$ be the Jacobian of f, and consider the associated fundamental matrix solution $\Phi(t)$, as in Definition 6.16, with $\Phi(0) = I$. Assume that the system $\dot{x} = f(x)$ has a periodic solution of period τ. Then, the eigenvalues of the matrix $M = \Phi(\tau)$ are called the *Floquet multipliers* of the periodic orbit. The matrix M is known as the *monodromy* matrix. Just as the eigenvalues of the Jacobian at an equilibrium point determine its stability properties and the asymptotic behavior of solutions around it, the eigenvalues of the monodromy matrix determine the stability properties of a periodic orbit and the asymptotic behavior solutions around it. Again, the central condition for linearization will be hyperbolicity.

Definition 7.20 *A periodic orbit is called* hyperbolic *if it has exactly one Floquet multiplier of magnitude 1.*

Remark 7.21 *For any given periodic orbit, one of the Floquet multipliers is always 1. Hyperbolicity guarantees that no other multiplier lies on the unit circle. On the other hand, multipliers are in general of the form $\mu = e^{\lambda\tau}$, where τ is the period, and λ, in general complex, is called the* Floquet exponent *and is not unique: It is only determined modulo $2\pi i$. However, they uniquely determine the magnitude of the Floquet multipliers, which ultimately determines the stability of the periodic orbits: Multipliers with magnitude less than 1 are called stable, and those with magnitude greater than 1 are unstable.*

■ **EXAMPLE 7.14**

The planar system

$$\begin{aligned} \dot{x} &= x(1 - x^2 - y^2) - y, \\ \dot{y} &= y(1 - x^2 - y^2) + x \end{aligned} \tag{7.36}$$

clearly has a solution of the form $(x(t), y(t)) = (\cos t, \sin t)$, which is a periodic orbit of period $\tau = 2\pi$. The details on Floquet multipliers and stability are given in Example 7.15.

The main question is whether we can apply a result like the Hartman–Grobman theorem for periodic orbits. The answer is yes, as long as the periodic orbit is hyperbolic. In other words, if the system $\dot{x} = f(x)$ is linearized around a hyperbolic periodic orbit, then locally, the nonlinear system is qualitatively equivalent to the linearized one. This result allows us to study a nonlinear system around a hyperbolic periodic orbit by studying the corresponding linearized system. Furthermore, just as we have the stable manifold theorem for equilibria, we also have one available for periodic orbits, with corresponding stable, center and unstable manifolds W^s, W^c, W^u, respectively, and with the same properties as for equilibria: They are invariant, and solutions starting on W^s will approach the periodic orbit as $t \to \infty$, solutions starting on W^c will stay around the periodic orbit for all time t, and solutions starting on W^u will move away from the periodic orbit as $t \to \infty$.

Theorem 7.22 Stable Manifold Theorem for Periodic Orbits *Let $E \subset \mathbb{R}^n$ be open containing a periodic orbit Γ, let $f \in C^1(E)$, and let ϕ_t be the flow of $\dot{x} = f(x)$. Suppose the periodic orbit is hyperbolic, and has k Floquet multipliers with magnitude strictly less than one and $n - k - 1$ with magnitude strictly greater than one. Then,*

(a) There exists a $(k+1)$-dimensional differentiable manifold S such that $\phi_t(S) \subset S$, $\forall t \geq 0$ and $\lim_{t\to\infty} \phi_t(x) = \Gamma$, $\forall x \in S$.

(b) There exists an $(n - k)$-dimensional differentiable manifold U such that $\phi_t(U) \subset U$, $\forall t \leq 0$ and $\lim_{t\to-\infty} \phi_t(x) = \Gamma$, $\forall x \in U$.

Remark 7.23 *If we denote the Floquet exponents with $\lambda_j = a_j + ib_j$, then one can define the stable, center, and unstable subspaces of the corresponding periodic orbit as in* (7.19), *using the associated generalized eigenvector of the monodromy matrix. These subspaces are tangent to the corresponding manifolds defined in Theorem 7.22.*

■ **EXAMPLE 7.15**

Consider the system

$$\begin{aligned} \dot{x} &= x(1 - x^2 - y^2) - y, \\ \dot{y} &= y(1 - x^2 - y^2) + x, \\ \dot{z} &= 3z. \end{aligned} \tag{7.37}$$

This system has a periodic orbit $\gamma(t) := (\cos t, \sin t, 0)$, and we want to linearize the system around γ. First, we need to compute the Jacobian:

$$Df(x, y, z) = \begin{bmatrix} 1 - 3x^2 - y^2 & -1 - 2xy & 0 \\ 1 - 2xy & 1 - x^2 - 3y^2 & 0 \\ 0 & 0 & 3 \end{bmatrix}. \tag{7.38}$$

By evaluating this at γ, we get

$$A(t) = Df(\gamma(t)) = \begin{bmatrix} -2\cos^2 t & -1 - \sin 2t & 0 \\ 1 - \sin 2t & -2\sin^2 t & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

Then, the linearization of (7.37) is

$$\dot{x} = A(t)\, x, \tag{7.39}$$

with fundamental matrix solution

$$\Phi(t) = \begin{bmatrix} e^{-2t}\cos t & -\sin t & 0 \\ e^{-2t}\sin t & \cos t & 0 \\ 0 & 0 & e^{3t} \end{bmatrix}.$$

The Floquet multipliers of γ are computed as the eigenvalues of the monodromy matrix

$$M = \Phi(2\pi) = \begin{bmatrix} e^{-4\pi} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & e^{6\pi} \end{bmatrix},$$

which are $\mu_1 = 1$, $\mu_2 = e^{-4\pi}$, and $\mu_3 = e^{6\pi}$. Then, γ is hyperbolic, because it has exactly one multiplier of magnitude one (it also has one stable multiplier, μ_2, and one unstable multiplier, μ_3). From Theorem 7.22, there is a two-dimensional stable manifold S, in this case coinciding with the xy plane (excluding the origin), and a two-dimensional unstable manifold W^u, which is a unit cylinder: Solutions spiral on the walls of the cylinder and away from the periodic orbit. See Figure 7.13.

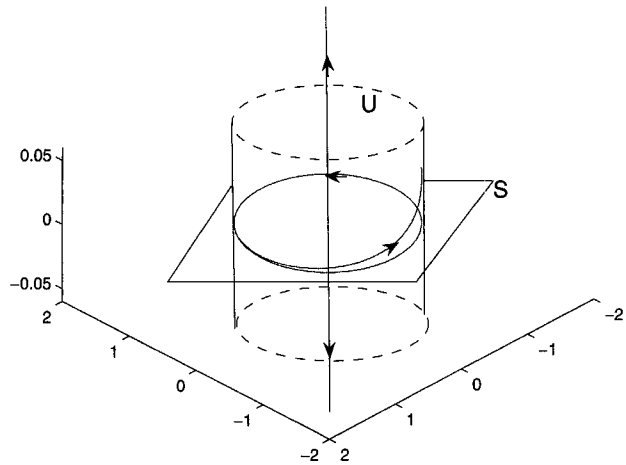


Figure 7.13 Solutions and invariant subspaces of (7.37).

We have introduced some ideas regarding the local analysis of nonlinear systems around hyperbolic equilibria and hyperbolic periodic orbits. When hyperbolicity is violated, study of nonlinear systems becomes more complex, and more advanced mathematical tools are needed, such as center manifolds, which is beyond the scope of this book. The reader is referred to books on dynamical systems such as [15], [17] and [27]. Here we want to introduce two special sets of solutions that illustrate some global behavior of nonlinear systems: connecting orbits and chaotic solutions.

7.2.3 Connecting orbits

In Example 7.13, we have already seen some global solutions of a nonlinear system. Those solutions connecting $(-\pi, 0)$ to $(\pi, 0)$,those connecting $(\pi, 0)$ to $(3\pi, 0)$, and so on, are very special solutions commonly known as heteroclinic orbits. In the particular case where a solution connects back to the same equilibrium point, it is called a homoclinic orbit. Homoclinic and heteroclinic are also known as connecting orbits. They not only connect equilibrium points, but they can also connect an equilibrium to a periodic orbit and two periodic orbits as well. This type of solutions are very important in the study of dynamical systems as they form what is known as separatrices: Solutions that serve as border curves between several solutions of the system. Observe in Figure 7.12 how the heteroclinic connections between $(-\pi, 0)$, $(\pi, 0)$, $(3\pi, 0)$, and so on, have closed solutions curves inside them, and other types of curves outside them, so that when certain separatrices are found, we can have a good understanding of the qualitative behavior of solutions, and we are not necessarily restricted to only a local setting anymore.

There is an increasing interest on connecting orbits mainly as a result of several important applications. For example, in [13] connecting orbits are identified in a water-wave model, in [40], and [41] the authors find homoclinic and heteroclinic connections in a model of celestial mechanics and they show how this may lead to space exploration with prescribed itineraries, and in [55] traveling waves from several applications are computed using connecting orbits. We want to give some basic ideas of this special type of solutions. For a detailed treatment of connecting orbits and applications, see e.g., [7], [19] and [20].

We need to start by considering a more general form of a nonlinear system, one that includes the presence of one or more real parameters in the vector field. Namely, we consider the parameter-dependent nonlinear systems

$$\dot{x} = f(x, \lambda), \qquad x(t) \in R^n, \quad \lambda \in R^p, \tag{7.40}$$

where $f : R^n \times R^p \longrightarrow R^n$ is assumed to be sufficiently smooth. Let $M_-(\lambda)$ be either a hyperbolic equilibrium $y_-(\lambda)$ or a hyperbolic periodic orbit $\gamma_-(\lambda)$ of (7.40), and let $M_+(\lambda)$ be a hyperbolic periodic orbit $\gamma_+(\lambda)$ of (7.40). We will use the notation $y_+(t, \lambda)$ to denote the periodic solution of period τ_+ corresponding to $\gamma_+(\lambda)$, and similarly $y_-(t, \lambda)$ will be the periodic solution of period τ_- relative to $\gamma_-(\lambda)$, if $M_-(\lambda) = \gamma_-(\lambda)$. As noted before, equilibrium points can be found by solving the system $f(x, \lambda) = 0$, and we will assume them to be known. However, the periodic orbits will need to be computed as part of the problem.

A solution $x(t, \lambda)$, $t \in R$ of (7.40) is called a *connecting orbit* from $M_-(\lambda)$ to $M_+(\lambda)$ if

$$\text{dist}\,(x(t, \lambda),\ M_\pm(\lambda)) \longrightarrow 0 \quad \text{as} \quad t \longrightarrow \pm\infty. \tag{7.41}$$

First, notice that now we are dealing with asymptotic solutions, those that are defined as t increases to infinity. Second, since the system (7.40) is autonomous (the vector field f does not explicitly depend on t), if $x(t, \lambda)$ is a solution, then also $x(t+\sigma, \lambda)$ is a solution, for any $\sigma \in R$. Then, an extra condition is necessary for the solution to be uniquely determined. Then, we impose a so–called *phase condition*

$$\psi(x, y_-, y_+, \lambda) = 0.$$

The central idea in all methods to compute connecting orbits is to truncate the real line to a finite and sufficiently large interval $[T_-,\ T_+]$, $T_- < 0 < T_+$, and to impose boundary conditions at $T_\pm$. A key observation is that the connecting orbit must leave $M_-(\lambda)$ along its unstable manifold and enter $M_+(\lambda)$ along its stable manifold. We know that these manifolds are tangent to the unstable subspace $E^u_-(\lambda)$ of $M_-(\lambda)$ and to the stable subspace $E^s_+(\lambda)$ of $M_+(\lambda)$, respectively.

By using the so-called *projection boundary conditions* [7], [19], the problem of computing the connecting orbit is transformed into the problem of solving the following boundary value problem:

$$\begin{aligned} &\dot{x} = f(x, \lambda), \quad T_- \le t \le T_+, \\ &L_-(\lambda)(x(T_-) - y_-(0)) = 0, \quad L_+(\lambda)(x(T_+) - y_+(0)) = 0, \\ &\psi(x, y_-, y_+, \lambda) = 0\,, \end{aligned} \tag{7.42}$$

where L_- and L_+ are smooth functions of λ, and span $E_-^{cs}(\lambda)$ and $E_+^{cu}(\lambda)$, the center-stable and center-unstable subspaces of $M_-(\lambda)$ and $M_+(\lambda)$, respectively. The vector $x(T_-) - y_-(0)$ lies on the unstable subspace E^u of $M_-(\lambda)$, and similarly $x(T_+) - y_+(0)$ lies on the stable subspace E^s of $M_+(\lambda)$. Thus, when they are multiplied by $L_-(\lambda)$ and $L_+(\lambda)$, respectively, their products in (7.42) vanish, by orthogonality. Also, in 7.42, ψ corresponds to the truncated version of the phase condition.

To find the matrix functions L_- and L_+ in (7.42), we make use of a very important matrix factorization introduced in Chapter 3: Schur factorization. In the case of an equilibrium, we perform the Schur factorization of the Jacobian evaluated at that point, and in the case of a periodic orbit, we compute the Schur factorization of the monodromy matrix. The starting point is the ordered block Schur factorization (3.39), but this time our matrix A depends on a parameter λ, and we want to perform a *smooth* ordered block Schur factorization of $A(\lambda)$, with factors as smooth as the matrix $A(\lambda)$. See [18].

Thus, assume $A(\lambda)$ is the $n \times n$ matrix function representing the Jacobian (or the monodromy matrix), at a hyperbolic equilibrium (or periodic orbit), and we are interested in finding $L_-(\lambda)$. Assume further that $A(\lambda)$ has n^s eigenvalues with negative real part (or $n^s + 1$ eigenvalues with magnitude less than or equal to one). Then, the (ordered block) Schur factorization gives

$$Q^T(\lambda)A(\lambda)\,Q(\lambda) \;=\; R(\lambda) = \begin{bmatrix} R_{11}(\lambda) & R_{12}(\lambda) \\ 0 & R_{22}(\lambda) \end{bmatrix}, \tag{7.43}$$

where $R_{11}(\lambda)$ is square of order n^s (or $n^s + 1$), and if μ is an eigenvalue of $R_{11}(\lambda)$, then $\mathrm{Re}(\mu) < 0$ (or $|\mu| \leq 1$). Then, we partition the orthogonal matrix $Q(\lambda)$ as

$$Q(\lambda) \;=\; [Q_1(\lambda) \quad Q_2(\lambda)],$$

where $Q_1(\lambda)$ is of order $n \times n^s$ (or $n \times (n^s + 1)$), and whose columns span the stable (or center-stable) subspace associated with $A(\lambda)$. Thus,

$$L_-(\lambda) := Q_1(\lambda).$$

We follow a similar idea to compute $L_+(\lambda)$. The main point is that the ordered Schur factorization allows us to place the eigenvalues we are interested in at the upper left block of R, and from there we can identify $Q_1(\lambda)$ and therefore $L_-(\lambda)$ or $L_+(\lambda)$.

Now let $W_-^{cu}(\lambda)$, respectively $W_+^{cs}(\lambda)$, be the center-unstable manifold of $M_-(\lambda)$, respectively the center-stable manifold of $M_+(\lambda)$. Suppose that there is a connecting orbit γ, connecting M_- and M_+; then, we must have $\gamma \subset W_-^{cu} \cap W_+^{cs}$. We expect the connecting orbit γ to be isolated if for the tangent spaces at $z(t)$ we have

$$T_{z(t)}W_-^{cu} \cap T_{z(t)}W_+^{cs} = T_{z(t)}\gamma = \text{ span } \{\dot{z}(t)\}, \qquad \forall\, t \in R, \tag{7.44}$$

and to persist if the intersection is *transversal*; i.e.,

$$T_{z(t)}W_-^{cu} + T_{z(t)}W_+^{cs} = R^{m+p}\,, \qquad \forall\, t \in R, \tag{7.45}$$

where $z = (x, \lambda)$.

For each given λ, let us write n_-^u, n_-^c, and n_-^s, for the dimensions of the unstable, center, and stable manifolds of $M_-(\lambda)$, and analogously we will write n_+^u, n_+^c, and n_+^s relatively to $M_+(\lambda)$. We will have that $n_\pm^c = 0$ for an equilibrium point, and $n_\pm^c = 1$ for a (hyperbolic) periodic orbit. With this notation, it is possible to establish a fundamental relation between number of parameters and dimensions of $W_\pm^{cs,cu}$ from (7.44) and (7.45), which reads [7]:

$$p = n_+^u - n_-^u - n_-^c + 1. \tag{7.46}$$

The main result in the theoretical setup is that the connecting orbit problem

$$\begin{aligned} &\dot{x} = f(x,\lambda)\,, \quad -\infty < t < \infty\,, \\ &\lim_{t\to-\infty} \operatorname{dist}(x(t,\lambda), M_-(\lambda)) = 0\,, \quad \lim_{t\to+\infty} \operatorname{dist}(x(t,\lambda), M_+(\lambda)) = 0\,, \\ &\psi(x, y_-, y_+, \lambda) = 0\,, \end{aligned} \tag{7.47}$$

is well posed if and only if the manifolds W_-^{cu} and W_+^{cs} intersect transversally in the sense of (7.44) and (7.45).

Connecting orbits can be efficiently computed by numerically solving the boundary value problem (7.42). Under appropriate conditions, the error, which results from truncating the interval of integration $(-\infty, \infty)$ to a finite interval $J := [T_-, T_+]$, decays exponentially. To state this result formally, we first give the following important result.

Lemma 7.24 *Let $F : B_\delta(w_0) \to Z$ be a C^1 mapping from some ball of radius δ in a Banach space W into some Banach space Z. Assume that $F'(w_0)$ is an homeomorphism and that for some constants c_1, c_2 we have*

$$\| F'(w) - F'(w_0) \| \le c_2 < c_1 \le \| F'(w_0)^{-1} \|^{-1}, \quad \forall\, w \in B_\delta(w_0)\,, \tag{7.48}$$

$$\| F(w_0) \| \le (c_1 - c_2)\,\delta. \tag{7.49}$$

Then F has a unique zero w_c in $B_\delta(w_0)$ and

$$\| w_0 - w_c \| \le (c_1 - c_2)^{-1} \| F(w_0) \|\,, \tag{7.50}$$

$$\| w_1 - w_2 \| \le (c_1 - c_2)^{-1} \| F(w_1) - F(w_2) \|\,, \qquad \forall\, w_1, w_2 \in B_\delta(w_0). \tag{7.51}$$

Next, we need to define the following spaces:

$$W := C^1(J, R^n) \times R^p, \quad Z := C(J, R^n) \times R^{n_-^c + n_-^s} \times R^{n_+^u + 1}.$$

For $\alpha, \beta > 0$, their norms are defined as

$$\begin{aligned} \|(x,\lambda)\|_W &= \sup_{t\in J_-} \|x(t)\| e^{\alpha t} + \sup_{t\in J_+} \|x(t)\| e^{-\beta t} + \|\lambda\|, \\ \|(y, r_-, r_+)\|_Z &= \|(y, r_-)\|_{Z_1} + \|(y, r_+)\|_{Z_2}, \quad \text{where} \end{aligned} \tag{7.52}$$

$$\begin{aligned} \|(y, r_-)\|_{Z_1} &= \sup_{t \in J_-} \|y(t)\| e^{\alpha t} + \|r_-\| \quad \text{and} \\ \|(y, r_+)\|_{Z_2} &= \sup_{t \in J_+} \|y(t)\| e^{-\beta t} + \|r_+\|, \qquad \|\cdot\| = \|\cdot\|_\infty, \end{aligned} \tag{7.53}$$

with $J_- = [T_-,\ 0]$ and $J_+ = [0,\ T_+]$.

With these norms, W and Z become *Banach spaces* (that is, complete normed vector spaces: Every Cauchy sequence is convergent). Anticipating the asymptotic convergence of $x(t)$ to $y(t)$ with rate $\epsilon > 0$ we impose the condition that for some constant C, $\|x(t) - y_\pm(t)\| \leq Ce^{-\epsilon|t|}$, as $t \to \pm\infty$.

The following theorem says that the error in approximating the connecting orbit is bounded by the projection boundary conditions.

Theorem 7.25 *Let* (7.44) *and* (7.45) *hold, and let* $(\bar{x}, \bar{\lambda})$ *be an orbit connecting either a hyperbolic equilibrium point* $y_-(\bar{\lambda})$ *or a hyperbolic periodic orbit* $\gamma_-(\bar{\lambda})$, *to a hyperbolic periodic orbit* $\gamma_+(\bar{\lambda})$. *Consider* (7.42) *and assume that* $f \in C^2(\mathbb{R}^{n+p}, \mathbb{R}^n)$, *and that* $L_\pm$ *are* C^1 *(in* λ*). Then, there exists* $\delta > 0$ *sufficiently small and* $C > 0$, *such that, for a sufficiently large interval of integration* $J = [T_-, T_+]$, *the boundary-value problem* (7.42) *has a unique solution* (x_J, λ_J) *in a ball of radius* δ *in* W. *Moreover, the following estimate holds:*

$$\begin{aligned} \|(x_J, \lambda) - (\bar{x}|_J, \bar{\lambda})\|_W \leq\ & C\,(\ \| L_-(\bar{\lambda})(\bar{x}(T_-) - y_-(s(T_-)))\| \\ & + \| L_+(\bar{\lambda})(x(T_+) - y_+(s(T_+)))\|\ ). \end{aligned} \tag{7.54}$$

The above theorem can be proved by applying Lemma 7.24 to $w_0 = (\bar{x}|_J, \bar{\lambda})$ and

$$F(x, \lambda) = (\dot{x} - f(x, \lambda),\ L_-(\bar{\lambda})(\bar{x}(T_-) - y_-(s(T_-))),\ L_+(\bar{\lambda})(x(T_+) - y_+(s(T_+)))\,).$$

Finally, we state the main theorem on connecting orbits, which establishes existence and uniqueness of solutions as well as exponentially decaying errors in the approximation.

Theorem 7.26 *Under the assumptions of Theorem 7.25, for* $J = [T_-,\ T_+]$ *sufficiently large, we have*

$$\|(x_J, \lambda) - (\bar{x}|_J, \bar{\lambda})\|_W \leq C\, e^{-2\min(\mu_-|T_-|,\ \mu_+ T_+)}. \tag{7.55}$$

In (7.55), $0 < \mu_- <$ Re μ, *for all unstable eigenvalues* μ *of the Jacobian* $f_x(y_-(\bar{\lambda}))$ *(if* $M_-(\bar{\lambda}) = y_-(\bar{\lambda})$*), or all unstable Floquet exponents of the monodromy relative to* $\gamma_-(\bar{\lambda})$ *(if* $M_-(\bar{\lambda}) = \gamma_-(\bar{\lambda})$*). Also,* $0 < \mu_+ <$ -Re μ, *for all stable Floquet exponents* μ *associated with the periodic orbit* $\gamma_+(\bar{\lambda})$.

Proof. The key tools to use are corollaries from the stable manifold theorems for equilibria and periodic orbits (see, e.g., [31]), which state that solutions starting in the corresponding

unstable or stable manifold, sufficiently near the equilibrium or the periodic orbit, approach them exponentially fast, as $t \to -\infty$ or $t \to \infty$, respectively. Moreover, in the case of a periodic orbit, the motion along the connecting orbit is synchronized with that on the periodic orbit (convergence in asymptotic phase). For the boundary condition at an equilibrium, there exists $T_1 < 0$, such that [6]

$$L_-(\bar{\lambda})(\bar{x}(t) - y_-(s(t))) = O(\, e^{-2\mu_- t}\,)\,, \quad \text{for} \quad t \le T_1\,. \tag{7.56}$$

Next we give the proof for the exponential decay of the error for the boundary condition at the periodic orbit γ_+. If $0 < \mu_+ < -\mathrm{Re}\,\mu$, for all characteristic exponents μ with negative real part of the periodic orbit $\gamma_+(\bar{\lambda})$, then there exists $T_2 > 0$ such that for all $t \ge T_2$,

$$\bar{x}(t) - \gamma_+(\bar{\lambda}) = O(e^{-\mu_+ t})\,. \tag{7.57}$$

Then, for any $t \ge T_2$, there is always a time shift $s(t)\ :\ 0 \le s(t) \le \tau_+$, such that $\bar{x}(t) - y_+(s(t)) = O(e^{-\mu_+ t})$. With this, by a Taylor expansion, we get

$$\begin{aligned} L_+(\bar{\lambda})(\bar{x}(t) - y_+(s(t))) \;&=\; L_+(\bar{\lambda})(y_+(s(t)) - y_+(s(t))) \\ &+\; L_+(\bar{\lambda})(y_+(s(t)) - y_+(s(t)))\,(\bar{x}(t) - y_+(s(t))) + O(\,\|\bar{x}(t) - \bar{y}(s(t))\|^2\,\|). \end{aligned}$$

Therefore,

$$L_+(\bar{\lambda})(\bar{x}(t) - y_+(s(t))) = O(\,\|\bar{x}(t) - y_+(s(t))\|^2\,), \tag{7.58}$$

and by (7.57),

$$L_+(\bar{\lambda})(\bar{x}(t) - y_+(s(t))) = O(\, e^{-2\mu_+ t}\,). \tag{7.59}$$

As for the periodic-to-periodic case, if $0 < \mu_- < -\mathrm{Re}\,\mu$ for all unstable Floquet exponents of the periodic orbit $\gamma_-(\bar{\lambda})$, then there exists a $T_1 < 0$ such that for all $t \le T_1$,

$$\bar{x}(t) - \gamma_-(\bar{\lambda}) = O(e^{-\mu_+ t}), \text{ and} \tag{7.60}$$

by proceeding in a similar way as we did for the periodic orbit γ_+, we can obtain (for a time shift $s(t)$))

$$L_-(\bar{\lambda})(\bar{x}(t) - \bar{y}_-(s(t))) = O(\, e^{-2\mu_- t}\,). \tag{7.61}$$

By combining (7.56), or (7.61), and (7.59) with inequality (7.54) from Theorem 7.25, we get the sought result.

□

With the theoretical background well established, algorithms for the numerical computation of connecting orbits can be constructed. The main part of such a code deals with solving the nonlinear system $\dot{x} = f(x, \lambda)$ as a boundary value problem:

$$\left\{ \begin{array}{rcl} \dot{x} &=& (T_+ - T_-)f(x,\lambda)\,, \quad 0 \le t \le 1\,, \\ L_-(\lambda)\big(x(0,\lambda) - y_-(\lambda)\big) &=& 0\,, \\ L_+(\lambda)\big(x(1,\lambda) - y_+(0,\lambda)\big) &=& 0\,, \end{array} \right. \tag{7.62}$$

and calling a subroutine to compute the periodic orbit, for the current given value of $\lambda = \hat{\lambda}$:

$$\left\{ \begin{array}{rcl} \dot{y}_+ &=& \tau_+ f(y_+, \hat{\lambda})\,, \quad 0 \le t \le 1\,, \\ y_+(0) &=& y_+(1)\,, \\ \sigma(y_+, \hat{\lambda}) &=& 0\,, \end{array} \right. \tag{7.63}$$

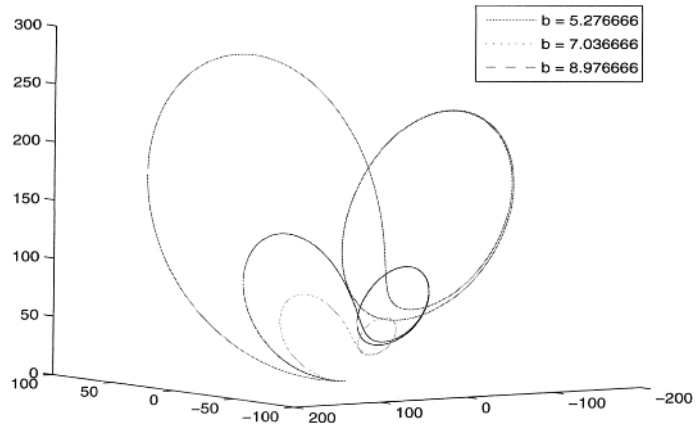


Figure 7.14 Connections in (7.64): $\lambda = 41.91,\ 70.12,\ 151.92$.

where $\sigma = 0$ is a phase condition for the periodic orbit, serving the role of $\psi = 0$ in (7.42), and the intervals on both systems are rescaled to $[0,\ 1]$. See [19] for details.

■ EXAMPLE 7.16

We consider the well-known Lorenz equations. This is a system that has been extensively studied because of the wide range of solution behaviors it shows, including chaotic solutions as well as homoclinic and heteroclinic orbits. Originally, the system was proposed by Lorenz as a simple model for weather prediction. The system is

$$\begin{aligned} \dot{x}_1 &= \sigma\,(x_2 - x_1), \\ \dot{x}_2 &= \lambda\, x_1 - x_2 - x_1 x_3, \\ \dot{x}_3 &= x_1 x_2 - b\, x_3, \end{aligned} \tag{7.64}$$

where we take $\sigma = 10$, $b = \frac{8}{3}$, and treat λ as a free parameter. There exists a connection from the origin to a periodic orbit as the result of the transversal intersection of the one-dimensional unstable manifold of the origin with the two-dimensional center-stable manifold of the periodic orbit for $\lambda = 24.057900$. The computed periodic orbit has period $\tau = 0.677171$ and Floquet multipliers $\mu_1 = 1$, $\mu_2 = 1.029332$, $\mu_3 = 0.000092$. Thus, it is an unstable periodic orbit, and therefore finding a connection to it is not an easy task.

