THE OXFORD SERIES IN ELECTRICAL AND COMPUTER ENGINEERING

Adel S. Sedra, Series Editor, Electrical Engineering
Michael R. Lightner, Series Editor, Computer Engineering

Allen and Holberg, *CMOS Analog Circuit Design*
Bobrow, *Elementary Linear Circuit Analysis, 2nd Ed.*
Bobrow, *Fundamentals of Electrical Engineering, 2nd Ed.*
Campbell, *The Science and Engineering of Microelectroinic Fabrication*
Chen, *Analog and Digital Control System Design*
Chen, *Linear System Theory and Design, 3rd Ed.*
Chen, *System and Signal Analysis, 2nd Ed.*
Comer, *Digital Logic and State Machine Design, 3rd Ed.*
Cooper and McGillem, *Probabilistic Methods of Signal and System Analysis, 3rd Ed.*
Fortney, *Principles of Electronics: Analog & Digital*
Franco, *Electric Circuits Fundamentals*
Granzow, *Digital Transmission Lines*
Guru and Hiziroglu, *Electric Machinery & Transformers, 2nd Ed.*
Hoole and Hoole, *A Modern Short Course in Engineering Electromagnetics*
Jones, *Introduction to Optical Fiber Communication Systems*
Krein, *Elements of Power Electronics*
Kuo, *Digital Control Systems, 3rd Ed.*
Lathi, *Modern Digital and Analog Communications Systems, 3rd Ed.*
McGillem and Cooper, *Continuous and Discrete Signal and System Analysis, 3rd Ed.*
Miner, *Lines and Electromagnetic Fields for Engineers*
Roberts and Sedra, *SPICE, 2nd Ed.*
Roulston, *An Introduction to the Physics of Semiconductor Devices*
Sadiku, *Elements of Electromagnetics, 2nd Ed.*
Santina, Stubberud, and Hostetter, *Digital Control System Design, 2nd Ed.*
Schwarz, *Electromagnetics for Engineers*
Schwarz and Oldham, *Electrical Engineering: An Introduction, 2nd Ed.*
Sedra and Smith, *Microelectronic Circuits, 4th Ed.*
Stefani, Savant, Shahian, and Hostetter, *Design of Feedback Control Systems, 3rd Ed.*
Van Valkenburg, *Analog Filter Design*
Warner and Grung, *Semiconductor Device Electronics*
Wolovich, *Automatic Control Systems*
Yariv, *Optical Electronics in Modern Communications, 5th Ed.*

LINEAR SYSTEM THEORY AND DESIGN

Third Edition

Chi-Tsong Chen
State University of New York at Stony Brook

New York Oxford
OXFORD UNIVERSITY PRESS
1999

OXFORD UNIVERSITY PRESS

Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Singapore Taipei Tokyo Toronto Warsaw

and associated companies in
Berlin Ibadan

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Chen, Chi-Tsong
Linear system theory and design / by Chi-Tsong Chen. — 3rd ed.
p. cm. — (The Oxford series in electrical and computer engineering)
Includes bibliographical references and index.
ISBN 0-19-511777-8 (cloth).
1. Linear systems. 2. System design. I. Title. II. Series.
QA402.C44 1998
629.8'32—dc21 97-35535
CIP

Printing (last digit): 9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

To
BIH-JAU

Contents

Preface

This text is intended for use in senior/first-year graduate courses on linear systems and multivariable system design in electrical, mechanical, chemical, and aeronautical departments. It may also be useful to practicing engineers because it contains many design procedures. The mathematical background assumed is a working knowledge of linear algebra and the Laplace transform and an elementary knowledge of differential equations.

Linear system theory is a vast field. In this text, we limit our discussion to the conventional approaches of state-space equations and the polynomial fraction method of transfer matrices. The geometric approach, the abstract algebraic approach, rational fractions, and optimization are not discussed.

We aim to achieve two objectives with this text. The first objective is to use simple and efficient methods to develop results and design procedures. Thus the presentation is not exhaustive. For example, in introducing polynomial fractions, some polynomial theory such as the Smith–McMillan form and Bezout identities are not discussed. The second objective of this text is to enable the reader to employ the results to carry out design. Thus most results are discussed with an eye toward numerical computation. All design procedures in the text can be carried out using any software package that includes singular-value decomposition, QR decomposition, and the solution of linear algebraic equations and the Lyapunov equation. We give examples using MATLAB®, as the package[1] seems to be the most widely available.

This edition is a complete rewriting of the book *Linear System Theory and Design*, which was the expanded edition of *Introduction to Linear System Theory* published in 1970. Aside from, hopefully, a clearer presentation and a more logical development, this edition differs from the book in many ways:

- The order of Chapters 2 and 3 is reversed. In this edition, we develop mathematical descriptions of systems before reviewing linear algebra. The chapter on stability is moved earlier.
- This edition deals only with real numbers and foregoes the concept of *fields*. Thus it is mathematically less abstract than the original book. However, all results are still stated as theorems for easy reference.
- In Chapters 4 through 6, we discuss first the time-invariant case and then extend it to the time-varying case, instead of the other way around.

1. MATLAB is a registered trademark of the MathWorks, Inc., 24 Prime Park Way, Natick, MA 01760-1500. Phone: 508-647-7000, fax: 508-647-7001, E-mail: info@mathworks.com, http://www.mathworks.com.

- The discussion of discrete-time systems is expanded.
- In state-space design, Lyapunov equations are employed extensively and multivariable canonical forms are downplayed. This approach is not only easier for classroom presentation but also provides an attractive method for numerical computation.
- The presentation of the polynomial fraction method is streamlined. The method is equated with solving linear algebraic equations. We then discuss pole placement using a one-degree-of-freedom configuration, and model matching using a two-degree-of-freedom configuration.
- Examples using MATLAB are given throughout this new edition.

This edition is geared more for classroom use and engineering applications; therefore, many topics in the original book are deleted, including strict system equivalence, deterministic identification, computational issues, some multivariable canonical forms, and decoupling by state feedback. The polynomial fraction design in the input/output feedback (controller/estimator) configuration is deleted. Instead we discuss design in the two-parameter configuration. This configuration seems to be more suitable for practical application. The eight appendices in the original book are either incorporated into the text or deleted.

The logical sequence of all chapters is as follows:

$$\textit{Chap. } 1\text{–}5 \Rightarrow \begin{cases} \textit{Chap. } 6 \Rightarrow \begin{cases} \textit{Chap. } 8 \\ \textit{Chap. } 7 \end{cases} \\ \textit{Sec. } 7.1\text{–}7.3 \Rightarrow \textit{Sec. } 9.1\text{–}9.3 \\ \qquad \Rightarrow \textit{Sec. } 7.6\text{–}7.8 \Rightarrow \textit{Sec. } 9.4\text{–}9.5 \end{cases}$$

In addition, the material in Section 7.9 is needed to study Section 8.6.4. However, Section 8.6.4 may be skipped without loss of continuity. Furthermore, the concepts of controllability and observability in Chapter 6 are useful, but not essential for the material in Chapter 7. All minimal realizations in Chapter 7 can be checked using the concept of degrees, instead of checking controllability and observability. Therefore Chapters 6 and 7 are essentially independent.

This text provides more than enough material for a one-semester course. A one-semester course at Stony Brook covers Chapters 1 through 6, Sections 8.1–8.5, 7.1–7.2, and 9.1–9.3. Time-varying systems are not covered. Clearly, other arrangements are also possible for a one-semester course. A solutions manual is available from the publisher.

I am indebted to many people in revising this text. Professor Imin Kao and Mr. Juan Ochoa helped me with MATLAB. Professor Zongli Lin and Mr. T. Anantakrishnan read the whole manuscript and made many valuable suggestions. I am grateful to Dean Yacov Shamash, College of Engineering and Applied Sciences, SUNY at Stony Brook, for his encouragement. The revised manuscript was reviewed by Professor Harold Broberg, EET Department, Indiana Purdue University; Professor Peyman Givi, Department of Mechanical and Aerospace Engineering, State University of New York at Buffalo; Professor Mustafa Khammash, Department of Electrical and Computer Engineering, Iowa State University; and Professor B. Ross Barmish, Department of Electrical and Computer Engineering, University of Wisconsin. Their detailed and critical comments prompted me to restructure some sections and to include a number of mechanical vibration problems. I thank them all.

I am indebted to Mr. Bill Zobrist of Oxford University Press who persuaded me to undertake this revision. The people at Oxford University Press, including Krysia Bebick, Jasmine Urmeneta, Terri O'Prey, and Kristina Della Bartolomea were most helpful in this undertaking. Finally, I thank my wife, Bih-Jau, for her support during this revision.

Chi-Tsong Chen

LINEAR SYSTEM THEORY AND DESIGN

Chapter 1

Introduction

1.1 Introduction

The study and design of physical systems can be carried out using empirical methods. We can apply various signals to a physical system and measure its responses. If the performance is not satisfactory, we can adjust some of its parameters or connect to it a compensator to improve its performance. This approach relies heavily on past experience and is carried out by trial and error and has succeeded in designing many physical systems.

Empirical methods may become unworkable if physical systems are complex or too expensive or too dangerous to be experimented on. In these cases, analytical methods become indispensable. The analytical study of physical systems consists of four parts: modeling, development of mathematical descriptions, analysis, and design. We briefly introduce each of these tasks.

The distinction between physical systems and models is basic in engineering. For example, circuits or control systems studied in any textbook are models of physical systems. A resistor with a constant resistance is a model; it will burn out if the applied voltage is over a limit. This power limitation is often disregarded in its analytical study. An inductor with a constant inductance is again a model; in reality, the inductance may vary with the amount of current flowing through it. Modeling is a very important problem, for the success of the design depends on whether the physical system is modeled properly.

A physical system may have different models depending on the questions asked. It may also be modeled differently in different operational ranges. For example, an electronic amplifier is modeled differently at high and low frequencies. A spaceship can be modeled as a particle in investigating its trajectory; however, it must be modeled as a rigid body in maneuvering. A spaceship may even be modeled as a flexible body when it is connected to a space station. In order to develop a suitable model for a physical system, a thorough understanding of the physical system and its operational range is essential. In this text, we will call a model of a physical system simply a *system*. Thus a physical system is a device or a collection of devices existing in the real world; a system is a model of a physical system.

Once a system (or model) is selected for a physical system, the next step is to apply various physical laws to develop mathematical equations to describe the system. For example, we apply Kirchhoff's voltage and current laws to electrical systems and Newton's law to mechanical systems. The equations that describe systems may assume many forms;

they may be linear equations, nonlinear equations, integral equations, difference equations, differential equations, or others. Depending on the problem under study, one form of equation may be preferable to another in describing the same system. In conclusion, a system may have different mathematical-equation descriptions just as a physical system may have many different models.

After a mathematical description is obtained, we then carry out analyses—quantitative and/or qualitative. In quantitative analysis, we are interested in the responses of systems excited by certain inputs. In qualitative analysis, we are interested in the general properties of systems, such as stability, controllability, and observability. Qualitative analysis is very important, because design techniques may often evolve from this study.

If the response of a system is unsatisfactory, the system must be modified. In some cases, this can be achieved by adjusting some parameters of the system; in other cases, compensators must be introduced. Note that the design is carried out on the model of the physical system. If the model is properly chosen, then the performance of the physical system should be improved by introducing the required adjustments or compensators. If the model is poor, then the performance of the physical system may not improve and the design is useless. Selecting a model that is close enough to a physical system and yet simple enough to be studied analytically is the most difficult and important problem in system design.

1.2 Overview

The study of systems consists of four parts: modeling, setting up mathematical equations, analysis, and design. Developing models for physical systems requires knowledge of the particular field and some measuring devices. For example, to develop models for transistors requires a knowledge of quantum physics and some laboratory setup. Developing models for automobile suspension systems requires actual testing and measurements; it cannot be achieved by use of pencil and paper. Computer simulation certainly helps but cannot replace actual measurements. Thus the modeling problem should be studied in connection with the specific field and cannot be properly covered in this text. In this text, we shall assume that models of physical systems are available to us.

The systems to be studied in this text are limited to linear systems. Using the concept of linearity, we develop in Chapter 2 that every linear system can be described by

$$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{G}(t, \tau)\mathbf{u}(\tau)\, d\tau \tag{1.1}$$

This equation describes the relationship between the input $\mathbf{u}$ and output $\mathbf{y}$ and is called the *input–output* or *external* description. If a linear system is lumped as well, then it can also be described by

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{1.2}$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \tag{1.3}$$

Equation (1.2) is a set of first-order differential equations and Equation (1.3) is a set of algebraic equations. They are called the *internal* description of linear systems. Because the vector $\mathbf{x}$ is called the *state*, the set of two equations is called the *state-space* or, simply, the *state* equation.

If a linear system has, in addition, the property of time invariance, then Equations (1.1) through (1.3) reduce to

$$\mathbf{y}(t) = \int_{0}^{t} \mathbf{G}(t - \tau)\mathbf{u}(\tau)\, d\tau \tag{1.4}$$

and

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{1.5}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{1.6}$$

For this class of linear time-invariant systems, the Laplace transform is an important tool in analysis and design. Applying the Laplace transform to (1.4) yields

$$\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}(s)\hat{\mathbf{u}}(s) \tag{1.7}$$

where a variable with a circumflex denotes the Laplace transform of the variable. The function $\hat{\mathbf{G}}(s)$ is called the *transfer matrix*. Both (1.4) and (1.7) are input–output or external descriptions. The former is said to be in the time domain and the latter in the frequency domian.

Equations (1.1) through (1.6) are called continuous-time equations because their time variable t is a continuum defined at every time instant in $(-\infty, \infty)$. If the time is defined only at discrete instants, then the corresponding equations are called discrete-time equations. This text is devoted to the analysis and design centered around (1.1) through (1.7) and their discrete-time counterparts.

We briefly discuss the contents of each chapter. In Chapter 2, after introducing the aforementioned equations from the concepts of lumpedness, linearity, and time invariance, we show how these equations can be developed to describe systems. Chapter 3 reviews linear algebraic equations, the Lyapunov equation, and other pertinent topics that are essential for this text. We also introduce the Jordan form because it will be used to establish a number of results. We study in Chapter 4 solutions of the state-space equations in (1.2) and (1.5). Different analyses may lead to different state equations that describe the same system. Thus we introduce the concept of equivalent state equations. The basic relationship between state-space equations and transfer matrices is also established. Chapter 5 introduces the concepts of bounded-input bounded-output (BIBO) stability, marginal stability, and asymptotic stability. Every system must be designed to be stable; otherwise, it may burn out or disintegrate. Therefore stability is a basic system concept. We also introduce the Lyapunov theorem to check asymptotic stability.

Chapter 6 introduces the concepts of controllability and observability. They are essential in studying the internal structure of systems. A fundamental result is that the transfer matrix describes only the controllable and observable part of a state equation. Chapter 7 studies minimal realizations and introduces coprime polynomial fractions. We show how to obtain coprime fractions by solving sets of linear algebraic equations. The equivalence of controllable and observable state equations and coprime polynomial fractions is established.

The last two chapters discuss the design of time-invariant systems. We use controllable and observable state equations to carry out design in Chapter 8 and use coprime polynomial fractions in Chapter 9. We show that, under the controllability condition, all eigenvalues of a system can be arbitrarily assigned by introducing state feedback. If a state equation is observable, full-dimensional and reduced-dimensional state estimators, with any desired

eigenvalues, can be constructed to generate estimates of the state. We also establish the separation property. In Chapter 9, we discuss pole placement, model matching, and their applications in tracking, disturbance rejection, and decoupling. We use the unity-feedback configuration in pole placement and the two-parameter configuration in model matching. In our design, no control performances such as rise time, settling time, and overshoot are considered; neither are constraints on control signals and on the degree of compensators. Therefore this is not a control text per se. However, all results are basic and useful in designing linear time-invariant control systems.

Chapter 2

Mathematical Descriptions of Systems

2.1 Introduction

The class of systems studied in this text is assumed to have some input terminals and output terminals as shown in Fig. 2.1. We assume that if an excitation or input is applied to the input terminals, a *unique* response or output signal can be measured at the output terminals. This unique relationship between the excitation and response, input and output, or cause and effect is essential in defining a system. A system with only one input terminal and only one output terminal is called a single-variable system or a single-input single-output (SISO) system. A system with two or more input terminals and/or two or more output terminals is called a multivariable system. More specifically, we can call a system a multi-input multi-output (MIMO) system if it has two or more input terminals and output terminals, a single-input multi-output (SIMO) system if it has one input terminal and two or more output terminals.