It is possible to perform what is known as continuation of solutions, that is, once a solution is located, allow the parameters to vary and compute other solutions for those values of the parameters. In this case, we do continuation on the b parameter (and λ changes correspondingly). In Figure 7.14, we show a few connecting orbits obtained by smooth continuation, and in Table 7.16, we show how the periods and the Floquet multipliers change as the parameters vary.

b	λ	Period	Floquet multipliers
2.7566666666	24.4872943345	0.6633333236	0.0001055321 1.0317331504
5.2766666666	41.9064116742	0.4301691358	0.0003875710 1.0945902025
8.9766662563	151.9179506324	0.2206728489	0.0081105622 1.5012230948
11.9766664585	861.7281515695	0.1304312721	0.0052148282 9.5770852838
13.4416464434	4829.6606254594	0.1000748120	0.0000049684 17715.58156289

Table 7.1 Periods and multipliers for theLorenz system (7.64).

For illustration, we show the Jacobian matrix J at the origin $(0,0,0)$ and its block Schur factorization involved in the computation of the connecting orbit for $\lambda = 2.7566666666$.

$$Q^T J Q = Q^T \begin{bmatrix} -10 & 10 & 0 \\ \lambda & -1 & 0 \\ 0 & 0 & -\frac{8}{3} \end{bmatrix} Q$$

$$= \left[\begin{array}{cc|c} -12.41495963 & 0.00000000 & -7.24333333 \\ 0.00000000 & -2.66666666 & 0.00000000 \\ \hline 0.00000000 & 0.00000000 & 1.41495963 \end{array}\right],$$

$$\text{where} \quad Q = [Q_1 \quad Q_2] = \left[\begin{array}{cc|c} -0.97205634 & 0.00000000 & 0.23474768 \\ 0.23474768 & 0.00000000 & 0.97205634 \\ 0.00000000 & 1.00000000 & 0.00000000 \end{array}\right].$$

Then, the matrix $L(\lambda) := Q_1$ spans the stable subspace of the equilibrium point (0,0,0) at $\lambda = 2.7566666666$.

■ EXAMPLE 7.17

To compute some periodic-to-periodic connections, we consider the following system:

$$\begin{cases} \dot{x} = (1-w)y + wx(1-x^2), \\ \dot{y} = (1-w)\big(-x + \lambda(1-x^2)y\big) + w(z-\gamma), \\ \dot{z} = (1-w)z\big((z^2-(1+\gamma)^2\big) + w\big[-y+\gamma+\lambda\big(1-(y-\gamma)^2\big)(z-\gamma)\big], \\ \dot{w} = w(1-w), \end{cases} \tag{7.65}$$

where $\gamma = 3 + \lambda$, and λ is a real positive parameter. This system is a homotopy from $w = 0$ to $w = 1$, which in essence takes us to two planar systems living in the (x, y) and (y, z) planes, respectively.

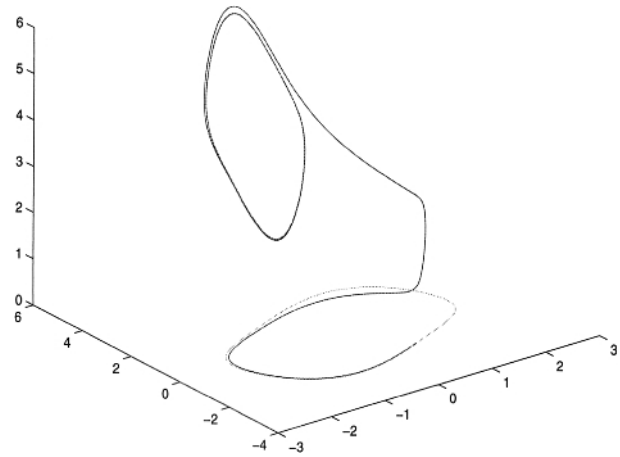


Figure 7.15 (7.65): Periodic-to-periodic, $\lambda = 0.5$.

In the (x, y) and (y, z) planes, the equations reduce to those of van der Pol oscillators. As it is well known, these oscillators have attracting periodic orbits (restricted to their respective planes). There are several heteroclinic connections between the two limit cycles of these van der Pol oscillators, and here we are interested in computing (and continuing) the one from $z = 0$, $w = 0$, call it γ_-, to that with $x = 0$, $w = 1$, call it γ_+. A simple computation shows that associated with both $\gamma_\pm$, there are two multipliers less than 1, one equal to 1, and one greater than 1. Therefore, we have $n^u_\pm = 1$, $n^c_\pm = 1$, and $n^s_\pm$=2. The balance (7.46) will give us $p = 0$; hence, there are no free parameters in the problem. In Figure 7.15, we show the connection for $\lambda = 1/2$, and in Figure 7.16, we show the second components of several of these connecting orbits obtained from continuation.

7.2.4 Chaos

In the area of differential equations and dynamical systems, chaos theory is a relatively new area of study. It has grown in importance not only because of the challenging mathematics behind it but also because of its wide range of important applications in areas such as physics, physiology, fractals, secure communication, and fluid dynamics. Recall that given the linear system

$$\dot{x} = A\,x, \quad x(0) = x_0, \tag{7.66}$$

its solution is explicitly given by

$$x(t) = e^{At}x_0. \tag{7.67}$$

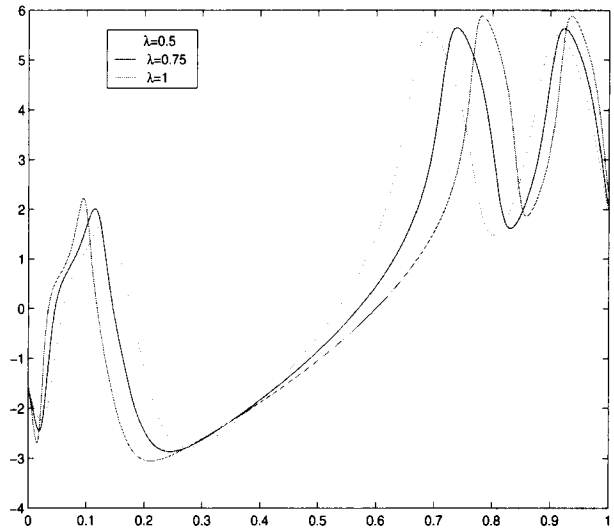


Figure 7.16 (7.65): Second component of connecting orbits.

This means that the solution is known for any time $t \geq 0$ and everything is predictable, so that no chaotic behavior is possible. However, when we consider nonlinear systems

$$\dot{x} = f(x), \quad x(0) = x_0, \tag{7.68}$$

there is no closed formula like the one in (7.67) for its solution. Thus, even if the vector field f is very smooth, the solution $x(t)$ may be unbounded or even unpredictable in general; some strange solutions may appear that apparently cannot be entirely and precisely described; this is one of the main differences between linear or nonlinear systems. Chaos theory tries to find and describe the underlying order in apparently random data.

Chaotic behavior is also related to sensitive dependence on initial conditions: The smallest change in the initial conditions can drastically change the long-term behavior of a system. This is exactly what Edward Lorenz experienced in the 1960s for the first time when he was studying a dynamical system for weather prediction. He realized that small errors in the input data (a finite number of digits of precision) caused the system to behave entirely differently; this made him conclude that it is simply impossible to predict the weather accurately. For a detailed study of the Lorenz system, see [61].

For a better illustration on the ideas of chaos, we briefly present some simple systems where a variety of solutions, including chaotic behavior, is observed.

Lorenz system. We study the system considered in Example 7.16

$$\begin{aligned} \dot{x}_1 &= \sigma\,(x_2 - x_1), \\ \dot{x}_2 &= \lambda\, x_1 - x_2 - x_1 x_3, \\ \dot{x}_3 &= x_1 x_2 - \beta\, x_3, \end{aligned} \tag{7.69}$$

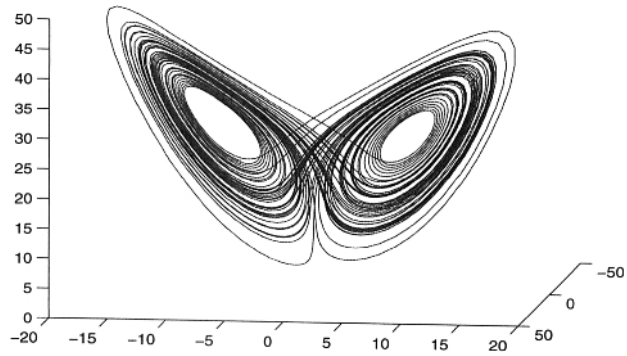


Figure 7.17 Strange attractor of (7.69).

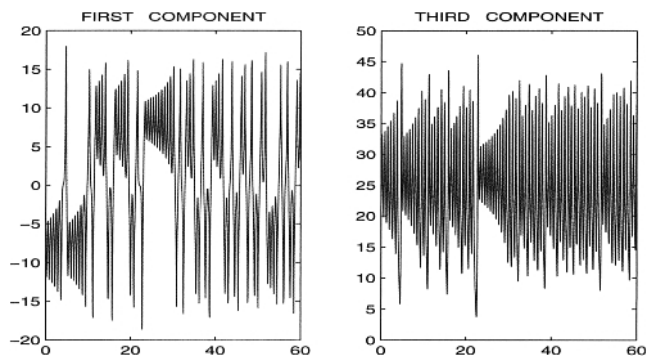


Figure 7.18 Solution components vs. time of (7.69).

with the parameter values $\sigma = 10, \lambda = 28$, and $\beta = 8/3$. The equations represent a simplified model of weather as fluid motion in the atmosphere driven by thermal buoyancy, known as convection. The variable x_1 measures the rate of convection overturning, the variable x_2 measures the horizontal temperature variation, and the variable x_3 measures the vertical temperature variation. The parameter σ represents the ratio of the viscous to thermal diffusivities, and λ represents the temperature difference and is the control parameter. There is a great variety of solutions to the system depending on the values of λ (see for instance Example 7.16). For $\lambda = 28$, solutions are pulled to what is known as a strange attractor. See Figures 7.17 and 7.18. Furthermore, the phase portrait will look very different if the initial condition is changed slightly. See Exercise 7.52.

The classic solutions that we know are the steady states or equilibrium points, in which the values of the coordinates never change after reaching those points, and the periodic solutions, where no matter how simple or complicated the trajectory is, the system goes

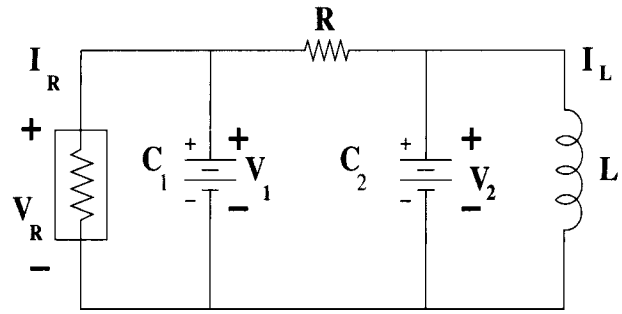


Figure 7.19 Chua's circuit.

into a loop repeating itself indefinitely. But in this new kind of solution, the trajectory does not settle down to a steady state and it is not periodic either. There is no way to predict exactly the path of the solution as time increases, but we know that it will stay around this strange set.

Chua's circuit. This is the simplest electronic device and model that exhibits complex behavior and a variety of chaotic phenomena. This is what makes it very popular, and it is also considered to be the paradigm for chaos. As Figure (7.19) shows, the circuit introduced by Leon Chua in the early 80s, consists of two capacitors C_1, C_2, one resistor R, one inductor L, and one non-linear resistor (the Chua's diode). Observe that the main difference between the Chua's circuit and the RLC circuit (see Section 6.1.9) is the presence of the nonlinear resistor.

If we let $x_1 = V_1, x_2 = V_2$ and $x_3 = I_L$, the system defining the Chua's circuit is

$$\begin{aligned} \dot{x}_1 &= \alpha(x_2 - h(x_1)), \\ \dot{x}_2 &= x_1 - x_2 + x_3, \\ \dot{x}_3 &= -\beta x_2, \end{aligned} \tag{7.70}$$

where $h(x) = \frac{2}{7}x - \frac{1}{28}(|x+1| - |x-1|)$. This function h is not smooth, but it can be approximated with a smooth function, by taking $|x| \approx \frac{2}{\pi}\arctan(10x)$.

The chaotic behavior of this system has been observed not only through mathematical analysis and computer simulation but also in laboratory experiments. As discussed before, it is a very useful system because it is quite simple and yet it exhibits several special solutions: homoclinic to the origin, torus breakdown, Hopf bifurcations, period doubling, stochastic resonance, strange attractors, and so on. In particular, a double-scroll attractor is shown in Figure 7.21. .

We remark that one interesting potential application of chaos is in secure communication. The idea is to encode a message within a (high dimensional) chaotic dynamics through small perturbations of a control parameter. Chaos is used to encrypt messages so that the transmitted signal is the sum of a chaotic signal and a given message, which can be reconstructed by the receiver, when synchronized with the sender.

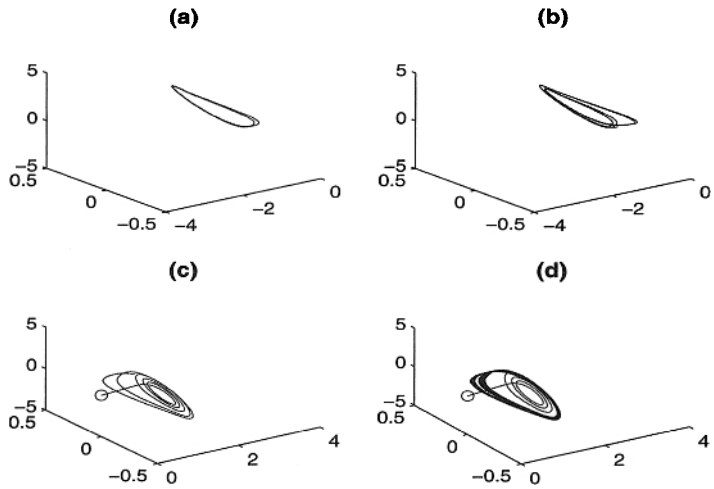


Figure 7.20 (a),(b) Period-doubling and (c),(d) strange attractor of (7.70).

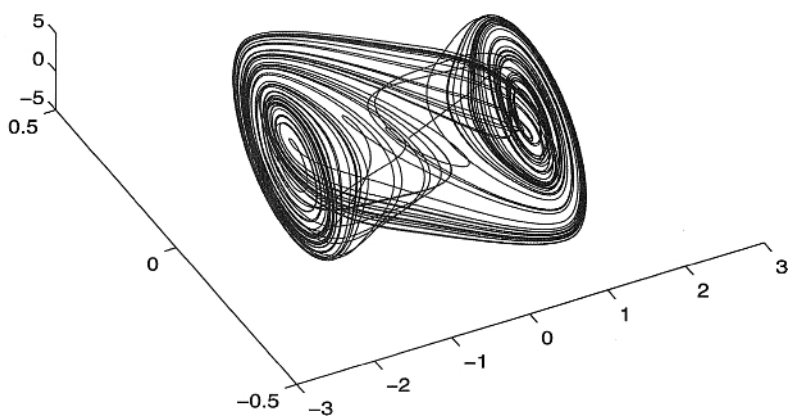


Figure 7.21 Chua's double-scroll attractor.

7.2.5 Bifurcations

In most applications, the vector field f depends on one or more parameters; that is, the dynamical systems take the form

$$\dot{x} = f(x, \mu), \tag{7.71}$$

with $\mu \in \mathbb{R}^p$, for some positive integer p. In general, solutions of (7.71) will vary as the parameters vary, but most importantly, there are special values of the parameters, say $\mu = \mu_0$, such that an arbitrarily small variation around μ_0 will cause drastic changes in the qualitative behavior of the solutions. This phenomenon known as bifurcation has grown in importance mainly as a result of its abundant applications in different fields, such as macroeconomic systems, physics, space exploration, biology, physiology, and so on. The purpose of this section is to present the basic ideas around such changes in qualitative behavior (in particular, stability) of solutions as a parameter varies.

The concept of bifurcation is closely related to that of stability and hyperbolicity. We want to start by introducing the concept of structural stability. To this end, for a vector field $f \in C^1(E)$, where E is an open subset of $\mathbb{R}^n$, we introduce the C^1 norm

$$\|f\|_1 = \sup_{x \in E} \|f(x)\| + \sup_{x \in E} \|Df(x)\|, \tag{7.72}$$

where on the right–hand side, the first is the Euclidean norm, and the second is a matrix norm.

Definition 7.27 *The dynamical system (7.71) is called* structurally stable *if there exists an $\epsilon > 0$ such that for all $g \in C^1(E)$, with*

$$\|f - g\|_1 < \epsilon,$$

f and g are equivalent in the sense that there is an homeomorphism that maps trajectories of (7.71) onto trajectories of $\dot{x} = g(x, \mu)$.

In simple words, this means that (7.71) is structurally stable if the qualitative behavior does not change for all nearby vector fields; i.e., small variations in f do not imply changes in such properties like stability or dimensions of stable or unstable manifolds. The homeomorphism property means that we can go from a trajectory of (7.71) to one of $\dot{x} = g(x, \mu)$ or back, by a continuous deformation.

One problem that has attracted much attention is on how to characterize or identify structurally stable systems. This is a very important question to solve, as such a stability property is a fundamental one to consider when studying systems modeling real world problems. For instance, when considering macroeconomic models, robustness of inferences about the behavior of the system becomes critically dependent on the sensitivity of the system to small changes in parameters.

For the case of two-dimensional systems, the famous Peixoto's theorem gives necessary and sufficient conditions for a dynamical system to be structurally stable in terms of hyperbolicity, wandering sets, and connecting orbits.

Unfortunately, there is no counterpart to Peixoto's theorem in higher dimensions. However, it is possible to give sufficient conditions for a system to be structurally stable for any dimensions. To this end, we need two definitions; first we define a nonwandering point as the one that stays inside an arbitrary neighborhood of it under the flow defined by (7.71). The nonwandering set of (7.71) is the set of all nonwandering points of (7.71). Common examples of nonwandering sets are equilibria and periodic orbits.

Definition 7.28 *Two differentiable manifolds M and N in $\mathbb{R}^n$ are said to* intersect transversally *if for every point $p \in M \cap N$, the tangent spaces satisfy*

$$T_pM \oplus T_pN = \mathbb{R}^n.$$

There is a set of sufficient conditions that a system $\dot{x} = f(x, \mu)$ has to satisfy for it to be structurally stable. A system satisfying those conditions is known as a Morse–Smale system, and the conditions are:

1. the number of equilibrium points and periodic orbits is finite and each is hyperbolic,
2. all stable and unstable manifolds that intersect do so transversally, and
3. the nonwandering set consists of equilibrium points and periodic orbits only.

Theorem 7.29 (Palis and Smale) *Every Morse-Smale system is structurally stable.*

The conditions for a Morse–Smale system are very similar to those in Peixotos's theorem, but more general, and Theorem 7.29 applies to dimensions higher than two. Unfortunately only, the converse of the theorem is not necessarily true for dimensions higher than two.

When systems are not structurally stable, say, the equilibrium points or periodic orbits are not hyperbolic, then we can expect bifurcation phenomena to happen, and in fact here we mainly consider nonhyperbolicity as a source of bifurcations. But again, this does not mean that systems with only hyperbolic equilibria and periodic orbits are structurally stable; all conditions above must be satisfied.

Now we need to give a definition of bifurcation. Consider again the system (7.71). We can define bifurcation as the phenomenon representing the sudden appearance of qualitatively different solutions as the parameter(s) is (are) slightly varied. For instance, an equilibrium point at a certain value of the parameter μ suddenly gives way to a set of periodic orbits, after an arbitrarily small change of the parameter.

More precisely, a parameter value $\mu = \mu_0$ for which the system (7.71) is not structurally stable is called a bifurcation value, and the corresponding (x_0, μ_0) is called a bifurcation point.

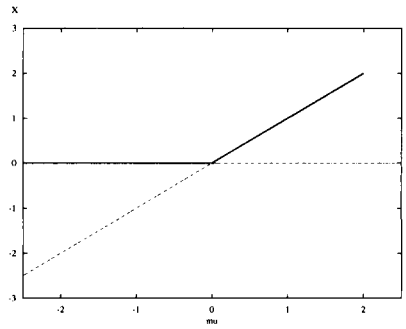


Figure 7.22 Transcritical bifurcation diagram of (7.73).

A great amount of work has been performed to devise algorithms to locate numerically bifurcations of nonlinear systems, and to identify the type of bifurcation. Probably the most reliable software available for the numerical study of bifurcations is **AUTO** [21], originally developed by E. Doedel. Also worth mentioning are XPPAUT [24], which is based on AUTO but a little more user-friendly, and more recently, Matcont [26].

Next, we present some of the most basic bifurcations of nonlinear systems. Here we illustrate such bifurcations by using the so-called *bifurcation diagrams*, where typically the horizontal axis is one of the parameters and the vertical axis is the norm of the solution. Thick solid curves represent stable equilibria, and dashed lines represent unstable equilibria. Stable periodic orbits are represented by small solid circles, whereas unstable periodic orbits are represented by empty circles. The bifurcation diagrams shown here were obtained using XPPAUT.

Transcritical bifurcations

This type of bifurcation is characterized by an exchange of stability at bifurcation values (a stable equilibrium becomes unstable and an unstable one becomes stable), and by the presence of nonhyperbolic equilibria. More generally, two different manifolds of equilibria cross each other, and at the crossing point, the equilibria exchange their stability properties. However, beyond the bifurcation point, the number of equilibria does not change.

■ **EXAMPLE 7.18**

Consider the system

$$\begin{aligned} \dot{x}_1 &= \mu x_1 - x_1^2, \\ \dot{x}_2 &= -x_2. \end{aligned} \tag{7.73}$$

The equilibrium points are $x^{(1)} = (0,0)$ and $x^{(2)} = (\mu, 0)$. The eigenvalues of the Jacobian at $x^{(1)}$ are $\lambda_{1,2} = -1,\ \mu$, and at $x^{(2)}$, they are $\lambda_{1,2} = -1, -\mu$. This means that for $\mu < 0$, $x^{(1)}$ is a stable node and $x^{(2)}$ is a saddle point, and that for $\mu > 0$, $x^{(1)}$ is a saddle, and $x^{(2)}$ is stable. An interchange of stability has ocurred, and we say a transcritical bifurcation has taken place at $\mu = 0$. See the bifurcation diagram in

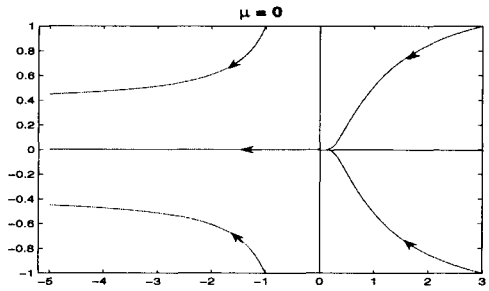


Figure 7.23 Phase portrait of (7.73) at the bifurcation value.

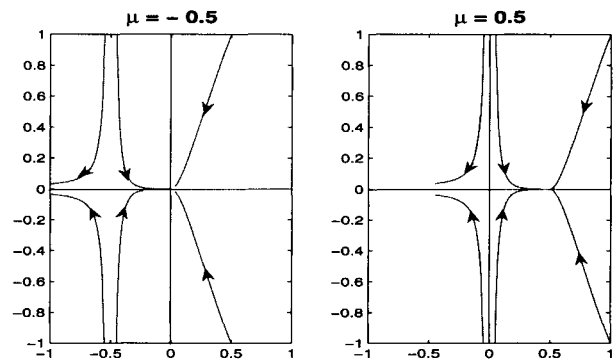


Figure 7.24 Phase portrait of (7.73): transcritical bifurcation.

Figure 7.22. In Figure 7.23 we show the phase portrait around $(0, 0)$, which is a (nonhyperbolic) unstable point for $\mu = 0$.

Compare this with Figure 7.24, where as the parameter μ goes from negative to positive, the equilibrium points interchange stability; e.g. ,the origin $(0, 0)$ is stable for $\mu = -0.5$ but it is a saddle point for $\mu = 0.5$; at the same time, the equilibrium $(\mu, 0)$ is a saddle point for $\mu = -0.5$, but it is a stable node for $\mu = 0.5$. The system is not structurally stable.

If we think of a macroeconomic model, one has to be extremely careful when performing inference; if a confidence region around parameter estimates includes a bifurcation point (x_0, λ_0), then various kinds of dynamics could be considered consistent, yet they may be in complete disagreement with the real problem. Although in the illustration in Example 7.18 we have used the values $\mu = -0.5$ and $\mu = 0.5$, this phenomenon is equally observed for arbitrarily small variations away from $\mu = 0$. Hence, making inferences in real-world problems in such a case is a very delicate issue.

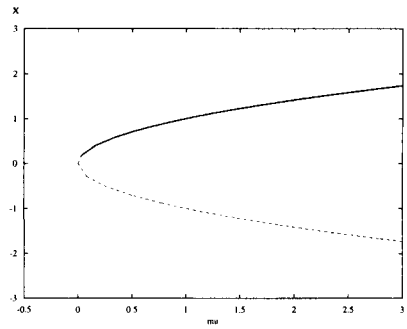


Figure 7.25 Saddle-node bifurcation diagram of (7.74).

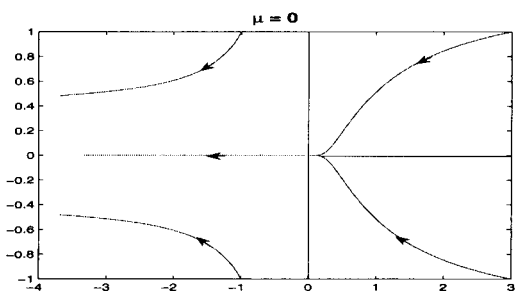


Figure 7.26 Phase portrait of (7.74) at the bifurcation value.

Saddle-node bifurcations

In this type of bifurcations, equilibria coalesce in such a way that the number of equilibria can go from two to one to none: Equilibria collide and annihilate one another. At the same time, stability properties change as the parameter passes through the bifurcation value. These bifurcations are also known as *fold bifurcations*.

■ EXAMPLE 7.19

Consider the system

$$\begin{aligned} \dot{x}_1 &= \mu - x_1^2 \\ \dot{x}_2 &= -x_2. \end{aligned} \tag{7.74}$$

The equilibrium points are $x^{(1)} = (\sqrt{\mu}, 0)$ and $x^{(2)} = (-\sqrt{\mu}, 0)$. The eigenvalues of the Jacobian at $x^{(1)}$ are $\lambda_{1,2} = -1, -2\sqrt{\mu}$, and at $x^{(2)}$, they are $\lambda_{1,2} = -1, 2\sqrt{\mu}$. The first thing we notice is that there are no equilibria for $\mu < 0$, there is only one equilibrium point for $\mu = 0$, and two for $\mu > 0$. See the bifurcation diagram in Figure 7.25. For $\mu > 0$, $x^{(1)}$ is a stable node and $x^{(2)}$ is a saddle point, and the system is not structurally stable. A saddle-node bifurcation has taken place at $\mu = 0$. In Figure 7.26, we show the phase portrait around $(0, 0)$, which is a (nonhyperbolic) unstable point for $\mu = 0$. In Figure 7.27, for μ negative there are no equilibria; for μ

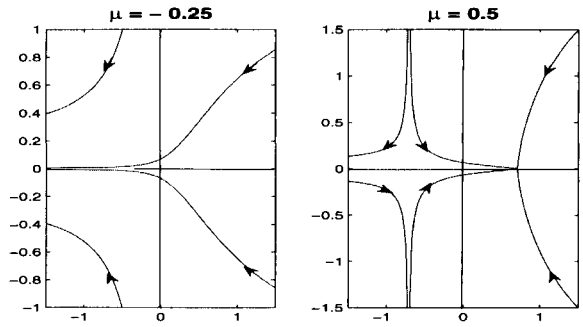


Figure 7.27 Phase portrait of (7.74): saddle-node bifurcation.

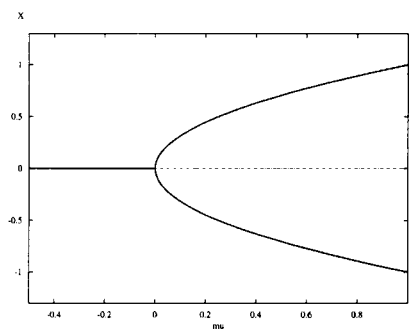


Figure 7.28 Pitchfork bifurcation diagram of (7.75).

positive, the equilibrium point $(\sqrt{\mu}, 0)$ is a stable node and the equilibrium $(-\sqrt{\mu}, 0)$ is a saddle point.

Pitchfork bifurcations

The main characteristic of a pitchfork bifurcation is that one equilibrium point bifurcates into three equilibria, one of them changing the stability property and the other two keeping stability behavior. The bifurcation happens at a nonhyperbolic equilibrium point and the system is not structurally stable.

■ **EXAMPLE 7.20**

Consider the system

$$\begin{aligned} \dot{x}_1 &= \mu x_1 - x_1^3 \\ \dot{x}_2 &= -x_2. \end{aligned} \tag{7.75}$$

The equilibrium points are $x^{(1)} = (0,0)$, $x^{(2)} = (\sqrt{\mu}, 0)$ and $x^{(3)} = (-\sqrt{\mu}, 0)$.F We observe that $x^{(1)}$ is the only equilibria for $\mu \leq 0$, and there are three equilibria

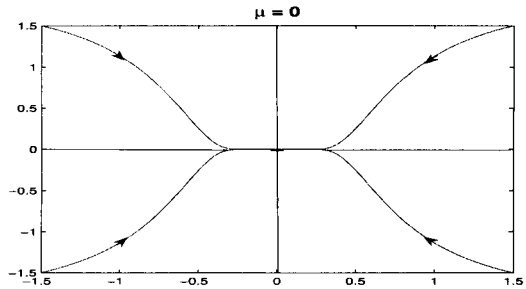


Figure 7.29 Phase portrait of (7.75) at bifurcation value.

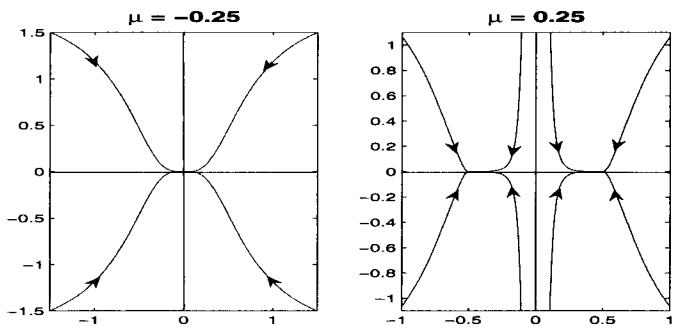


Figure 7.30 Phase portrait of (7.75): pitchfork bifurcation.

for $\mu > 0$. Thus, a pitchfork bifurcation happens at $\mu = 0$, where one equilibrium bifurcates into three as the parameter μ increases. See bifurcation diagram in Figure 7.28.

The eigenvalues of the Jacobian at $x^{(1)}$ are $\lambda_{1,2} = -1, \mu$; at $x^{(2)}$ and at $x^{(3)}$, they are $\lambda_{1,2} = -1, -2\mu$. Thus, for $\mu < 0$, $x^{(1)}$ is a stable node; for $\mu > 0$, $x^{(1)}$ is a saddle, while $x^{(2)}$ and $x^{(3)}$ are always stable nodes. For $\mu = 0$, $x^{(1)} = (0, 0)$ is a (nonhyperbolic) stable equilibrium point. See the behavior of solutions around it in Figure 7.29. Something similar happens for μ negative as we can see in Figure 7.30, where we also show solutions for μ positive. Observe how $x^{(1)}$ has now turned into a saddle point, and we go from one to three equilibrium points.

We insist on the fact that even though the illustration above was performed for $\mu \pm 0.25$, this radical change in the qualitative behavior of solutions happens for values of the parameter μ arbitrarily close to zero. This shows that in real–world applications, it becomes crucial to know on which side of the origin the system is operating, even when the system's dynamics are observed to be stable.

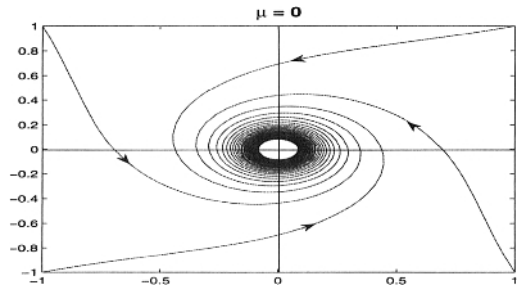


Figure 7.31 Phase portrait of (7.76) at bifurcation value.

Hopf bifurcations

The bifurcations presented so far can be found even in a one-dimensional setting. However, for a Hopf bifurcation, we need the system to be at least two-dimensional. This type of bifurcation is very special because a periodic orbit is born from an equilibrium point as the parameter varies, when such equilibrium changes its stability properties. The amplitude of the periodic orbit at birth is zero, and it increases with the parameter. We show a Hopf bifurcation diagram later in Section 7.3.

■ **EXAMPLE 7.21**

Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1(\mu - x_1^2 - x_2^2) - x_2, \\ \dot{x}_2 &= x_2(\mu - x_1^2 - x_2^2) + x_1. \end{aligned} \tag{7.76}$$

For any parameter μ, the only equilibrium point is $(x_1, x_2) = (0, 0)$, and similarly to Example 7.15, the system has a stable periodic orbit $\sqrt{\mu}\,(\cos t, \sin t)$, for $\mu > 0$. The Jacobian at $(0, 0)$ has eigenvalues $\lambda_{1,2} = \mu \pm \mathrm{i}$, which implies that $(0, 0)$ is a stable focus for $\mu < 0$, and it is an unstable focus of $\mu > 0$. Thus, the equilibirum $(0, 0)$ changes from stable to unstable as the parameter goes from negative to positive, and at $\mu = 0$, a stable periodic orbit bifurcates from $(0, 0)$. This is what characterizes a Hopf bifurcation.

Also observe in Figures 7.31 and 7.32 that for $\mu \leq 0$, solutions spiral toward the stable equilibrium $(0, 0)$, although for $\mu = 0$, solutions first go around the origin describing some kind of smaller and smaller circles, and for $\mu < 0$, solutions spiral toward the origin almost directly. On the other hand, for $\mu > 0$, solutions are attracted toward the stable periodic orbit, from outside and inside.

Note: The origin in Example 7.21 is nonhyperbolic and could be either a stable focus or a center for $\mu = 0$. Also recall that the Hartman–Grobman theorem cannot be applied to

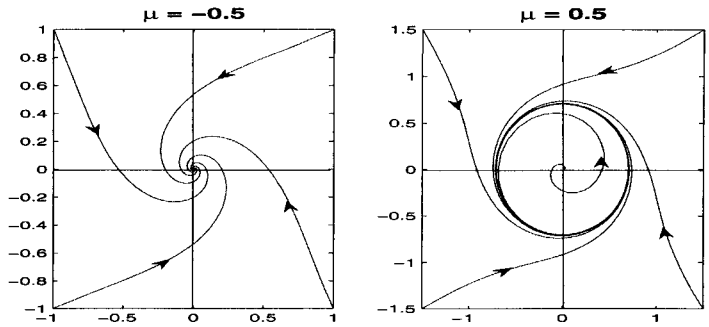


Figure 7.32 Phase portrait of (7.76): Hopf bifurcation.

this equilibrium point. By changing the system (7.76) to polar coordinates, one can show that the origin is a stable focus for $\mu = 0$.

Singularity bifurcations

This type of bifurcations is less known than the ones introduced above, but they are a very important phenomenon in applications, in particular, macroeconomic models [32, 33], and they are also very illustrative of the drastic changes in qualitative behavior that solutions show at a bifurcation value. The systems involved are related to differential algebraic equations, and are a generalization of the systems considered in the bifurcations above, namely

$$B(\mu)\,\dot{x} = f(x, \mu). \tag{7.77}$$

Thus, all cases considered before are a particular case with $B(\mu) = I$. In this new setting, the matrix $B(\mu)$ may be singular, but its singularity itself does not necessarily imply the presence of a singularity-induced bifurcation, which occurs when the rank of $B(\mu)$ changes as the parameter μ varies.

What makes this type of bifurcation different from the cases considered before is the drastic change in the dimension of the phase portrait as the parameter moves through a bifurcation value. This is mainly a consequence of moving from a differential to a differential-algebraic system as the parameter varies, so that the algebraic equation imposes a restriction on the paths to be followed by the solutions.

■ **EXAMPLE 7.22**

We consider the system

$$\begin{aligned} \dot{x}_1 &= x_1(1 - x_1), \\ \mu\,\dot{x}_2 &= x_1 - x_2^2. \end{aligned} \tag{7.78}$$

This corresponds to $B(\mu) = \begin{bmatrix} 1 & 0 \\ 0 & \mu \end{bmatrix}$ in (7.77). There are three equilibria: $x^{(1)} =$

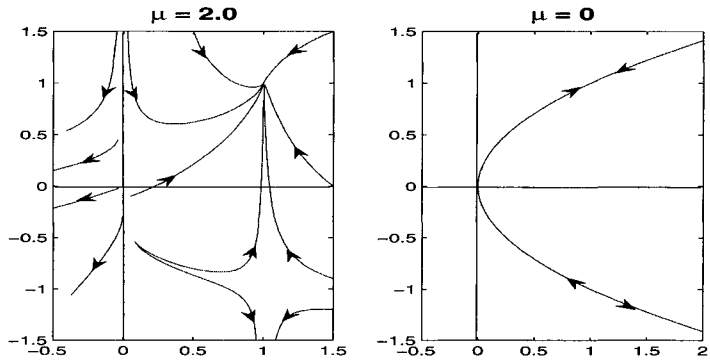


Figure 7.33 Phase portrait of (7.78) for a singularity bifurcation.

$(0,0)$, $x^{(2)} = (1,1)$, and $x^{(3)} = (1,-1)$. For any μ, $(0,0)$ is unstable; for $\mu < 0$, $(1,1)$ is unstable and $(1,-1)$ is stable, while for $\mu > 0$, $(1,1)$ is stable and $(1,-1)$ is unstable. Thus, $x^{(2)}$ and $x^{(3)}$ interchange stability as μ crosses the value zero. At this bifurcation value $\mu = 0$, when the rank of $B(\mu)$ decreases from two to one, something more remarkable happens: The second equation in (7.78) becomes an algebraic constraint, forcing the the behavior of the system to degenerate into motion only along the parabola $x_1 = x_2^2$. This results into a dramatic drop in the dimension of the dynamics from a two-dimensional state space to a one-dimensional curve. This is illustrated in Figure 7.33. For a negative value of μ, the solutions are somehow similar to those for positive μ, except of course that the stability properties of $x^{(2)}$ and $x^{(3)}$ are interchanged.

In practical applications, the potential presence of a singularity bifurcation signals important implications for robustness of dynamic inferences, not only because of the change of stability properties on both sides of the bifurcation values, but also because the dynamics of the system may drop into a "black hole" lower dimensional state space. Thus, in these cases a very high precision of parameter estimates is required, as the dynamics can be dramatically different within different subsets of the parameter estimates' confidence region.

7.3 PREDATOR–PREY MODELS WITH HARVESTING

To illustrate several of the concepts on dynamical systems introduced in this chapter, we study some predator–prey models with harvesting. Although the models introduced here are only two-dimensional systems, their dynamics are very rich as they exhibit several types of equilibria, periodic orbits, bifurcations, connecting orbits, and so on. On the other hand, we remark that predator–prey models in general are an active and important area of

research in applied mathematics, and they play a central role in areas like biology, ecology, and economics.

The inclusion of a harvesting component in a predator–prey system brings new interesting mathematical results, and its importance in resource management cannot be overstressed. Researchers study the exploitation of resources including fish stocks and other species, and governments use the results to enact policy with the intention of avoiding overexploitation of certain species or the degradation of particular land areas, while commercial harvesting companies adjust their actions to maximize profits.

Let $x(t)$ and $y(t)$ denote the prey and predator populations respectively, at time t. The following is a generalization of the classical Lotka–Volterra predator–prey rescaled model:

$$\begin{aligned} \dot{x} &= x(1-x) - \tfrac{axy}{1+mx}, \\ \dot{y} &= y\left(-d + \tfrac{bx}{1+mx}\right) - H(y), \end{aligned} \tag{7.79}$$

where H is a function that defines harvesting, in this case on the predator. The parameters a, b, d, and m are all positive. The parameter a represents the capture rate of the prey, m represents the half-saturation constant, d is the natural death rate of the predator, and b represents the prey conversion rate (how much the predator gains by capturing prey).

The simplest case of harvesting is to assume that the function H is constant; that is, $H(y) = h$, for some positive constant h. A little more realistic assumption is to take $H(y) = cy$, for some positive constant c, that is, as the predator population increases, so does its harvesting, to keep the population in check. Mathematically, these two harvesting functions provide the system with interesting dynamics; however, they are not satisfactory from the point of view of practical applications. For instance, as more predators become available, harvesting more at a linear rate might not be profitable and eventually impossible because of constrained resources for catching the predator.

Instead, here we consider the so–called (continuous) *threshold policy harvesting*:

$$H(y) = \begin{cases} 0 & \text{if } y < T \\ \frac{h(y-T)}{h+(y-T)} & \text{if } y \geq T, \end{cases} \tag{7.80}$$

where T is the threshold population size that determines when harvesting starts or stops and h is the rate-of-harvesting limit. In simple words, when a population is above a certain level or threshold, harvesting occurs; when the population falls below that level, harvesting stops. See Figure 7.34.

For simplicity, $H(y)$ in (7.80) can also be defined as a piecewise constant function, jumping from no harvesting to some positive (relatively large) harvesting values. But that type of harvesting is impractical in real–world application [51] because it would be difficult for managers to harvest immediately at a certain rate once $y \geq T$ because of time delays and capital constraints. Instead, the continuous harvesting (7.80) allows managers to increase the harvesting rate smoothly as the population increases. See Figure 7.34.

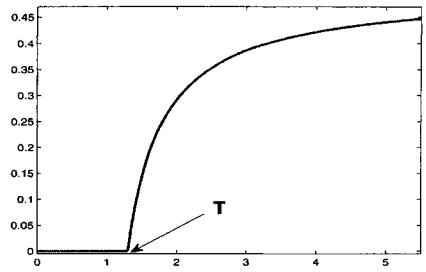


Figure 7.34 Harvesting function $H(y)$.

7.3.1 Boundedness of solutions

We are mainly interested in analyzing only the solutions in the closed first quadrant $\mathbb{R}^2_+$ because any solutions outside of the first quadrant are not biologically interpretable. Our first step into the analysis is to ensure that solutions starting on this first quadrant stay bounded. Accordingly, we have the following theorem.

Theorem 7.30 *All solutions of the system (7.79) starting in the positive first quadrant $\mathbb{R}^2_+$ are uniformly bounded.*

Proof. Let $w = x + \frac{a}{b}y$. Then, $\dot{w} = x(1-x) - \frac{ad}{b}y - \frac{a}{b}H(y)$. For each $k > 0$, we have

$$\begin{aligned} \dot{w} + kw &= x(1-x+k) + a\left(\tfrac{k-d}{b}\right)y - \tfrac{a}{b}H(y) \\ &\leq \tfrac{(1+k)^2}{4} + a\left(\tfrac{k-d}{b}\right)y - \tfrac{a}{b}H(y). \end{aligned}$$

Choose $k < d$. Since $0 \leq \frac{a}{b}H(y) \leq \frac{ah}{b}$ for $y \geq 0$, there exists $B > 0$ such that $\dot{w} + kw \leq B$, or $\dot{w} \leq B - kw$. Now consider the differential equation $\dot{r} = B - kw$, with $r(0) = w(0) = w_0$, whose solution $r(t) = \frac{B}{k}(1 - e^{-kt}) + w_0 e^{-kt}$ is bounded for $t \geq 0$. By using a differential inequality [28] we get $0 < w(t) \leq r(t) \leq \frac{B}{k}$, as $t \to \infty$. That is, solutions stay in

$$S = \{(x,y) \in \mathbb{R}^2_+ : x + \frac{a}{b}y = \frac{B}{k} + \gamma, \text{ with } \gamma > 0\}.$$

□

7.3.2 Equilibrium point analysis

We find the equilibrium points of (7.79) and analyze their properties, by first taking $m = b$ and by setting the growth rate of the prey to a. The following two equilibria always exist:

$$P_1 = (x_1, y_1) = (0,0) \tag{7.81}$$

and

$$P_2 = (x_2, y_2) = (1, 0). \tag{7.82}$$

When $y < T$, we also have the point

$$P_3 = (x_3, y_3) = \left(\frac{d}{b(1-d)}, \frac{b(1-d)-d}{b(1-d)^2} \right). \tag{7.83}$$

This equilibrium point $P_3 \in \mathbb{R}^2_+$ when $d < \frac{b}{b+1}$, which means that the predator growth parameter b must be sufficiently large relative to the predator death parameter d, for a coexistence equilibrium to exist.

If we denote

$$N_y = \left\{ (x, y) : x = \frac{dy^2 + (dh - dT + h)y - hT}{b[(1-d)y^2 + (dT - dh - T)y + hT]} \right\}$$

and

$$N_x = \{ (x, y) : y = (1-x)(1+bx) \},$$

then one can show that for $y \geq T$, $P_+ = (x_4, y_4)$ are equilibrium points of (7.79), if they are in the set $N_x \cap N_y$. These points $P_+ \in \mathbb{R}^2_+$ when $x_4 \in (0, 1)$.

Next, we want to find out the stability properties of the equilibria. The Jacobian of (7.79) is

$$J(x, y) = \begin{bmatrix} a\left[1 - 2x - \frac{y}{(1+bx)^2}\right] & -\frac{ax}{1+bx} \\ \frac{by}{(1+bx)^2} & \frac{bx}{1+bx} - d - \psi \end{bmatrix}, \tag{7.84}$$

where ψ is given by

$$\psi = \begin{cases} 0 & \text{if } y \leq T \\ \frac{h^2}{(h+y-T)^2} & \text{if } y > T. \end{cases}$$

The trace (T) and determinant (D) of (7.84) are then given by

$$T = a\left[1 - 2x - \frac{y}{(1+2x)^2}\right] + \frac{bx}{1+bx} - d \tag{7.85}$$

and

$$D = a\left[1 - 2x - \frac{y}{(1+bx)^2}\right]\left(\frac{bx}{1+bx} - d\right) + \frac{abxy}{(1+bx)^3}. \tag{7.86}$$

We evaluate these two quantities at the equilibrium points and use the trace-determinant Theorem 7.5 to obtain:

a) At the point (x_1, y_1), the trace and determinant are
$T = a - d$ and $D = -ad < 0$. Therefore, P_1 is always a saddle point.

This tells us that it is rarely the case that the predator and the prey simultaneously go extinct. Only when the initial condition starts along the stable branch of the origin does this happen.

b) At the point (x_2, y_2), the trace and determinant are

$T = \frac{b}{1+b} - a - d$ and $D = a\left[d - \frac{b}{1+b}\right]$. Therefore, P_2 has the following stability conditions:

1. If $\frac{b}{1+b} > d$, then (x_2, y_2) is a saddle point.
2. If $\frac{b}{1+b} < d$ and $\left(\frac{b}{1+b} - a - d\right)^2 > 4a\left(d - \frac{b}{1+b}\right)$, then (x_2, y_2) is a stable node.
3. If $\frac{b}{1+b} < d$ and $\left(\frac{b}{1+b} - a - d\right)^2 > 4a\left(d - \frac{b}{1+b}\right)$, then (x_2, y_2) is a stable spiral.

This means that if the death rate of the predator is sufficiently large, then the prey axis fixed point is stable. If the predator conversion rate b is sufficiently large, however, then the prey axis fixed point becomes unstable as a saddle. Notice that no matter how large a value of b, if $d > 1$, then the predator cannot reproduce quickly enough to survive and the prey survives alone.

c) At the point (x_3, y_3), the trace and determinant are
$T = \frac{ad[1+b(d-1)+d]}{b(d-1)}$ and $D = a\left[d - \frac{(1+b)d^2}{b}\right]$. Therefore, P_3 has the following stability conditions:

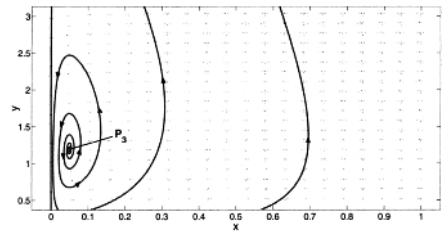
1. If $b = \frac{1+d}{1-d}$, then (x_3, y_3) is of center-type.
2. If $d < \frac{b}{1+b}$ and $b < \frac{1+d}{1-d}$, then (x_3, y_3) is either a stable focus or a stable node.
3. If $d < \frac{b}{1+b}$ and $b > \frac{1+d}{1-d}$, then (x_3, y_3) is either an unstable focus or an unstable node.

Notice that the condition $P_3 \in \mathbb{R}^2_{++}$ implies $d < \frac{b}{b+1}$, and on the other hand for P_3 to be a saddle, we need $D < 0$, or $d > \frac{b}{b+1}$. Thus, the coexistence equilibrium P_3 cannot be a saddle when it is biologically feasible.

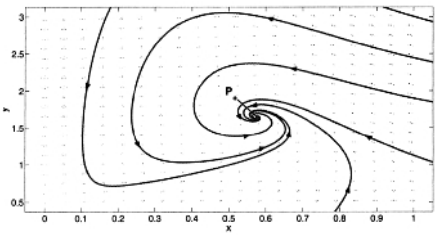
(d) At the point (x_4, y_4), the trace and determinant are
$T = -d - \frac{x(b+a\lambda)}{1+bx} + \varphi$ and $D = \frac{ax[b-bx+\lambda(1+bx)(-d+\frac{bx}{1+bx}+\varphi)]}{(1+bx)^2}$,
where $\lambda = -1 + b - 2bx$ and $\varphi = \frac{h^2}{[T-h+(x-1)(1+bx)]^2}$.

As we can see, the mathematical expressions for the trace and determinant of the Jacobian at $P_+ = (x_4, y_4)$ are more complicated, and in the expressions above, x is the solution of a polynomial equation of fifth degree with coefficients depending on the parameters. Thus, the inequalities obtained from trace-determinant analysis are much more involved, and we do not show them here, although they can be explicitly found. See [47] for details. Numerically, however, we can still obtain conclusions about the stability behavior for all the equilibria. More recent results in the same direction can be found in [57]

Numerical computations. To better understand the impact of continuous threshold policy (CTP) harvesting on a predator–prey ecosystem, we rely on numerical computations. While analytical derivations have helped to shed some light on the properties of CTP, computational simulation can further determine whether CTP can improve the conditions of the ecosystem.



(a) Unharvested.



(b) Harvested.

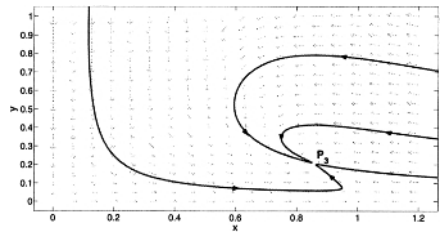
Figure 7.35 $P_3 = (0.05, 1.1875)$, $P_+ = (0.532, 1.65)$.

The first case involves a high predation conversion rate parameter value b. Without harvesting, the coexistence equilibrium is an unstable focus. The harvesting agent can use CTP to stabilize the system by choosing a sufficiently high harvesting limit h. See Figure 7.35. Notice that in this case, CTP increases the long-term populations of both the predator and the prey while simultaneously stabilizing the populations. The parameters used are $a = 0.5, b = 5, d = 0.2, h = 4, T = 0.5$.

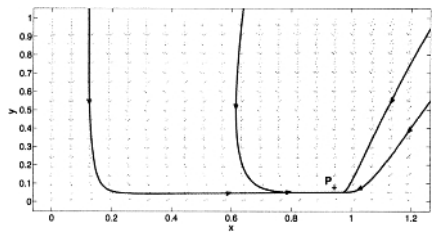
The second case is evidence for the notion that CTP cannot damage the ecosystem when it is in a stable state, i.e. enacting CTP does not cause the coexistence equilibrium to become unstable. Suppose the harvesting agent enacts CTP but chooses large parameter values for h and T to determine whether CTP can negatively affect the coexistence equilibrium. See Figure 7.36. The predator is now being harvested under extreme effort with a low threshold (T) and a high harvesting limit h. Nevertheless, the coexistence equilibrium is still stable. We have used the parameter values $a = 0.25, b = 0.5, d = 0.3, h = 10, T = 0.05$.

7.3.3 Bifurcations

As discussed before, given a dynamical system, one of the central questions is how the given model depends on the choice of the parameters, or, more precisely, how the qualitative behavior of solutions changes as one or more parameters are allowed to vary. In this section, we track for possible bifurcations as they represent drastic changes in the predator–prey system. Our model (7.79) exhibits several bifurcations including fold and Hopf bifurcations when we allow parameters to vary continuously.



(a) Unharvested.



(b) Harvested.

Figure 7.36 $P_3 = (0.8571, 0.2041)$, $P_+ = (0.9654, 0.0513)$.

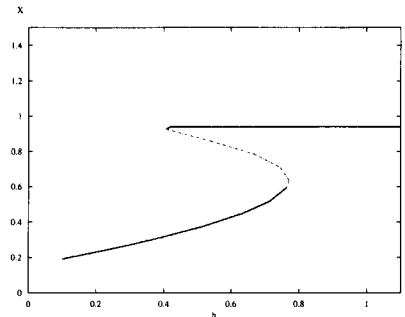
In Figure 7.37(a), we show the basic bifurcation diagram, when the parameter h varies continuously. Two fold bifurcations are detected at $h \approx 0.4055193$ and $h \approx 0.7701848$. At these values of the parameter h, solutions change their properties from stable to unstable and back to stable. This bifurcation diagram also allows to see the number of equilibria: e.g. ,the system has three equilibria for $0.4055194 \leq h \leq 0.7701848$, two of them stable and one unstable.

Starting at the fold bifurcation for $h \approx 0.7701848$, we can perform two-parameter continuation by allowing both h and m to vary at the same time. This gives the bifurcation diagram shown in Figure 7.37(b), which represents a curve of fold points; this means that along this curve in the (h, m) space, the system (7.79) has a fold.

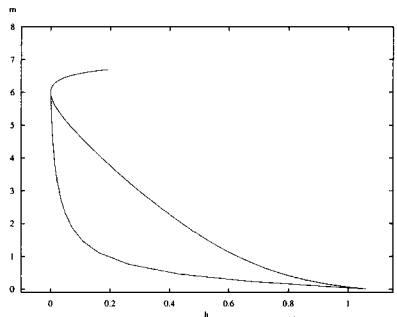
If instead of h we set m as the main parameter of continuation, then a Hopf bifurcation is detected. Indeed, for $m \approx 1.7$, such bifurcation is numerically detected, where the equilibrium (x_2, y_2) is a center. From here, a branch of stable periodic orbits is born. See the bifurcation diagram in Figure 7.38.

7.3.4 Connecting orbits

We can see from Figure 7.38 that periodic solutions are born at an equilibrium point as the parameter passes through the Hopf bifurcation value. Here, we first explicitly compute one such periodic solution, and then we allow the parameter h to vary to compute a branch of stable periodic orbits. Later we compute heteroclinic connections, more precisely, point-to-periodic connections. We start with $h = 0.1$, and we follow the periodic orbits down



(a) Basic bifurcation diagram.



(b) Two-parameter bifurcation diagram.

Figure 7.37 Bifurcation diagrams of predator–prey model.

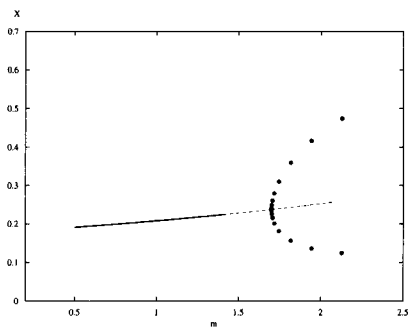


Figure 7.38 Hopf bifurcation gives out periodic solutions.

to smaller values of h down to $h = 0$ (no harvesting), and also to larger values of h until the corresponding stable Floquet multiplier approaches the value 1, and the periodic orbit collapses to a point. See Figure 7.39 and Table 7.2.

Each periodic solution shown in Figure 7.39 represents long-term coexistence of predator and prey for different harvesting efforts h. Observe that for larger values of h, we have relatively smaller values of both populations,although coexistence is preserved.

Clearly, if an initial value for the system (7.79) is on one of its periodic orbits, then long-term coexistence is guaranteed. The question is on how one can arrive to such solutions

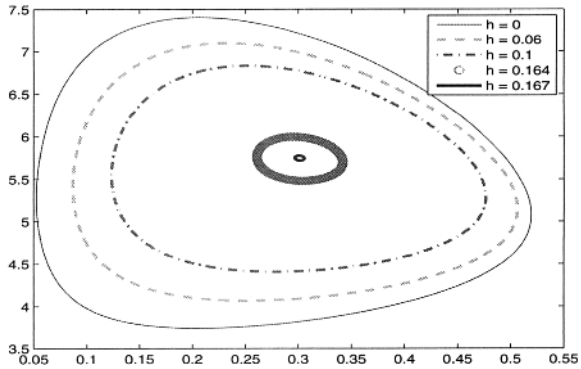


Figure 7.39 A branch of periodic orbits of (7.79).

h	Period	Floquet multiplier
.0	29.3147399	.3303902374914
.08	28.2861845	.4676316116910
.13	27.6349319	.6953439617105
.16	27.2112798	.9301526888350
.167	27.1082952	.9996358112280

Table 7.2 Period and Floquet multipliers.

from a given equilibrium point. Point-to-periodic solutions represent in this case paths that explicitly describe the evolution of predator and prey populations from, say, an unstable state, to a (long-term) stable one. We compute such heteroclinic orbits allowing the parameter h to vary and compute a branch of point-to-periodic connections. Given an equilibrium point, we start in the direction of its unstable manifold to arrive through the stable manifold of the periodic orbit. This tells us how from an equilibrium point of the system (7.79) one can manipulate or perturb the predator and prey populations to force the system to arrive to a long-term stable periodic solution. See Figure 7.40.