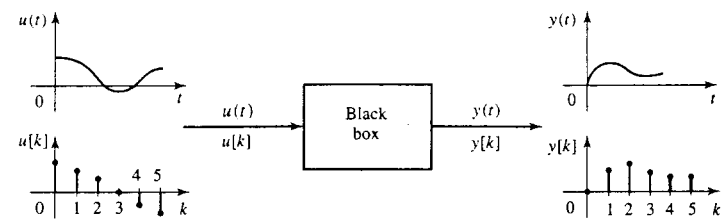


Figure 2.1 System.

A system is called a *continuous-time* system if it accepts continuous-time signals as its input and generates continuous-time signals as its output. The input will be denoted by lowercase italic $u(t)$ for single input or by boldface $\mathbf{u}(t)$ for multiple inputs. If the system has p input terminals, then $\mathbf{u}(t)$ is a $p \times 1$ vector or $\mathbf{u} = [u_1 \ u_2 \ \cdots \ u_p]'$, where the prime denotes the transpose. Similarly, the output will be denoted by $y(t)$ or $\mathbf{y}(t)$. The time t is assumed to range from $-\infty$ to ∞.

A system is called a *discrete-time* system if it accepts discrete-time signals as its input and generates discrete-time signals as its output. All discrete-time signals in a system will be assumed to have the same sampling period T. The input and output will be denoted by $u[k] := u(kT)$ and $y[k] := y(kT)$, where k denotes discrete-time instant and is an integer ranging from $-\infty$ to ∞. They become boldface for multiple inputs and multiple outputs.

2.1.1 Causality and Lumpedness

A system is called a *memoryless system* if its output $\mathbf{y}(t_0)$ depends only on the input applied at t_0; it is independent of the input applied before or after t_0. This will be stated succinctly as follows: current output of a memoryless system depends only on current input; it is independent of past and future inputs. A network that consists of only resistors is a memoryless system.

Most systems, however, have memory. By this we mean that the output at t_0 depends on $\mathbf{u}(t)$ for $t < t_0$, $t = t_0$, and $t > t_0$. That is, current output of a system with memory may depend on past, current, and future inputs.

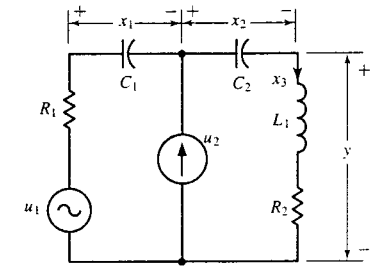
A system is called a *causal* or *nonanticipatory* system if its current output depends on past and current inputs but not on future input. If a system is not causal, then its current output will depend on future input. In other words, a noncausal system can *predict* or *anticipate* what will be applied in the future. No physical system has such capability. Therefore every physical system is causal and causality is a necessary condition for a system to be built or implemented in the real world. This text studies only causal systems.

Current output of a causal system is affected by past input. How far back in time will the past input affect the current output? Generally, the time should go all the way back to minus infinity. In other words, the input from $-\infty$ to time t has an effect on $\mathbf{y}(t)$. Tracking $\mathbf{u}(t)$ from $t = -\infty$ is, if not impossible, very inconvenient. The concept of state can deal with this problem.

> ***Definition 2.1*** *The state $\mathbf{x}(t_0)$ of a system at time t_0 is the information at t_0 that, together with the input $\mathbf{u}(t)$, for $t \geq t_0$, determines uniquely the output $\mathbf{y}(t)$ for all $t \geq t_0$.*

By definition, if we know the state at t_0, there is no more need to know the input $\mathbf{u}(t)$ applied before t_0 in determining the output $\mathbf{y}(t)$ after t_0. Thus in some sense, the state summarizes the effect of past input on future output. For the network shown in Fig. 2.2, if we know the voltages $x_1(t_0)$ and $x_2(t_0)$ across the two capacitors and the current $x_3(t_0)$ passing through the inductor, then for any input applied on and after t_0, we can determine uniquely the output for $t \geq t_0$. Thus the state of the network at time t_0 is

Figure 2.2 Network with 3 state variables.



$$\mathbf{x}(t_0) = \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \\ x_3(t_0) \end{bmatrix}$$

It is a 3×1 vector. The entries of $\mathbf{x}$ are called *state variables*. Thus, in general, we may consider the initial state simply as a set of initial conditions.

Using the state at t_0, we can express the input and output of a system as

$$\left.\begin{matrix} \mathbf{x}(t_0) \\ \mathbf{u}(t), \ t \geq t_0 \end{matrix}\right\} \rightarrow \mathbf{y}(t), \ t \geq t_0 \tag{2.1}$$

It means that the output is partly excited by the initial state at t_0 and partly by the input applied at and after t_0. In using (2.1), there is no more need to know the input applied before t_0 all the way back to $-\infty$. Thus (2.1) is easier to track and will be called a state-input–output pair.

A system is said to be *lumped* if its number of state variables is finite or its state is a finite vector. The network in Fig. 2.2 is clearly a lumped system; its state consists of three numbers. A system is called a *distributed* system if its state has infinitely many state variables. The transmission line is the most well known distributed system. We give one more example.

EXAMPLE 2.1 Consider the unit-time delay system defined by

$$y(t) = u(t - 1)$$

The output is simply the input delayed by one second. In order to determine $\{y(t), \ t \geq t_0\}$ from $\{u(t), \ t \geq t_0\}$, we need the information $\{u(t), \ t_0 - 1 \leq t < t_0\}$. Therefore the initial state of the system is $\{u(t), \ t_0 - 1 \leq t < t_0\}$. There are infinitely many points in $\{t_0 - 1 \leq t < t_0\}$. Thus the unit-time delay system is a distributed system.

2.2 Linear Systems

A system is called a *linear* system if for every t_0 and any two state-input–output pairs

$$\left.\begin{matrix} \mathbf{x}_i(t_0) \\ \mathbf{u}_i(t), \ t \geq t_0 \end{matrix}\right\} \rightarrow \mathbf{y}_i(t), \ t \geq t_0$$

for $i = 1, \ 2$, we have

$$\left.\begin{matrix}\mathbf{x}_1(t_0)+\mathbf{x}_2(t_0)\\ \mathbf{u}_1(t)+\mathbf{u}_2(t),\quad t\geq t_0\end{matrix}\right\}\rightarrow \mathbf{y}_1(t)+\mathbf{y}_2(t),\quad t\geq t_0 \quad \text{(additivity)}$$

and

$$\left.\begin{matrix}\alpha\mathbf{x}_1(t_0)\\ \alpha\mathbf{u}_1(t),\quad t\geq t_0\end{matrix}\right\}\rightarrow \alpha\mathbf{y}_1(t),\quad t\geq t_0 \quad \text{(homogeneity)}$$

for any real constant α. The first property is called the *additivity* property, the second, the *homogeneity* property. These two properties can be combined as

$$\left.\begin{matrix}\alpha_1\mathbf{x}_1(t_0)+\alpha_2\mathbf{x}_2(t_0)\\ \alpha_1\mathbf{u}_1(t)+\alpha_2\mathbf{u}_2(t),\quad t\geq t_0\end{matrix}\right\}\rightarrow \alpha_1\mathbf{y}_1(t)+\alpha_2\mathbf{y}_2(t),\quad t\geq t_0$$

for any real constants α_1 and α_2, and is called the *superposition property*. A system is called a nonlinear system if the superposition property does not hold.

If the input $\mathbf{u}(t)$ is identically zero for $t \geq t_0$, then the output will be excited exclusively by the initial state $\mathbf{x}(t_0)$. This output is called the *zero-input response* and will be denoted by $\mathbf{y}_{zi}$ or

$$\left.\begin{matrix}\mathbf{x}(t_0)\\ \mathbf{u}(t)\equiv\mathbf{0},\quad t\geq t_0\end{matrix}\right\}\rightarrow \mathbf{y}_{zi}(t),\quad t\geq t_0$$

If the initial state $\mathbf{x}(t_0)$ is zero, then the output will be excited exclusively by the input. This output is called the *zero-state response* and will be denoted by $\mathbf{y}_{zs}$ or

$$\left.\begin{matrix}\mathbf{x}(t_0)=\mathbf{0}\\ \mathbf{u}(t),\quad t\geq t_0\end{matrix}\right\}\rightarrow \mathbf{y}_{zs}(t),\quad t\geq t_0$$

The additivity property implies

$$\text{Output due to } \left\{\begin{matrix}\mathbf{x}(t_0)\\ \mathbf{u}(t),\quad t\geq t_0\end{matrix}\right. = \text{output due to } \left\{\begin{matrix}\mathbf{x}(t_0)\\ \mathbf{u}(t)\equiv\mathbf{0},\quad t\geq t_0\end{matrix}\right. + \text{output due to } \left\{\begin{matrix}\mathbf{x}(t_0)=\mathbf{0}\\ \mathbf{u}(t),\quad t\geq t_0\end{matrix}\right.$$

or

$$\textit{Response} = \textit{zero-input response} + \textit{zero-state response}$$

Thus the response of every linear system can be decomposed into the zero-state response and the zero-input response. Furthermore, the two responses can be studied separately and their sum yields the complete response. For nonlinear systems, the complete response can be very different from the sum of the zero-input response and zero-state response. Therefore we cannot separate the zero-input and zero-state responses in studying nonlinear systems.

If a system is linear, then the additivity and homogeneity properties apply to zero-state responses. To be more specific, if $\mathbf{x}(t_0) = \mathbf{0}$, then the output will be excited exclusively by the input and the state-input–output equation can be simplified as $\{\mathbf{u}_i \rightarrow \mathbf{y}_i\}$. If the system is linear, then we have $\{\mathbf{u}_1 + \mathbf{u}_2 \rightarrow \mathbf{y}_1 + \mathbf{y}_2\}$ and $\{\alpha\mathbf{u}_i \rightarrow \alpha\mathbf{y}_i\}$ for all α and all $\mathbf{u}_i$. A similar remark applies to zero-input responses of any linear system.

Input–output description We develop a mathematical equation to describe the zero-state response of linear systems. In this study, the initial state is assumed implicitly to be zero and the

output is excited exclusively by the input. We consider first SISO linear systems. Let $\delta_\Delta(t - t_1)$ be the pulse shown in Fig. 2.3. It has width Δ and height $1/\Delta$ and is located at time t_1. Then every input $u(t)$ can be approximated by a sequence of pulses as shown in Fig. 2.4. The pulse in Fig. 2.3 has height $1/\Delta$; thus $\delta_\Delta(t - t_i)\Delta$ has height 1 and the left-most pulse in Fig. 2.4 with height $u(t_i)$ can be expressed as $u(t_i)\delta_\Delta(t - t_1)\Delta$. Consequently, the input $u(t)$ can be expressed symbolically as

$$u(t) \approx \sum_i u(t_i)\delta_\Delta(t - t_i)\Delta$$

Let $g_\Delta(t, t_i)$ be the output at time t excited by the pulse $u(t) = \delta_\Delta(t - t_i)$ applied at time t_i. Then we have

$$\delta_\Delta(t - t_i) \rightarrow g_\Delta(t, t_i)$$

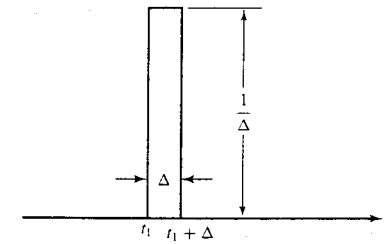
$$\delta_\Delta(t - t_i)u(t_i)\Delta \rightarrow g_\Delta(t, t_i)u(t_i)\Delta \quad \text{(homogeneity)}$$

$$\sum_i \delta_\Delta(t - t_i)u(t_i)\Delta \rightarrow \sum_i g_\Delta(t, t_i)u(t_i)\Delta \quad \text{(additivity)}$$

Thus the output $y(t)$ excited by the input $u(t)$ can be approximated by

$$y(t) \approx \sum_i g_\Delta(t, t_i)u(t_i)\Delta \tag{2.2}$$

Figure 2.3 Pulse at t_1.



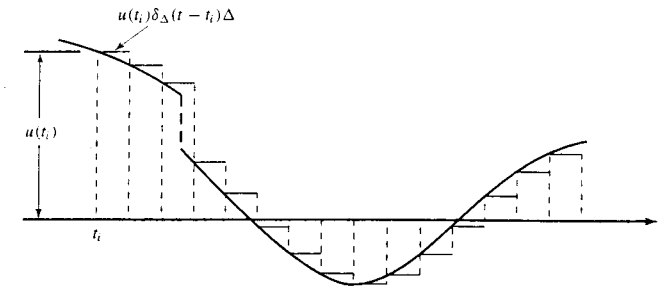


Figure 2.4 Approximation of input signal.

Now if Δ approaches zero, the pulse $\delta_\Delta(t-t_i)$ becomes an *impulse* at t_i, denoted by $\delta(t-t_i)$, and the corresponding output will be denoted by $g(t, t_i)$. As Δ approaches zero, the approximation in (2.2) becomes an equality, the summation becomes an integration, the discrete t_i becomes a continuum and can be replaced by τ, and Δ can be written as $d\tau$. Thus (2.2) becomes

$$y(t) = \int_{-\infty}^{\infty} g(t, \tau)u(\tau)\, d\tau \tag{2.3}$$

Note that $g(t, \tau)$ is a function of two variables. The second variable denotes the time at which the impulse input is applied; the first variable denotes the time at which the output is observed. Because $g(t, \tau)$ is the response excited by an impulse, it is called the *impulse response*.

If a system is causal, the output will not appear before an input is applied. Thus we have

$$\text{Causal} \iff g(t, \tau) = 0 \quad \text{for } t < \tau$$

A system is said to be *relaxed* at t_0 if its initial state at t_0 is $\mathbf{0}$. In this case, the output $y(t)$, for $t \geq t_0$, is excited exclusively by the input $u(t)$ for $t \geq t_0$. Thus the lower limit of the integration in (2.3) can be replaced by t_0. If the system is causal as well, then $g(t, \tau) = 0$ for $t < \tau$. Thus the upper limit of the integration in (2.3) can be replaced by t. In conclusion, every linear system that is causal and relaxed at t_0 can be described by

$$y(t) = \int_{t_0}^{t} g(t, \tau)u(\tau)\, d\tau \tag{2.4}$$

In this derivation, the condition of lumpedness is not used. Therefore any lumped or distributed linear system has such an input–output description. This description is developed using only the additivity and homogeneity properties; therefore every linear system, be it an electrical system, a mechanical system, a chemical process, or any other system, has such a description.

If a linear system has p input terminals and q output terminals, then (2.4) can be extended to

$$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{G}(t, \tau)\mathbf{u}(\tau)\, d\tau \tag{2.5}$$

where

$$\mathbf{G}(t, \tau) = \begin{bmatrix} g_{11}(t, \tau) & g_{12}(t, \tau) & \cdots & g_{1p}(t, \tau) \\ g_{21}(t, \tau) & g_{22}(t, \tau) & \cdots & g_{2p}(t, \tau) \\ \vdots & \vdots & & \vdots \\ g_{q1}(t, \tau) & g_{q2}(t, \tau) & \cdots & g_{qp}(t, \tau) \end{bmatrix}$$

and $g_{ij}(t, \tau)$ is the response at time t at the ith output terminal due to an impulse applied at time τ at the jth input terminal, the inputs at other terminals being identically zero. That is, $g_{ij}(t, \tau)$ is the impulse response between the jth input terminal and the ith output terminal. Thus $\mathbf{G}$ is called the *impulse response matrix* of the system. We stress once again that if a system is described by (2.5), the system is linear, relaxed at t_0, and causal.

State-space description Every linear lumped system can be described by a set of equations of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{2.6}$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \tag{2.7}$$

where $\dot{\mathbf{x}} := d\mathbf{x}/dt$.[1] For a p-input q-output system, $\mathbf{u}$ is a $p \times 1$ vector and $\mathbf{y}$ is a $q \times 1$ vector. If the system has n state variables, then $\mathbf{x}$ is an $n \times 1$ vector. In order for the matrices in (2.6) and (2.7) to be compatible, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ must be $n \times n$, $n \times p$, $q \times n$, and $q \times p$ matrices. The four matrices are all functions of time or time-varying matrices. Equation (2.6) actually consists of a set of n first-order differential equations. Equation (2.7) consists of q algebraic equations. The set of two equations will be called an n-dimensional *state-space* equation or, simply, *state* equation. For distributed systems, the dimension is infinity and the two equations in (2.6) and (2.7) are not used.

The input–output description in (2.5) was developed from the linearity condition. The development of the state-space equation from the linearity condition, however, is not as simple and will not be attempted. We will simply accept it as a fact.