7.3.5 Other models

A long list of other predator–prey models can be studied, with different interactions between predator and prey and with different harvesting functions. For instance, one can study the so-called ratio-dependent models, whereby the rate of prey consumption does not just depend on the prey population, but instead on the ratio of predator and prey population. In such models, the interaction of predator and prey is represented as

$$\frac{axy}{my + x}.$$

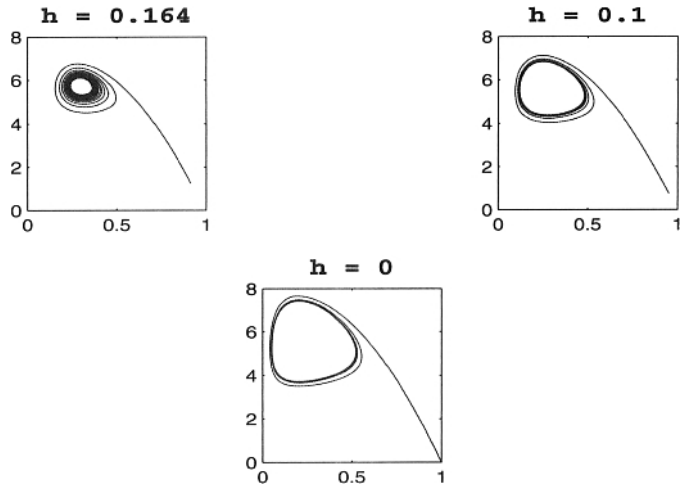


Figure 7.40 Branch of point-to-periodic connections of (7.79).

One can also consider higher–dimensional systems to study the cases of, say, two or more prey and one predator, or two or more predator and one prey. Higher–dimensional systems are also obtained when considering a disease acting on one or both species, as one has to add equations to consider infected populations.

If one considers the motion of one or more of the species, then diffusion terms have to be added to the model, leading to a system of partial differential equations. This kind of systems can be transformed into a classic system of ordinary differential equations by certain change of coordinates. This is done for instance when studying the corresponding traveling wave solutions of the system.

Finally, one can consider different harvesting functions, e.g., piecewise continuous functions, or periodic functions to represent seasonal or rotational harvesting, and harvesting could be applied to either or both, predator and prey.

For illustration, one can slightly modify the original model (7.79) to the following one:

$$\begin{aligned} \dot{x} &= x(1-x) - \tfrac{axy}{1+mx} - H(x), \\ \dot{y} &= y\left(-d + \tfrac{bx}{1+mx}\right), \end{aligned} \tag{7.87}$$

where now the harvesting is on the prey and it is given by the following piecewise linear function (that follows about the same shape as that in (7.80)).

$$H(x) = \begin{cases} 0 & \text{if } x < T_1, \\ \frac{h(x-T_1)}{T_2-T_1} & \text{if } T_1 \le x < T_2, \\ h & \text{if } x \ge T_2. \end{cases} \tag{7.88}$$

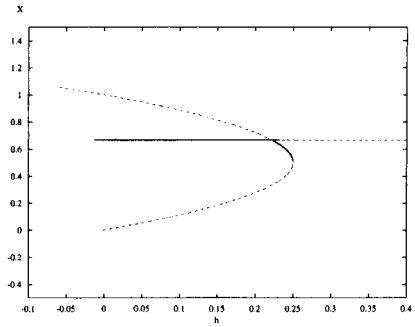


Figure 7.41 Fold and transcritical bifurcations of (7.87), and (7.88).

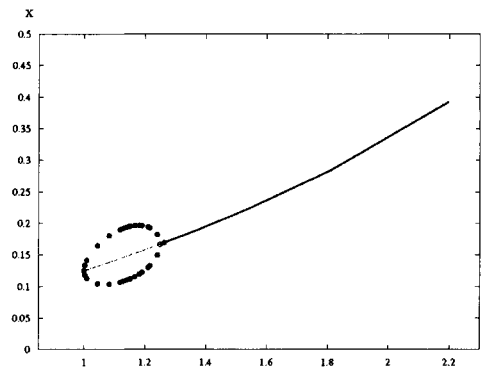


Figure 7.42 Hopf bifurcation of (7.87), and (7.88)

Here, we have two threshold values, T_1 and T_2, and one has to study separately the equilibrium points for the three regions: $x < T_1$, $T_1 \leq x < T_2$, and $x \geq T_2$. Then, one can follow a similar study to the one introduced above, to find possible periodic solutions and their stability properties, and one can also track for bifurcations and special solutions.

We have computed two bifurcations that are clearly shown in Figure 7.41. There is a transcritical bifurcation at $h \approx 2.2$, where two equilibria interchange stability, and a fold bifurcation at $h \approx 2.5$, where one stable equilibria collides with an unstable one. Figure 7.42 shows a Hopf bifurcation, giving out a branch of periodic orbits. These bifurcation diagrams of model (7.87) already show the existence of several interesting solutions with different stability properties and bifurcations. In general, a rich dynamics of the system should be found. We leave most computations and analysis of this and other models as exercises for the reader. Further details are also given in [8].

7.4 FINAL REMARKS AND FURTHER READING

We have presented a short introduction to some basic topics on dynamical systems. We started with linear systems $\dot{x} = Ax$ and the study of their solutions through matrix exponentials and similarity. The concept of stability and the asymptotic behavior of solutions as well as their corresponding invariant subspaces were studied in some detail, using two-dimensional systems as an introductory step.

We have studied nonlinear dynamical systems $\dot{x} = f(x)$ locally and globally, mostly from the theoretical point of view, but we have taken care of introducing some numerical techniques to compute some special solutions. Recall that in general it is not possible to have an explicit or analytical solution of a nonlinear dynamical system, but the theory introduced allows us to study the qualitative behavior of solutions, without explicitly knowing them. At this point, some numerical software becomes essential to help understand the general behavior of solutions.

Linearization has been the central tool for studying nonlinear systems locally, thanks especially to two main theorems: the stable manifold theorem and the Hartman–Grobman theorem, which in both cases, take hyperbolicity as the main assumption. We then have briefly studied bifurcations (mainly as a consequence of losing hyperbolicity), chaos, and connecting orbits, in an effort to understand a nonlinear system beyond their local behavior.

Although we have limited ourselves to continuous dynamical systems, a similar study can be done for discrete systems, as well as for systems with delay and other special properties. There is a long list of excellent references to study these and other very important areas of dynamical systems in great detail, from both, the theoretical and numerical points of view.

For an excellent introduction into these topics, the books by Perko [54] and Wiggins [67] are a great start. For a little more advanced study, besides the article by Holmes [37] and the articles cited within this chapter, see the books by Guckenheimer and Holmes [27], and by Chow et al. [15]. A great theoretical as well as numerical study of dynamical systems can be found in the books by Dieci and Eirola [17], Hale and Koçak [29], and Kuznetsov [43]. For material specifically related to numerical methods for bifurcation problems, see [22].

EXERCISES

7.1 True or False? The origin $x = 0$ is the only equilibrium point of a given linear system $\dot{x} = Ax$.

7.2 Let A be an $n \times n$ matrix. Show that $\|e^A\| \leq e^{\|A\|}$.

7.3 Let A be an $n \times n$, matrix and let $\tau > 0$. Show that the series

$$\sum_{k=0}^{\infty} \frac{A^k t^k}{k!}$$

is uniformly and absolutely convergent for any $|t| \leq \tau$.

7.4 Let v be an eigenvector of a matrix A corresponding to an eigenvalue λ. Show that v is also an eigenvector of e^A, with eigenvalue e^λ.

7.5 Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of a matrix $A_{n\times n}$, and let a real number γ satisfy $\gamma > \max_{j=1,\ldots,n} \operatorname{Re}(\lambda_j)$. Show that

$$\|e^{At}\| \leq Ce^{\gamma t},$$

for some $C > 0$.

Hint: Use Exercise 6.52.

7.6 Assume that all eigenvalues of a matrix $A_{n\times n}$ have negative real parts. Show that there exist positive constants K and α such that every solution $x(t)$ of $\dot{x} = Ax$ satisfies

$$\|x(t)\| \leq Ke^{-\alpha t}, \qquad t \geq 0.$$

That is, all solutions approach zero as $t \to \infty$.

Hint: Use Exercise 7.5.

7.7 Consider the system $\dot{x} = Ax + f(t)$, with f continuous, and assume that there exist constants α, $C > 0$, $t_0 \geq 0$, such that $|f(t)| \leq Ce^{\alpha t}$, for $t \geq t_0$. Show that every solution $x(t)$ of the system satisfies

$$|x(t)| \leq Ke^{\beta t}, \quad t \geq t_0,$$

for some constants β and $K > 0$.

7.8 The following matrices have the same eigenvectors. Sketch the phase portrait of the corresponding system $\dot{x} = Ax$.

(a) $A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$, (b) $A = \begin{bmatrix} 3 & -2 \\ 0 & 1 \end{bmatrix}$, (c) $A = \begin{bmatrix} -3 & 2 \\ 0 & -1 \end{bmatrix}$,

(d) $A = \begin{bmatrix} -1 & 4 \\ 0 & 3 \end{bmatrix}$, (e) $A = \begin{bmatrix} 3 & -4 \\ 0 & -1 \end{bmatrix}$, (f) $A = \begin{bmatrix} 0 & -3 \\ 0 & -3 \end{bmatrix}$.

7.9 Let A and B be two matrices such that $AB = BA$. Show that $e^{A+B} = e^A e^B$.

7.10 Let $A = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}$. Show that

$$e^{At} = e^{\lambda t} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}.$$

Hint: Let $B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, write $A = \lambda I + B$, and use Exercise 7.9.

7.11 Let $A = \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. Show that

$$e^{At} = e^{at} \begin{bmatrix} \cos bt & -\sin bt \\ \sin bt & \cos bt \end{bmatrix}.$$

Hint: First use induction to show that $A^k = \begin{bmatrix} \mathrm{Re}(\lambda^k) & -\mathrm{Im}(\lambda^k) \\ \mathrm{Im}(\lambda^k) & \mathrm{Re}(\lambda^k) \end{bmatrix}$, where $\lambda = a + ib$.

7.12 Let $u(t)$ and $v(t)$ be solutions of the system $\dot{x} = Ax$. Show that any linear combination of $u(t)$ and $v(t)$ is also a solution of the system.

7.13 For the following matrices

$$\text{(a) } A = \begin{bmatrix} 3 & 4 \\ 1 & 3 \end{bmatrix}, \qquad \text{(b) } A = \begin{bmatrix} 5 & 4 \\ 3 & 1 \end{bmatrix},$$

$$\text{(c) } A = \begin{bmatrix} 2 & -2 \\ 4 & 6 \end{bmatrix}, \qquad \text{(d) } A = \begin{bmatrix} 0 & 4 \\ -1 & 0 \end{bmatrix}$$

solve the corresponding linear system $\dot{x} = Ax$, and sketch the phase portraits on both, the (x_1, x_2) and the (y_1, y_2) planes, where as usual $y = P^{-1}x$. Indicate whether the origin is stable, unstable, a node, a focus, or a center.

7.14 Solve the IVP $\dot{x} = Ax$, $x(0) = [c_1 \quad c_2]^T$, and sketch the phase portraits corresponding to the matrices

$$A = \begin{bmatrix} -3 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Indicate whether the origin is stable, unstable, a node, a focus, or a center.

7.15 Consider the system

$$\begin{aligned} \dot{x}_1 &= ax_1 - x_2, \\ \dot{x}_2 &= x_1 + bx_2. \end{aligned}$$

Under which conditions on the parameters a and b is the origin a saddle, a center, or a stable/unstable focus?

7.16 Let A be an $n \times n$ matrix and λ an eigenvalue of A. We define the generalized eigenspace of A corresponding to λ as the set E_λ of all generalized eigenvectors of A associated with λ, together with the zero vector. Show that E_λ is invariant under A, that is, $AE_\lambda \subset E_\lambda$.

7.17 Let A be an $n \times n$ matrix. Show that the subspaces E^s, E^c, E^u are invariant with respect to the system $\dot{x} = Ax$, and that

$$\mathbb{R}^n = E^s \oplus E^c \oplus E^u.$$

Hint: For the first part, use Exercise 7.16.

7.18 Consider the system $\dot{x} = Ax$, where $A = \begin{bmatrix} -4 & 0 & -1 \\ 2 & 6 & 73 \\ 4 & 0 & -4 \end{bmatrix}$. Find the invariant subspaces E^s, E^c, E^u of the system.

7.19 Consider the system $\dot{x} = Ax$, where $A = \begin{bmatrix} -1 & -10 & -10 \\ 0 & 10 & 0 \\ 10 & -11 & -1 \end{bmatrix}$. Find the invariant subspaces E^s, E^c, E^u of the system, and then sketch the phase portrait.

7.20 Find the subspace E^c, and plot the solution curves of the system $\dot{x} = Ax$, where $A = \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix}$.

7.21 Find the subspaces E^s, E^c, E^u, and plot the solution curves of the system $\dot{x} = Ax$, where

$$\text{(a) } A = \begin{bmatrix} -1 & -1 & 2 \\ 1 & -1 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \qquad \text{(b) } A = \begin{bmatrix} -2 & 1 & 2 \\ 0 & 5 & 1 \\ 0 & 5 & 1 \end{bmatrix}.$$

7.22 **(Floquet's theorem).** Consider the linear system $\dot{x} = A(t)x$, where $A(t)$ is a continuous periodic matrix function of period τ, and let $\Phi(t)$ be any fundamental matrix solution of such a system. Show that there exists a constant matrix B and a nonsingular periodic matrix function $P(t)$ of period τ such that

$$\Phi(t) = P(t)\, e^{Bt}.$$

Hint: Use the facts that $\Phi(t+\tau)$ is also a fundamental matrix solution of the same system, so that $\Phi(t+\tau) = \Phi(t)C$, for some constant matrix C, and that there exists a constant matrix B such that $C = e^{B\tau}$.

7.23 Under the hypotheses of Exercise 7.22, show that $x(t)$ is a solution of $\dot{x} = A(t)x$ if and only if $y(t) = P^{-1}(t)x(t)$ is a solution of $\dot{y} = By$.

7.24 Consider the nonhomogeneous system $\dot{x} = A(t)x + f(t)$, with A and f continuous and periodic, with the same period τ. Show that a solution $x(t)$ of such a system is periodic of period τ if $x(\tau) = x(0)$.

7.25 Consider the nonhomogeneous system $\dot{x} = A(t)x + f(t)$, with $A = \begin{bmatrix} 0 & 1 \\ -2 & -4 \end{bmatrix}$ and $f(t) = \begin{bmatrix} 0 \\ \cos t \end{bmatrix}$. Show that the matrix $e^{2\pi A} - I$ is invertible.

7.26 For the matrix A and function f of Exercise 7.25, find a vector $z \in \mathbb{R}^2$ such that

$$z = e^{2\pi A}v + e^{2\pi A}\int_0^{2\pi} e^{-sA} f(s)\, ds.$$

7.27 **(Gronwall's inequality.)** Let $u(t)$ be a nonnegative continuous function such that for positive constants K and C, we have

$$u(t) \le K + C\int_0^t u(s)\, ds,$$

for all $t \in [0,\ a]$. Show that

$$u(t) \le Ke^{Ct},$$

for all $t \in [0,\ a]$.

7.28 Let $E \subset \mathbb{R}^n$ be open and $f : E \to \mathbb{R}^n$. Show that if $f \in C^1(E)$, then f is locally Lipschitz on E.

7.29 Let $E = [0,\ 1] \times [0,\ 1]$. Show that the function $f : E \to \mathbb{R}^2$ defined by $f(x, y) = (x + y,\ xy)$ is Lipschitz.

7.30 Find the stable and unstable manifolds S and U, respectively, of the system

$$\begin{aligned} \dot{x}_1 &= 3x_1 \\ \dot{x}_2 &= -x_1^2 - x_2 \end{aligned}$$

by first finding the solution explicitly.

7.31 Find the stable and unstable manifolds S and U, respectively of the system of Exercise 7.30 by using successive approximations.

7.32 Let $H = I - F$, where $F : \mathbb{R}^n \to \mathbb{R}^n$. Show that if F is a contraction, then H is a homeomorphism from $\mathbb{R}^n$ to $\mathbb{R}^n$.

7.33 Let a and b be positive real numbers, and consider the systems

$$\dot{x} = ax \qquad \text{and} \qquad \dot{x} = bx.$$

Show that these two systems are equivalent by finding a homeomorphism h that maps trajectories of the first system onto trajectories of the second one.

7.34 Consider the system

$$\begin{aligned} \dot{x}_1 &= -2x_1 + x_2^2, \\ \dot{x}_2 &= -x_1 + x + 2. \end{aligned}$$

Show that the system has two equilibria, and find the linearizations of this system around those points. Compute and plot the solutions of both, the linear and nonlinear systems.

7.35 Consider the system

$$\begin{aligned} \dot{x}_1 &= 2x_1 - x_1x_2 + \tfrac{1}{2}x_1^2 \\ \dot{x}_2 &= \tfrac{1}{2}x_2 - x_1x_2. \end{aligned}$$

Show that the origin is hyperbolic and find the linearization of this system around the origin. Compute and plot the solutions of both, the linear and nonlinear systems.

7.36 Consider the following equation for a pendulum with damping:

$$y'' + 2ay' + b^2 \sin y = 0.$$

Show that the origin is asymptotically stable for any $a > 0, b > 0$, by first transforming the equation into a two-dimensional system of first order $\dot{x} = f(x, a, b)$, and then by linearizing around the origin.

7.37 Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -ax_1^2x_2 - x_1. \end{aligned}$$

Find the linearization of this system around the origin, and show that the eigenvalues of the corresponding matrix are purely imaginary for any $a \in \mathbb{R}$. Plot the solutions of the nonlinear system and verify that the origin is attracting for $a > 0$ and repelling for $a < 0$. Does this contradict Hartman–Grobman theorem?

7.38 Consider the Lorenz system (7.64), with $b = 1$ but σ and λ free. Find all the equilibrium points of the system. What stability properties has the origin when $\lambda < 1$? Show that the other two nontrivial equilibria are sinks as long as $1 < \lambda < \frac{\sigma(\sigma+4)}{\sigma-2}$.

7.39 Consider the system

$$\begin{aligned} \dot{x}_1 &= \mu x_1(x_1^2 + x_2^2) + x_2, \\ \dot{x}_2 &= \mu x_2(x_1^2 + x_2^2) - x_1. \end{aligned}$$

First, verify that the eigenvalues of the Jacobian of the system at the origin are purely imaginary for all $\mu \in \mathbb{R}$. Show however that for $\mu < 0$, the origin is asymptotically stable and that for $\mu > 0$ it is unstable.

Hint: Compute the derivative of the square of the distance of any solution $(x_1(t), x_2(t))$ to the origin.

7.40 Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1(1 - x_1^2 - x_2^2) + x_2, \\ \dot{x}_2 &= x_2(1 - x_1^2 - x_2^2) - x_1 \end{aligned}$$

Use polar coordinates (r, θ) to rewrite the system as

$$\begin{aligned} \dot{r} &= r(1 - r^2), \\ \dot{\theta} &= -1. \end{aligned}$$

With $r(0) = r_0$ and $\theta(0) = \theta_0$, show that the solution is given by

$$r(t) = \frac{r_0}{\sqrt{r_0^2 + (1 - r_0^2)e^{-2t}}}, \qquad \theta(t) = -t + \theta_0$$

and that the trajectories approach the unit circle as $t \to \infty$.

7.41 Consider the following predator–prey model with diffusion:

$$\begin{cases} (u_1)_t = a u_1(c - u_1) - \frac{u_1 u_2}{u_1+1} \\ (u_2)_t = (u_2)_{xx} - u_2 + b\frac{u_1 u_2}{u_1+1}. \end{cases} \tag{7.89}$$

Introduce the traveling wave coordinates

$$u_1(t,x) = v_1(x + st) = v_1(z), \qquad u_2(t,x) = v_2(x + st) = v_2(z)$$

and $v_3 = \dot{v}_2$, where the dot denotes derivative with respect to z, to convert (7.89) into a 3-dimensional system $\dot{v} = f(v)$. Compute a branch of periodic orbits (periodic traveling wave solutions) for $a = 2, b = 4, c = 3$ by setting the speed wave s as a free parameter.

7.42 Show that $\Gamma(t) = (0, \cos t, \sin t)$ is a periodic solution of

$$\begin{aligned} \dot{x}_1 &= -x_1 x_2^2 - x_1 x_3^2, \\ \dot{x}_2 &= x_1^2 x_2 - x_3, \\ \dot{x}_3 &= x_2 + x_1^2 x_3 \end{aligned}$$

and find the Floquet multipliers of the periodic orbit.

7.43 Let D be a closed and bounded set in $\mathbb{R}^n$ and for f defined and continuous on $[0,\ \tau] \times D$, for some $\tau > 0$, consider the system

$$\begin{aligned} \dot{x} &= f(t,x), \\ f(t,x) &= f(t+\tau,x). \end{aligned}$$

Assume that for sufficiently small $\epsilon > 0$, $x + \epsilon f(t,x) \in D$, for all t and all x on the boundary of D. Show that the system has a periodic solution of period τ.

7.44 Consider the system

$$\begin{aligned} \dot{x} &= A(t)x + f(t), \qquad a \le t \le b \\ B_a x(a) &+ B_b\, x(b) = c, \end{aligned} \tag{7.90}$$

with solution

$$x(t) = \Phi(t)Q^{-1}c + \int_a^b G(t,s)f(t)\,ds$$

satisfying $\|x\|_\infty \le k_1\|c\|_\infty + k_2\|f\|_1$, where

$$k_1 = \|\Phi(t)Q^{-1}\|_\infty, \qquad \text{and} \qquad k_2 = \sup_{a\le t,\, s\le b} \|G(t,s)\|_\infty.$$

Show that k_1, k_2 and the function G are independent of the fundamental matrix solution $\Phi(t)$.

7.45 Consider the following perturbation of the system (7.90):

$$\begin{aligned} \dot{y} &= A(t)y + f(t) + g(t), \qquad a \le t \le b, \\ B_a y(a) &+ B_b\, y(b) = c + d. \end{aligned} \tag{7.91}$$

Define $e(t) = x(t) - y(t)$ and show that $\|e\|_\infty \le k_1\|d\| + k_2\|g\|_\infty$.

Hint: First consider the system satisfied by $e(t)$ by combining (7.90) and (7.91).

7.46 Consider the Kuramoto–Sivashinsky equation

$$u_t + u_{xxxx} + u_{xx} + uu_x = 0,$$

which models spatio-temporal evolution of a flame front. Take the change of variables $u(x,t) = v(x) - s^2 t,\ y = v'$ to get

$$y''' + y' + \frac{1}{2}y^2 - s^2 = 0.$$

Now, as usual, write this last equation as a system $\dot{u} = f(u)$. For the traveling wave speed $s = 1$, the system has two (hyperbolic) periodic orbits. Compute those and find the dimension of the corresponding stable and unstable manifolds.

7.47 Prove Theorem 7.25.

7.48 Consider the system

$$\begin{aligned} \dot{x}_1 &= 0.2x_1 + x_2, \\ \dot{x}_2 &= -x_1 - x_3, \\ \dot{x}_3 &= x_2x_3 - \mu x_3 + 0.2. \end{aligned}$$

For $\mu = 2.2$, compute the solution starting at $(0.6472, 3.7669, 1.9206)$ to get a periodic orbit of period $\tau_1 \approx 5.765$. For $\mu = 3.1$, start the solution at $(-1.9341, 4.6617, 0.3610)$ and now you will see a periodic orbit of period $\tau_2 = 2\tau_1$. This process repeats as μ increases. Such a phenomenon is known as period doubling bifurcation.

7.49 Consider the predator–prey system (7.87), (7.88). Find the equilibrium points and their stability properties. Then compute a branch of periodic orbits that appear at the Hopf bifurcation.

7.50 Consider the following ratio-dependent predator–prey model with linear harvesting on the prey:

$$\begin{aligned} \dot{x} &= x(1-x) - \tfrac{axy}{y+x} - hx, \\ \dot{y} &= y\left(-d + \tfrac{bx}{y+x}\right). \end{aligned} \tag{7.92}$$

(a) Find the equilibrium points of the system and determine the values of the parameters to guarantee that these points represent coexistence of both species

(b) Set $a = 2, b = 1, d = 0.75$, and for increasing values of h, starting at $h = 0.3126$, compute a branch of connecting orbits from the equilibrium $(1 - h, 0)$ to a branch of (stable) periodic orbits.

7.51 Consider the system

$$\begin{aligned} \dot{u}_1 &= u_1 + \beta v_1 - u_1^3 - 3u_1u_2^2 - u_1(v_1^2 + v_2^2) - 2v_1v_2u_2, \\ \dot{v}_1 &= -\beta u_1 + v_1 - v_1^3 - 3v_1v_2^2 - v_1(u_1^2 + u_2^2) - 2u_1u_2v_2, \\ \dot{u}_2 &= (1-2\lambda)u_2 + (\beta - 2\lambda)v_2 - u_2^3 - 3u_1^2u_2 - u_2(v_1^2 + v_2^2) - 2v_1v_2u_1, \\ \dot{v}_2 &= -(\beta + 2\lambda)u_2 + (1-2\lambda)v_2 - v_2^3 - 3v_1^2v_2 - v_2(u_1^2 + u_2^2) - 2u_1u_2v_1, \end{aligned}$$

where $\beta = 0.55$ and λ is a free parameter. This system has the following periodic orbits:

$$y_-(t) = (\,0,\ 0,\ \rho(t)\cos\theta(t),\ \rho(t)\sin\theta(t)\,), \qquad y_+(t) = (\cos\beta t,\ -\sin\beta t,\ 0,\ 0),$$

where

$$\dot{\rho} = (1 - 2\lambda - 2\lambda\sin 2\theta)\rho - \rho^3\,, \quad \dot{\theta} = -\beta - 2\lambda\cos 2\theta\,.$$

Compute a branch of connecting orbits from y_- to y_+ starting at $\lambda = 0.1$, and find the dimension of the corresponding stable and unstable manifolds. To plot the solutions in three dimensions, use the coordinates $(u_2, v_2, \sqrt{u_1^2 + v_1^2})$.

7.52 Experiment the sensitive dependence of solutions on the initial conditions for the Lorenz system (7.69), with the same value parameters $\sigma = 10, \lambda = 28, b = 8/3$, by first using the initial condition $x(0) = (0,\ 0.01,\ 0)$ and then $x(0) = (-0.01,\ -0.01, 0)$. Plot and compare both solutions.

7.53 Consider the system

$$\begin{aligned} \dot{x}_1 &= -x_2 - x_3, \\ \dot{x}_2 &= x_1 + ax_2, \\ \dot{x}_3 &= b + x_3(x_1 - c). \end{aligned}$$

Take $a = 0.15, b = 0.2$, and $c = 10$, and compute the corresponding solution starting at $x(0) = (10,\ 0,\ 0)$. You should obtain the so–called Rössler attractor.

7.54 Consider the one-dimensional dynamical system $\dot{x} = f(x, \lambda)$, with $\lambda \in \mathbb{R}$, and assume that for some $(x_0, \lambda_0) \in \mathbb{R} \times \mathbb{R}$, the following conditions hold:

$$f(x_0, \lambda_0) = 0, \qquad \frac{\partial f}{\partial x}(x_0, \lambda_0) = 0, \qquad \frac{\partial^2 f}{\partial x^2}(x_0, \lambda_0) \neq 0, \qquad \frac{\partial f}{\partial \lambda}(x_0, \lambda_0) \neq 0.$$

Show that $\dot{x} = f(x, \lambda)$ undergoes a saddle-node bifurcation at $\lambda = \lambda_0$.

Hint: Using the first and last conditions above, apply the implicit function theorem to show that there is a smooth curve $\lambda = \lambda(x)$ of equilibria, then show that its graph is concave.

7.55 Consider the one-dimensional equation

$$\dot{x} = f(x, a, b) = a + bx - x^3.$$

(a) Fix $a = 0$ and study the number of equilibria and their stability for $b > 0$ and for $b \leq 0$. Verify that this gives a pitchfork bifurcation.

(b) Solve simultaneously the equations $f(x, a, b) = 0$ and $\frac{\partial f}{\partial x} f(x, a, b) = 0$ to get both a and b in terms of x and then combine the expressions for a and b to get $27a^2 = 4d^3$. Plot this on the (a, b) plane. The curve you obtain represents a cusp bifurcation.