2.3 Linear Time-Invariant (LTI) Systems

A system is said to be *time invariant* if for every state-input–output pair

$$\left.\begin{matrix} \mathbf{x}(t_0) \\ \mathbf{u}(t), \quad t \geq t_0 \end{matrix}\right\} \rightarrow \mathbf{y}(t), \quad t \geq t_0$$

and any T, we have

$$\left.\begin{matrix} \mathbf{x}(t_0 + T) \\ \mathbf{u}(t - T), \quad t \geq t_0 + T \end{matrix}\right\} \rightarrow \mathbf{y}(t - T), \quad t \geq t_0 + T \quad \text{(time shifting)}$$

It means that if the initial state is shifted to time $t_0 + T$ and the same input waveform is applied from $t_0 + T$ instead of from t_0, then the output waveform will be the same except that it starts to appear from time $t_0 + T$. In other words, if the initial state and the input are the same, no matter at what time they are applied, the output waveform will always be the same. Therefore, for time-invariant systems, we can always assume, without loss of generality, that $t_0 = 0$. If a system is not time invariant, it is said to be *time varying*.

Time invariance is defined for systems, not for signals. Signals are mostly time varying. If a signal is time invariant such as $u(t) = 1$ for all t, then it is a very simple or a trivial signal. The characteristics of time-invariant systems must be independent of time. For example, the network in Fig. 2.2 is time invariant if R_i, C_i, and L_i are constants.

Some physical systems must be modeled as time-varying systems. For example, a burning rocket is a time-varying system, because its mass decreases rapidly with time. Although the performance of an automobile or a TV set may deteriorate over a long period of time, its characteristics do not change appreciable in the first couple of years. Thus a large number of physical systems can be modeled as time-invariant systems over a limited time period.

1. We use $A := B$ to denote that A, by definition, equals B. We use $A =: B$ to denote that B, by definition, equals A.

Input–output description The zero-state response of a linear system can be described by (2.4). Now if the system is time invariant as well, then we have[2]

$$g(t, \tau) = g(t + T, \tau + T) = g(t - \tau, 0) = g(t - \tau)$$

for any T. Thus (2.4) reduces to

$$y(t) = \int_0^t g(t - \tau)u(\tau)\, d\tau = \int_0^t g(\tau)u(t - \tau)\, d\tau \tag{2.8}$$

where we have replaced t_0 by 0. The second equality can easily be verified by changing the variable. The integration in (2.8) is called a convolution integral. Unlike the time-varying case where g is a function of two variables, g is a function of a single variable in the time-invariant case. By definition $g(t) = g(t - 0)$ is the output at time t due to an impulse input applied at time 0. The condition for a linear time-invariant system to be causal is $g(t) = 0$ for $t < 0$.

EXAMPLE 2.2 The unit-time delay system studied in Example 2.1 is a device whose output equals the input delayed by 1 second. If we apply the impulse $\delta(t)$ at the input terminal, the output is $\delta(t - 1)$. Thus the impulse response of the system is $\delta(t - 1)$.

EXAMPLE 2.3 Consider the unity-feedback system shown in Fig. 2.5(a). It consists of a multiplier with gain a and a unit-time delay element. It is a SISO system. Let $r(t)$ be the input of the feedback system. If $r(t) = \delta(t)$, then the output is the impulse response of the feedback system and equals

$$g_f(t) = a\delta(t - 1) + a^2\delta(t - 2) + a^3\delta(t - 3) + \cdots = \sum_{i=1}^{\infty} a^i\delta(t - i) \tag{2.9}$$

Let $r(t)$ be any input with $r(t) \equiv 0$ for $t < 0$; then the output is given by

$$y(t) = \int_0^t g_f(t - \tau)r(\tau)\, d\tau = \sum_{i=1}^{\infty} a^i \int_0^t \delta(t - \tau - i)r(\tau)\, d\tau$$

$$= \sum_{i=1}^{\infty} a^i r(\tau)\Big|_{\tau=t-i} = \sum_{i=1}^{\infty} a^i r(t - i)$$

Because the unit-time delay system is distributed, so is the feedback system.

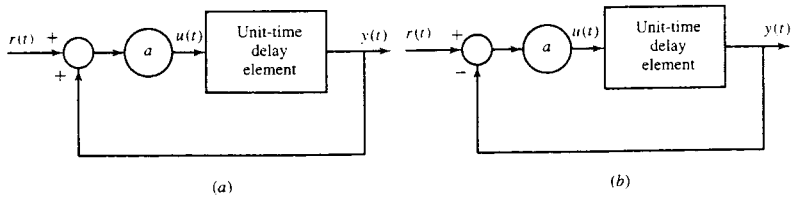


Figure 2.5 Positive and negative feedback systems.

2. Note that $g(t, \tau)$ and $g(t - \tau)$ are two different functions. However, for convenience, the same symbol g is used.

Transfer-function matrix The Laplace transform is an important tool in the study of linear time-invariant (LTI) systems. Let $\hat{y}(s)$ be the Laplace transform of $y(t)$, that is,

$$\hat{y}(s) = \int_0^\infty y(t)e^{-st}dt$$

Throughout this text, we use a variable with a circumflex to denote the Laplace transform of the variable. For causal systems, we have $g(t) = 0$ for $t < 0$ or $g(t - \tau) = 0$ for $\tau > t$. Thus the upper integration limit in (2.8) can be replaced by ∞. Substituting (2.8) and interchanging the order of integrations, we obtain

$$\hat{y}(s) = \int_{t=0}^{\infty} \left(\int_{\tau=0}^{\infty} g(t - \tau)u(\tau)\, d\tau \right) e^{-s(t-\tau)}e^{-s\tau}dt$$

$$= \int_{\tau=0}^{\infty} \left(\int_{t=0}^{\infty} g(t - \tau)e^{-s(t-\tau)}\, dt \right) u(\tau)e^{-s\tau}d\tau$$

which becomes, after introducing the new variable $v = t - \tau$,

$$\hat{y}(s) = \int_{\tau=0}^{\infty} \left(\int_{v=-\tau}^{\infty} g(v)e^{-sv}\, dv \right) u(\tau)e^{-s\tau}\, d\tau$$

Again using the causality condition to replace the lower integration limit inside the parentheses from $v = -\tau$ to $v = 0$, the integration becomes independent of τ and the double integrations become

$$\hat{y}(s) = \int_{v=0}^{\infty} g(v)e^{-sv}\, dv \int_{\tau=0}^{\infty} u(\tau)e^{-s\tau}\, d\tau$$

or

$$\hat{y}(s) = \hat{g}(s)\hat{u}(s) \tag{2.10}$$

where

$$\hat{g}(s) = \int_0^\infty g(t)e^{-st}dt$$

is called the *transfer function* of the system. Thus the transfer function is the Laplace transform of the impulse response and, conversely, the impulse response is the inverse Laplace transform of the transfer function. We see that the Laplace transform transforms the convolution integral in (2.8) into the algebraic equation in (2.10). In analysis and design, it is simpler to use algebraic equations than to use convolutions. Thus the convolution in (2.8) will rarely be used in the remainder of this text.

For a p-input q-output system, (2.10) can be extended as

$$\begin{bmatrix} \hat{y}_1(s) \\ \hat{y}_2(s) \\ \vdots \\ \hat{y}_q(s) \end{bmatrix} = \begin{bmatrix} \hat{g}_{11}(s) & \hat{g}_{12}(s) & \cdots & \hat{g}_{1p}(s) \\ \hat{g}_{21}(s) & \hat{g}_{22}(s) & \cdots & \hat{g}_{2p}(s) \\ \vdots & \vdots & & \vdots \\ \hat{g}_{q1}(s) & \hat{g}_{q2}(s) & \cdots & \hat{g}_{qp}(s) \end{bmatrix} \begin{bmatrix} \hat{u}_1(s) \\ \hat{u}_2(s) \\ \vdots \\ \hat{u}_p(s) \end{bmatrix}$$

or

$$\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}(s)\hat{\mathbf{u}}(s) \tag{2.11}$$

where $\hat{g}_{ij}(s)$ is the transfer function from the jth input to the ith output. The $q \times p$ matrix $\hat{\mathbf{G}}(s)$ is called the *transfer-function matrix* or, simply, *transfer matrix* of the system.

EXAMPLE 2.4 Consider the unit-time delay system studied in Example 2.2. Its impulse response is $\delta(t-1)$. Therefore its transfer function is

$$\hat{g}(s) = \mathcal{L}[\delta(t-1)] = \int_0^\infty \delta(t-1)e^{-st}dt = e^{-st}\big|_{t=1} = e^{-s}$$

This transfer function is an irrational function of s.

EXAMPLE 2.5 Consider the feedback system shown in Fig. 2.5(a). The transfer function of the unit-time delay element is e^{-s}. The transfer function from r to y can be computed directly from the block diagram as

$$\hat{g}_f(s) = \frac{ae^{-s}}{1-ae^{-s}} \tag{2.12}$$

This can also be obtained by taking the Laplace transform of the impulse response, which was computed in (2.9) as

$$g_f(t) = \sum_{i=1}^{\infty} a^i \delta(t-i)$$

Because $\mathcal{L}[\delta(t-i)] = e^{-is}$, the Laplace transform of $g_f(t)$ is

$$\hat{g}_f(s) = \mathcal{L}[g_f(t)] = \sum_{i=1}^{\infty} a^i e^{-is} = ae^{-s}\sum_{i=0}^{\infty}(ae^{-s})^i$$

Using

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1-r}$$

for $|r| < 1$, we can express the infinite series in closed form as

$$\hat{g}_f(s) = \frac{ae^{-s}}{1-ae^{-s}}$$

which is the same as (2.12).

The transfer function in (2.12) is an irrational function of s. This is so because the feedback system is a distributed system. If a linear time-invariant system is lumped, then its transfer function will be a rational function of s. We study mostly lumped systems; thus the transfer functions we will encounter are mostly rational functions of s.

Every rational transfer function can be expressed as $\hat{g}(s) = N(s)/D(s)$, where $N(s)$ and $D(s)$ are polynomials of s. Let us use deg to denote the degree of a polynomial. Then $\hat{g}(s)$ can be classified as follows:

- $\hat{g}(s)$ proper $\Leftrightarrow$ deg $D(s) \geq$ deg $N(s) \Leftrightarrow \hat{g}(\infty) =$ zero or nonzero constant.

- $\hat{g}(s)$ strictly proper $\Leftrightarrow$ deg $D(s) >$ deg $N(s) \Leftrightarrow \hat{g}(\infty) = 0$.
- $\hat{g}(s)$ biproper $\Leftrightarrow$ deg $D(s) =$ deg $N(s) \Leftrightarrow \hat{g}(\infty) =$ nonzero constant.
- $\hat{g}(s)$ improper $\Leftrightarrow$ deg $D(s) <$ deg $N(s) \Leftrightarrow |\hat{g}(\infty)| = \infty$.

Improper rational transfer functions will amplify high-frequency noise, which often exists in the real world; therefore improper rational transfer functions rarely arise in practice.

A real or complex number λ is called a *pole* of the proper transfer function $\hat{g}(s) = N(s)/D(s)$ if $|\hat{g}(\lambda)| = \infty$; a *zero* if $\hat{g}(\lambda) = 0$. If $N(s)$ and $D(s)$ are *coprime*, that is, have no common factors of degree 1 or higher, then all roots of $N(s)$ are the zeros of $\hat{g}(s)$, and all roots of $D(s)$ are the poles of $\hat{g}(s)$. In terms of poles and zeros, the transfer function can be expressed as

$$\hat{g}(s) = k\frac{(s-z_1)(s-z_2)\cdots(s-z_m)}{(s-p_1)(s-p_2)\cdots(s-p_n)}$$

This is called the *zero-pole-gain* form. In MATLAB, this form can be obtained from the transfer function by calling `[z,p,k]= tf2zp(num,den)`.

A rational matrix $\hat{\mathbf{G}}(s)$ is said to be proper if its every entry is proper or if $\hat{\mathbf{G}}(\infty)$ is a zero or nonzero constant matrix; it is strictly proper if its every entry is strictly proper or if $\hat{\mathbf{G}}(\infty)$ is a zero matrix. If a rational matrix $\hat{\mathbf{G}}(s)$ is square and if both $\hat{\mathbf{G}}(s)$ and $\hat{\mathbf{G}}^{-1}(s)$ are proper, then $\hat{\mathbf{G}}(s)$ is said to be biproper. We call λ a pole of $\hat{\mathbf{G}}(s)$ if it is a pole of some entry of $\hat{\mathbf{G}}(s)$. Thus every pole of every entry of $\hat{\mathbf{G}}(s)$ is a pole of $\hat{\mathbf{G}}(s)$. There are a number of ways of defining zeros for $\hat{\mathbf{G}}(s)$. We call λ a *blocking zero* if it is a zero of every nonzero entry of $\hat{\mathbf{G}}(s)$. A more useful definition is the *transmission zero*, which will be introduced in Chapter 9.

State-space equation Every linear time-invariant lumped system can be described by a set of equations of the form

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)\\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t)\end{aligned} \tag{2.13}$$

For a system with p inputs, q outputs, and n state variables, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are, respectively, $n \times n$, $n \times p$, $q \times n$, and $q \times p$ constant matrices. Applying the Laplace transform to (2.13) yields

$$\begin{aligned}s\hat{\mathbf{x}}(s) - \mathbf{x}(0) &= \mathbf{A}\hat{\mathbf{x}}(s) + \mathbf{B}\hat{\mathbf{u}}(s)\\ \hat{\mathbf{y}}(s) &= \mathbf{C}\hat{\mathbf{x}}(s) + \mathbf{D}\hat{\mathbf{u}}(s)\end{aligned}$$

which implies

$$\hat{\mathbf{x}}(s) = (s\mathbf{I}-\mathbf{A})^{-1}\mathbf{x}(0) + (s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{u}}(s) \tag{2.14}$$

$$\hat{\mathbf{y}}(s) = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{x}(0) + \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{u}}(s) + \mathbf{D}\hat{\mathbf{u}}(s) \tag{2.15}$$

They are algebraic equations. Given $\mathbf{x}(0)$ and $\hat{\mathbf{u}}(s)$, $\hat{\mathbf{x}}(s)$ and $\hat{\mathbf{y}}(s)$ can be computed algebraically from (2.14) and (2.15). Their inverse Laplace transforms yield the time responses $\mathbf{x}(t)$ and $\mathbf{y}(t)$. The equations also reveal the fact that the response of a linear system can be decomposed

as the zero-state response and the zero-input response. If the initial state $\mathbf{x}(0)$ is zero, then (2.15) reduces to

$$\hat{\mathbf{y}}(s) = [\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}]\hat{\mathbf{u}}(s)$$

Comparing this with (2.11) yields

$$\hat{\mathbf{G}}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \tag{2.16}$$

This relates the input–output (or transfer matrix) and state-space descriptions.

The functions `tf2ss` and `ss2tf` in MATLAB compute one description from the other. They compute only the SISO and SIMO cases. For example, `[num,den] = ss2tf(a,b,c,d,1)` computes the transfer matrix from the first input to all outputs or, equivalently, the first column of $\hat{\mathbf{G}}(s)$. If the last argument 1 in `ss2tf(a,b,c,d,1)` is replaced by 3, then the function generates the third column of $\hat{\mathbf{G}}(s)$.

To conclude this section, we mention that the Laplace transform is not used in studying linear time-varying systems. The Laplace transform of $g(t, \tau)$ is a function of two variables and $\mathcal{L}[\mathbf{A}(t)\mathbf{x}(t)] \neq \mathcal{L}[\mathbf{A}(t)]\mathcal{L}[\mathbf{x}(t)]$; thus the Laplace transform does not offer any advantage and is not used in studying time-varying systems.

2.3.1 Op-Amp Circuit Implementation

Every linear time-invariant (LTI) state-space equation can be implemented using an operational amplifier (op-amp) circuit. Figure 2.6 shows two standard op-amp circuit elements. All inputs are connected, through resistors, to the inverting terminal. Not shown are the grounded noninverting terminal and power supply. If the feedback branch is a resistor as shown in Fig. 2.6(a), then the output of the element is $-(ax_1 + bx_2 + cx_3)$. If the feedback branch is a capactor with capacitance C and $RC = 1$ as shown in Fig. 2.6(b), and if the output is assigned as x, then $\dot{x} = -(av_1 + bv_2 + cv_3)$. We call the first element an *adder*; the second element, an *integrator*. Actually, the adder functions also as multipliers and the integrator functions also as multipliers and adder. If we use only one input, say, x_1, in Fig. 2.6(a), then the output equals $-ax_1$, and the element can be used as an *inverter* with gain a. Now we use an example to show that every LTI state-space equation can be implemented using the two types of elements in Fig. 2.6.