7.56 Consider the predator–prey system

$$\begin{aligned} \dot{x}_1 &= 2x_1(1 - x_1) - \tfrac{2x_1x_2}{\mu + x_1}, \\ \dot{x}_2 &= -x_2 + \tfrac{2x_1x_2}{\mu + x_1}. \end{aligned}$$

Show that $(\mu,\ 2\mu(1 - \mu))$ is an equilibrium point, and that for $\mu = 1/3$, the Jacobian at that equilibrium has purely imaginary eigenvalues. Then plot the solutions for $\mu < 1/3$, $\mu = 1/3$, and $\mu > 1/3$ to verify that a Hopf bifurcation occurs.

7.57 Consider the system

$$\begin{aligned} \dot{x}_1 &= x_1(1 - x_1), \\ \mu\dot{x}_2 &= x_1 - x_2. \end{aligned}$$

Find the equilibrium points and their stability properties. Plot the phase portraits for $\mu = 0$ and $\mu = 2$ to see a drop on the dimensions. This is a singularity bifurcation.

7.58 Consider the system

$$\begin{aligned} \dot{x}_1 &= -x_1 \sin\mu - x_2 \cos\mu + (1 - x_1^2 - x_2^2)^2(x_1 \cos\mu - x_2 \sin\mu), \\ \dot{x}_2 &= x_1 \cos\mu - x_2 \sin\mu + (1 - x_1^2 - x_2^2)^2(x_1 \sin\mu + x_2 \cos\mu). \end{aligned}$$

Transform this system to polar coordinates and verify that for $\mu < 0$ there are no periodic orbits because $\dot{r} > 0$, that for $\mu = 0$, the unit circle is an unstable (saddle) periodic orbit, and that for sufficiently small $\mu > 0$, there are two periodic orbits, one stable, with radius less than 1, and one unstable, with radius greater than 1. Plot the solutions for these three cases. This phenomenon is known as saddle-node bifurcation of periodic orbits.

7.59 Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_1 - x_1^2 + \mu x_2. \end{aligned}$$

Compute and plot the solutions for $\mu = -1, \mu = 0$, and $\mu = 1$. You will see a so-called homoclinic bifurcation: For $\mu = 0$ there is a homoclinic orbit to the origin, and a set of periodic orbits around $(1,\ 0)$, but the loop and the orbits are broken for $\mu \neq 0$.

7.60 Consider the system

$$\begin{aligned} \dot{x}_1 &= 1 - x_1^2 + \mu x_1 x_2. \\ \dot{x}_2 &= x_1 x_2 + \mu(1 - x_1^2). \end{aligned}$$

The system has two equilibria, $(-1, 0)$ and $(1, 0)$. Are these equilibria hyperbolic? Compute and plot the solutions for $\mu = -1$, $\mu = 0$, and $\mu = 1$ around the equilibrium points. A so–called heteroclinic bifurcation happens at $\mu = 0$, when the two equilibria are connected.

CHAPTER 8

MATHEMATICAL MODELS

In the previous chapters, we have illustrated the importance and applicability of several of the mathematical concepts and methods introduced in this book by studying some mathematical models, in some cases briefly, and in some other cases, in great detail. Those applications and mathematical models include composite Bezier curves and postscript fonts, Google's PageRank, information retrieval, image compression, and predator–prey models with harvesting.

In this chapter, we introduce three mathematical models to further illustrate the process and importance of mathematical modeling to solve concrete problems. The first and second models, although quite different, fall within the area of optimization, the third model deals with an infection disease as a dynamical system, and finally on our last model, we consider a generalization of two interacting species models, where the interspecific relationships can change from mutualism to competition to host-parasite as certain parameters change.

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

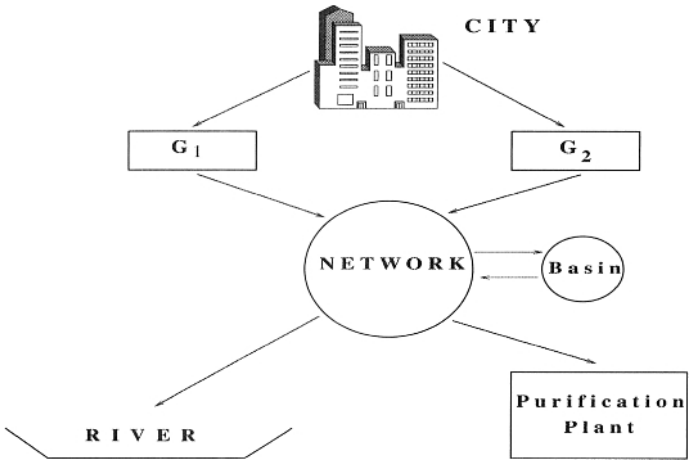


Figure 8.1 Classic waste water management.

8.1 OPTIMIZATION OF A WASTE MANAGEMENT SYSTEM

In every city, region, or country of the world, authorities always pay great attention to dealing with waste water problem. They look for optimal ways of collecting and transporting waste water into purification plants in order to avoid the polluted water flowing into the rivers, lakes, or the sea.

We develop a linear programming model to optimize a waste water management system, which aims to minimize the outflow of waste water to a river or lake, to keep a constant inflow to a purification plant, and to use storage pipes uniformly. We use computer simulations to show the behaviors of the system with various rainfall inputs. The method described in this section could be a starting point and a useful reference for the design of a waste water management system.

8.1.1 Background

We assume a classic system of waste water management is in place, and we aim at developing a more efficient system in terms of cost and lower quantities of waster water flowing into the river.

The Classic Method.

Figure 8.1 shows a typical method of handling waste water. G_1 and G_2 are two regions where waste water is collected. Waste water flows into a network of pipes and finally outflows either to a purification plant or to a river. When the amount of water in the network is more than the total capacities of the pipes, part of the water is sent to a storage

basin, which is usually located near the network. When the inflow of water decreases, the water collected in the storage basin is then sent back to the system.

This simple system works satisfactorily, but there are some disadvantages:

(1) The pipes are only transport pipes. Their capacities are not used for any other purposes.

(2) The storage basin itself is in fact a problem, especially for two reasons: first, it is expensive to build such a basin, and the costs of maintenance are very high; second, such a basin is usually open to the environment, which may cause environmental pollution.

(3) The system as such does not allow control of the flow of waste water into the river.

The New Method

The problems mentioned above motivate us to look for an alternative method. This new method should aim at controlling the flow of water within the network in such a way that the storage basin is no longer needed, and it is possible to control outflow to the river. The main idea is to install branching points and gates at specific points of the pipes, which are regulated appropriately so that the flow in the system can be controlled. In this way, some of the pipes become able to store water on one side of the gates, and their capacities are supposed to be used properly for that purpose.

8.1.2 Description of the system

We assume we can divide the city or the region into two regions, D_1 and D_2, where waste water is collected separately, and the volume of waste water collected in one of the regions is typically higher than that of the other one, say, volume of water collected in D_1 is higher than that in D_2.

Next, we introduce some notation. We will denote the pipes of the system as K_i, with determined capacities C_i, $i = 3, \ldots, 7$. We assume that K_3, K_4, K_5 are simply transport pipes, and that K_6, K_7 can also be used as storage pipes. We remark that each of these pipes $K_3, \ldots, K_7$ could in practice represent a set of pipes. With Z_i, , $i = 1, 2, 3$, we will denote the branching points of the system. The gates installed in the pipes are denoted with T_1 and T_2. W will denote the water purification plant and R the river.

Waste water enters the network system through the pipes K_3 and K_4. Under normal conditions, water from K_3 flows into pipe K_6, but if there is more water in K_6 than in K_7, some amount of water in K_3 must be sent to K_7 via pipe K_5, by regulating the branching point Z_1. Water from K_4 flows directly into K_7. See Figure 8.2. At the other end of the network, by regulating the gates T_1 and T_2, it becomes possible to control the volume of water out of the system, and especially we aim at minimizing the flow of waste water into the river.

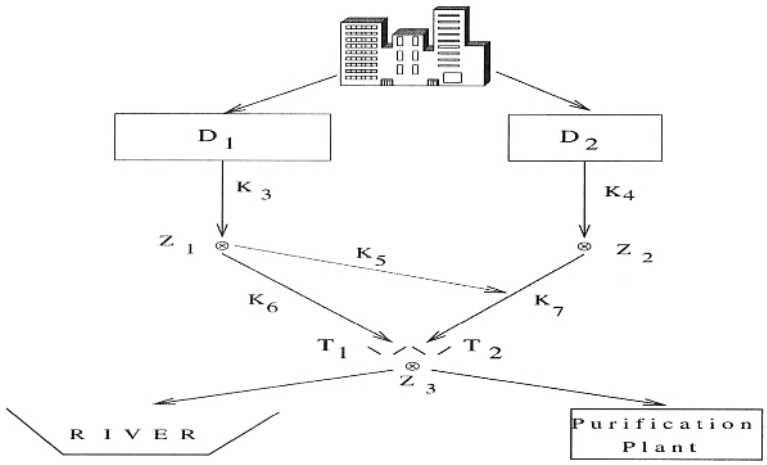


Figure 8.2 Proposed waste water management.

The main objectives on this model are:

(1) The flow of waste water into the river Q_R should be minimal.

(2) The flow of waste water into the purification plant Q_W should be constant.

(3) The capacities of the storage pipes K_6 and K_7 should be used uniformly.

8.1.3 Development of the mathematical model

We are assuming that region D_1 collects more water than D_2, due for instance to the fact that the area of D_1 is greater than that of D_2, or that region D_1 is mainly an industrial zone. Each of these regions receives industrial and domestic waste water as well as rainwater. For the system it is assumed that industrial and domestic water volumes are constant with respect to time, whereas the rainfall is a function that does depend on time.

Let us further introduce the following notations, with volume of water measured in cubic meters:

$V_i(t)$, $i = 1, 2$, the total water collected (industrial, domestic, and rain water) in region R_i at time t, say, for every 5 minutes.

$Q_i(t)$, $i = 3, \ldots, 7$, the total amount of water in pipe K_i at time t, for every 5 minutes.

$Q_R(t)$, the amount of water into the river at time t, for every 5 minutes.

Q_W, the amount of water into the purification plant at time t, for every 5 minutes,

δ_i = time delay of water flow for pipe K_i, in minutes, where without loss of generality, we can take

$$\delta_j = n_j \Delta t, \quad n_j \in \mathbb{N}, \quad \Delta t \in \mathbb{R}^+.$$

Assume for instance that a certain amount of water collected at D_1 needs 20 minutes to reach gate T_1 through K_6, and similarly for water collected at D_2 to reach gate T_2 through K_7. Here we choose the following time delays, in minutes:

$$\delta_4 = \Delta t = 5, \quad \delta_3 = \delta_5 = \delta_6 = 2\Delta t, \quad \delta_7 = 3\Delta t.$$

We also impose the condition that

$$Z_1(t), T_1(t), T_2(t) \in [0, 1]. \tag{8.1}$$

More clearly, $Z_1(t) = 0$ means all of the water in K_3 at time t flows into the transport pipe K_5, and in other extreme ($Z_1(t) = 1$), all of the water flows down to the storage pipe K_6. Also, $T_i(t) = 0$, $i = 1, 2$ means that gate T_i is completely closed, and completely open if $T_i(t) = 1$. Thus, $T_i(t) = \alpha \in [0, 1]$ means that 100α percent of the water will pass the gate, e.g., when $T_i(t) = \frac{1}{2}$, half of the water stored in the corresponding pipes will pass the gate.

8.1.4 Building the objective function

With the assumptions and notations above, we can now introduce the functions that will allow us to define our objective function.

At Z_1:

$$\begin{aligned} Q_3(t) &= V_1(t - \delta_3) = V_1(t - 2). \\ Q_5(t) &= [1 - Z_1(t)]V_1(t - \delta_3) = [1 - Z_1(t)]V_1(t - 2). \end{aligned}$$

At Z_2:

$$Q_4(t) = V_2(t - \delta_4) = V_2(t - 1).$$

At T_1:

$$\begin{aligned} Q_6(t) &= Z_1(t - \delta_6)Q_3(t - \delta_6) + [1 - T_1(t - 1)]Q_6(t - 1). \\ &= Z_1(t - 2)V_1(t - 4) + [1 - T_1(t - 1)]Q_6(t - 1). \end{aligned} \tag{8.2}$$

At T_2:

$$\begin{aligned} Q_7(t) &= Q_5(t - \delta_5 - \delta_7) + Q_4(t - \delta_7) + [1 - T_2(t - 1)]Q_7(t - 1). \\ &= [1 - Z_1(t - 5)]V_1(t - 7) + A_2(t - 4) + [1 - T_2(t - 1)]Q_7(t - 1). \end{aligned} \tag{8.3}$$

Finally, at Z_3 we have

$$Q_R(t) = T_1(t)Q_6(t) + T_2(t)Q_7(t) - Q_W. \tag{8.4}$$

Equation (8.4) defines the objective function of the system, which has to be minimized.

In such an objective function, $Q_6(t)$ and $Q_7(t)$ are obtained from equations (8.2) and (8.3). This implies that $Z_1(t)$, needs to be computed. To this end, recall that the storage pipes K_6 and K_7 should be used uniformly. Then, we have

$$\left| \frac{Q_7(t)}{Q_6(t)} - \frac{C_7}{C_6} \right| = \min. \tag{8.5}$$

From equation (8.5), it is possible to compute a unique $Z_1(t) \in [0, 1]$. As a result, equation (8.4) gives a linear objective function in terms of $T_1(t)$ and $T_2(t)$.

8.1.5 Building the constraints

The given system must satisfy a certain set of constraints, namely,

(1) $T_1(t), T_2(t)$ must satisfy relation (8.1); that is,

$$0 \leq T_1(t) \leq 1, \qquad 0 \leq T_2(t) \leq 1. \tag{8.6}$$

(2) Accumulations of water in storage pipe K_6 should not exceed its maximum capacity; that is,

$$Q_6(t) - T_1(t)Q_6(t) \leq C_6.$$

or,

$$[1 - T_1(t)]Q_6(t) \leq C_6. \tag{8.7}$$

(3) Similarly, for storage pipe K_7:

$$[1 - T_2(t)]Q_7(t) \leq C_7. \tag{8.8}$$

(4) The system output should not be greater than the sum of system input and water storage:

$$\begin{aligned} T_1(t)Q_6(t) + T_2(t)Q_7(t) \quad & \leq V_1(t-4) + V_2(t-4) + [1 - T_1(t-1)]Q_6(t-1) \\ & +[1 - T_2(t-1)]Q_7(t-1). \end{aligned} \tag{8.9}$$

Therefore, the mathematical model of the system can be summarized as follows:

$$\text{Minimize} \quad Q_R = Q_6(t)T_1(t) + Q_7(t)T_2(t) - Q_W \tag{8.10}$$

subject to

$$\begin{aligned} & 0 \leq T_1(t) \leq 1, && 0 \leq T_2(t) \leq 1, \\ & [1 - T_1(t)]\, Q_6(t) \leq C_6, && [1 - T_2(t)]Q_7(t) \leq C_7, \\ & T_1(t)Q_6(t) + T_2(t)Q_7(t) \quad \leq V_1(t-4) + V_2(t-4) + [1 - T_1(t-1)]Q_6(t-1) \\ & \qquad\qquad\qquad\qquad\qquad\qquad +[1 - T_2(t-1)]Q_7(t-1). \end{aligned} \tag{8.11}$$

8.1.6 Numerical experiments

Using some realistic data, numerical computations were run to test our model. We necessarily give the storage pipes K_6 and K_7 certain capacities that eventually should not be able to hold all the waste water coming into the network, especially because of the unexpected

heavy rains. This way, our model shows moments when some waste water will flow to the river.

Let a_1, a_2 denote the total industrial and domestic waste water that can be considered as constant inputs, in regions D_1 and D_2, respectively, and let $b_1(t)$, $b_2(t)$ denote the precipitations in regions D_1 and D_2 respectively. Then, the total water inputs are

$$V_1 = a_1 + b_1(t), \quad V_2 = a_2 + b_2(t).$$

Precipitations with respect to the rainfall depth in the 2 regions D_1 and D_2 are usually given as

$$b_1 = 300\rho h_1, \qquad b_2 = 200\rho h_2,$$

where h_i, $i = 1, 2$ is the depth of rainfall (mm), and $\rho \in [0,1]$ describes the loss of water due to evaporation, absorption, and so on. A typical value is $\rho = 0.6$. We can use the approximations:

$$b_1(t) = 540 \exp\left(\frac{-(t-10)^2}{2 \times 1^2}\right), \quad t \in [0, 19],$$

$$b_1(t) = 136 \exp\left(\frac{-(t-30)^2}{2 \times 4^2}\right), \quad t \in [20, 40],$$

and

$$b_2(t) = 360 \exp\left(\frac{-(t-20)^2}{2 \times 2.5^2}\right), \quad t \in [0, 40].$$

The whole simulation is carried out for 40 time steps, i.e. we have a 200-minute experiment. In region D_1, a very heavy rainfall occurs at the 10-th time step, and this degree of rain lasts for 10 minutes; another degree of rainfall is around the 30th time step lasting for 40 minutes. In region D_2, the very heavy rainfall comes at the 20th time step lasting for 25 minutes.

Tables 1 and 2 give some data used in the computations.

a_1 (m^3/5 min)	a_2 (m^3/5 min)	Q_{KA} (m^3/5 min)	C_3 (m^3)	C_4 (m^3)	C_5 (m^3)	C_6 (m^3)	C_7 (m^3)
50	25	150	700	500	100	700	600

Table 1. Constant water input and capacities

degrees of rain	depth (mm/5 min)	b_1 (m^3/5 min)	b_2 (m^3/5 min)	time duration (min)
very heavy	3	540	360	5
heavy	1.5	270	180	15
weak	0.2	36	24	

Table 2. Precipitation and rainfall depth.

The main observations regarding this experiment are:

(1) For $t \in [1, 11]$, there is not enough water in the system to keep the full load of the purification plant (as both $Q_{R\min}$ and $Q_{R\max}$ are negative, the actual outflow Q_W is less than $150\,m^3/5\,\text{min}$). In this case, we let both gates T_1 and T_2 completely open, i.e., $T_1 = T_2 = 1$.

(2) For $t \in [12, 22]$, although there is a short, heavy rainfall in region D_1, and it also starts raining in region D_2, there is a time delay, and afterward this amount of water is partly stored in the storage pipes K_6 and K_7, which does not result in overflow to the river (Q_R is still zero). Meanwhile, since K_6 and K_7 are gradually full of water, gates T_1 and T_2 are regulated from completely open state to partly close state in order to keep a constant outflow to the purification plant.

(3) For $t \in [23, 30]$, water stored in the system has been out of the maximum capacities of the storage pipes K_6 and K_7 (especially in K_7; therefore, some water has to be overflowed to the river ($Q_R > 0$), besides a constant outflow to the purification plant. This is the case we do not expect. As there is more water in K_7 than in K_6, gate T_1 has to be closed for a while in order to release the load of K_7.

(4) For $t \in [31, 34]$, even though there is a light rain in region D_1, the input delay and the storage in K_6 and K_7 play a role in keeping no overflow to the river.

(5) For $t \in [35, 37]$, the delayed rainfall in region D_1 causes the storage in K_6 and K_7 out of their capacities, so some water flows to the river.

(6) For $t \in [38, 40]$, the case is similar to that in (4).

See Tables 3 and 4. In particular, observe in Table 3 how the value of $Z_1(t)$ varies as the relationship between the volumes in K_6 and K_7 changes. Similarly observe in Table 4 how the values of T_1 and T_2 vary from zero to one according to the collected volumes of water. Also observe that some waste water flows to the river ($Q_R > 0$), at the time the maximum capacity of one of the storage pipes is reached. For more details on this model, see [30].

$No.$	Z_1	Q_5	Q_6	Q_7	$Q_{R\max}$	$Q_{R\min}$
1	.7500	12.5000	100.0000	75.0000	25.0000	−150.0000
3	1.0000	.0000	37.5000	50.0000	−62.5000	−150.0000
4	1.0000	.0000	50.0000	50.0000	−50.0000	−150.0000
5	.5962	20.1891	50.0000	25.5522	−74.4478	−150.0000
6	.5962	20.1891	50.0000	38.0522	−61.9478	−150.0000
7	.5962	20.1899	29.8109	25.5522	−94.6370	−150.0000
8	1.0000	.0000	29.8109	25.5522	−94.6370	−150.0000
9	1.0000	.0000	29.8121	25.5522	−94.6358	−150.0000
10	1.0000	.0000	50.1811	45.7413	−54.0775	−150.0000
11	.7351	100.0000	55.9989	45.7413	−48.2598	−150.0000
12	.8305	100.0000	123.0811	45.7421	18.8232	−150.0000
13	.7351	100.0000	294.0654	27.8365	171.9019	−150.0000
14	.1875	100.0000	659.3483	28.1057	537.4540	−150.0000
15	.0000	55.9989	812.1209	28.4118	690.5327	−37.8791
16	.0000	50.1811	710.6873	130.0778	690.7651	−139.3127
17	.0000	50.0020	674.9150	147.9930	672.9080	−150.0000
18	.0000	50.0000	654.2357	163.8808	668.1165	−150.0000
19	.0000	50.0000	646.6573	195.1798	691.8372	−150.0000
20	.0000	50.0000	646.6573	226.2721	722.9294	−150.0000
21	.0000	50.0000	646.6573	326.6841	823.3414	−150.0000
22	.0000	55.9754	646.6573	513.0997	1009.7571	−150.0000
23	1.0000	.0000	646.6573	770.4216	1267.0790	20.4216
24	1.0000	.0000	646.6573	1035.0000	1531.6573	285.0000
25	1.0000	.0000	707.4774	1007.3219	1564.7993	264.7993
26	1.0000	.0000	768.4056	936.4137	1554.8193	254.8193
27	1.0000	.0000	779.4120	856.2062	1485.6183	185.6183
28	1.0000	.0000	794.1527	725.0934	1369.2461	69.2462
29	.7661	35.7054	812.2653	673.7207	1335.9861	35.9860
30	.4118	100.0000	832.4882	645.2085	1327.6967	27.6967
31	.4500	100.0000	816.9528	632.1428	1299.0957	−.9044
32	.4624	100.0000	770.0196	626.2470	1246.2666	−53.7334
33	.8488	27.4859	737.3718	616.2627	1203.6345	−96.3655
34	.9611	6.6071	710.5706	640.3215	1200.8921	−99.1079
35	1.0000	.0000	755.2219	725.5522	1330.7742	30.7741
36	.3639	84.2698	863.4125	725.5522	1438.9647	138.9647
37	.2505	84.1409	852.6582	725.5522	1428.2104	128.2104
38	.2250	72.9676	748.2184	653.0381	1251.2565	−48.7435
39	1.0000	.0000	697.4211	614.1190	1161.5402	−135.8810
40	1.0000	.0000	639.6386	568.6388	1058.2773	−150.0000

Table 3. Numerical results, part I.

No.	V_1	V_2	T_1	T_2	Q_R
1	50.0000	25.5522	1.0000	1.0000	.0000
2	50.0000	25.5522	1.0000	1.0000	.0000
3	50.0000	25.5522	1.0000	1.0000	.0000
4	50.0000	25.5522	1.0000	1.0000	.0000
5	50.0020	25.5522	1.0000	1.0000	.0000
6	50.1811	25.5522	1.0000	1.0000	.0000
7	55.9989	25.5522	1.0000	1.0000	.0000
8	123.0811	25.5522	1.0000	1.0000	.0000
9	377.5266	25.5522	1.0000	1.0000	.0000
10	590.0000	25.5522	1.0000	1.0000	.0000
11	377.5266	25.5522	1.0000	1.0000	.0000
12	123.0811	27.1514	.8656	.9501	.0000
13	55.9989	32.1428	.4241	.9083	.0000
14	50.1811	45.2085	.1892	.8983	.0000
15	50.0020	73.7207	.1533	.8970	.0000
16	50.0000	125.0934	.0503	.8781	.0000
17	50.0000	200.2308	.0306	.8738	.0000
18	50.0000	286.4137	.0116	.8691	.0000
19	50.0000	357.3219	.0000	.7685	.0000
20	55.9754	385.0000	.0000	.6629	.0000
21	60.8201	357.3219	.0000	.4592	.0000
22	68.4056	286.4137	.0000	.2923	.0000
23	79.4121	200.2308	.0000	.2212	20.4216
24	94.1527	125.0934	.0000	.4203	285.0000
25	112.2653	73.7207	.0106	.4044	264.7993
26	132.4882	45.2085	.0890	.3593	254.8193
27	152.6582	32.1428	.1019	.2992	185.6183
28	170.0196	27.1514	.1186	.1725	69.2462
29	181.8157	25.5522	.1382	.1094	35.9860
30	186.0000	25.5522	.1591	.0701	27.6967
31	181.8157	25.5522	.1432	.0523	.0000
32	170.0196	25.5522	.1487	.0567	.0000
33	152.6582	25.5522	.1530	.0604	.0000
34	132.4882	25.5522	.1544	.0630	.0000
35	112.2653	25.5522	.0731	.1730	30.7741
36	94.1527	25.5522	.1893	.1730	138.9647
37	79.4121	25.5522	.1790	.1730	128.2104
38	68.4056	25.5522	.1055	.1088	.0000
39	60.8201	25.5522	.1132	.1157	.0000
40	55.9754	25.5522	.1235	.1248	.0000

Table 4. Numerical results, part II.

8.2 GROUPING PROBLEM IN NETWORKS

A common layout in factories of different kinds, is to have a set of automatic machines located in a very large room, grouped according to certain properties or rules, and a set of workers that are in charge of one or several machines, making sure that the machines are running properly and being ready to fix any sudden problem, so that the idle time is minimal.

In this section, we consider an actual setup of 159 machines that are grouped in sets, and we are interested in finding ways to regroup the machines so that the total idle time of machines

in each group should be minimal. We need to take into consideration that each machine performs a given task with a certain degree of difficulty, which varies from machine to machine.

8.2.1 Background

The case study in view comes from a manufacturing company involved in some kind of weaving to produce items needed at hospitals. They have a network of 159 machines that must be visited by the workers during a shift period. The machines are already fixed to their places, and production has long begun. Relocation of machines in the network is of no interest. The machines are currently organized into eight groups, and each group is manned by one worker. The grouping up to now is not based on any standard analytical format.

To every machine is assigned a type of material. This assignment is associated with some difficulty called the *lastgrad* (LG), which the worker experiences. This LG number incorporates all the predictable problems the worker encounters while doing his work, and such a number has been determined by a standard organization. It is interesting to note that the sum of the LGs in a group determines how much the worker should earn: the higher the LG, the higher the worker's earning. Ironically however, the higher the LG, the less efficient the worker becomes. In addition, to justify a period's pay, there is a minimum LG allowable in a group.

As the material moves on to different machines in the production process , the material (e.g. ,thread) may get entangled in one of the machines thereby causing defective production. An electromechanical device is installed to alert about this problem. The device gives a signal of faults and stops the machine automatically, and this obviously affects the production in the entire group. The major task of the worker is to continuously move around the machines in his/her assigned group during the shift period to offset the faults and keep the machines running as efficiently as possible.

The objective is to develop a mathematical model that:

- Regroups the machines, without actually moving them, to ensure that total LG in the groups is homogeneous.
- Minimizes the distance traveled .

We consider two possible approaches to achieving the objective:

- The N-median Approach
- The Probabilistic Approach

8.2.2 The N-median approach

The N-median problem is the problem of locating N "facilities" relative to a set of "customers" such that the sum of the shortest demand weighted distance between "customers" and "facilities" is minimized. In our case, the N-median problem is that of locating N facilities, called medians (the workers), with respect to m existing facilities (the machines), so as to minimize a sum of weighted distances between each existing facility and a closest median.

We make the following assumptions:

- The 8 workers are the new facilities (medians) to be located .
- The 159 machines are the m existing facilities .
- Every machine's position is represented by a vertex of the network .
- The LGs are the weights on the machines

8.2.2.1 Building the objective function Let $Y = \{y_1, y_2, ..., y_8\}$ be a set of 8 medians to be located. For $i = 1, \ldots, 159$, let V_i and L_i denote the vertex corresponding to machine i, and its associated LG, respectively. Let also $d(Y, V_i)$ be the shortest distance between vertex V_i and the nearest median.

Then the problem is

$$\text{Minimize} \quad f(Y) = \sum_{i=1}^{159} L_i d(Y, V_i).$$

We remark that it suffices to consider only vertex locations in order to find an absolute N-median. There are two aspects of the N-median :

- Selecting vertices to be median location
- Assigning vertices to the medians

This implies that we must define two decision variables .

$$y_j = \begin{cases} 1 & \text{if vertex } V_j \text{ is selected as median,} \\ 0 & \text{otherwise,} \end{cases}$$

$$X_{ij} = \begin{cases} 1 & \text{if vertex } V_i \text{ is assigned to median } V_j, \\ 0 & \text{otherwise.} \end{cases}$$

8.2.2.2 Building the constraints There are some constraints associated with this problem. Since we cannot have more than eight medians in our problem, we must have

$$\sum_{j=1}^{159} y_j \leq 8.$$

Also, we can only assign vertices to medians that have been selected; that is,

$$X_{ij} \leq y_j \quad \forall i,j, \quad \text{or} \quad -X_{ij} + y_j \geq 0, \quad \forall i,j.$$

To ensure that every vertex is assigned to only one median, we have

$$\sum_{j=1}^{159} X_{ij} = 1, \qquad \forall i.$$

Putting everything together, we obtain the optimization problem formulation :

$$\text{Minimize} \qquad Z = \sum_{j=1}^{159} \sum_{i=1}^{159} L_i d_{ij} X_{ij}$$

subject to

$$\sum_{j=1}^{159} X_{ij} = 1, \qquad i = 1, ...159$$

$$-X_{ij} + y_j \geq 0 \qquad i = 1, ...159, \qquad j = 1, ...159$$

$$\sum_{j=1}^{159} y_j \leq 8,$$

$$y_j \in \{0, 1\},$$

where $d_{ij} = d(V_i, V_j)$ is the shortest distance between vertex V_i and vertex V_j.

8.2.3 The probabilistic approach

The idea behind the probabilistic approach is to normalize the LGs to 1 such that we can talk about probabilities, and then we compute the expected weighted distance, which can be minimized. We make the same assumptions as in the N-median approach, except that here LGs represent weights on the probabilities .

8.2.3.1 Building the objective function In order to define our objective function, we introduce the following notations:

- Let $m_1, \ldots, m_{159}$ denote the machines to be grouped.

- Let $G_1, \ldots, G_8$ denote the groups.
- Let L_i denote the lastgrad for machine m_i.
- Let $X_{kj} = \begin{cases} 1 & \text{if machine } m_j \text{ is in group } G_k \\ 0 & \text{otherwise.} \end{cases}$
- Let $P_i = \frac{L_i}{\sum_{j=1}^{159} L_j}$ be the probability of going to machine m_i per given unit of time.
- let d_{ij} be the shortest distance between machines m_i and m_j.

As a result of the independence of the LGs, we can define $P_{ij} = P_i P_j$ as the probability of going from machine m_i to machine m_j.
Considering d_{ij} as the event that the distance d_{ij} is traveled, we can speak of the expected weighted distance traveled as

$$E[D] = \sum_{k=1}^{8} \sum_{i=1}^{159} \sum_{j=1}^{159} X_{kj} P_{ij} d_{ij}.$$

8.2.3.2 Building the constraints We have some requirements to fulfill:

- The sum of all machines in the 8 groups must add up to 159

$$\sum_{k=1}^{8} \sum_{j=1}^{159} X_{kj} = 159.$$

- The sum of LGs in each group fulfills some bounding constraint

$$a \leq \sum_{j=1}^{159} X_{kj} L_j \leq b, \qquad k = 1, ..., 8.$$

- No one machine belongs to more than one group

$$\sum_{j=1}^{8} X_{kj} = 1, \qquad j = 1, ..., 159.$$

Putting everything together, we have the following optimization problem:

$$\text{Minimize} \qquad E[D] = \sum_{k=1}^{8} \sum_{i=1}^{159} \sum_{j=1}^{159} X_{kj} P_{ij} d_{ij},$$

subject to

$$\sum_{k=1}^{8}\sum_{j=1}^{159} X_{kj} = 159,$$

$$a \leq \sum_{j=1}^{159} X_{kj} L_j \leq b, \qquad k = 1, ..., 8,$$

$$\sum_{j=1}^{8} X_{kj} = 1, \qquad k = 1, ..., 159.$$

8.2.4 Numerical experiments

The package AMPL was used to implement the algorithms corresponding to both the N-median and the probabilistic approach. Unfortunately, the available AMPL version at the moment the experiments were run, allowed us to solve problems only up to a maximum of 300 variables. This corresponds to only 12 machines in our problem. For this reason, we demonstrate only a toy example to test our algorithms using only 10 machines and 3 groups. We take arbitrary locations of machines, arbitrary LGs, and arbitrary distances as shown in the layout drawings.

The result of the N-median algorithm produces a fine grouping of the machines while neglecting the homogeneity of the total LG of the groups. See Figures 8.3 and 8.4. On the other hand, the probabilistic algorithm takes good care of the homogeneity of the total LG of the groups but may give bad grouping. See Figure 8.5. In the subsequent figures, the number on the upper right corner denotes the machine number, whereas the number on the lower left corner indicates the LG.

A consideration of an eight-traveling salesman approach in which the eight workers are eight traveling salesmen making a tour of the machine network might give a good result. Here we have only considered distance as a measure of idle time. This idea does not measure idle time correctly if more than one machine is idle at the same time and one machine must be fixed before the other. Any approach that considers time to fix the first machine plus distance between the first and second machine and so on might be a better measure of idle time. However, this mathematical modeling can serve as a starting point for more sophisticated approaches. For more details on the model just discussed, see [1].

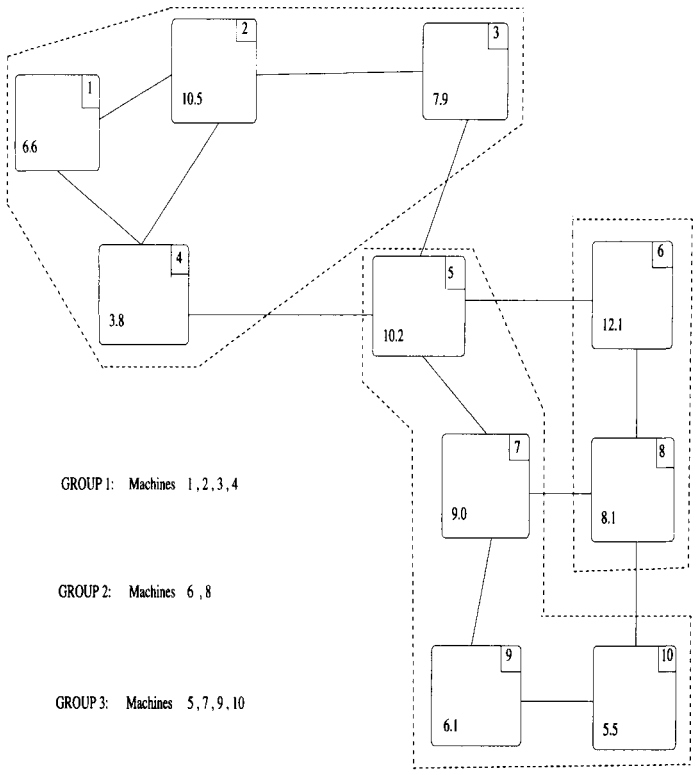


Figure 8.3 N-median approach, I.

8.3 AMERICAN CUTANEOUS LEISHMANIASIS

Here we study a simple model of an infectious disease as a further illustration of the applicability of the concepts studied in previous chapters. The methods used can be applied to other problems on infectious diseases and other dynamical systems. As usual, we study the existence equilibrium points and possible existence or nonexistence of periodic solutions, their stability properties, and some special solutions such as heteroclinic orbits.

8.3.1 Background

Several infectious diseases have been thoroughly studied, but this is not the case with a zoonosis that has affected several areas in Brazil and Venezuela and that is in fact a parasitic disease occurring throughout the Americas from Texas to Argentina. It is the so–called American Cutaneous Leishmaniasis (ACL), a resurgent parasitic disease transmitted by the byte of phlebotomine sand flies ("vectors") from reservoir hosts. Although the study of such disease has a few similarities with that of other infectious diseases like malaria, there are some differences that make an independent approach valuable.

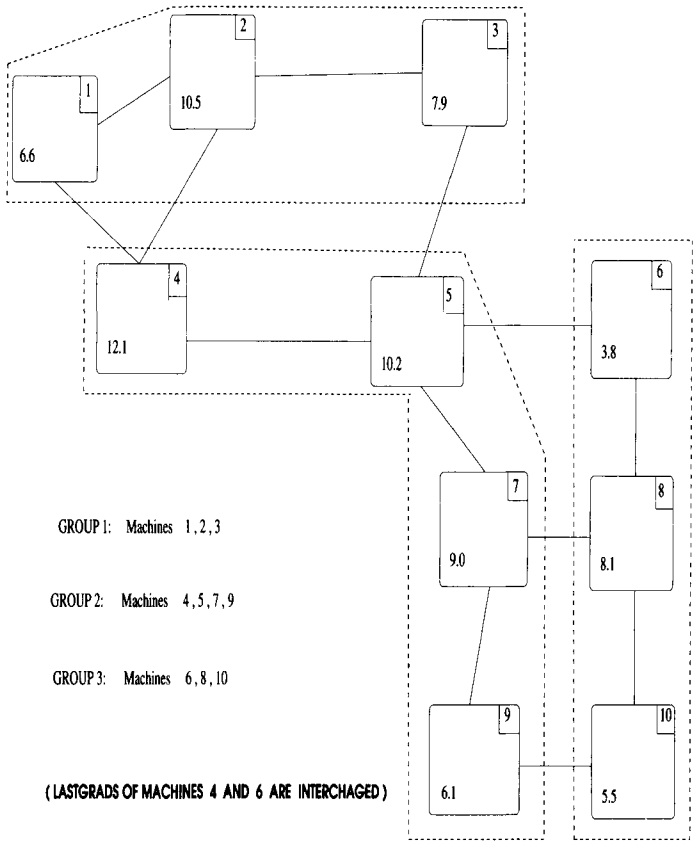


Figure 8.4 N-median approach, II.

A classic model for malaria considers mainly mosquitoes and human population, where all individuals are assumed to be both sink and source of infection, that is, *reservoir hosts*. However, in a model for ACL, it is assumed that humans are only sinks of infection; that is, they are only *incidental hosts*, because humans are not an important source of infection toward vectors (mosquitoes) and neither toward other humans.

The main assumptions of our model are:

(a) The dynamics of infection follows the mass action principle, which means that net rate of spread of the infection is proportional to the product of the density of susceptible individuals and the density of infectious individuals.

(b) The total population of hosts and vectors is constant.

(c) The infection is regulated by the recovery of the (vertebrate) hosts and the mortality of the vectors. Recovered hosts become susceptible (no immunity), and dead vectors are replaced by new individuals.

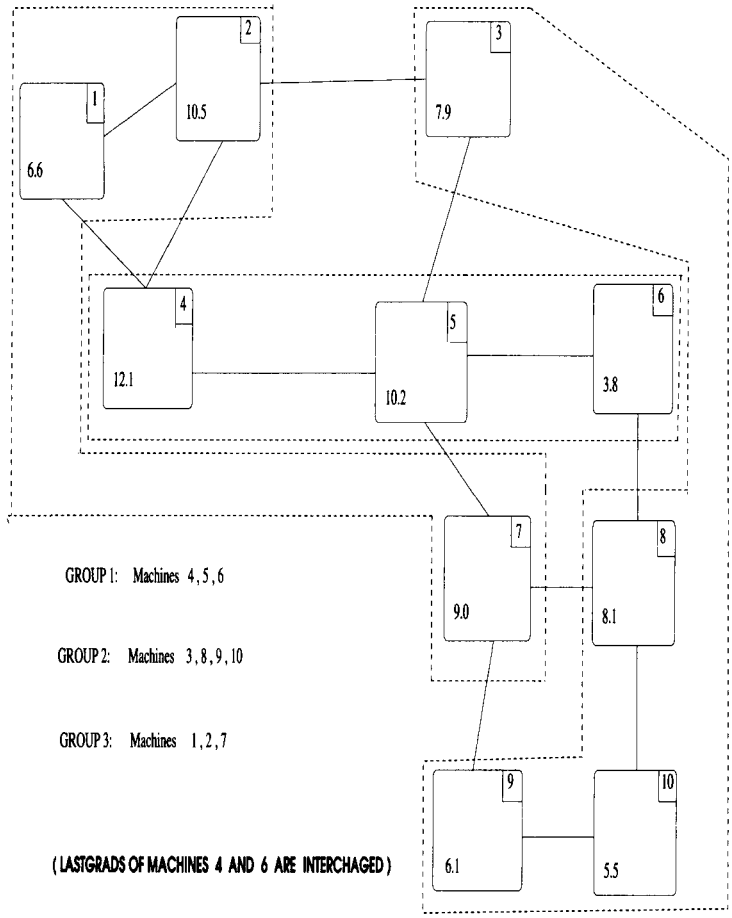


Figure 8.5 Probabilistic approach.

8.3.2 Development of the mathematical model

Several species can be considered as reservoirs: dogs, donkeys, horses, opossums, and so on. Based on the concrete experience at differences small rural areas, here we consider equines and dogs as reservoirs. Accordingly, we define

$H(t)$: Infected human population size	A : Human population size
$Q(t)$: Infected equine population size	B : Equine population size
$G(t)$: Infected dog population size	C : Dog population size
$V(t)$: Infected vector population size	D : Vector population size

The parameters are:

β_H : rate of infection per individual in vector–human encounters.

β_Q : rate of infection per individual in vector–equine encounters.

β_G : rate of infection per individual in vector–dog encounters.

γ_H : rate of recovery per unit time of infected humans.

γ_Q : rate of recovery per unit time of infected equines.

γ_G : rate of recovery per unit time of infected dogs.

μ : rate of mortality per unit time of vectors.

With the considerations above, we have the following model:

$$\begin{aligned} \dot{H} &= \beta_H V(A-H) - \gamma_H H, \\ \dot{Q} &= \beta_Q V(B-Q) - \gamma_Q Q, \\ \dot{G} &= \beta_G V(C-G) - \gamma_G G, \\ \dot{V} &= \beta_G G(D-V) + \beta_Q Q(D-V) - \mu V. \end{aligned} \tag{8.12}$$

It is known that μ is about 1.26 per cycle of three days, which means we take $\mu = 0.42$/day. The rate of recovery γ is defined as the inverse of the average lifetime of the infection. Taking the average infection lifetime of 6 weeks for humans, 4 years for equines and 11 months for dogs, we obtain

$$\gamma_H = 0.024, \quad \gamma_Q = 0.0007, \quad \text{and} \quad \gamma_G = 0.003. \tag{8.13}$$

For the simulations below, we take $A = 124$, $B = 28, C = 43, D = 4807$, and we consider the number of infected individuals as $H = 16$, $Q = 6$, and $G = 4$.

With these values, the infection rates β can be calculated by solving a system derived from (8.12) itself. For instance, to find β_Q and the corresponding value of V (when the equines are considered to be reservoir), we solve

$$\begin{aligned} \beta_Q V(B-Q) - \gamma_Q Q &= 0, \\ \beta_Q Q(D-V) - \mu V &= 0, \end{aligned} \tag{8.14}$$

from which we obtain $\beta_Q = 0.0000527$ and $V = 3.62$, which is rounded up to $V = 4$. A similar system is solved to find β_G. Since humans are not considered reservoir but incidental hosts, for the calculation of β_H, we solve from the first equation of (8.12):

$$\beta_H = \frac{\gamma_H H}{V(A-H)}.$$

Reservoir	β_H	β_Q	β_G	V	R_0
Humans	0.000139	0.00000734	0.0000118	24	1.143
Equines	0.000982	0.00005270	0.0000850	4	1.271
Dogs	0.000948	0.00005090	0.0000821	4	1.106
Equines and Dogs	0.000982	0.00005270	0.0000850	8	2.456

Table 8.1 Parameter values and threshold value.

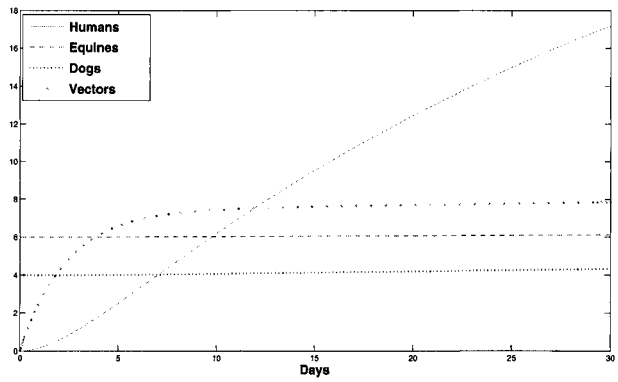


Figure 8.6 Time evolution of solution in (8.12).

In Table 8.1, we show values of the infection rates β for the different situations of reservoir hosts [14].

To have a first view of the system's behavior, let us assume that humans and vectors are susceptible individuals; i.e., they are not infected ($H = V = 0$), and let us take $Q = 6$ and $G = 4$. With these initial conditions, we solve the system (8.12), and the results for the evolution of the zoonosis in the first month are shown in Figure 8.6. It is important to observe that the four functions are increasing, with the one representing the vectors increasing much faster than the rest in the first five days.

8.3.3 Equilibria and periodic orbits

The equilibria for the system (8.12) are given by

$$\begin{aligned} H^0 &= \frac{\beta_H A V^0}{\beta_H V^0 + \gamma_H}, \\ Q^0 &= \frac{\beta_Q B V^0}{\beta_Q V^0 + \gamma_Q}, \\ G^0 &= \frac{\beta_G C V^0}{\beta_G V^0 + \gamma_G}, \\ V^0 &= \frac{(\beta_Q Q^0 + \beta_G G^0) D}{\beta_Q Q^0 + \beta_G G^0 + \mu}, \end{aligned} \tag{8.15}$$

where one such equilibria is the trivial solution $(H^0, Q^0, G^0, V^0) = (0, 0, 0, 0)$.

We now want to show that under certain condition, it is not possible to have periodic solutions. To this end, we first say a general system $\dot{x} = f(x)$ is *K-cooperative* (or simply, cooperative) if f is differentiable and

$$\frac{\partial f_i}{\partial x_j}(x) \geq 0, \qquad \text{for all} \quad i \neq j,$$

A classic result on dynamical system says that if a system $\dot{x} = f(x)$ is cooperative, then it does not have any attracting periodic orbits. Now, if we consider that dogs are not part of the system, we have the model

$$\begin{aligned} \dot{H} &= \beta_H V(A - H) - \gamma_H H, \\ \dot{Q} &= \beta_Q V(B - Q) - \gamma_Q Q, \\ \dot{V} &= \beta_Q Q(D - V) - \mu V. \end{aligned} \tag{8.16}$$

Then, one can show that the system (8.16) does not have any attracting periodic solution. Indeed, let $x = (H, Q, V)$ and $f(x) = (f_1(x), f_2(x), f_3(x))$ with

$$\begin{aligned} f_1(x) &= \beta_H V(A - H) - \gamma_H H, \\ f_2(x) &= \beta_Q V(B - Q) - \gamma_Q Q, \\ f_3(x) &= \beta_Q Q(D - V) - \mu V. \end{aligned} \tag{8.17}$$

so that our system has the general form $\dot{x} = f(x)$. Then, we have

$$\begin{aligned} \frac{\partial f_1}{\partial x_2} &= \frac{\partial f_1}{\partial Q} = 0, & \frac{\partial f_1}{\partial x_3} &= \frac{\partial f_1}{\partial V} = \beta_H(A - H), \\ \frac{\partial f_2}{\partial x_1} &= \frac{\partial f_2}{\partial H} = 0, & \frac{\partial f_2}{\partial x_3} &= \frac{\partial f_2}{\partial V} = \beta_Q(B - Q), \\ \frac{\partial f_3}{\partial x_1} &= \frac{\partial f_3}{\partial H} = 0, & \frac{\partial f_3}{\partial x_2} &= \frac{\partial f_3}{\partial Q} = \beta_Q(D - V), \end{aligned}$$

so that $\frac{\partial f_i}{\partial x_j} \geq 0$ for $i \neq j$, and $i, j = 1, 2, 3$. That is, system (8.16) is cooperative in D, and therefore, it cannot have attracting periodic orbits. Similar conclusions can be arrived at when changing the reservoir hosts, and when dealing with system (8.12).

One approach to study the nonexistence of any periodic solutions is to apply a generalization of Bendixon's criterion. The idea is to compute the Jacobian J of the system and to define the symmetric matrix

$$S = \frac{1}{2}\left(J^T + J\right).$$

A theorem in [48] says that if S has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ such that $\lambda_1 + \lambda_2 < 0$, then no periodic solution can exist for the system. If we assume that the entire population of vectors, dogs, and equines are infected, then we can arrive to a conclusion about periodic solutions.

Consider, then, the system (8.12) with the entire population of vectors, dogs and donkeys as infected. The matrix S in such a case takes the form

$$S = \begin{bmatrix} -\beta_H V - \gamma_H & 0 & 0 & \frac{1}{2}\beta_H(A-H) \\ 0 & -\beta_Q V - \gamma_Q & 0 & 0 \\ 0 & 0 & -\beta_G V - \gamma_G & 0 \\ \frac{1}{2}\beta_H(A-H) & 0 & 0 & -\beta_G G - \beta_Q Q - \mu \end{bmatrix}.$$

Then, the eigenvalues of S are (in order of magnitude)

$$\begin{aligned} \lambda_1 &= \tfrac{1}{2}\left(a+g-\sqrt{(a-g)^2+4b^2}\right), & \lambda_3 &= -\beta_G V - \gamma_G, \\ \lambda_2 &= \tfrac{1}{2}\left(a+g+\sqrt{(a-g)^2+4b^2}\right), & \lambda_4 &= -\beta_Q V - \gamma_Q, \end{aligned}$$

where $a = -\beta_H V - \gamma_H$, $b = \frac{1}{2}\beta_H(A-H)$, $g = -\beta_G G - \beta_Q Q - \mu$.

One can readily verify that we always have $\lambda_1 + \lambda_2 < 0$, and therefore, there are in fact no periodic solutions to such a system.

8.3.4 Stability properties

We want to study a little more closely the stability properties of the equilibria. Let us consider again the system (8.16). We want to prove two things:

(a) the origin is a saddle, and

(b) The nontrivial equilibrium is a saddle if $\alpha < \beta_Q^2 BC$, and depending on the sign of certain discriminant, it is either a stable node or a stable focus if $\alpha > \beta_Q^2 BC$, where $\alpha = \beta_Q(\beta_Q BV + \beta_Q CQ + \gamma_Q Q + \mu V) + \mu\gamma_Q$.

First, one can easily find out that the critical points of (8.16) are the origin $x_0 = (0,0,0)$ and a non trivial one

$$x_1 = \left(\frac{A\beta_H(\beta_Q^2 BC - \mu\gamma_Q)}{\beta_H(\beta_Q^2 BC - \mu\gamma_Q) + \beta_R\gamma_H(\beta_Q B + \mu)}, \frac{\beta_Q^2 BC - \mu\gamma_Q}{\beta_Q(\beta_Q C + \gamma_Q)}, \frac{\beta_Q^2 BC - \mu\gamma_Q}{\beta_Q(\beta_Q B + \mu)}\right),$$

where we require that (the inequality below is called the threshold condition)

$$R_0 := \frac{\beta_Q^2 BC}{\mu} > \gamma_Q, \tag{8.18}$$

so that the equilibrium x_1 lies in the positive octant. The general Jacobian of the system is

$$J = \left[\begin{array}{c|cc} -\beta_H V - \gamma_H & 0 & \beta_H(A-H) \\ \hline 0 & -\beta_Q V - \gamma_Q & \beta_Q(B-Q) \\ 0 & \beta_Q(C-V) & -\beta_Q Q - \mu \end{array}\right]. \tag{8.19}$$

At the origin, this reduces to

$$J_0 = \left[\begin{array}{c|cc} -\gamma_H & 0 & \beta_H A \\ \hline 0 & -\gamma_Q & \beta_Q B \\ 0 & \beta_Q C & -\mu \end{array}\right],$$

and its eigenvalues can be computed directly. They are

$$\lambda_1 = -\gamma_H, \quad \lambda_{2,3} = \frac{-(\gamma_Q + \mu) \pm \sqrt{(\gamma_Q + \mu)^2 + 4(\beta_Q^2 BC - \mu\gamma_Q)}}{2},$$

and it is clear that with the condition (8.18), the origin is a saddle point, and therefore, unstable. Also observe that if had $\frac{\beta_Q^2 BC}{\mu} = \gamma_Q$, then the origin would be nonhyperbolic and if $\frac{\beta_Q^2 BC}{\mu} < \gamma_Q$, then it would be stable.

To study the eigenvalues of the Jacobian (8.19) at the nontrivial equilibrium x_1, we see that one of the eigenvalues is $\lambda_1 = -\beta_H V - \gamma_H < 0$. The other two eigenvalues come from the 2×2 lower right block. If we denote with δ the determinant of that block and with τ its trace, then the two eigenvalues are given by

$$\lambda = \frac{\tau \pm \sqrt{\tau^2 - 4\delta}}{2}, \tag{8.20}$$

where

$$\delta = (\beta_Q V + \gamma_Q)(\beta_Q Q + \mu) - \beta_Q^2 (C - V)(B - Q),$$

$$\tau = -(\beta_Q V + \gamma_Q + \beta_Q Q + \mu) < 0, \text{ and}$$

$$\tau^2 - 4\delta = (\beta_Q V + \gamma_Q + \beta_Q Q + \mu)^2 - 4[(\beta_Q V + \gamma_Q)(\beta_Q Q + \mu) - \beta_Q^2 (C - V)(B - Q)].$$

Observe that δ can be written as

$$\delta = (\beta_Q^2 BV + \beta_Q^2 CQ + \beta_Q \gamma_Q Q + \mu\beta_Q V + \mu\gamma_Q) - \beta_Q^2 BC = \alpha - \beta_Q^2 BC.$$

Thus, if $\alpha < \beta_Q^2 BC$, then $\delta < 0$, and therefore, there are two eigenvalues of opposite sign and x_1 is saddle. Similarly, if $\alpha > \beta_Q^2 BC$, then $\delta > 0$, and since we know that $\tau < 0$, we can conclude that x_1 is a stable node if $\tau^2 - 4\delta \geq 0$ or a stable focus if $\tau^2 - 4\delta < 0$, by trace-determinant analysis.

Several observations are in order:

(a) Biologically, the inequality (8.18) says that the ratio of the rate of infection of the main reservoir hosts (donkeys) to the rate of mortality of the vectors (scaled by the population product BC) will have to be greater than the rate of recovery of the infected reservoir host, so that there is some infected population. This is in fact a threshold condition for infection.

(b) Under the threshold condition (8.18), the origin is unstable. This means that a state with no infected individuals can be easily perturbed into a state with some infected individuals.

(c) Although it is not easy to determine the sign of $\tau^2 - 4\delta$ in general, for the application in mind, the parameter μ is many times larger that the other parameters. This makes the term τ^2 much larger than 4δ, and therefore, $\tau^2 > 4\delta$.

(d) The inequalities above depend much on the quantity $\beta_Q^2 BC$, that is, the rate of infection in vector-equine encounters, the total population of donkeys, and the total population of vectors, in other words, the very carriers of the disease.

8.3.5 Numerical computations

Some numerical computations can help understand further the general model (8.12) and can help find some solutions that would be too difficult to find analytically. If we fix all parameters except β_H, then besides the trivial solution, there are two more equilibria. One of them has some negative components, so that it is not of practical interest. The third equilibrium is given by

$$Y = \left(\frac{-0.13817\beta_H + 124\beta_H^2}{(0.000766193 + \beta_H)(-0.00111427 + \beta_H)}, 19.6623, 20.2186, 31.3237 \right). \tag{8.21}$$

The general Jacobian is

$$\left[\begin{array}{c|ccc} -\beta_H V - \gamma_H & 0 & 0 & \beta_H(A-H) \\ \hline 0 & -\beta_Q V - \gamma_Q & 0 & \beta_Q(B-Q) \\ 0 & 0 & -\beta_G V - \gamma_G & \beta_G(D-G) \\ 0 & \beta_Q(C-V) & \beta_G(C-V) & -\beta_G G - \beta_Q Q - \mu \end{array} \right] \tag{8.22}$$

As a result of the block triangular form of the Jacobian, we observe that $-\beta_H V - \gamma_H < 0$ is one of the eigenvalues and is the only one that depends on β_H. For $\beta_H = 0.000982$ and the fixed values of the other parameters, the origin is unstable and Y is stable. In fact, the eigenvalues of the Jacobian at the origin are

$$\lambda_1 = -0.02400000, \lambda_2 = -0.42442592, \lambda_3 = -0.00141438, \lambda_4 = 0.00214030.$$

This also tells us that the the unstable manifold of the origin is one-dimensional. Similary, the eigenvalues of the Jacobian at Y are

$$\mu_1 = -0.05475986, \mu_2 = -0.42489150, \mu_3 = -0.00183740, \mu_4 - 0.00403915,$$

which indicates that the stable manifold of Y is four-dimensional.

We are also interested in computing a possible connecting orbit from the origin to Y. This is of great importance because such a curve could explicitly describe the path from the non infected population to the stable state of infected individuals. Once such an orbit is located, one can compute as usual a branch of connecting orbits through smooth continuation, in this case for β_H in the interval $[0.00001,\ 0.05]$. See Figures 8.7 and 8.8, where we plot the time evolution of the infected humans. Connecting orbits are very important in the global study of dynamical systems, acting as separatrices of solutions in the phase portrait, giving a better understanding of the behavior of nearby solutions.

For more details on this mathematical model of American cutaneous leishmaniasis, see [12] and [14].