Consider the state equation

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 2 & -0.3 \\ 1 & -8 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} -2 \\ 0 \end{bmatrix} u(t) \tag{2.17}$$

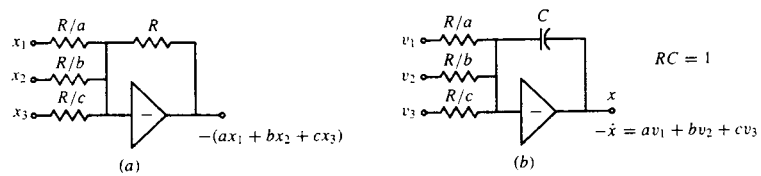


Figure 2.6 Two op-amp circuit elements.

$$y(t) = [-2 \;\; 3]\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + 5u(t) \tag{2.18}$$

It has dimension 2 and we need two integrators to implement it. We have the freedom in choosing the output of each integrator as $+x_i$ or $-x_i$. Suppose we assign the output of the left-hand-side (LHS) integrator as x_1 and the output of the right-hand-side (RHS) integrator as $-x_2$ as shown in Fig. 2.7. Then the input of the LHS integrator should be, from the first equation of (2.17), $-\dot{x}_1 = -2x_1 + 0.3x_2 + 2u$ and is connected as shown. The input of the RHS integrator should be $\dot{x}_2 = x_1 - 8x_2$ and is connected as shown. If the output of the adder is chosen as y, then its input should equal $-y = 2x_1 - 3x_2 - 5u$, and is connected as shown. Thus the state equation in (2.17) and (2.18) can be implemented as shown in Fig. 2.7. Note that there are many ways to implement the same equation. For example, if we assign the outputs of the two integrators in Fig. 2.7 as x_1 and x_2, instead of x_1 and $-x_2$, then we will obtain a different implementation.

In actual operational amplifier circuits, the range of signals is limited, usually 1 or 2 volts below the supplied voltage. If any signal grows outside the range, the circuit will saturate or burn out and the circuit will not behave as the equation dictates. There is, however, a way to deal with this problem, as we will discuss in Section 4.3.1.

2.4 Linearization

Most physical systems are nonlinear and time varying. Some of them can be described by the nonlinear differential equation of the form

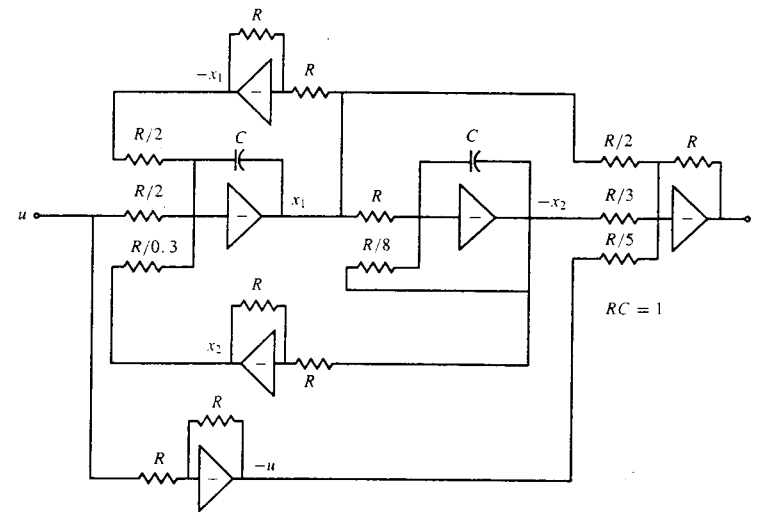


Figure 2.7 Op-amp implementation of (2.17) and (2.18).

$$\dot{\mathbf{x}}(t) = \mathbf{h}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{2.19}$$
$$\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t)$$

where $\mathbf{h}$ and $\mathbf{f}$ are nonlinear functions. The behavior of such equations can be very complicated and its study is beyond the scope of this text.

Some nonlinear equations, however, can be approximated by linear equations under certain conditions. Suppose for some input function $\mathbf{u}_o(t)$ and some initial state, $\mathbf{x}_o(t)$ is the solution of (2.19); that is,

$$\dot{\mathbf{x}}_o(t) = \mathbf{h}(\mathbf{x}_o(t), \mathbf{u}_o(t), t) \tag{2.20}$$

Now suppose the input is perturbed slightly to become $\mathbf{u}_o(t) + \bar{\mathbf{u}}(t)$ and the initial state is also perturbed only slightly. For some nonlinear equations, the corresponding solution may differ from $\mathbf{x}_o(t)$ only slightly. In this case, the solution can be expressed as $\mathbf{x}_o(t) + \bar{\mathbf{x}}(t)$ with $\bar{\mathbf{x}}(t)$ small for all t.[3] Under this assumption, we can expand (2.19) as

$$\begin{aligned}\dot{\mathbf{x}}_o(t) + \dot{\bar{\mathbf{x}}}(t) &= \mathbf{h}(\mathbf{x}_o(t) + \bar{\mathbf{x}}(t), \mathbf{u}_o(t) + \bar{\mathbf{u}}(t), t) \\ &= \mathbf{h}(\mathbf{x}_o(t), \mathbf{u}_o(t), t) + \frac{\partial \mathbf{h}}{\partial \mathbf{x}}\bar{\mathbf{x}} + \frac{\partial \mathbf{h}}{\partial \mathbf{u}}\bar{\mathbf{u}} + \cdots\end{aligned} \tag{2.21}$$

where, for $\mathbf{h} = [h_1\ h_2\ h_3]'$, $\mathbf{x} = [x_1\ x_2\ x_3]'$, and $\mathbf{u} = [u_1\ u_2]'$,

$$\mathbf{A}(t) := \frac{\partial \mathbf{h}}{\partial \mathbf{x}} := \begin{bmatrix} \partial h_1/\partial x_1 & \partial h_1/\partial x_2 & \partial h_1/\partial x_3 \\ \partial h_2/\partial x_1 & \partial h_2/\partial x_2 & \partial h_2/\partial x_3 \\ \partial h_3/\partial x_1 & \partial h_3/\partial x_2 & \partial h_3/\partial x_3 \end{bmatrix}$$

$$\mathbf{B}(t) := \frac{\partial \mathbf{h}}{\partial \mathbf{u}} := \begin{bmatrix} \partial h_1/\partial u_1 & \partial h_1/\partial u_2 \\ \partial h_2/\partial u_1 & \partial h_2/\partial u_2 \\ \partial h_3/\partial u_1 & \partial h_3/\partial u_2 \end{bmatrix}$$

They are called *Jacobians*. Because $\mathbf{A}$ and $\mathbf{B}$ are computed along the two time functions $\mathbf{x}_o(t)$ and $\mathbf{u}_o(t)$, they are, in general, functions of t. Using (2.20) and neglecting higher powers of $\bar{\mathbf{x}}$ and $\bar{\mathbf{u}}$, we can reduce (2.21) to

$$\dot{\bar{\mathbf{x}}}(t) = \mathbf{A}(t)\bar{\mathbf{x}}(t) + \mathbf{B}(t)\bar{\mathbf{u}}(t)$$

This is a linear state-space equation. The equation $\mathbf{y}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t)$ can be similarly linearized. This linearization technique is often used in practice to obtain linear equations.

2.5 Examples

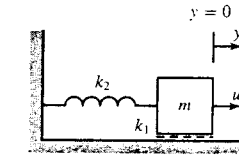
In this section we use examples to illustrate how to develop transfer functions and state-space equations for physical systems.

EXAMPLE 2.6 Consider the mechanical system shown in Fig. 2.8. It consists of a block with mass m connected to a wall through a spring. We consider the applied force u to be the input

3. This is not true in general. For some nonlinear equations, a very small difference in initial states will generate completely different solutions, yielding the phenomenon of *chaos*.

and displacement y from the equilibrium to be the output. The friction between the floor and the block generally consists of three distinct parts: static friction, Coulomb friction, and viscous friction as shown in Fig. 2.9. Note that the horizontal coordinate is velocity $\dot{y} = dy/dt$. The friction is clearly not a linear function of the velocity. To simplify analysis, we disregard the static and Coulomb frictions and consider only the viscous friction. Then the friction becomes linear and can be expressed as $k_1\dot{y}(t)$, where k_1 is the viscous friction coefficient. The characteristics of the spring are shown in Fig. 2.10; it is not linear. However, if the displacement is limited to $(y_1,\ y_2)$ as shown, then the spring can be considered to be linear and the spring force equals $k_2 y$, where k_2 is the spring constant. Thus the mechanical system can be modeled as a linear system under linearization and simplification.

Figure 2.8 Mechanical system.



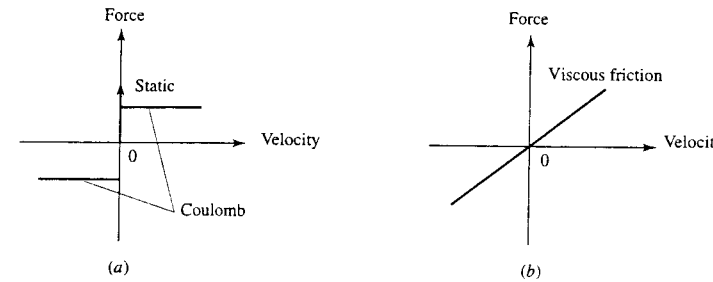
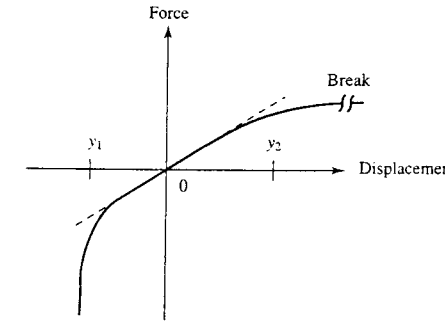


Figure 2.9 Mechanical system.(a) Static and Coulomb frictions. (b) Viscous friction.

Figure 2.10 Characteristic of spring.



We apply Newton's law to develop an equation to describe the system. The applied force u must overcome the friction and the spring force. The remainder is used to accelerate the block. Thus we have

$$m\ddot{y} = u - k_1\dot{y} - k_2 y \tag{2.22}$$

where $\ddot{y} = d^2y(t)/dt^2$ and $\dot{y} = dy(t)/dt$. Applying the Laplace transform and assuming zero initial conditions, we obtain

$$ms^2\hat{y}(s) = \hat{u}(s) - k_1 s\hat{y}(s) - k_2\hat{y}(s)$$

which implies

$$\hat{y}(s) = \frac{1}{ms^2 + k_1 s + k_2}\hat{u}(s)$$

This is the input–output description of the system. Its transfer function is $1/(ms^2 + k_1 s + k_2)$.

If $m = 1$, $k_1 = 3$, $k_2 = 2$, then the impulse response of the system is

$$g(t) = \mathcal{L}^{-1}\left[\frac{1}{s^2 + 3s + 2}\right] = \mathcal{L}^{-1}\left[\frac{1}{s+1} - \frac{1}{s+2}\right] = e^{-t} - e^{-2t}$$

and the convolution description of the system is

$$y(t) = \int_0^t g(t-\tau)u(\tau)\,d\tau = \int_0^t (e^{-(t-\tau)} - e^{-2(t-\tau)})u(\tau)\,d\tau$$

Next we develop a state-space equation to describe the system. Let us select the displacement and velocity of the block as state variables; that is, $x_1 = y$, $x_2 = \dot{y}$. Then we have, using (2.22),

$$\dot{x}_1 = x_2 \qquad m\dot{x}_2 = u - k_1 x_2 - k_2 x_1$$

They can be expressed in matrix form as

$$\begin{bmatrix}\dot{x}_1(t)\\ \dot{x}_2(t)\end{bmatrix} = \begin{bmatrix}0 & 1\\ -k_2/m & -k_1/m\end{bmatrix}\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix} + \begin{bmatrix}0\\ 1/m\end{bmatrix}u(t)$$

$$y(t) = [1\ \ 0]\begin{bmatrix}x_1(t)\\ x_2(t)\end{bmatrix}$$

This state-space equation describes the system.

EXAMPLE 2.7 Consider the system shown in Fig. 2.11. It consists of two blocks, with masses m_1 and m_2, connected by three springs with spring constants k_i, $i = 1, 2, 3$. To simplify the discussion, we assume that there is no friction between the blocks and the floor. The applied force u_1 must overcome the spring forces and the remainder is used to accelerate the block, thus we have

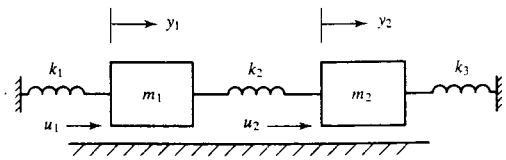


Figure 2.11 Spring-mass system.

$$u_1 - k_1 y_1 - k_2(y_1 - y_2) = m_1\ddot{y}_1$$

or

$$m_1\ddot{y}_1 + (k_1 + k_2)y_1 - k_2 y_2 = u_1 \tag{2.23}$$

For the second block, we have

$$m_2\ddot{y}_2 - k_2 y_1 + (k_1 + k_2)y_2 = u_2 \tag{2.24}$$

They can be combined as

$$\begin{bmatrix}m_1 & 0\\ 0 & m_2\end{bmatrix}\begin{bmatrix}\ddot{y}_1\\ \ddot{y}_2\end{bmatrix} + \begin{bmatrix}k_1 + k_2 & -k_2\\ -k_2 & k_1 + k_2\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix} = \begin{bmatrix}u_1\\ u_2\end{bmatrix}$$

This is a standard equation in studying vibration and is said to be in normal form. See Reference [18]. Let us define

$$x_1 := y_1 \qquad x_2 := \dot{y}_1 \qquad x_3 := y_2 \qquad x_4 := \dot{y}_2$$

Then we can readily obtain

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\\ \dot{x}_4\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 0\\ \dfrac{-(k_1+k_2)}{m_1} & 0 & \dfrac{k_2}{m_1} & 0\\ 0 & 0 & 0 & 1\\ \dfrac{k_2}{m_2} & 0 & \dfrac{-(k_1+k_2)}{m_1} & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\end{bmatrix} + \begin{bmatrix}0 & 0\\ \dfrac{1}{m_1} & 0\\ 0 & 0\\ 0 & \dfrac{1}{m_2}\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix}$$

$$\mathbf{y} := \begin{bmatrix}y_1\\ y_2\end{bmatrix} = \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix}\mathbf{x}$$

This two-input two-output state equation describes the system in Fig. 2.11.

To develop its input–output description, we apply the Laplace transform to (2.23) and (2.24) and assume zero initial conditions to yield

$$m_1 s^2\hat{y}_1(s) + (k_1 + k_2)\hat{y}_1(s) - k_2\hat{y}_2(s) = \hat{u}_1(s)$$

$$m_2 s^2\hat{y}_2(s) - k_2\hat{y}_1(s) + (k_1 + k_2)\hat{y}_2(s) = \hat{u}_2(s)$$

From these two equations, we can readily obtain

$$\begin{bmatrix}\hat{y}_1(s)\\ \hat{y}_2(s)\end{bmatrix} = \begin{bmatrix}\dfrac{m_2 s^2 + k_1 + k_2}{d(s)} & \dfrac{k_2}{d(s)}\\ \dfrac{k_2}{d(s)} & \dfrac{m_1 s^2 + k_1 + k_2}{d(s)}\end{bmatrix}\begin{bmatrix}\hat{u}_1(s)\\ \hat{u}_2(s)\end{bmatrix}$$

where

$$d(s) := (m_1 s^2 + k_1 + k_2)(m_2 s^2 + k_1 + k_2) - k_2^2$$

This is the transfer-matrix description of the system. Thus what we will discuss in this text can be applied directly to study vibration.

EXAMPLE 2.8 Consider a cart with an inverted pendulum hinged on top of it as shown in Fig. 2.12. For simplicity, the cart and the pendulum are assumed to move in only one plane, and the friction, the mass of the stick, and the gust of wind are disregarded. The problem is to maintain the pendulum at the vertical position. For example, if the inverted pendulum is falling in the direction shown, the cart moves to the right and exerts a force, through the hinge, to push the pendulum back to the vertical position. This simple mechanism can be used as a model of a space vehicle on takeoff.

Let H and V be, respectively, the horizontal and vertical forces exerted by the cart on the pendulum as shown. The application of Newton's law to the linear movements yields

$$M\frac{d^2y}{dt^2} = u - H$$

$$H = m\frac{d^2}{dt^2}(y + l\sin\theta) = m\ddot{y} + ml\ddot{\theta}\cos\theta - ml(\dot{\theta})^2\sin\theta$$

$$mg - V = m\frac{d^2}{dt^2}(l\cos\theta) = ml[-\ddot{\theta}\sin\theta - (\dot{\theta})^2\cos\theta]$$

The application of Newton's law to the rotational movement of the pendulum around the hinge yields

$$mgl\sin\theta = ml\ddot{\theta}\cdot l + m\ddot{y}l\cos\theta$$

They are nonlinear equations. Because the design objective is to maintain the pendulum at the vertical position, it is reasonable to assume θ and $\dot{\theta}$ to be small. Under this assumption, we can use the approximation $\sin\theta = \theta$ and $\cos\theta = 1$. By retaining only the linear terms in θ and $\dot{\theta}$ or, equivalently, dropping the terms with θ^2, $(\dot{\theta})^2$, $\theta\dot{\theta}$, and $\theta\ddot{\theta}$, we obtain $V = mg$ and

$$M\ddot{y} = u - m\ddot{y} - ml\ddot{\theta}$$

$$g\theta = l\ddot{\theta} + \ddot{y}$$

which imply

$$M\ddot{y} = u - mg\theta \tag{2.25}$$

$$Ml\ddot{\theta} = (M+m)g\theta - u \tag{2.26}$$

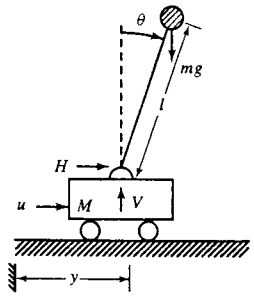


Figure 2.12 Cart with inverted pendulum.

Using these linearized equations, we now can develop the input–output and state-space descriptions. Applying the Laplace transform to (2.25) and (2.26) and assuming zero initial conditions, we obtain

$$Ms^2\hat{y}(s) = \hat{u}(s) - mg\hat{\theta}(s)$$

$$Mls^2\hat{\theta}(s) = (M+m)g\hat{\theta}(s) - \hat{u}(s)$$

From these equations, we can readily compute the transfer function $\hat{g}_{yu}(s)$ from u to y and the transfer function $\hat{g}_{\theta u}(s)$ from u to θ as

$$\hat{g}_{yu}(s) = \frac{s^2 - g}{s^2[Ms^2 - (M+m)g]}$$

$$\hat{g}_{\theta u}(s) = \frac{-1}{Ms^2 - (M+m)g}$$

To develop a state-space equation, we select state variables as $x_1 = y$, $x_2 = \dot{y}$, $x_3 = \theta$, and $x_4 = \dot{\theta}$. Then from this selection, (2.25), and (2.26) we can readily obtain

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\\ \dot{x}_4\end{bmatrix} = \begin{bmatrix}0 & 1 & 0 & 0\\ 0 & 0 & -mg/M & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & (M+m)g/Ml & 0\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\end{bmatrix} + \begin{bmatrix}0\\ 1/M\\ 0\\ -1/Ml\end{bmatrix}u$$

$$y = [1\ 0\ 0\ 0]\mathbf{x} \tag{2.27}$$

This state equation has dimension 4 and describes the system when θ and $\dot{\theta}$ are very small.

EXAMPLE 2.9 A communication satellite of mass m orbiting around the earth is shown in Fig. 2.13. The altitude of the satellite is specified by $r(t)$, $\theta(t)$, and $\phi(t)$ as shown. The orbit can be controlled by three orthogonal thrusts $u_r(t)$, $u_\theta(t)$, and $u_\phi(t)$. The state, input, and output of the system are chosen as

$$\mathbf{x}(t) = \begin{bmatrix}r(t)\\ \dot{r}(t)\\ \theta(t)\\ \dot{\theta}(t)\\ \phi(t)\\ \dot{\phi}(t)\end{bmatrix} \qquad \mathbf{u}(t) = \begin{bmatrix}u_r(t)\\ u_\theta(t)\\ u_\phi(t)\end{bmatrix} \qquad \mathbf{y}(t) = \begin{bmatrix}r(t)\\ \theta(t)\\ \phi(t)\end{bmatrix}$$

Then the system can be shown to be described by

$$\dot{\mathbf{x}} = \mathbf{h}(\mathbf{x},\mathbf{u}) = \begin{bmatrix}\dot{r}\\ r\dot{\theta}^2\cos^2\phi + r\dot{\phi}^2 - k/r^2 + u_r/m\\ \dot{\theta}\\ -2\dot{r}\dot{\theta}/r + 2\dot{\theta}\dot{\phi}\sin\phi/\cos\phi + u_\theta/mr\cos\phi\\ \dot{\phi}\\ -\dot{\theta}^2\cos\phi\sin\phi - 2\dot{r}\dot{\phi}/r + u_\phi/mr\end{bmatrix} \tag{2.28}$$

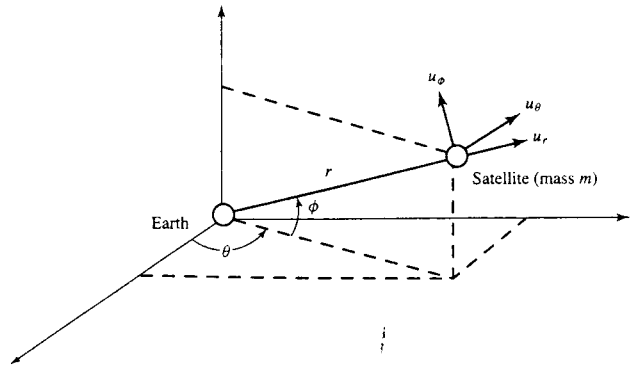


Figure 2.13 Satellite in orbit.

One solution, which corresponds to a circular equatorial orbit, is given by

$$\mathbf{x}_o(t) = [r_o \;\; 0 \;\; \omega_o t \;\; \omega_o \;\; 0 \;\; 0]' \qquad \mathbf{u}_o \equiv \mathbf{0}$$

with $r_o^3\omega_o^2 = k$, a known physical constant. Once the satellite reaches the orbit, it will remain in the orbit as long as there are no disturbances. If the satellite deviates from the orbit, thrusts must be applied to push it back to the orbit. Define

$$\mathbf{x}(t) = \mathbf{x}_o(t) + \tilde{\mathbf{x}}(t) \qquad \mathbf{u}(t) = \mathbf{u}_o(t) + \tilde{\mathbf{u}}(t) \qquad \mathbf{y}(t) = \mathbf{y}_o + \tilde{\mathbf{y}}(t)$$

If the perturbation is very small, then (2.28) can be linearized as

$$\dot{\tilde{\mathbf{x}}}(t) = \left[\begin{array}{cccc:cc} 0 & 1 & 0 & 0 & 0 & 0 \\ 3\omega_o^2 & 0 & 0 & 2\omega_o r_o & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & \dfrac{-2\omega_o}{r_o} & 0 & 0 & 0 & 0 \\ \hdots & \hdots & \hdots & \hdots & \hdots & \hdots \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -\omega_o^2 & 0 \end{array}\right] \tilde{\mathbf{x}}(t) + \left[\begin{array}{cc:c} 0 & 0 & 0 \\ \dfrac{1}{m} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \dfrac{1}{mr_o} & 0 \\ \hdots & \hdots & \hdots \\ 0 & 0 & 0 \\ 0 & 0 & \dfrac{1}{mr_o} \end{array}\right] \tilde{\mathbf{u}}(t)$$

$$\tilde{\mathbf{y}}(t) = \left[\begin{array}{cccc:cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hdots & \hdots & \hdots & \hdots & \hdots & \hdots \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right] \tilde{\mathbf{x}}(t) \tag{2.29}$$

This six-dimensional state equation describes the system. In this equation, **A**, **B**, and **C** happen to be constant. If the orbit is an elliptic one, then they will be time varying. We note that the three matrices are all block diagonal. Thus the equation can be decomposed into two uncoupled parts, one involving r and θ, the other ϕ. Studying these two parts independently can simplify analysis and design.

EXAMPLE 2.10 In chemical plants, it is often necessary to maintain the levels of liquids. A simplified model of a connection of two tanks is shown in Fig. 2.14. It is assumed that under normal operation, the inflows and outflows of both tanks all equal Q and their liquid levels equal H_1 and H_2. Let u be inflow perturbation of the first tank, which will cause variations in liquid level x_1 and outflow y_1 as shown. These variations will cause level variation x_2 and outflow variation y in the second tank. It is assumed that

$$y_1 = \frac{x_1 - x_2}{R_1} \quad \text{and} \quad y = \frac{x_2}{R_2}$$

where R_i are the flow resistances and depend on the normal height H_1 and H_2. They can also be controlled by the valves. Changes of liquid levels are governed by

$$A_1\, dx_1 = (u - y_1)\, dt \quad \text{and} \quad A_2\, dx_2 = (y_1 - y)\, dt$$

where A_i are the cross sections of the tanks. From these equations, we can readily obtain

$$\dot{x}_1 = \frac{u}{A_1} - \frac{x_1 - x_2}{A_1 R_1}$$

$$\dot{x}_2 = \frac{x_1 - x_2}{A_2 R_1} - \frac{x_2}{A_2 R_2}$$

Thus the state-space description of the system is given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1/A_2R_2 & 1/A_1R_1 \\ 1/A_2R_1 & -(1/A_2R_1 + 1/A_2R_2) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/A_1 \\ 0 \end{bmatrix} u$$

$$y = [0 \;\; 1/R_2]\,\mathbf{x}$$

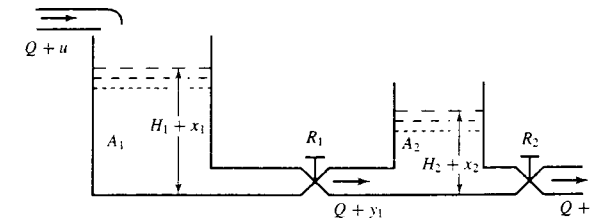


Figure 2.14 Hydraulic tanks.

Its transfer function can be computed as

$$\hat{g}(s) = \frac{1}{A_1A_2R_1R_2s^2 + (A_1R_1 + A_1R_2 + A_2R_2)s + 1}$$

2.5.1 *RLC* networks

In *RLC* networks, capacitors and inductors can store energy and are associated with state variables. If a capacitor voltage is assigned as a state variable x, then its current is $C\dot{x}$, where C is its capacitance. If an inductor current is assigned as a state variable x, then its voltage is $L\dot{x}$, where L is its inductance. Note that resistors are memoryless elements, and their currents or voltages should not be assigned as state variables. For most simple *RLC* networks, once state variables are assigned, their state equations can be developed by applying Kirchhoff's current and voltage laws, as the next example illustrates.

EXAMPLE 2.11 Consider the network shown in Fig. 2.15. We assign the C_i-capacitor voltages as x_i, $i = 1, 2$ and the inductor current as x_3. It is important to specify their polarities. Then their currents and voltage are, respectively, $C_1\dot{x}_1$, $C_2\dot{x}_2$, and $L\dot{x}_3$ with the polarities shown. From the figure, we see that the voltage across the resistor is $u - x_1$ with the polarity shown. Thus its current is $(u - x_1)/R$. Applying Kirchhoff's current law at node A yields $C_2\dot{x}_2 = x_3$; at node B it yields

$$\frac{u - x_1}{R} = C_1\dot{x}_1 + C_2\dot{x}_2 = C_1\dot{x}_1 + x_3$$

Thus we have

$$\dot{x}_1 = -\frac{x_1}{RC_1} - \frac{x_3}{C_1} + \frac{u}{RC_1}$$

$$\dot{x}_2 = \frac{1}{C_2}x_3$$

Appling Kirchhoff's voltage law to the right-hand-side loop yields $L\dot{x}_3 = x_1 - x_2$ or

$$\dot{x}_3 = \frac{x_1 - x_2}{L}$$

The output y is given by

$$y = L\dot{x}_3 = x_1 - x_2$$

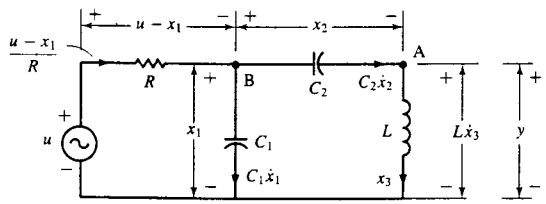


Figure 2.15 Network.

They can be combined in matrix form as

$$\dot{\mathbf{x}} = \begin{bmatrix} -1/RC_1 & 0 & -1/C_1 \\ 0 & 0 & 1/C_2 \\ 1/L & -1/L & 0 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1/RC_1 \\ 0 \\ 0 \end{bmatrix}u$$

$$y = [1 \quad -1 \quad 0]\mathbf{x} + 0 \cdot u$$

This three-dimensional state equation describes the network shown in Fig. 2.15.

The procedure used in the preceding example can be employed to develop state equations to describe simple *RLC* networks. The procedure is fairly simple: assign state variables and then use branch characteristics and Kirchhoff's laws to develop state equations. The procedure can be stated more systematically by using graph concepts, as we will introduce next. The procedure and subsequent Example 2.12, however, can be skipped without loss of continuity.

First we introduce briefly the concepts of tree, link, and cutset of a network. We consider only connected networks. Every capacitor, inductor, resistor, voltage source, and current source will be considered as a branch. Branches are connected at nodes. Thus a network can be considered to consist of only branches and nodes. A loop is a connection of branches starting from one point and coming back to the same point without passing any point twice. *The algebraic sum of all voltages along every loop is zero* (Kirchhoff's voltage law). The set of all branches connect to a node is called a *cutset*. More generally, a cutset of a connected network is any minimal set of branches so that the removal of the set causes the remaining network to be unconnected. For example, removing all branches connected to a node leaves the node unconnected to the remaining network. *The algebraic sum of all branch currents in every cutset is zero* (Kirchhoff's current law).

A *tree* of a network is defined as any connection of branches connecting all the nodes but containing no loops. A branch is called a *tree branch* if it is in the tree, a *link* if it is not. With respect to a chosen tree, every link has a unique loop, called the *fundamental loop*, in which the remaining loop branches are all tree branches. Every tree branch has a unique cutset, called the *fundamental cutset*, in which the remaining cutset branches are all links. In other words, a fundamental loop contains only one link and a fundamental cutset contains only one tree branch.

Procedure for developing state-space equations[4]

1. Consider an *RLC* network. We first choose a *normal tree*. The branches of the normal tree are chosen in the order of voltage sources, capacitors, resistors, inductors, and current sources.
2. Assign the capacitor voltages in the normal tree and the inductor currents in the links as state variables. Capacitor voltages in the links and inductor currents in the normal tree are not assigned.
3. Express the voltage and current of every branch in terms of the state variables and, if necessary, the inputs by applying Kirchhoff's voltage law to fundamental loops and Kirchhoff's current law to fundamental cutsets.

4. The reader may skip this procedure and go directly to Example 2.13.

4. Apply Kirchhoff's voltage or current law to the fundamental loop or cutset of every branch that is assigned as a state variable.

EXAMPLE 2.12 Consider the network shown in Fig. 2.16. The normal tree is chosen as shown with heavy lines; it consists of the voltage source, two capacitors, and the 1-Ω resistor. The capacitor voltages in the normal tree and the inductor current in the link will be assigned as state variables. If the voltage across the 3-F capacitor is assigned as x_1, then its current is $3\dot{x}_1$. The voltage across the 1-F capacitor is assigned as x_2 and its current is $\dot{x}_2$. The current through the 2-H inductor is assigned as x_3 and its voltage is $2\dot{x}_3$. Because the 2-Ω resistor is a link, we use its fundamental loop to find its voltage as $u_1 - x_1$. Thus its current is $(u_1 - x_1)/2$. The 1-Ω resistor is a tree branch. We use its fundamental cutset to find its current as x_3. Thus its voltage is $1 \cdot x_3 = x_3$. This completes Step 3.