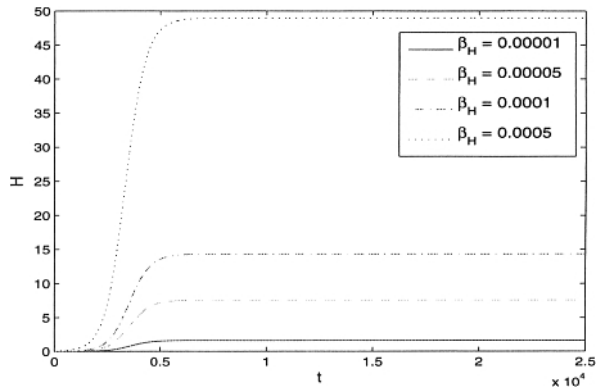


Figure 8.7 Branch of connecting orbits for (8.12).

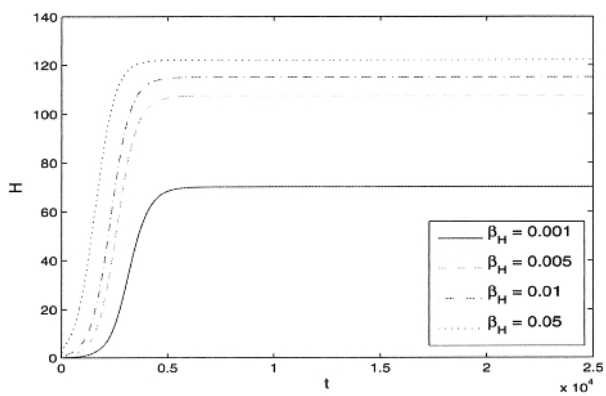


Figure 8.8 Branch of connecting orbits for (8.12).

8.4 VARIABLE POPULATION INTERACTIONS

In Section 7.3, we studied a particular dynamical system, where the two species had a fixed predator–prey relationship. That is, the first species was always the prey and the second one was the predator. However, there are cases where two or more species can be found coexisting under different kinds of associations at different times or environments: the relationship between given species can change from host-parasite to mutualism to competition.

The static classification of interactions like predator–prey, or competition among others, is in those cases inadequate as such relation may vary depending on population densities, age or size of individuals, as well as on several environmental parameters. In this context, population models with conditional (not fixed) interactions, represented via nonconstant functions have been the focus of several relatively recent studies, including [34, 35, 36, 68]. Besides the rigorous mathematical work, they clearly highlight the practical importance and the relevance of studying these models. One important point is that the presence of a species involves cost and benefits to the second species sharing the same environment, and therefore their interspecific relationship needs to be represented by continuous functions that can assume positive and negative values, not but just one parameter.

Nature provides us with several examples for such variable interactions. For instance, ants benefit from their interactions with aphis, because the latter provide certain secretions which are rich in sugars and aminoacids; ants at the same time provide protection to aphis from their natural predators. However, the magnitude of these benefits depends on the relative densities of the two populations involved: at low aphid densities, benefits for them are high, but at higher densities, such benefits are low, none, or even negative. There are even examples where predator and prey may interchange role, as in the case of lobsters and whelks in the islands of Malgas and Markus, in South Africa [34, 35], all depending again on population densities and environmental parameters.

In this section, we study the dynamics of a two-species model that incorporates two rational α- functions representing the variable interaction between the species. We study the existence and stability of equilibria, possible existence of periodic solutions, and bifurcation phenomena, and we explicitly establish threshold values that help determine whether given equilibrium points represent mutualism, competition, or host–parasite behavior. Several numerical examples are given to illustrate the main results on the number equilibria and their stability properties, bifurcations, and how the interaction between the species varies with population densities and environmental parameters.

8.4.1 Model formulation

We consider the following model:

$$\begin{aligned} \frac{dx}{dt} &= x(1-\frac{x}{K_1}) + \alpha_1(x,y)\frac{xy}{K_1}, \\ \frac{dy}{dt} &= y(1-\frac{y}{K_2}) + \alpha_2(x,y)\frac{xy}{K_2}, \end{aligned} \tag{8.23}$$

where $x(t) \geq 0$, $y(t) \geq 0$ represent the population densities, and K_1, K_2 are the carrying capacities of species 1 and 2, respectively. Classically, one considers α_1 and α_2 as fixed parameters, and their signs determine the interspecific interaction (functional response). More precisely, we have:

Mutualism, if $\alpha_1 > 0$ and $\alpha_2 > 0$, competition, if $\alpha_1 < 0$ and $\alpha_2 < 0$, host-parasite, if $\alpha_1 > 0$ and $\alpha_2 < 0$, or $\alpha_1 < 0$ and $\alpha_2 > 0$.

Here, we consider functions

$$\alpha_1(y) = \frac{a_1 - b_1 y}{c_1 + d_1 y}, \qquad \alpha_2(x) = \frac{a_2 - b_2 x}{c_2 + d_2 x}, \tag{8.24}$$

to represent such specific interactions (including saturation effects) between both species, where the parameters $a_{1,2}, b_{1,2}, c_{1,2}, d_{1,2}$, representing changes in environmental conditions are all positive. Thus, the interspecific interactions are not fixed but vary with the environmental parameters and the system state, and they can take positive or negative values. Note that cooperative interaction happens at low population densities, but the interaction becomes competitive or host-parasitic at higher densities.

Note: Observe that if $c_1 = c_2 = 1$ and $d_1 = d_2 = 0$, then this model reduces to the one studied in [68], where the authors, among other results, give a detailed analysis of existence and stability of equilibria of the corresponding model.

8.4.2 Local stability of equilibria

The system (8.23), (8.24) has four equilibria:

$$P_0 = (0,0), \quad P_1 = (K_1, 0), \quad P_2 = (0, K_2), \quad P_3 = (x_3, y_3),$$

where (x_3, y_3) is the solution of

$$\begin{aligned} K_1 - x + \tfrac{(a_1 - b_1 y)y}{c_1 + d_1 y} &= 0, \\ K_2 - y + \tfrac{(a_2 - b_2 x)x}{c_2 + d_2 x} &= 0. \end{aligned} \tag{8.25}$$

Denote $A = K_1 d_1 + a_1$, and $B = K_2 d_2 + a_2$. Then, by solving for x in the first equation of the system (8.25), we get $x = \frac{K_1 c_1 + Ay - b_1 y^2}{c_1 + d_1 y}$. The parabola in the numerator has vertex on the first quadrant, and one positive root

$$\hat{y} = (A + \sqrt{A^2 + 4b_1 c_1 K_1}\,)/2b_1. \tag{8.26}$$

Therefore, $x = x_3$ is positive if and only if $0 < y < \hat{y}$. Similarly, by solving the second equation of (8.25) for y, we conclude that $y = y_3$ is positive if and only if $0 < x < \hat{x}$, where $\hat{x}$ is the positive root of the parabola $K_2 c_2 + Bx - b_2 x^2$, and

$$\hat{x} = (B + \sqrt{B^2 + 4b_2 c_2 K_2}\,)/2b_2. \tag{8.27}$$

Therefore, there is no coexistence equilibrium P_3 if either $x \geq \hat{x}$ or $y \geq \hat{y}$. In Figure 8.9, $\hat{x}$ is the intersection of the solid nullcline with the x–axis, and $\hat{y}$ is the intersection of the dashed nullcline with the y–axis.

The general Jacobian of (8.23) is

$$J(x,y) = \begin{bmatrix} 1 - \frac{2}{K_1}x + \frac{(a_1 - b_1 y)y}{K_1(c_1 + d_1 y)} & \frac{(a_1 c_1 - 2b_1 c_1 y - b_1 d_1 y^2)x}{K_1(c_1 + d_1 y)^2} \\ \frac{(a_2 c_2 - 2b_2 c_2 x - b_2 d_2 x^2)y}{K_2(c_2 + d_2 x)^2} & 1 - \frac{2}{K_2}y + \frac{(a_2 - b_2 x)x}{K_2(c_2 + d_2 x)} \end{bmatrix}. \tag{8.28}$$

Theorem 8.1 *The boundary equilibrium points of* (8.23) *have the following local stability properties:*

(a) $P_0 = (0,0)$ *is always an unstable node.*

(b) $P_1 = (K_1, 0)$ *is a saddle if* $a_2 - b_2 K_1 > -\frac{K_2}{K_1}(c_2 + d_2 K_1)$, *and a stable node, if the inequality is reversed.* P_1 *is never a focus nor of center type.*

(c) $P_2 = (0,\ K_2)$ *is a saddle if* $a_1 - b_1 K_2 > -\frac{K_1}{K_2}(c_1 + d_1 K_2)$, *and a stable node, if the inequality is reversed.* P_2 *is never a focus nor of center type.*

Proof.

(a) At P_0, the eigenvalues of the Jacobian are $\lambda_{1,2} = 1$, and therefore, the origin is an unstable node.

(b) At P_1 the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 1 + \frac{K_1}{K_2}\left(\frac{a_2 - b_2 K_1}{c_2 + d_2 K_1}\right)$. Therefore, P_1 is saddle if $a_2 - b_2 K_1 > -\frac{K_2}{K_1}(c_2 + d_2 K_1)$ and stable node if the inequality is reversed. It is clear that P_1 is never a focus nor a center, as none of the eigenvalues is complex.

(c) At P_2, the eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 1 + \frac{K_2}{K_1}\left(\frac{a_1 - b_1 K_2}{c_1 + d_1 K_2}\right)$. Therefore, P_2 is saddle if $a_1 - b_1 K_2 > -\frac{K_1}{K_2}(c_1 + d_1 K_2)$ and stable node if the inequality is reversed. P_2 is never a focus nor a center, as none of the eigenvalues is complex.

□

Remark 8.2 *That the origin* P_0 *is always an unstable node means that mutual extinction is never possible. Good news.*

Interspecific relationships. For the coexistence equilibrium $P_3 = (x_3, y_3)$, with $x_3 > 0$ and $y_3 > 0$, the nullclines determined by the equations (8.25) can intersect at zero, one, two, or three points, determining the number of coexistence equilibria (see Figure 8.9). The sign of the slope of the nullclines at the corresponding equilibrium point indicates the type of relationship: Denoting with m_1 such slope of the first nullcline and with m_2 the slope of the second nullcline in (8.25), we have mutualism if $m_1 > 0$ and $m_2 > 0$, competition if $m_1 < 0$ and $m_2 < 0$, and host-parasite if $m_1 > 0$ and $m_2 < 0$, or $m_1 < 0$ and $m_2 > 0$.

More precisely, with $(x, y) = (x_3, y_3)$, we find from (8.25) that

$$m_1 = \frac{(c_1 + d_1 y)^2}{a_1 c_1 - 2b_1 c_1 y - b_1 d_1 y^2}, \qquad m_2 = \frac{a_2 c_2 - 2b_2 c_2 x - b_2 d_2 x^2}{(c_2 + d_2 x)^2} \tag{8.29}$$

represent the slopes of the nullclines at P_3.

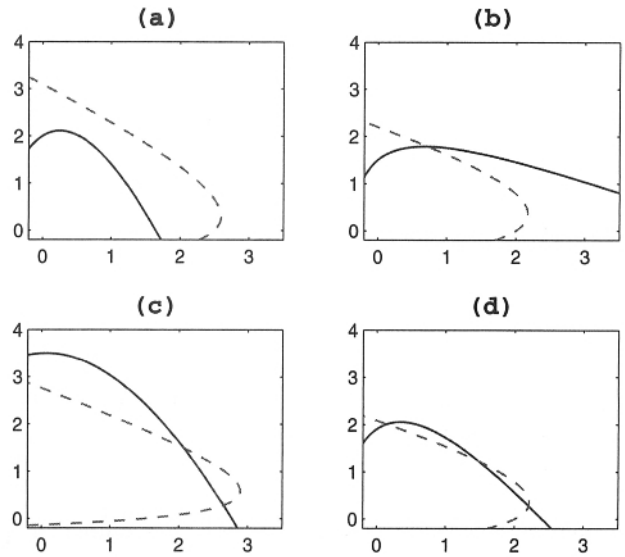


Figure 8.9 Sample nullclines of system (8.23).

Denote $D_1 = a_1c_1 - 2b_1c_1y - b_1d_1y^2$, $D_2 = a_2c_2 - 2b_2c_2x - b_2d_2x^2$, and let us also denote $E_1 = \sqrt{4b_1c_1(b_1c_1 + a_1d_1)}$, $E_2 = \sqrt{4b_2c_2(b_2c_2 + a_2d_2)}$. Then, there is mutualism at P_3 when $D_1 > 0$ and $D_2 > 0$, that is, when $0 < y < y^+$ and $0 < x < x^+$, where

$$x^+ = \frac{E_1}{2b_1d_1} - \frac{c_1}{d_1} \qquad \text{and} \qquad y^+ = \frac{E_2}{2b_2d_2} - \frac{c_2}{d_2}. \tag{8.30}$$

Similarly, there is competition at P_3 when $x > x^+$ and $y > y^+$. Finally, there is a host-parasite relationship at P_3 when either $x > x^+$ and $0 < y < y^+$, or $0 < x < x^+$ and $y > y^+$.

Remark 8.3 *These values x^+, y^+ are very important threshold values that determine the kind of relationship between the species.*

For general stability results on P_3, we use again the nullcline equations (8.25) and take a closer look at the Jacobian in (8.28). Then, we have the following theorem.

Theorem 8.4 *The coexistence equilibrium $P_3 = (x_3, y_3)$ of system* (8.23), (8.24) *is*

(a) locally asymptotically stable if and only if $\frac{m_2}{m_1} < 1$, and a saddle if the inequality is reversed.

(b) a stable focus if $(\frac{x}{K_1} - \frac{y}{K_2})^2 < -\frac{4m_2xy}{m_1k_1k_2}$.

(c) never of center-type.

Proof.

(a) The Jacobian in (8.28) at $(x, y) = (x_3, y_3)$ can be written as

$$J(x,y) = \begin{bmatrix} -\frac{x}{K_1} & \frac{x}{K_1 m_1} \\ \frac{m_2 y}{K_2} & -\frac{y}{K_2} \end{bmatrix}.$$

Thus, the eigenvalues are

$$\lambda_{1,2} = \frac{-(\frac{x}{K_1} + \frac{y}{K_2}) \pm \sqrt{\left(\frac{x}{K_1} + \frac{y}{K_2}\right)^2 - \frac{4}{K_1 K_2}\left(1 - \frac{m_2}{m_1}\right) xy}}{2}. \tag{8.31}$$

Hence, a necessary and sufficient condition for both eigenvalues to be negative (or equivalently, for P_3 to be locally asymptotically stable) is that $m_2/m_1 < 1$. Clearly, if the inequality is reversed, then P_3 is a saddle.

(b) For the eigenvalues in (8.31) to be complex, we need to require that $\left(\frac{x}{K_1} + \frac{y}{K_2}\right)^2 < \frac{4}{K_1 K_2}\left(1 - \frac{m_2}{m_1}\right) xy$, or equivalently, that $\frac{x^2}{K_1^2} + \frac{2xy}{K_1 K_2} + \frac{y^2}{K_2^2} < \frac{4xy}{K_1 K_2} - \frac{4m_2 xy}{m_1 K_1 K_2}$, which can be rewritten as $(\frac{x}{K_1} - \frac{y}{K_2})^2 < -\frac{4m_2 xy}{m_1 k_1 k_2}$. This also shows that one necessary condition is that $\frac{m_2}{m_1} < 0$ (only host-parasite case). We also see that the real part of the eigenvalue is always negative.

(c) From (8.31) it is clear that no eigenvalue of the Jacobian at the coexistence equilibrium $(x, y) = (x_3, y_3)$ can be purely imaginary.

□

Note: A sample set of parameters to get a focus is $a_1 = 1$, $b_1 = 1.3$, $c_1 = 0.1$, $d_1 = 1.5$, $k_1 = 0.3$, $a_2 = 0.7$, $b_2 = 1.4$, $c_2 = 0.2$, $d_2 = 0.3$, $k_2 = 0.5$, and the equilibrium point is $(x, y) = (0.1281338,\ 1.0098408)$.

No limit cycles. We can get some formal sufficient conditions under which the original system (8.23), (8.24) has no limit cycles (periodic solutions) in the positive quadrant $\mathbb{R}^2_+$. To this end, we first denote such a system as $\dot{x} = F_1(x, y)$, $\dot{y} = F_2(x, y)$. If we let $R(x, y) = 1/(xy)$, then we have:

$$Q := \frac{\partial(RF_1)}{\partial x} + \frac{\partial RF_2}{\partial y} = -\left(\frac{1}{K_1 y} + \frac{1}{K_2 x}\right) < 0$$

for all $x > 0, y > 0$. By using Dulac's criterion [54] this implies that no limit cycles can exist under these conditions. Thus, we have proved the following theorem.

Theorem 8.5 *The system* (8.23) *has no limit cycles in the positive quadrant* $\mathbb{R}^2_+$.

The statement in Theorem 8.5 implies that in our system no periodic or oscillatory behavior of population densities of both species for $t \geq 0$ is possible.

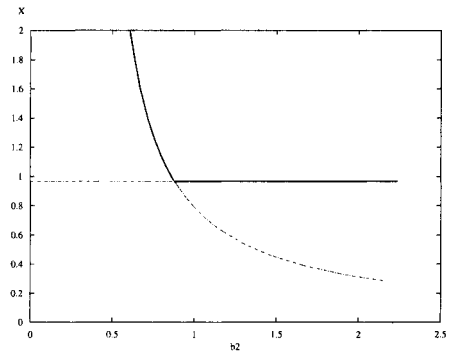


Figure 8.10 Transcritical bifurcation for (8.23) with $c_1 = c_2 = 0$.

8.4.3 Bifurcations

As discussed before, given a dynamical system, one of the central questions is how the given model depends on the choice of the parameters, or more precisely, on how the qualitative behavior of solutions changes as one or more parameters are allowed to vary. In this section we track for possible bifurcations as they represent drastic changes in the dynamical system for very small changes in parameter values. That is, we are interested in the changes in the species interactions, as the ecological conditions change. Our model exhibits several bifurcations including transcritical and saddle-node bifurcations. We start showing the existence of a transcritical bifurcation for our model when $c_1 = c_2 = 0$.

Transcritical bifurcation. One can show that there is a transcritical bifurcation at the critical point $P_1 = (K_1 + \frac{a_1}{d_1},\ 0)$ when b_2 is allowed to vary continuously. More precisely, at a certain value of b_2, one stable and one unstable branch of solutions collide and then interchange stability properties.

Theorem 8.6 *The system* (8.23) *undergoes a transcritical bifurcation at the equilibrium* P_1 *when* $b_2 = d_1 B/A$, *and* $c_1 = c_2 = 0$, *where as usual,* $A = K_1 d_1 + a_1$ *and* $B = K_2 d_2 + a_2$.

Proof. For $c_1 = c_2 = 0$, the Jacobian at P_1 has eigenvalues

$$\lambda_1 = -\frac{A}{K_1 d_1} \neq 0 \quad \text{and} \quad \lambda_2 = \frac{1}{k_2 d_1 d_2}(d_1 B - b_2 A).$$

Therefore, the Jacobian at P_1 has one zero eigenvalue precisely when $b_2 = d_1 B/A$. For such eigenvalue $\lambda_2 = 0$, an eigenvector is $v^T = [1 \quad -\frac{d_1}{b_1}]$ and a left eigenvector is $w^T = [0 \quad 1]$. If we write the right–hand side of (8.23) as $f(x, y)$, then one gets $w^T f_{b_2}(P_1) = 0$,

$$w^T D f_{b_2}(P_1) v = \frac{b_1 A^2}{(c_2 + \frac{d_2 A}{d_1}) K_2 d_1^3} \neq 0 \text{ and } w^T D^2 f(P_1)(v, v) = \frac{2 d_1 (b_2 + d_2)}{K_2 b_1 d_2} \neq 0.$$

Thus, by using Sotomayor's theorem [54], one concludes that under the conditions above, the system (8.23) undergoes a transcritical bifurcation at P_1. □

In Figure 8.10, we show a transcritical bifurcation at $b_2^* = 0.8793104$. This means that for $b_2 < b_2^*$, there exists a branch of stable coexistence equilibria and a branch of unstable species 1-only equilibria, and that for $b_2 > b_2^*$, those branches of equilibria interchange stability.

Saddle-node bifurcations. For the general model (8.23), (8.24), one can show the presence of saddle-node (fold) bifurcations for the coexistence equilibrium $P_3 = (x, y) = (x_3, y_3)$, which means that for a given value of $b_2 = b_2^*$ and near P_3, there are no equilibria for $b_2 < b_2^*$ and there are two equilibria for $b_2 > b_2^*$. Here we denote

$$F = (c_1 + d_1 y)^2, \quad G = (c_2 + d_2 x)^2, \quad H = 2c_2 x + d_2 x^2,$$

and D_1 is as in (8.29).

Theorem 8.7 *The general system* (8.23) *undergoes a saddle-node bifurcation at the coexistence equilibrium* P_3 *when*

$$b_2 = \frac{a_2 c_2 D_1 - FG}{D_1 H}.$$

Proof. Recall that the eigenvalues of the Jacobian of (8.23) at the coexistence equilibrium P_3 are given by (8.31). This implies that one of the eigenvalues is zero when $m_1 = m_2$. By using (8.29), this condition is equivalent to require that $Hb_2 = a_2c_2 - FG/D_1$; that is, $b_2 = \frac{a_2c_2D_1 - FG}{D_1H}$. An eigenvector of the Jacobian associated with the zero eigenvalue is $v = [1 \quad m_1]^T$, and a left eigenvector is $w = [1 \quad \frac{K_2 x}{m_1 K_1 y}]^T$. Denoting the right–hand side of (8.23) as $f(x, y)$, and recalling that $m_1 = m_2$, we have

$$f_{b_2}(P_3) = \begin{bmatrix} 0 \\ \\ -\frac{x^2 y}{K_2(c_2+d_2x)} \end{bmatrix}, \qquad D^2 f(P_3)(v, v) = \begin{bmatrix} -\frac{2c_1x(b_1c_1+a_1d_1)m_1^2}{K_1(c_1+d_1y)^3)} \\ \\ -\frac{2c_2y(b_2c_2+a_2d_2)}{K_2(c_2+d_2x)^3} \end{bmatrix}.$$

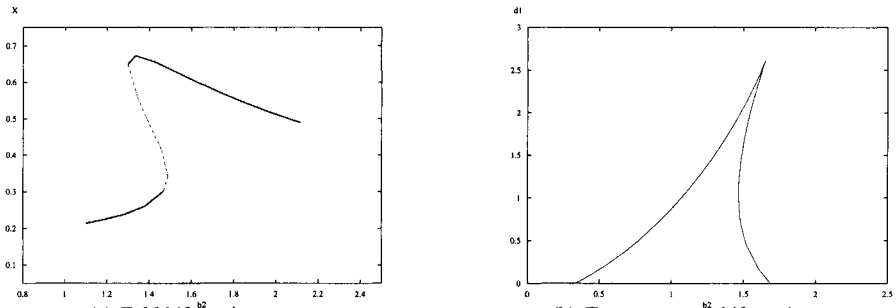
Therefore, we get $w^T f_{b_2} = -\frac{x^3}{m_1 K_1(c_2+d_2x)} \neq 0$ and

$$w^T D^2 f(P_3)(v, v) = -\frac{2x}{K_1}\left(\frac{c_1(b_1c_1 + a_1d_1)m_1^2}{(c_1 + d_1y)^3} + \frac{c_2(b_2c_2 + a_2d_2)}{m_1(c_2 + d_2x)^3}\right) \neq 0.$$

Thus, for the given value of b_2, Sotomayor's theorem [54] implies the existence of a saddle-node bifurcation for (8.23).

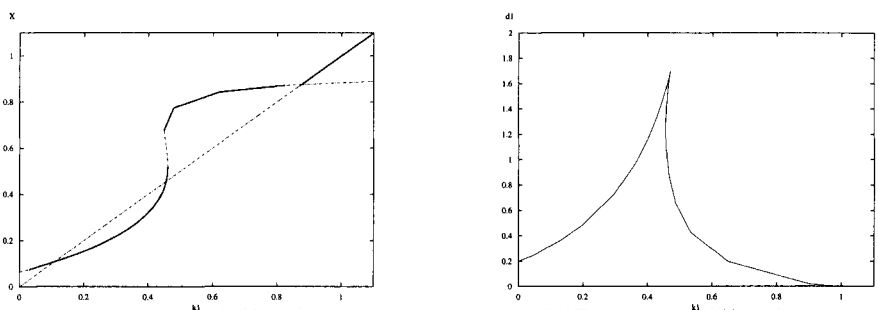
□

In Figure 8.11(a), we show two saddle-nodes bifurcations, at $b_2^- = 1.298441$, and at $b_2^+ = 1.487304$. Locally, this means that, say for $b_2 > b_2^+$, there are no equilibria, but that for $b_2 < b_2^+$, there is a branch of stable and a branch of unstable coexistence equilibria. A little more globally, observe in the same figure that for $b_2^- < b_2 < b_2+$, there are always three coexistence equilibria, two stable and one unstable, but outside that interval, there is only one. Here, d_1 is fixed at $d_1 = 1.5$.



(a) Fold bifurcations. (b) Two-parameter bifurcation.

Figure 8.11 Bifurcation diagrams for the general system (8.23)



(a) Fold and transcritical bifurcations. (b) Two-parameter bifurcation.

Figure 8.12 Bifurcation diagrams for the general system (8.23).

Starting at $b_2 = b_2^+$, we now let both b_2 and d_1 free (two-parameter bifurcation), to get a more global picture, the cusp in Figure 8.11(b). This is a curve of fold (saddle-node) points, which means that along this curve, the system (8.23) has a fold. The cases $b_2 = b_2^-$ and $b_2 = b_2^+$ were particular cases. Furthermore, this also means that inside the cusp, there are three coexistence equilibria, and only one (stable) equilibrium outside it, for each b_2, d_1.

Remark 8.8 *We need to point out again that at bifurcation points, arbitrarily small changes in the parameters (environmental fluctuations) lead to radical changes in the qualitative behavior of solutions. In particular, jumps between stable and unstable branches, and changes in the number of equilibria.*

This bifurcation analysis can be performed for parameters other than b_2 and d_1, depending on what is the focus of study (say, intrinsic or extrinsic ecological factors). For completeness, we show in Figure 8.12(a), a bifurcation diagram with K_1 as free parameter. It shows two fold points at $K_1^- = 0.4498474$ and $K_1^+ = 0.4602927$, as well as transcritical bifurcation at $K_1 = 0.8765586$. Starting at K_1^+, one does two-parameter bifurcation with both K_1 and d_1 free, to get a curve of fold points: the cusp of Figure 8.12(b). The analytical proof of existence of such bifurcations, and statements about the number of stable/unstable equilibria is similar to the discussion above. One can also arrive at conclusions as to

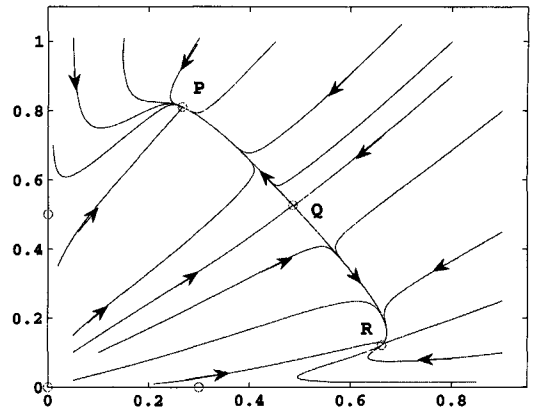


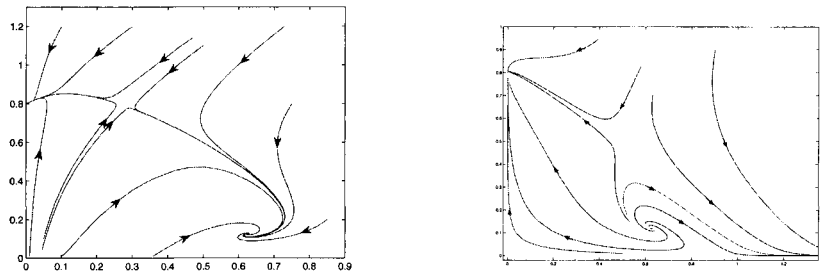
Figure 8.13 Sample phase portrait of the general system (8.23), (8.24).

whether equilibria represent mutualism, competition, or host-parasite interaction, based on our results above.

It is very illustrative to plot the phase portrait for a given set of parameters to summarize some ideas. See Figure 8.13. The given system has three boundary equilibria: (0, 0), (0.3, 0), (0, 0.5), that is, mutual extinction, species 1-only and species 2-only, as well as three coexistence equilibria: $P = (0.2663224,\ 0.8112827)$, $Q = (0.4860977,\ 0.5273574)$, and $R = (0.6626026,\ 0.1217539)$. Note that the three coexistence equilibria fall within the limits established according to (8.26) and (8.27), where $\hat{x} = 0.7080$, and $\hat{y} = 1.1357$. In this case, we also see that the three boundary equilibria are unstable; in particular, for the species 2-only equilibrium, we have that $0.35 = a_1 - b_1K_2 > -\frac{K_1}{K_2}(c_1 + d_1K_2) = -.051$, in agreement with part c) of Theorem 8.1, indicating it is a saddle.