The 3-F capacitor is a tree branch and its fundamental cutset is as shown. The algebraic sum of the cutset currents is 0 or

$$\frac{u_1 - x_1}{2} - 3\dot{x}_1 + u_2 - x_3 = 0$$

which implies

$$\dot{x}_1 = -\tfrac{1}{6}x_1 - \tfrac{1}{3}x_3 + \tfrac{1}{6}u_1 + \tfrac{1}{3}u_2$$

The 1-F capacitor is a tree branch, and from its fundamental cutset we have $\dot{x}_2 - x_3 = 0$ or

$$\dot{x}_2 = x_3$$

The 2-H inductor is a link. The voltage along its fundamental loop is $2\dot{x}_3 + x_3 - x_1 + x_2 = 0$ or

$$\dot{x}_3 = \tfrac{1}{2}x_1 - \tfrac{1}{2}x_2 - \tfrac{1}{2}x_3$$

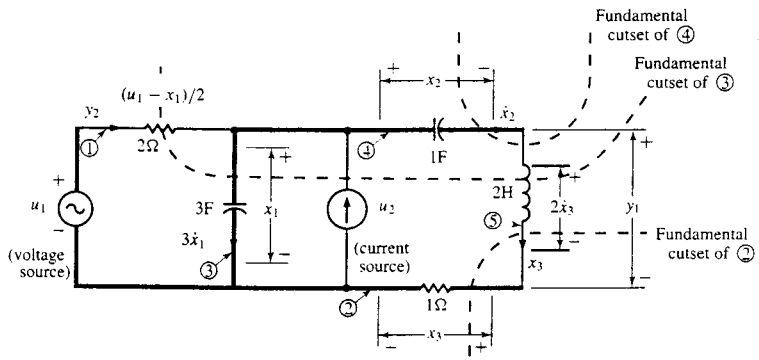


Figure 2.16 Network with two inputs.

They can be expressed in matrix form as

$$\dot{\mathbf{x}} = \begin{bmatrix} -\frac{1}{6} & 0 & -\frac{1}{3} \\ 0 & 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \frac{1}{6} & \frac{1}{3} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}\mathbf{u} \tag{2.30}$$

If we consider the voltage across the 2-H inductor and the current through the 2-Ω resistor as the outputs, then we have

$$y_1 = 2\dot{x}_3 = x_1 - x_2 - x_3 = [1 \quad -1 \quad -1]\mathbf{x}$$

and

$$y_2 = 0.5(u_1 - x_1) = [-0.5 \;\; 0 \;\; 0]\mathbf{x} + [0.5 \;\; 0]\mathbf{u}$$

They can be written in matrix form as

$$\mathbf{y} = \begin{bmatrix} 1 & -1 & -1 \\ -0.5 & 0 & 0 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 & 0 \\ 0.5 & 0 \end{bmatrix}\mathbf{u} \tag{2.31}$$

Equations (2.30) and (2.31) are the state-space description of the network.

The transfer matrix of the network can be computed directly from the network or using the formula in (2.16):

$$\hat{\mathbf{G}}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

We will use MATLAB to compute this equation. We type

```
a=[-1/6 0 -1/3;0 0 1;0.5 -0.5 -0.5];b=[1/6 1/3;0 0;0 0];
c=[1 -1 -1;-0.5 0 0];d=[0 0;0.5 0];
[N1,d1]=ss2tf(a,b,c,d,1)
```

which yields

```
N1=
    0.0000    0.1667   -0.0000   -0.0000
    0.5000    0.2500    0.3333   -0.0000

d1=
    1.0000    0.6667    0.7500    0.0833
```

This is the first column of the transfer matrix. We repeat the computation for the second input. Thus the transfer matrix of the network is

$$\hat{\mathbf{G}}(s) = \begin{bmatrix} \dfrac{0.1667s^2}{s^3 + 0.6667s^2 + 0.75s + 0.083} & \dfrac{0.3333s^2}{s^3 + 0.6667s^2 + 0.75s + 0.0833} \\ \dfrac{0.5s^3 + 0.25s^2 + 0.3333s}{s^3 + 0.6667s^2 + 0.75s + 0.083} & \dfrac{-0.1667s^2 - 0.0833s - 0.0833}{s^3 + 0.6667s^2 + 0.75s + 0.0833} \end{bmatrix}$$

EXAMPLE 2.13 Consider the network shown in Fig. 2.17(a), where T is a tunnel diode with the characteristics shown in Fig. 2.17(b). Let x_1 be the voltage across the capacitor and x_2 be the current through the inductor. Then we have $v = x_1$ and

$$x_2(t) = C\dot{x}_1(t) + i(t) = C\dot{x}_1(t) + h(x_1(t))$$
$$L\dot{x}_2(t) = E - Rx_2(t) - x_1(t)$$

They can be arranged as

$$\dot{x}_1(t) = \frac{-h(x_1(t))}{C} + \frac{x_2(t)}{C}$$
$$\dot{x}_2(t) = \frac{-x_1(t) - Rx_2(t)}{L} + \frac{E}{L} \tag{2.32}$$

This set of nonlinear equations describes the network. Now if $x_1(t)$ is known to lie only inside the range (a, b) shown in Fig. 2.17(b), then $h(x_1(t))$ can be approximated by $h(x_1(t)) = x_1(t)/R_1$. In this case, the network can be reduced to the one in Fig. 2.17(c) and can be described by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1/CR_1 & 1/C \\ -1/L & -R/L \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/L \end{bmatrix} E$$

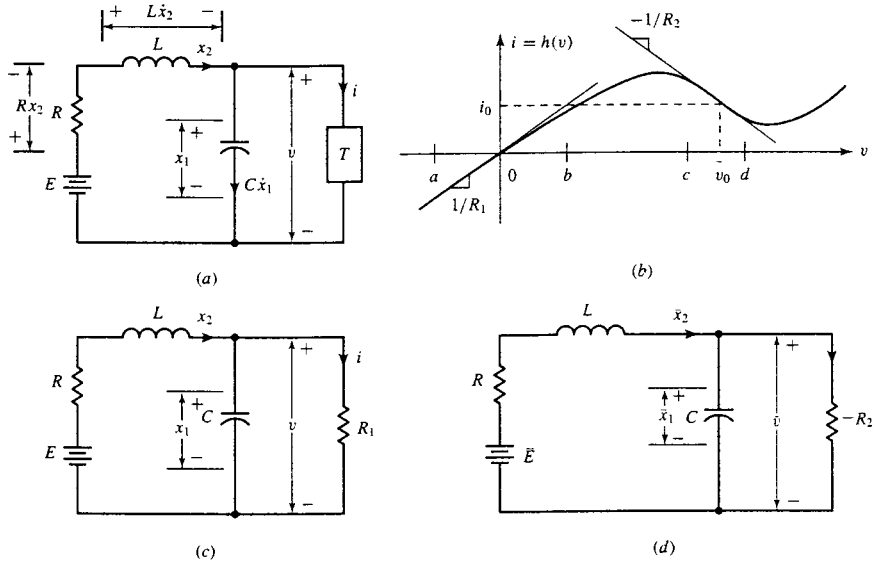


Figure 2.17 Network with a tunnel diode.

This is an LTI state-space equation. Now if $x_1(t)$ is known to lie only inside the range (c, d) shown in Fig. 2.17(b), we may introduce the variables $\bar{x}_1(t) = x_1(t) - v_o$, and $\bar{x}_2(t) = x_2(t) - i_o$ and approximate $h(x_1(t))$ as $i_o - \bar{x}_1(t)/R_2$. Substituting these into (2.32) yields

$$\begin{bmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \end{bmatrix} = \begin{bmatrix} 1/CR_2 & 1/C \\ -1/L & -R/L \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/L \end{bmatrix} \bar{E}$$

where $\bar{E} = E - v_o - Ri_o$. This equation is obtained by shifting the operating point from $(0, 0)$ to (v_o, i_o) and by linearization at (v_o, i_o). Because the two linearized equations are identical if $-R_2$ is replaced by R_1 and $\bar{E}$ by E, we can readily obtain its equivalent network shown in Fig. 2.17(d). Note that it is not obvious how to obtain the equivalent network from the original network without first developing the state equation.

2.6 Discrete-Time Systems

This section develops the discrete counterpart of continuous-time systems. Because most concepts in continuous-time systems can be applied directly to the discrete-time systems, the discussion will be brief.

The input and output of every discrete-time system will be assumed to have the same sampling period T and will be denoted by $\mathbf{u}[k] := \mathbf{u}(kT)$, $\mathbf{y}[k] := \mathbf{y}(kT)$, where k is an integer ranging from $-\infty$ to $+\infty$. A discrete-time system is causal if current output depends on current and past inputs. The state at time k_0, denoted by $\mathbf{x}[k_0]$, is the information at time instant k_0, which together with $\mathbf{u}[k]$, $k \geq k_0$, determines uniquely the output $\mathbf{y}[k]$, $k \geq k_0$. The entries of $\mathbf{x}$ are called state variables. If the number of state variables is finite, the discrete-time system is lumped; otherwise, it is distributed. Every continuous-time system involving time delay, as the ones in Examples 2.1 and 2.3, is a distributed system. In a discrete-time system, if the time delay is an integer multiple of the sampling period T, then the discrete-time system is a lumped system.

A discrete-time system is linear if the additivity and homogeneity properties hold. The response of every linear discrete-time system can be decomposed as

$$\textit{Response} = \textit{zero-state response} + \textit{zero-input response}$$

and the zero-state responses satisfy the superposition property. So do the zero-input responses.

Input–output description Let $\delta[k]$ be the *impulse sequence* defined as

$$\delta[k - m] = \begin{cases} 1 & \text{if } k = m \\ 0 & \text{if } k \neq m \end{cases}$$

where both k and m are integers, denoting sampling instants. It is the discrete counterpart of the impulse $\delta(t - t_1)$. The impulse $\delta(t - t_1)$ has zero width and infinite height and cannot be generated in practice; whereas the impulse sequence $\delta[k - m]$ can easily be generated. Let $u[k]$ be any input sequence. Then it can be expressed as

$$u[k] = \sum_{m=-\infty}^{\infty} u[m]\delta[k - m]$$

Let $g[k, m]$ be the output at time instant k excited by the impulse sequence applied at time instant m. Then we have

$$\delta[k-m] \to g[k,m]$$
$$\delta[k-m]u[m] \to g[k,m]u[m] \quad \text{(homogeneity)}$$
$$\sum_m \delta[k-m]u[m] \to \sum_m g[k,m]u[m] \quad \text{(additivity)}$$

Thus the output $y[k]$ excited by the input $u[k]$ equals

$$y[k] = \sum_{m=-\infty}^{\infty} g[k,m]u[m] \tag{2.33}$$

This is the discrete counterpart of (2.3) and its derivation is considerably simpler. The sequence $g[k,m]$ is called the *impulse response sequence.*

If a discrete-time system is causal, no output will appear before an input is applied. Thus we have

$$\text{Causal} \iff g[k,m] = 0, \quad \text{for } k < m$$

If a system is relaxed at k_0 and causal, then (2.33) can be reduced to

$$y[k] = \sum_{m=k_0}^{k} g[k,m]u[m] \tag{2.34}$$

as in (2.4).

If a linear discrete-time system is time invariant as well, then the time shifting property holds. In this case, the initial time instant can always be chosen as $k_0 = 0$ and (2.34) becomes

$$y[k] = \sum_{m=0}^{k} g[k-m]u[m] = \sum_{m=0}^{k} g[m]u[k-m] \tag{2.35}$$

This is the discrete counterpart of (2.8) and is called a *discrete convolution.*

The z-transform is an important tool in the study of LTI discrete-time systems. Let $\hat{y}(z)$ be the z-transform of $y[k]$ defined as

$$\hat{y}(z) := Z[y[k]] := \sum_{k=0}^{\infty} y[k]z^{-k} \tag{2.36}$$

We first replace the upper limit of the integration in (2.35) to ∞,[5] and then substitute it into (2.36) to yield

$$\begin{aligned}\hat{y}(z) &= \sum_{k=0}^{\infty}\left(\sum_{m=0}^{\infty} g[k-m]u[m]\right)z^{-(k-m)}z^{-m} \\ &= \sum_{m=0}^{\infty}\left(\sum_{k=0}^{\infty} g[k-m]z^{-(k-m)}\right)u[m]z^{-m} \\ &= \left(\sum_{l=0}^{\infty} g[l]z^{-l}\right)\left(\sum_{m=0}^{\infty} u[m]z^{-m}\right) =: \hat{g}(z)\hat{u}(z)\end{aligned}$$

5. This is permitted under the causality assumption.

where we have interchanged the order of summations, introduced the new variable $l = k - m$, and then used the fact that $g[l] = 0$ for $l < 0$ to make the inner summation independent of m. The equation

$$\hat{y}(z) = \hat{g}(z)\hat{u}(z) \tag{2.37}$$

is the discrete counterpart of (2.10). The function $\hat{g}(z)$ is the z-transform of the impulse response sequence $g[k]$ and is called the *discrete transfer function.* Both the discrete convolution and transfer function describe only zero-state responses.

EXAMPLE 2.14 Consider the unit-sampling-time delay system defined by

$$y[k] = u[k-1]$$

The output equals the input delayed by one sampling period. Its impulse response sequence is $g[k] = \delta[k-1]$ and its discrete transfer function is

$$\hat{g}(z) = Z[\delta[k-1]] = z^{-1} = \frac{1}{z}$$

It is a rational function of z. Note that every continuous-time system involving time delay is a distributed system. This is not so in discrete-time systems.

EXAMPLE 2.15 Consider the discrete-time feedback system shown in Fig. 2.18(a). It is the discrete counterpart of Fig. 2.5(a). If the unit-sampling-time delay element is replaced by its transfer function z^{-1}, then the block diagram becomes the one in Fig. 2.18(b) and the transfer function from r to y can be computed as

$$\hat{g}(z) = \frac{az^{-1}}{1-az^{-1}} = \frac{a}{z-a}$$

This is a rational function of z and is similar to (2.12). The transfer function can also be obtained by applying the z-transform to the impulse response sequence of the feedback system. As in (2.9), the impulse response sequence is

$$g_f[k] = a\delta[k-1] + a^2\delta[k-2] + \cdots = \sum_{m=1}^{\infty} a^m\delta[k-m]$$

The z-transform of $\delta[k-m]$ is z^{-m}. Thus the transfer function of the feedback system is

$$\begin{aligned}\hat{g}_f(z) &= Z[g_f[k]] = az^{-1} + a^2z^{-2} + a^3z^{-3} + \cdots \\ &= az^{-1}\sum_{m=0}^{\infty}(az^{-1})^m = \frac{az^{-1}}{1-az^{-1}}\end{aligned}$$

which yields the same result.

The discrete transfer functions in the preceding two examples are all rational functions of z. This may not be so in general. For example, if

$$g[k] = \begin{cases} 0 & \text{for } m \le 0 \\ 1/k & \text{for } k = 1, 2, \ldots \end{cases}$$

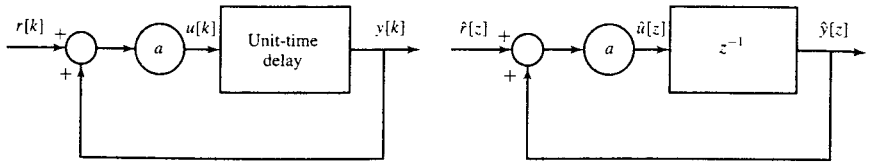


Figure 2.18 Discrete-time feedback system.

Then we have

$$\hat{g}(z) = z^{-1} + \tfrac{1}{2}z^{-2} + \tfrac{1}{3}z^{-3} + \cdots = -\ln(1 - z^{-1})$$

It is an irrational function of z. Such a system is a distributed system. We study in this text only lumped discrete-time systems and their discrete transfer functions are all rational functions of z.

Discrete rational transfer functions can be proper or improper. If a transfer function is improper such as $\hat{g}(z) = (z^2 + 2z - 1)/(z - 0.5)$, then

$$\frac{\hat{y}(z)}{\hat{u}(z)} = \frac{z^2 + 2z - 1}{z - 0.5}$$

which implies

$$y[k + 1] - 0.5y[k] = u[k + 2] + 2u[k + 1] - u[k]$$

or

$$y[k + 1] = 0.5y[k] + u[k + 2] + 2u[k + 1] - u[k]$$

It means that the output at time instant $k + 1$ depends on the input at time instant $k + 2$, a future input. Thus a discrete-time system described by an improper transfer function is not causal. We study only causal systems. Thus all discrete rational transfer functions will be proper. We mentioned earlier that we also study only proper rational transfer functions of s in the continuous-time case. The reason, however, is different. Consider $\hat{g}(s) = s$ or $y(t) = du(t)/dt$. It is a pure differentiator. If we define the differentiation as

$$y(t) = \frac{du(t)}{dt} = \lim_{\Delta \to 0} \frac{u(t + \Delta) - u(t)}{\Delta}$$

where $\Delta > 0$, then the output $y(t)$ depends on future input $u(t + \Delta)$ and the differentiator is not causal. However, if we define the differentiation as

$$y(t) = \frac{du(t)}{dt} = \lim_{\Delta \to 0} \frac{u(t) - u(t - \Delta)}{\Delta}$$

then the output $y(t)$ does not depend on future input and the differentiator is causal. Therefore in continuous-time systems, it is open to argument whether an improper transfer function represents a noncausal system. However, improper transfer functions of s will amplify high-frequency noise, which often exists in the real world. Therefore improper transfer functions are avoided in practice.

State-space equations Every linear lumped discrete-time system can be described by

$$\begin{aligned} \mathbf{x}[k + 1] &= \mathbf{A}[k]\mathbf{x}[k] + \mathbf{B}[k]\mathbf{u}[k] \\ \mathbf{y}[k] &= \mathbf{C}[k]\mathbf{x}[k] + \mathbf{D}[k]\mathbf{u}[k] \end{aligned} \tag{2.38}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are functions of k. If the system is time invariant as well, then (2.38) becomes

$$\begin{aligned} \mathbf{x}[k + 1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] \\ \mathbf{y}[k] &= \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k] \end{aligned} \tag{2.39}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are constant matrices. Let $\hat{\mathbf{x}}(z)$ be the z-transform of $\mathbf{x}[k]$ or

$$\hat{\mathbf{x}}(z) = Z[\mathbf{x}[k]] := \sum_{k=0}^{\infty} \mathbf{x}[k]z^{-k}$$

Then we have

$$\begin{aligned} Z[\mathbf{x}[k + 1]] &= \sum_{k=0}^{\infty} \mathbf{x}[k + 1]z^{-k} = z \sum_{k=0}^{\infty} \mathbf{x}[k + 1]z^{-(k+1)} \\ &= z\left[\sum_{l=1}^{\infty} \mathbf{x}[l]z^{-l} + \mathbf{x}[0] - \mathbf{x}[0]\right] = z(\hat{\mathbf{x}}(z) - \mathbf{x}[0]) \end{aligned}$$

Applying the z-transform to (2.39) yields

$$\begin{aligned} z\hat{\mathbf{x}}(z) - z\mathbf{x}[0] &= \mathbf{A}\hat{\mathbf{x}}(z) + \mathbf{B}\hat{\mathbf{u}}(z) \\ \hat{\mathbf{y}}(z) &= \mathbf{C}\hat{\mathbf{x}}(z) + \mathbf{D}\hat{\mathbf{u}}(z) \end{aligned}$$

which implies

$$\hat{\mathbf{x}}(z) = (z\mathbf{I} - \mathbf{A})^{-1}z\mathbf{x}[0] + (z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{u}}(z) \tag{2.40}$$

$$\hat{\mathbf{y}}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}z\mathbf{x}[0] + \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{u}}(z) + \mathbf{D}\hat{\mathbf{u}}(z) \tag{2.41}$$

They are the discrete counterparts of (2.14) and (2.15). Note that there is an extra z in front of $\mathbf{x}[0]$. If $\mathbf{x}[0] = \mathbf{0}$, then (2.41) reduces to

$$\hat{\mathbf{y}}(z) = [\mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}]\hat{\mathbf{u}}(z) \tag{2.42}$$

Comparing this with the MIMO case of (2.37) yields

$$\hat{\mathbf{G}}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} \tag{2.43}$$

This is the discrete counterpart of (2.16). If the Laplace transform variable s is replaced by the z-transform variable z, then the two equations are identical.

EXAMPLE 2.16 Consider a money market account in a brokerage firm. If the interest rate depends on the amount of money in the account, it is a nonlinear system. If the interest rate is the same no matter how much money is in the account, then it is a linear system. The account is a time-varying system if the interest rate changes with time; a time-invariant system if the interest rate is fixed. We consider here only the LTI case with interest rate $r = 0.015\%$ per day and compounded daily. The input $u[k]$ is the amount of money deposited into the account on the kth day and the output $y[k]$ is the total amount of money in the account at the end of the kth day. If we withdraw money, then $u[k]$ is negative.

If we deposit one dollar on the first day (that is, $u[0] = 1$) and nothing thereafter ($u[k] = 0$, $k = 1, 2, \ldots$), then $y[0] = u[0] = 1$ and $y[1] = 1 + 0.00015 = 1.00015$. Because the money is compounded daily, we have

$$y[2] = y[1] + y[1] \cdot 0.00015 = y[1] \cdot 1.00015 = (1.00015)^2$$

and, in general,

$$y[k] = (1.00015)^k$$

Because the input $\{1, 0, 0, \ldots\}$ is an impulse sequence, the output is, by definition, the impulse response sequence or

$$g[k] = (1.00015)^k$$

and the input–output description of the account is

$$y[k] = \sum_{m=0}^{k} g[k-m]u[m] = \sum_{m=0}^{k} (1.00015)^{k-m} u[m] \tag{2.44}$$

The discrete transfer function is the z-transform of the impulse response sequence or

$$\hat{g}(z) = Z[g[k]] = \sum_{k=0}^{\infty} (1.00015)^k z^{-k} = \sum_{k=0}^{\infty} (1.00015z^{-1})^k$$

$$= \frac{1}{1 - 1.00015z^{-1}} = \frac{z}{z - 1.00015} \tag{2.45}$$

Whenever we use (2.44) or (2.45), the initial state must be zero, or there is initially no money in the account.

Next we develop a state-space equation to describe the account. Suppose $y[k]$ is the total amount of money at the end of the kth day. Then we have

$$y[k+1] = y[k] + 0.00015y[k] + u[k+1] = 1.00015y[k] + u[k+1] \tag{2.46}$$

If we define the state variable as $x[k] := y[k]$, then

$$\begin{aligned} x[k+1] &= 1.00015x[k] + u[k+1] \\ y[k] &= x[k] \end{aligned} \tag{2.47}$$

Because of $u[k+1]$, (2.47) is not in the standard form of (2.39). Thus we cannot select $x[k] := y[k]$ as a state variable. Next we select a different state variable as

$$x[k] := y[k] - u[k]$$

Substituting $y[k+1] = x[k+1] + u[k+1]$ and $y[k] = x[k] + u[k]$ into (2.46) yields

$$\begin{aligned} x[k+1] &= 1.00015x[k] + 1.00015u[k] \\ y[k] &= x[k] + u[k] \end{aligned} \tag{2.48}$$

This is in the standard form and describes the money market account.

The linearization discussed for the continuous-time case can also be applied to the discrete-time case with only slight modification. Therefore its discussion will not be repeated.

2.7 Concluding Remarks

We introduced in this chapter the concepts of causality, lumpedness, linearity, and time invariance. Mathematical equations were then developed to describe causal systems, as summarized in the following.

System type	Internal description	External description
Distributed, linear		$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{G}(t, \tau)\mathbf{u}(\tau)\, d\tau$
Lumped, linear	$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}$ $\mathbf{y} = \mathbf{C}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u}$	$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{G}(t, \tau)\mathbf{u}(\tau)\, d\tau$
Distributed, linear, time-invariant		$\mathbf{y}(t) = \int_{0}^{t} \mathbf{G}(t-\tau)\mathbf{u}(\tau)\, d\tau$ $\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}(s)\hat{\mathbf{u}}(s)$, $\hat{\mathbf{G}}(s)$ irrational
Lumped, linear, time-invariant	$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$ $\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}$	$\mathbf{y}(t) = \int_{0}^{t} \mathbf{G}(t-\tau)\mathbf{u}(\tau)\, d\tau$ $\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}(s)\hat{\mathbf{u}}(s)$, $\hat{\mathbf{G}}(s)$ rational

Distributed systems cannot be described by finite-dimensional state-space equations. External description describes only zero-state responses; thus whenever we use the description, systems are implicitly assumed to be relaxed or their initial conditions are assumed to be zero.

We study in this text mainly lumped linear time-invariant systems. For this class of systems, we use mostly the time-domain description $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ in the internal description and the frequency-domain (Laplace-domain) description $\hat{\mathbf{G}}(s)$ in the external description. Furthermore, we will express every rational transfer matrix as a fraction of two polynomial matrices, as we will develop in the text. By so doing, all designs in the SISO case can be extended to the multivariable case.

The class of lumped linear time-invariant systems constitutes only a very small part of nonlinear and linear systems. For this small class of systems, we are able to give a complete treatment of analyses and syntheses. This study will form a foundation for studying more general systems.

PROBLEMS

2.1 Consider the memoryless systems with characteristics shown in Fig. 2.19, in which u denotes the input and y the output. Which of them is a linear system? Is it possible to introduce a new output so that the system in Fig. 2.19(b) is linear?

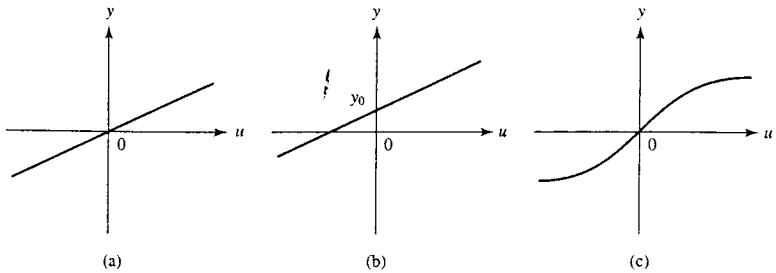


Figure 2.19

2.2 The impulse response of an ideal lowpass filter is given by

$$g(t) = 2\omega \frac{\sin 2\omega(t - t_0)}{2\omega(t - t_0)}$$

for all t, where ω and t_0 are constants. Is the ideal lowpass filter causal? Is it possible to build the filter in the real world?

2.3 Consider a system whose input u and output y are related by

$$y(t) = (P_\alpha u)(t) := \begin{cases} u(t) & \text{for } t \leq \alpha \\ 0 & \text{for } t > \alpha \end{cases}$$

where α is a fixed constant. The system is called a *truncation operator*, which chops off the input after time α. Is the system linear? Is it time-invariant? Is it causal?

2.4 The input and output of an initially relaxed system can be denoted by $y = Hu$, where H is some mathematical operator. Show that if the system is causal, then

$$P_\alpha y = P_\alpha H u = P_\alpha H P_\alpha u$$

where P_α is the truncation operator defined in Problem 2.3. Is it true $P_\alpha H u = H P_\alpha u$?

2.5 Consider a system with input u and output y. Three experiments are performed on the system using the inputs $u_1(t)$, $u_2(t)$, and $u_3(t)$ for $t \geq 0$. In each case, the initial state $\mathbf{x}(0)$ at time $t = 0$ is the same. The corresponding outputs are denoted by y_1, y_2, and y_3. Which of the following statements are correct if $\mathbf{x}(0) \neq \mathbf{0}$?

1. If $u_3 = u_1 + u_2$, then $y_3 = y_1 + y_2$.
2. If $u_3 = 0.5(u_1 + u_2)$, then $y_3 = 0.5(y_1 + y_2)$.
3. If $u_3 = u_1 - u_2$, then $y_3 = y_1 - y_2$.

Which are correct if $\mathbf{x}(0) = \mathbf{0}$?

2.6 Consider a system whose input and output are related by

$$y(t) = \begin{cases} u^2(t)/u(t-1) & \text{if } u(t-1) \neq 0 \\ 0 & \text{if } u(t-1) = 0 \end{cases}$$

for all t. Show that the system satisfies the homogeneity property but not the additivity property.

2.7 Show that if the additivity property holds, then the homogeneity property holds for all rational numbers α. Thus if a system has some "continuity" property, then additivity implies homogeneity.

2.8 Let $g(t, \tau) = g(t+\alpha, \tau+\alpha)$ for all t, τ, and α. Show that $g(t, \tau)$ depends only on $t - \tau$. [*Hint:* Define $x = t + \tau$ and $y = t - \tau$ and show that $\partial g(t, \tau)/\partial x = 0$.]

2.9 Consider a system with impulse response as shown in Fig. 2.20(a). What is the zero-state response excited by the input $u(t)$ shown in Fig. 2.20(b).

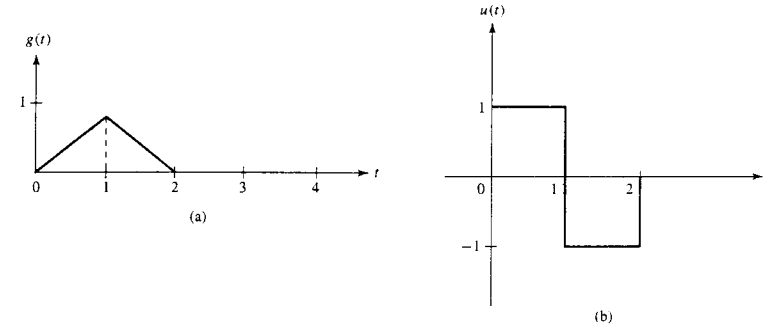


Figure 2.20

2.10 Consider a system described by

$$\ddot{y} + 2\dot{y} - 3y = \dot{u} - u$$

What are the transfer function and the impulse response of the system?

2.11 Let $\bar{y}(t)$ be the unit-step response of a linear time-invariant system. Show that the impulse response of the system equals $d\bar{y}(t)/dt$.

2.12 Consider a two-input and two-output system described by

$$D_{11}(p)y_1(t) + D_{12}(p)y_2(t) = N_{11}(p)u_1(t) + N_{12}(p)u_2(t)$$

$$D_{21}(p)y_1(t) + D_{22}(p)y_2(t) = N_{21}(p)u_1(t) + N_{22}(p)u_2(t)$$

where N_{ij} and D_{ij} are polynomials of $p := d/dt$. What is the transfer matrix of the system?

2.13 Consider the feedback systems shown in Fig. 2.5. Show that the unit-step responses of the positive-feedback system are as shown in Fig. 2.21(a) for $a = 1$ and in Fig. 2.21(b) for $a = 0.5$. Show also that the unit-step responses of the negative-feedback system are as shown in Figs. 2.21(c) and 2.21(d), respectively, for $a = 1$ and $a = 0.5$.

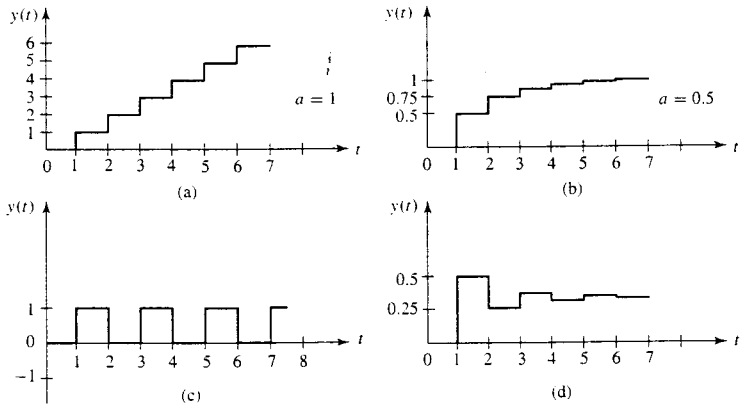


Figure 2.21

2.14 Draw an op-amp circuit diagram for

$$\dot{\mathbf{x}} = \begin{bmatrix} -2 & 4 \\ 0 & 5 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 2 \\ -4 \end{bmatrix}u$$

$$y = [3 \;\; 10]\mathbf{x} - 2u$$

2.15 Find state equations to describe the pendulum systems in Fig. 2.22. The systems are useful to model one- or two-link robotic manipulators. If θ, θ_1, and θ_2 are very small, can you consider the two systems as linear?

2.16 Consider the simplified model of an aircraft shown in Fig. 2.23. It is assumed that the aircraft is in an equilibrium state at the pitched angle θ_0, elevator angle u_0, altitude h_0, and cruising speed v_0. It is assumed that small deviations of θ and u from θ_0 and u_0 generate forces $f_1 = k_1\theta$ and $f_2 = k_2u$ as shown in the figure. Let m be the mass of the aircraft, I the moment of inertia about the center of gravity P, $b\dot{\theta}$ the aerodynamic damping, and h the deviation of the altitude from h_0. Find a state equation to describe the system. Show also that the transfer function from u to h, by neglecting the effect of I, is

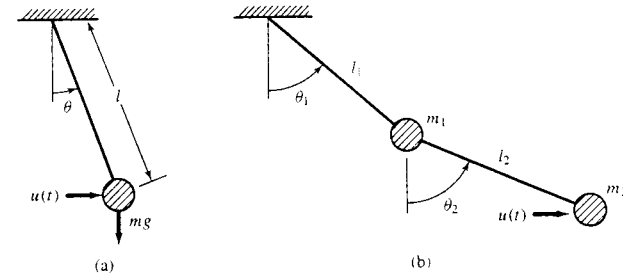


Figure 2.22

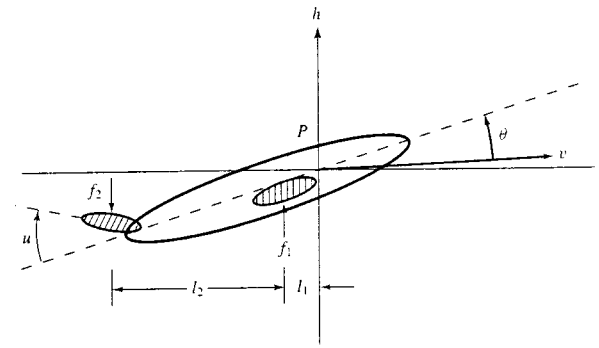


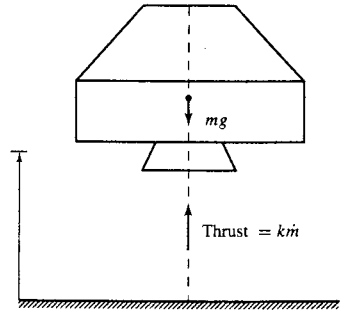
Figure 2.23

$$\hat{g}(s) = \frac{\hat{h}(s)}{\hat{u}(s)} = \frac{k_1k_2l_2 - k_2bs}{ms^2(bs + k_1l_1)}$$

2.17 The soft landing phase of a lunar module descending on the moon can be modeled as shown in Fig. 2.24. The thrust generated is assumed to be proportional to $\dot{m}$, where m is the mass of the module. Then the system can be described by $m\ddot{y} = -k\dot{m} - mg$, where g is the gravity constant on the lunar surface. Define state variables of the system as $x_1 = y$, $x_2 = \dot{y}$, $x_3 = m$, and $u = \dot{m}$. Find a state-space equation to describe the system.