Since we have used $b_2 = 1.4$ and $d_1 = 1.5$, we are inside the cusp of Figure 8.11(b), hence the existence of three coexistence equilibria, and we know from Figure 8.11(a) that two of them must be stable and the third one unstable, and this is exactly what we see in Figure 8.13: two stable nodes (P and R) and one saddle (Q). This fact is also a consequence of Theorem 8.4: For P, Q, R we have $m_2/m_1 = 0.3995,\ 1.4124,\ -1.2898$, respectively. Note also that the only equilibrium, Q, where both species reach equilibrium densities greater than their carrying capacities, is unstable.

For the given parameter values of the system illustrated in Figure 8.13, we have $x^+ = 0.1694$ and $y^+ = 0.2153$, the threshold values in (8.30). This immediately implies that P and Q are competitive equilibria and R is a host-parasite equilibrium. See Remark 8.3. The same can be verified using the definition of the slopes in (8.29). Indeed, for P we have $m_1 = -1.2438$, $m_2 = -0.4969$, for Q these slopes are $m_1 = -1.3702$, $m_2 = -1.9353$, and finally, for R, we get $m_1 = 2.0255$, $m_2 = -2.6125$, confirming our statements on interspecific relationships for the coexistence equilibria.



(a) Five equilibria.

(b) Six equilibria.

Figure 8.14 Sample phase portraits of system (8.32).

The coexistent species may change their type of interaction when the parameters vary. For instance, if we take, say, $b_2 = 0.75$ and $d_1 = 0.4$ (keeping all other parameters the same), then the new $P = (0.13456,\ 0.83537)$ goes from competitive to host-parasite equilibrium, and if we take $b_2 = 1.55$ and $d_1 = 2.2$, then the new $R = (0.568072,\ 0.22317)$ goes from host-parasite to competitive equilibrium. This can be readily verified as above, using either the corresponding threshold values x^+ and y^+, or the slopes m_1 and m_2.

Finally, we also observe that a) the only way to reach Q is going along its stable manifold, and b) from a perturbation, say, of the trivial equilibrium, or from Q, we can bring the system to one of the two stable coexistence equilibria P or R, following their unstable manifolds. These are special cases of connecting orbits: solution paths that go from one equilibrium to another one, from an equilibrium to a periodic orbit, or from one periodic orbit to another one, and they help determine separatrices in dynamical systems. For further details on this model, see [46].

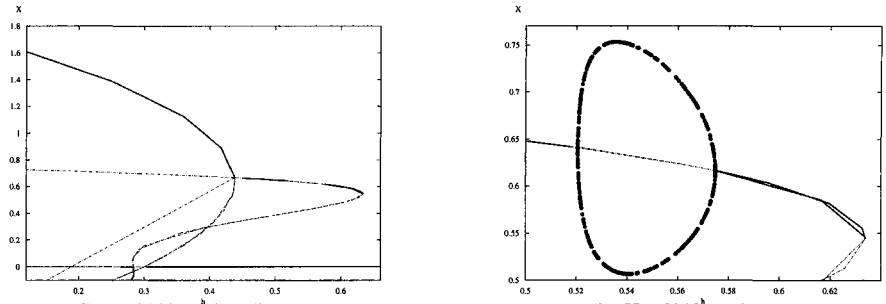
Final Remarks. Classic models in variable population interactions, including those in [34, 35, 36, 68], have detected equilibrium points that are only either nodes or saddle points. No Hopf bifurcations or periodic solutions have been found on such models. The work in [46] gives conditions for an equilibrium point to be a focus (Theorem 8.4). More recently, in [58], a model of variable population interactions has equilibria that can be nodes, saddles, foci, or centers. Furthermore, Hopf bifurcations and periodic orbits have also been found.

The model studied in [58] is

$$\begin{aligned} \frac{dx}{dt} &= x(r_1 - k_1 x) + (a - by)\,xy - \frac{hx}{e+x}, \\ \frac{dy}{dt} &= y(r_2 - k_2 y) + (c - dx)\,xy, \end{aligned} \tag{8.32}$$

which includes a harvesting function on the first species, represented by the rational function

$$H(x) = \frac{hx}{e+x}. \tag{8.33}$$



(a) General bifurcation diagram. (b) Hopf bifurcation.

Figure 8.15 Bifurcation diagrams for system (8.23)

This harvesting function increases smoothly with the size x of the population, h represents the rate-of-harvesting limit, and e is the number of species 1 it takes to reach one half of the maximum harvesting rate.

Among other results, the work in [58] shows that the model (8.32) has foci at a coexistence equilibrium point $P(x, y)$, for some $x > 0$, $y > 0$, transcritical bifurcations at $(0, 0)$ and at $(0, \frac{r_2}{k_2})$, and saddle node bifurcations at a coexistence equilibrium as well as at a boundary equilibrium $(x^*, 0)$, for some $x^* > 0$. It also shows the existence of Hopf bifurcations from a coexistence equilibrium, which gives out several closed orbits or periodic solutions. See Figures 8.14 and 8.15. The study of this kind of models is relatively new, and provides us with excellent illustrations of the theory and applications of dynamical systems introduced here.

REFERENCES

1. R. Agu, H. Nghiem, J. Rebaza. *Grouping problem in networks.* Technical Report Universität Kaiserlautern, Technomathematik (1994).
2. R. Andersen, Y. Peres. *Finding sparse cuts locally using evolving sets.* Proceedings of 41 ACM Symposium on Theory of Computing (2009), 235–244.
3. P. Berkhin. *Bookmark-coloring algorithm for personalized PageRank computing.* Internet Mathematics **3** (2005), 41–62.
4. A. Berman, R. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Classics in Applied Mathematics, SIAM (1994).
5. M. Berry, Z. Drmac, E. Jessup. *Matrices, vector spaces and information retrieval.* SIAM Review **41** (1999), 335–362.
6. W.J. Beyn. *The numerical computation of connecting orbits in dynamical systems.* IMA J. of Numerical Analysis **9** (1990), 379–405.
7. W. Beyn. *On well-posed problems for connecting orbits in dynamical systems.* Contemporary Mathematics **172** (1994), 131–168.
8. J. Bohn, J. Rebaza, K. Speer. *Continuous threshold prey harvesting in predator–prey models.* Int. J. of Computational and Mathematical Sciences **5** (2011), 111–118.
9. W. Boyce, R. Di Prima. *Elementary Differential Equations and Boundary Value Problems.* Wiley (2009).
10. F. Brauer, J. Nohel. *The Qualitative Theory of Ordinary Differential Equations, an Introduction.* Dover Publications (1989).
11. S. Brin, L. Page, T. Winograd. *The PageRank citation ranking: bringing order to the web.* Technical Report 1999-0120, Stanford University (1999).
12. J. Carreño, T. Lara, J. Rebaza. *On a model of American cutaneous leishmaniasis.* G. J. of Pure and Applied Mathematics **5** (2009), 263–275.

Copyright © 2012 John Wiley & Sons, Inc.

13. A.R. Champneys, G. J. Lord. *Computation of homoclinic solutions to periodic orbits in a reduced water-wave problem.* Physica D **102** (1997), 101–124.

14. L. F. Chaves, M. J. Hernandez, S. Ramos. *Simulación de modelos matemáticos como herramienta para el estudio de los reservorios de leishmaniasis cutánea americana.* Divulg. Matemáticas **16** (2008), 125–154.

15. S.-N. Chow, C. Li, D. Wang. *Normal Forms and Bifurcation of Planar Vector Fields.* Cambridge University Press (2009).

16. G. del Corso, A. Gulli, F. Romani. *Exploiting web matrix permutations to speedup PageRank computation.* Technical Report TR-04/2004 Istituto di Informatica e Telematica, CNR (2004).

17. L. Dieci, T. Eirola. *Numerical Dynamical Sytems.* Book manuscript (2011).

18. L. Dieci, T. Eirola. *On smooth decompositions of matrices.* SIAM J. of Matrix Analysis and Applications **20** (1999), 800–819.

19. L. Dieci, J. Rebaza. *Point-to-periodic and periodic-to-periodic connections.* BIT Numerical Mathematics **44** (2004), 41–62.

20. E. Doedel, B. Kooi, G. Van Voorn, Y. Kuznetsov. *Continuation of connecting orbits in 3D-ODEs (II) Point-to-cycle connections.* Int. J. of Bifurcation and Chaos **18** (2008), 1889–1903.

21. E. Doedel, A. Champneys, T. Fairgrieve, R. Paffenroth, Y. Kuznetsov, B. Sandstede, X.J. Wang, *AUTO 2000: Continuation and bifurcation software for ordinary differential equations (with HomCont).* Technical Report, (2001).

22. E. Doedel, L. Tuckerman (Editors). *Numerical Methods for Bifurcation Problems and Large-Scale Dynamical Systems.* IMA Volumes in Mathematics and its Applications. Springer Verlag (2000).

23. C. Edwards, D. Penney. *Elementary Differential Equations.* Prentice Hall (2007).

24. B. Ermentrout. *Simulating, Analyzing, and Animating Dynamical Systems.* SIAM (2002).

25. G. Golub, C. Van Loan. *Matrix Computations.* Johns Hopkins University Press (1996).

26. W. Govaerts, Y. A. Kuznetsov. *MATCONT: A MATLAB package for numerical bifurcation analysis of ODEs.* ACM Transactions on Mathematical Software **29** (2003), 141-164.

27. J. Guckenheimer, P. Holmes. *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields.* Springer Verlag (2002).

28. J. Hale. *Ordinary Differential Equations.* Dover Publications (2009).

29. J. Hale, H. Koçak. *Dynamics and Bifurcations.* TAM Springer Verlag (1996).

30. K. Han, S. Mathema, J. Rebaza. *Optimization of a waste water management system.* Technical report Universität Kaiserlautern, Technomathematik (1993).

31. P. Hartman. *Ordinary Differential Equations.* SIAM (2002).

32. Y. He, W. Barnett. *Singularity bifurcations.* J. of Macroeconomics **28** (2006), 5–22.

33. S. He, W. Barnett. *Existence of singularity bifurcation in an Euler equations model of the United States economy: Grandmont was right.* MPRA paper No 12803 (2009).

34. M. Hernandez, I. Barradas. *Variation in the outcome of population interactions: bifurcations and catastrophes.* Mathematical Biology **46** (2003), 571–594.

35. M. Hernandez. *Spatiotemporal dynamics in variable population interaction with density-dependent interaction coefficients.* Ecological Modelling **214** (2008), 3–16.

36. J. Holland and D. DeAngelis, *Consumer-resource theory predicts dynamic transitions between outcomes of interspecific interactions.* Ecology Letters **12** (2009), 1357–1366.

37. P. Holmes. *Ninety plus thirty years of nonlinear dynamics: less is more and more is different.* Int. J. of Bifurcations and Chaos **15** (2005), 2703–2716.

38. R. Horn, C. Johnson. *Matrix Analysis*. Cambridge University Press (1990).

39. I. Ipsen, S. Kirkland. *Convergence analysis of a PageRank algorithm by Langville and Meyer*. SIAM J. Matrix Analysis and Applications **27** (2006), 952–967.

40. W. S. Koon, M. W. Lo, J.E. Marsden. *Dynamical Systems, the Three-Body Problem and Space Mission Design*. Springer Verlag (2007).

41. W. S. Koon, M. W. Lo, J.E. Marsden, S. D. Ross. *Heteroclinic connections between periodic orbits and resonance transitions in celestial mechanics*. Chaos, **10** (2000), 427–469.

42. D. Kreider, R. Kuller, D. Ostberg. *Elementary Differential Equations*. Addison-Wesley (1969).

43. Y. Kuznetsov. *Elements of Applied Bifurcation Theory*. TAM Springer Verlag (2004).

44. A. Langville, C. D. Meyer. *A survey of eigenvector methods for web information retrieval*. SIAM Review **47** (2005), 135–161.

45. A. N. Langville, C. D. Meyer. *Deeper inside PageRank*. Internet Mathematics **1** (2004), 335–380.

46. T. Lara, J. Rebaza. *Dynamics of transitions in population interactions*. To appear in Nonl. Analysis: Real World Appl. (2012).

47. B. Leard, J. Rebaza. *Analysis of predator–prey models with continuous threshold harvesting*. Applied Mathematics and Computation **217** (2011), 5265–5278.

48. Y. Li, J. Muldowney. *On Bendixon's criterion*. J. Differential Equations **106** (1993), 27–39.

49. C. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM (2000).

50. C. D. Meyer. *Stochastic complementation, uncoupling Markov chains and the theory of nearly reducible systems*. SIAM Review **31** (1989), 240–272.

51. M. E. M. Meza, A. Bhaya, Kaszkurewiczk, M. S. Costa. *Threshold policies control for predatorprey systems using a control Liapunov function approach*. Theoretical Population Biology **67** (2005), 273–284.

52. C. Moler, D. Morrison. *Singular value analysis of cryptograms*. American Mathematical Monthly **90** (1983), 78–87.

53. S. Osborne, J. Rebaza, E. Wiggins. *On accelerating PageRank computation*. Internet Mathematics **6** (2009), 29–43.

54. L. Perko. *Differential Equations and Dynamical Systems*. Springer Verlag (2006).

55. J. Rebaza. *Computation of traveling waves via an algorithm for connecting orbits*. Communications on Applied Nonlinear Analysis **13** (2006), 43–59.

56. J. Rebaza. *Smooth Schur factorizations in the continuation of separatrices*. Linear Algebra and its Applications **421** (2007), 138–156.

57. J. Rebaza *Dynamics of prey threshold harvesting and refuge*. J. Comp. & Appl. Math. (2011) doi:10.1016/j.cam.2011.10.005.

58. J. Rebaza. *Bifurcations and periodic orbits in variable population interactions*. Preprint (2011).

59. D. Salomon. *Data Compression: The Complete Reference*. Springer Verlag (2006).

60. T. Sauer. *Numerical analysis*. Addison–Wesley (2005).

61. C. Sparrow. *The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors*. Dover Publications (2005).

62. G. Stewart. *Matrix Algorithm Vol I: Basic Decompositions*. SIAM (1998).

63. G. Stewart. *Matrix Algorithm Vol II: Eigensystems*. SIAM (2001).

64. G. Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press (1986).

65. K. S. Thyagarajan. *Digital Image Processing with Application to Digital Cinema.* Focal Press (2006).

66. D. Watkins. *Fundamentals of Matrix Computations.* Wiley-Interscience (2002).

67. S. Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos, 2nd Ed.* TAM Springer Verlag (2003).

68. B. Zhang, Z. Zhang, Z. Li, Y. Tao, *Stability analysis of a two-species model with transitions between population interactions.* J. Theoretical Biology **248** (2007), 145–153.

INDEX

C^1 norm, 365
LU factorization, 132
    and permutations, 139
    existence and uniqueness, 136
N-median problem, 405
QR factorization, 142, 153
    and least squares, 224
    and linear systems, 154
    Householder, 153
    thin, 144
    uniqueness, 147
SVD compression, 267

A-invariant, 166
AC coefficients, 262, 264
Adjacency matrix, 86
Adjoint of a matrix, 3
Aggregated matrix, 112, 129
Aggregation/disaggregation
    exact, 113
Aggregation/dissagregation
    approximate, 115
Algebraic multiplicity, 54
American cutaneous leishmaniasis, 410
Approximate aggregation matrix, 114
Approximate eigenvector, 75
Asymptotically stable equilibrium, 328

Backward substitution, 133
Banach space, 356
Baseline JPEG, 257
Basis, 27
    canonical, 27, 28
    orthonormal, 175
Beats, 295
Bernstein polynomials, 29, 59
Best teacher, 197
Bezier curve, 61, 76
    composite, 66
    midpoint property, 64
Bezier spline, 66
Bifurcaion
    saddle-node, 430
Bifurcation, 365, 379
    cusp, 393
    fold, 369, 380
    heteroclinic, 394
    homoclinic, 393
    Hopf, 372, 380, 384, 430
    period doubling, 363, 392
    pitchfork, 370
    saddle-node, 369, 392, 426
    saddle-node of periodic orbits, 393
    singularity, 373
    transcritical, 367, 384, 425, 430
Bifurcation diagram, 367, 380
Bit size, 263
Block diagonalization, 180
Block Schur factorization, 170, 175, 359
Block triangular matrix, 10, 167, 169
    eigenvalues of, 51

A First Course in Applied Mathematics. By Jorge Rebaza
Copyright © 2012 John Wiley & Sons, Inc.

Boundedness of solutions, 376

Canonical basis, 27, 28
Cauchy sequence, 281
Cauchy–Schwarz inequality, 69
Center, 329, 332, 335
Center subspace, 338
Chaos, 360
Characteristic equation, 48, 288
Choleski, 13
Choleski factorization, 206
Chrominance matrix, 258
Chua's circuit, 363
Cipher, 195
Circuits, 302
Clustering, 192
Codomain, 42
Column space, 34
    orthonormal basis, 158
    orthonormal basis of, 143
Competition, 286, 421
Complementary spaces, 32
Compression of color images, 257
Compression with SVD, 267
Condition number, 18, 73, 154
Connecting orbits, 352, 380, 418
Consistent norm, 8
Contraction, 280, 341
Control points, 61
Convergent sequence, 281
Convex hull, 64
Cooperative systems, 415
Cosine of an angle, 187
Cryptogram, 194
Cusp bifurcation, 393

Damped equations, 298
Damping, 288
Damping constant, 288
Dangling node, 95, 109, 118, 128
    reordering, 109
DC coefficient, 262
DCT, 236
    1–dimensional, 242
    1-dimensional, 236
DCT coefficients, 252
DCT interpolation theorem, 237
DCT least squares approximation, 240, 245
Defect of an eigenvalue, 311
Defective eigenvalue, 54, 310
Degenerate equilibriumm point, 334
Dependence on initial conditions, 340
Determinant, 11, 58
Diagonal matrix, 9
Diagonalizable matrix, 57, 76
Differentiable manifold, 343
Digram frequency matrix, 195
Dimension formula, 39, 42
Dimension of a vector space, 29
Direct sum, 32, 338
    basis of, 75
Discrete cosine transform, 236, 242
Domain, 42
Dominant eigenspace, 124
Dominant eigenvector, 81
Dot product, 8
Double-scroll attractor, 363
Dynamical system, 327

Eigenbasis, 177
Eigenblock, 177
Eigenpair, 177
Eigenspace, 54
    dominant, 124
    generalized, 387
Eigenvalue, 48, 307
    complex, 309
    defective, 54, 310
    distinct, 308
    of a symmetric matrix, 52, 75
    of transpose, 50
    repeated, 309
    simple, 54
Eigenvector, 48, 307
    approximate, 75
    dominant, 81
    generalized, 54, 310, 311, 322
    left, 179
Electrical resonance, 306
Elementary row operations, 133
Emigration, 284
    time-varying, 285
Entropy, 260
Equilibrium equations, 229
Equilibrium point, 327, 341
    asymptotically stable, 328
    stable, 328
    unstable, 328
Exchanging eigenblocks, 177
Existence and uniqueness, 278, 283, 340

Fixed point, 280
Fixed point iteration, 281
Flame fronts, 391
Floquet exponent, 350
Floquet multiplier, 349
Floquet's theorem, 388
Flow of the system, 342
Focus, 329, 331, 335
Fold bifurcation, 369, 380
Frobenius norm, 6, 7, 69, 70, 209
Full rank matrix, 143
Fundamental matrix solution, 311
Fundamental theorem of linear algebra, 39

Gaussian elimination, 133
Generalized eigenspace, 387
Generalized eigenvector, 54, 310, 311, 322

Geometric multiplicity, 54
Gershgorin circle theorem, 75, 211
Google, 92
Google matrix, 97
Gram–Schmidt, 31
Gram-Schmidt, 72, 143
Graph of a matrix, 86
Gronwall's inequality, 388
Grouping in networks, 404

Haar matrix, 273
Hadamard matrix, 273
Hartman–Grobman theorem, 347
Harvesting, 375, 430
Hermitian matrix, 13
Hessenberg matrix, 10, 151
    properly, 211
Heteroclinic bifurcation, 394
Heteroclinic orbit, 352
Homeomorphism, 343, 389
Homoclinic bifurcation, 393
Homoclinic orbit, 352
Hooke's law, 288
Hopf bifurcation, 372, 380, 384, 430
Host-parasite, 421
Householder
    QR factorization, 153
Householder matrix, 19, 151, 152
Hue, 246
Huffman coding, 260
Huffman coding tree, 261
Human visual system, 246
Hyperbolic equilibrium, 342
Hyperbolic periodic orbit, 349

IAD algorithm, 115
Identity matrix, 11
Image, 35, 42
Impedance, 306
Induced norm, 7, 69
Information retrieval, 186
Initial value problem, 278
Inner product, 8
Invariant set, 338
Invariant subspace, 48, 124, 166, 172, 173
Inverse discrete cosine transform, 238, 242
Inverse matrix, 11
    computation of, 138
Invertible matrix, 11
Irreducible matrix, 86, 96
Isolated equilibrium, 334
Iterative aggregation/disaggregation, 111, 115

JPEG, 257
JPEG standard, 254

Kernel, 42
Kirchhoff's law, 304
Krylov subspace, 211
Kuramoto-Sivashinsky equations, 391

Lagrange multipliers, 228
Lastgrad, 405
Leading principal submatrix, 14, 136
Least squares, 218
    and QR, 224
    rank deficient, 226
Left eigenvector, 55, 74, 76, 179
Left inverse matrix, 72
Level shifting, 251
Linear combination, 24
Linear independence, 26
Linear operator, 41
Linear regression, 219
Linear system, 40, 48, 73, 106
    overdetermined, 216
Linear systems, 132, 137, 141
Linear transformation, 40
Linearization, 342
    around a periodic orbit, 348
    around an equilibrium, 342
Link matrix, 86
Lipschitz function, 279, 341
Locally Lipschitz, 341, 389
Logistic equation, 286
Lorenz system, 358, 361
Lossless compression, 246
Lossy compression, 246
Lotka–Volterra model, 375
Low-pass filtering, 250
Low-rank approximation, 161, 163, 197, 199
Lower triangular matrix, 9
Luminance quantization matrix, 254

M-matrix, 127
Manifold, 343
Markov chain, 111
Matrix, 1
    addition, 2
    adjacency, 86
    adjoint, 3
    aggregated, 112, 129
    block triangular, 10, 167
    chrominance, 258
    diagonal, 9
    diagonalizable, 57, 76
    digram frequency, 195
    exponential, 14
    full rank, 143
    Haar, 273
    Hadamard, 273
    Hermitian, 13
    Hessenberg, 10, 151
    Householder, 19, 151, 152
    indentity, 11
    inverse, 11
    invertible, 11
    irreducible, 86, 96

link, 86
lower triangular, 70
M-, 127
nilpotent, 212
nonnegative, 84
nonsingular, 11
orthogonal, 17, 72, 73
orthogonal projection, 44
permutation, 15
positive, 86
positive definite, 12, 14
primitive, 91
product, 2
projection, 43, 76, 221, 233
pseudoinverse, 45, 209
rank-1, 159
reducible, 86
skew hermitian, 34
skew symmetric, 34, 75
square root, 210
stochastic, 85, 95
strictly diagonal dominant, 71
symmetric, 12
symmetric positive definite, 13
term-document, 186
transition, 111
transpose, 3
trapezoidal, 10
triangular, 9, 70
tridiagonal, 70
unitary, 17
Matrix exponential, 72, 326
Matrix norm, 5, 59
spectral, 165
submultiplicative, 70
Monodromy matrix, 349
Morse–Smale system, 366
Multiplicity, 54
Multipliers, 134
Mutualism, 421

Natural frequency, 293
Network grouping, 404
Newton's law, 288
Nilpotent matrix, 212
Node, 329, 330, 335, 336
Nonhomogeneous equations, 292
Nonhomogeneous systems, 314
Nonnegative matrix, 84
Nonsingular matrix, 11
Nonwandering set, 366
Norm
C^1, 365
Euclidean, 5
Frobenius, 6, 7, 209
induced, 7, 69
of a matrix, 5
of a vector, 4
Normal equations, 44, 83, 218
Null space, 38
Nullity, 42

Ordered block Schur factorization, 170, 175
Orthogonal complement, 33, 73
Orthogonal matrix, 17, 72, 73
Orthogonal projection, 44, 46, 72, 234
Orthogonal vectors, 9
Orthogonality, 8
Orthogonalization, 31
Orthogonally invariant, 72
Orthonormal basis, 31, 46, 175
Orthonormal extension, 20, 178
Orthonormal vectors, 9
Orthonormalization, 144
Overdetermined system, 216

PageRank vector, 93
Particular solution, 293
Peixoto's theorem, 366
Period doubling bifurcation, 363, 392
Periodic orbit, 349
Periodic solution, 348
Permutation matrix, 15
Perp, 33
Perron–Frobenius theorem, 89
Personalization vector, 97, 99
Phase angle, 294
Phase condition, 353
Phase portrait, 327
Phase space, 327
Pitchfork bifurcation, 370
Pixel, 246
Population model, 283
Positive definite matrix, 12, 14
Positive matrix, 86
Postscript fonts, 59, 67
Power method, 80, 103
rate of convergence, 82
Practical resonance, 301
Predator–prey models, 374, 390, 392, 393
Primitive matrix, 91
Principal submatrix, 14
Probabilistic vector, 90
Projection, 32, 72
Projection boundary conditions, 353, 356
Projection matrix, 43, 76, 221, 233
Pseudoinverse, 45, 209

QR algorithm, 149
Quadratic form, 216
Quantization, 254
Query matching, 187

Rössler attractor, 392
Radio tuning, 306
Range, 35, 42
Rank, 35, 42, 76
Rank deficient least squares, 226

Rank-1 approximation, 197
Rank-1 matrix, 159
    modification, 108
Rank-2 approximation, 199
Rayleigh quotient, 121
Reducible matrix, 86
Reordering, 109, 128
Residual, 184
Resonance, 297
    electrical, 306
    practical, 301
Resonance frequency, 306
RGB, 257
RLC circuits, 302
Row space, 36
    orthonormal basis, 158

Saddle, 328, 329, 335
Saddle-node bifurcation, 369, 392
    of periodic orbits, 393
Saturation, 246
Schur factorization, 166, 174, 354
    block, 170, 175, 359
    complex, 171
    ordered, 176, 180, 354
    real, 168
    uniqueness, 170
Separatrices, 352
Sherman–Morrison formula, 108, 128
Similarity, 56, 148, 327
    invariant properties, 58
Simple eigenvalue, 54
Singular value decomposition, 155, 163
    uniqueness, 158
Singular values, 155, 160
Singular vectors, 156
Singularity bifurcation, 393
Singularity bifurcations, 373
Skew hermitian matrix, 34
Skew symmetric matrix, 34, 75
Span, 25
Spd, 13
Spectral radius, 70, 123
Spectrum of a matrix, 51
Spetral norm, 165
Spring systems, 287
Square root of a matrix, 210
Stable equilibrium, 328
Stable manifold theorem, 343
    for periodic orbits, 350
Stable subspace, 338
Stationary distribution, 111
Steady solution, 300
Stochastic complement, 112
Stochastic matrix, 85, 95
Strange attractor
    Chua, 363
    Lorenz, 362
    Rössler, 392
Strictly diagonal dominant, 71
Strongly connected graph, 86
Structural stability, 365
Subdominant eigenvalue, 82
Submatrix, 14
    leading principal, 14, 136
    principal, 14
Submultiplicative inequality, 70
Subspace, 22, 73
    invariant, 166, 172, 173
Successive approximations, 341, 344
Sum of subspaces, 24
    dimension of, 30
Superposition principle, 313
Sylvester equation, 180
Symmetric matrix, 12
    eigenvalues, eigenvectors of, 75
    inverse of, 75
    product of, 75
Symmetric positive definite, 13

Teaching, 197
Term–term comparison, 192
Term-document matrix, 186
Threshold condition, 416
Threshold policy harvesting, 375
Trace, 3, 69
Trace-determinant analysis, 335, 377, 417
Transcritical bifurcation, 367, 384
Transient solution, 299
Transition matrix, 111
Transpose, 3, 69
Transversal intersection, 354, 366
Trapezoidal matrix, 10
Traveling waves, 390
Triangular matrix, 9, 70
    eigenvalues of, 49
Tridiagonal matrix, 70
Tuning instruments, 296

Undamped equations, 292
Unit ball, 5
Unit lower triangular, 134
Unit vector, 9
Unitary matrix, 17
Unstable equilibrium, 328
Unstable subspace, 338
Upper triangular matrix, 9

Van der Pol oscillators, 360
Variable population interactions, 420
Variation of constants formula, 323
Vector, 4
    orthonormal, 9
    probabilistic, 90
    unit, 9
Vector norm, 4
Vector space, 21
Vector subspace, 22, 73
Vfc rule, 195, 196

Waste water management, 395
Weakly diagonal dominant, 71
Wronskian, 319

Zigzag pattern, 265