2.18 Find the transfer functions from u to y_1 and from y_1 to y of the hydraulic tank system shown in Fig. 2.25. Does the transfer function from u to y equal the product of the two transfer functions? Is this also true for the system shown in Fig. 2.14? [*Answer:* No, because of the *loading* problem in the two tanks in Fig. 2.14. The loading problem is an important issue in developing mathematical equations to describe composite systems. See Reference [7].]

Figure 2.24



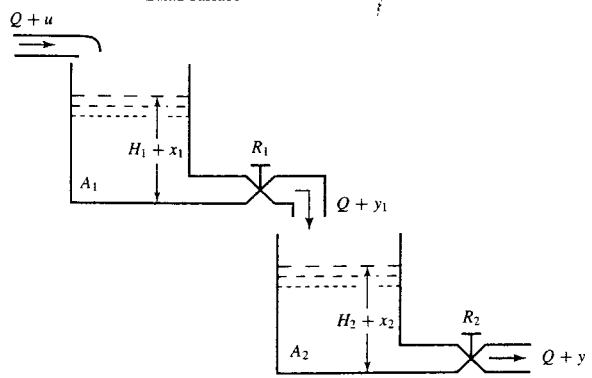
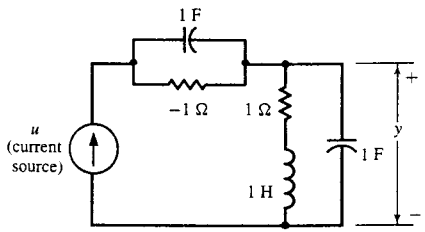


Figure 2.25

2.19 Find a state equation to describe the network shown in Fig. 2.26. Find also its transfer function.

Figure 2.26

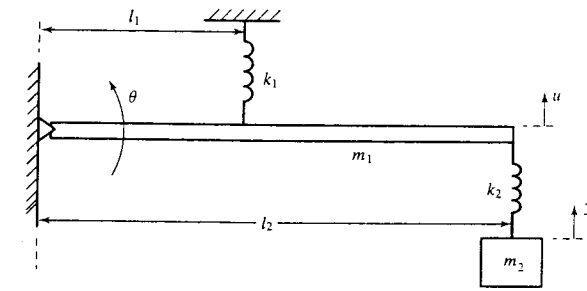


2.20 Find a state equation to describe the network shown in Fig. 2.2. Compute also its transfer matrix.

2.21 Consider the mechanical system shown in Fig. 2.27. Let I denote the moment of inertia of the bar and block about the hinge. It is assumed that the angular displacement θ is very small. An external force u is applied to the bar as shown. Let y be the displacement

of the block, with mass m_2, from equilibrium. Find a state-space equation to describe the system. Find also the transfer function from u to y.

Figure 2.27



Chapter 3

Linear Algebra

3.1 Introduction

This chapter reviews a number of concepts and results in linear algebra that are essential in the study of this text. The topics are carefully selected, and only those that will be used subsequently are introduced. Most results are developed intuitively in order for the reader to better grasp the ideas. They are stated as theorems for easy reference in later chapters. However, no formal proofs are given.

As we saw in the preceding chapter, all parameters that arise in the real world are real numbers. Therefore we deal only with real numbers, unless stated otherwise, throughout this text. Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ be, respectively, $n \times m$, $m \times r$, $l \times n$, and $r \times p$ real matrices. Let $\mathbf{a}_i$ be the ith column of $\mathbf{A}$, and $\mathbf{b}_j$ the jth row of $\mathbf{B}$. Then we have

$$\mathbf{AB} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_m] \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix} = \mathbf{a}_1\mathbf{b}_1 + \mathbf{a}_2\mathbf{b}_2 + \cdots + \mathbf{a}_m\mathbf{b}_m \tag{3.1}$$

$$\mathbf{CA} = \mathbf{C}[\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_m] = [\mathbf{Ca}_1 \ \mathbf{Ca}_2 \ \cdots \ \mathbf{Ca}_m] \tag{3.2}$$

and

$$\mathbf{BD} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix} \mathbf{D} = \begin{bmatrix} \mathbf{b}_1\mathbf{D} \\ \mathbf{b}_2\mathbf{D} \\ \vdots \\ \mathbf{b}_m\mathbf{D} \end{bmatrix} \tag{3.3}$$

These identities can easily be verified. Note that $\mathbf{a}_i\mathbf{b}_i$ is an $n \times r$ matrix; it is the product of an $n \times 1$ column vector and a $1 \times r$ row vector. The product $\mathbf{b}_i\mathbf{a}_i$ is not defined unless $n = r$; it becomes a scalar if $n = r$.

3.2 Basis, Representation, and Orthonormalization

Consider an n-dimensional real linear space, denoted by $\mathcal{R}^n$. Every vector in $\mathcal{R}^n$ is an n-tuple of real numbers such as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

To save space, we write it as $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_n]'$, where the prime denotes the transpose.

The set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ in $\mathcal{R}^n$ is said to be *linearly dependent* if there exist real numbers $\alpha_1, \alpha_2, \ldots, \alpha_m$, not all zero, such that

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \cdots + \alpha_m\mathbf{x}_m = \mathbf{0} \tag{3.4}$$

If the only set of α_i for which (3.4) holds is $\alpha_1 = 0, \alpha_2 = 0, \ldots, \alpha_m = 0$, then the set of vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ is said to be *linearly independent*.

If the set of vectors in (3.4) is linearly dependent, then there exists at least one α_i, say, α_1, that is different from zero. Then (3.4) implies

$$\begin{aligned} \mathbf{x}_1 &= -\frac{1}{\alpha_1}[\alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3 + \cdots + \alpha_m\mathbf{x}_m] \\ &=: \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \cdots + \beta_m\mathbf{x}_m \end{aligned}$$

where $\beta_i = -\alpha_i/\alpha_1$. Such an expression is called a linear combination.

The *dimension* of a linear space can be defined as the maximum number of linearly independent vectors in the space. Thus in $\mathcal{R}^n$, we can find at most n linearly independent vectors.

Basis and representation A set of linearly independent vectors in $\mathcal{R}^n$ is called a *basis* if every vector in $\mathcal{R}^n$ can be expressed as a unique linear combination of the set. In $\mathcal{R}^n$, any set of n linearly independent vectors can be used as a basis. Let $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ be such a set. Then every vector $\mathbf{x}$ can be expressed uniquely as

$$\mathbf{x} = \alpha_1\mathbf{q}_1 + \alpha_2\mathbf{q}_2 + \cdots + \alpha_n\mathbf{q}_n \tag{3.5}$$

Define the $n \times n$ square matrix

$$\mathbf{Q} := [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n] \tag{3.6}$$

Then (3.5) can be written as

$$\mathbf{x} = \mathbf{Q}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n\end{bmatrix} =: \mathbf{Q}\bar{\mathbf{x}} \tag{3.7}$$

We call $\bar{\mathbf{x}} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_n]'$ the *representation* of the vector $\mathbf{x}$ with respect to the basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$.

We will associate with every $\mathcal{R}^n$ the following *orthonormal basis*:

$$\mathbf{i}_1 = \begin{bmatrix}1\\0\\0\\\vdots\\0\\0\end{bmatrix}, \ \mathbf{i}_2 = \begin{bmatrix}0\\1\\0\\\vdots\\0\\0\end{bmatrix}, \ \cdots, \ \mathbf{i}_{n-1} = \begin{bmatrix}0\\0\\0\\\vdots\\1\\0\end{bmatrix}, \ \mathbf{i}_n = \begin{bmatrix}0\\0\\0\\\vdots\\0\\1\end{bmatrix} \tag{3.8}$$

With respect to this basis, we have

$$\mathbf{x} := \begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix} = x_1\mathbf{i}_1 + x_2\mathbf{i}_2 + \cdots + x_n\mathbf{i}_n = \mathbf{I}_n\begin{bmatrix}x_1\\x_2\\\vdots\\x_n\end{bmatrix}$$

where $\mathbf{I}_n$ is the $n \times n$ unit matrix. In other words, the representation of any vector $\mathbf{x}$ with respect to the orthonormal basis in (3.8) equals itself.

EXAMPLE 3.1 Consider the vector $\mathbf{x} = [1 \ 3]'$ in $\mathcal{R}^2$ as shown in Fig. 3.1, The two vectors $\mathbf{q}_1 = [3 \ 1]'$ and $\mathbf{q}_2 = [2 \ 2]'$ are clearly linearly independent and can be used as a basis. If we draw from $\mathbf{x}$ two lines in parallel with $\mathbf{q}_2$ and $\mathbf{q}_1$, they intersect at $-\mathbf{q}_1$ and $2\mathbf{q}_2$ as shown. Thus the representation of $\mathbf{x}$ with respect to $\{\mathbf{q}_1, \mathbf{q}_2\}$ is $[-1 \ 2]'$. This can also be verified from

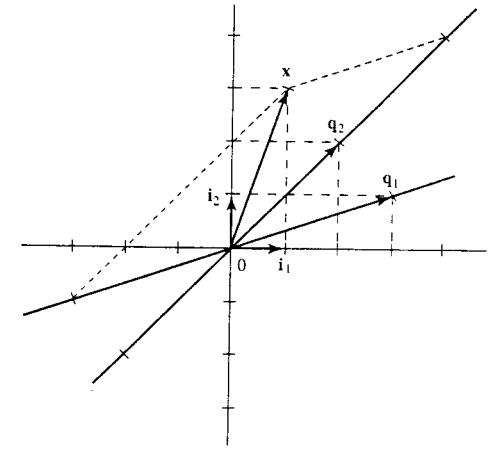
$$\mathbf{x} = \begin{bmatrix}1\\3\end{bmatrix} = [\mathbf{q}_1 \ \mathbf{q}_2]\begin{bmatrix}-1\\2\end{bmatrix} = \begin{bmatrix}3 & 2\\1 & 2\end{bmatrix}\begin{bmatrix}-1\\2\end{bmatrix}$$

To find the representation of $\mathbf{x}$ with respect to the basis $\{\mathbf{q}_2, \mathbf{i}_2\}$, we draw from $\mathbf{x}$ two lines in parallel with $\mathbf{i}_2$ and $\mathbf{q}_2$. They intersect at $0.5\mathbf{q}_2$ and $2\mathbf{i}_2$. Thus the representation of $\mathbf{x}$ with respect to $\{\mathbf{q}_2, \mathbf{i}_2\}$ is $[0.5 \ 2]'$. (Verify.)

Norms of vectors The concept of *norm* is a generalization of length or magnitude. Any real-valued function of $\mathbf{x}$, denoted by $||\mathbf{x}||$, can be defined as a norm if it has the following properties:

1. $||\mathbf{x}|| \geq 0$ for every $\mathbf{x}$ and $||\mathbf{x}|| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
2. $||\alpha\mathbf{x}|| = |\alpha|||\mathbf{x}||$, for any real α.
3. $||\mathbf{x}_1 + \mathbf{x}_2|| \leq ||\mathbf{x}_1|| + ||\mathbf{x}_2||$ for every $\mathbf{x}_1$ and $\mathbf{x}_2$.

Figure 3.1 Different representations of vector $\mathbf{x}$.



The last inequality is called the *triangular inequality*.

Let $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_n]'$. Then the norm of $\mathbf{x}$ can be chosen as any one of the following:

$$||\mathbf{x}||_1 := \sum_{i=1}^{n} |x_i|$$

$$||\mathbf{x}||_2 := \sqrt{\mathbf{x}'\mathbf{x}} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

$$||\mathbf{x}||_\infty := \max_i |x_i|$$

They are called, respectively, 1-norm, 2- or Euclidean norm, and infinite-norm. The 2-norm is the length of the vector from the origin. We use exclusively, unless stated otherwise, the Euclidean norm and the subscript 2 will be dropped.

In MATLAB, the norms just introduced can be obtained by using the functions `norm(x,1)`, `norm(x,2)=norm(x)`, and `norm(x,inf)`.

Orthonormalization A vector $\mathbf{x}$ is said to be normalized if its Euclidean norm is 1 or $\mathbf{x}'\mathbf{x} = 1$. Note that $\mathbf{x}'\mathbf{x}$ is scalar and $\mathbf{x}\mathbf{x}'$ is $n \times n$. Two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are said to be *orthogonal* if $\mathbf{x}_1'\mathbf{x}_2 = \mathbf{x}_2'\mathbf{x}_1 = 0$. A set of vectors $\mathbf{x}_i$, $i = 1, 2, \ldots, m$, is said to be *orthonormal* if

$$\mathbf{x}_i'\mathbf{x}_j = \begin{cases}0 & \text{if } i \neq j\\ 1 & \text{if } i = j\end{cases}$$

Given a set of linearly independent vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m$, we can obtain an orthonormal set using the procedure that follows:

$$\mathbf{u}_1 := \mathbf{e}_1 \qquad \mathbf{q}_1 := \mathbf{u}_1/\|\mathbf{u}_1\|$$
$$\mathbf{u}_2 := \mathbf{e}_2 - (\mathbf{q}_1'\mathbf{e}_2)\mathbf{q}_1 \qquad \mathbf{q}_2 := \mathbf{u}_2/\|\mathbf{u}_2\|$$
$$\vdots$$
$$\mathbf{u}_m := \mathbf{e}_m - \sum_{k=1}^{m-1}(\mathbf{q}_k'\mathbf{e}_m)\mathbf{q}_k \qquad \mathbf{q}_m := \mathbf{u}_m/\|\mathbf{u}_m\|$$

The first equation normalizes the vector $\mathbf{e}_1$ to have norm 1. The vector $(\mathbf{q}_1'\mathbf{e}_2)\mathbf{q}_1$ is the projection of the vector $\mathbf{e}_2$ along $\mathbf{q}_1$. Its subtraction from $\mathbf{e}_2$ yields the vertical part $\mathbf{u}_2$. It is then normalized to 1 as shown in Fig. 3.2. Using this procedure, we can obtain an orthonormal set. This is called the *Schmidt orthonormalization procedure*.

Let $\mathbf{A} = [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_m]$ be an $n \times m$ matrix with $m \leq n$. If all columns of $\mathbf{A}$ or $\{\mathbf{a}_i,\ i = 1, 2, \ldots, m\}$ are orthonormal, then

$$\mathbf{A}'\mathbf{A} = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_m' \end{bmatrix} [\mathbf{a}_1 \ \mathbf{a}_2 \ \cdots \ \mathbf{a}_m] = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}_m$$

where $\mathbf{I}_m$ is the unit matrix of order m. Note that, in general, $\mathbf{A}\mathbf{A}' \neq \mathbf{I}_n$. See Problem 3.4.

3.3 Linear Algebraic Equations

Consider the set of linear algebraic equations

$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{3.9}$$

where $\mathbf{A}$ and $\mathbf{y}$ are, respectively, $m \times n$ and $m \times 1$ real matrices and $\mathbf{x}$ is an $n \times 1$ vector. The matrices $\mathbf{A}$ and $\mathbf{y}$ are given and $\mathbf{x}$ is the unknown to be solved. Thus the set actually consists of m equations and n unknowns. The number of equations can be larger than, equal to, or smaller than the number of unknowns.

We discuss the existence condition and general form of solutions of (3.9). The *range space* of $\mathbf{A}$ is defined as all possible linear combinations of all columns of $\mathbf{A}$. The *rank* of $\mathbf{A}$ is

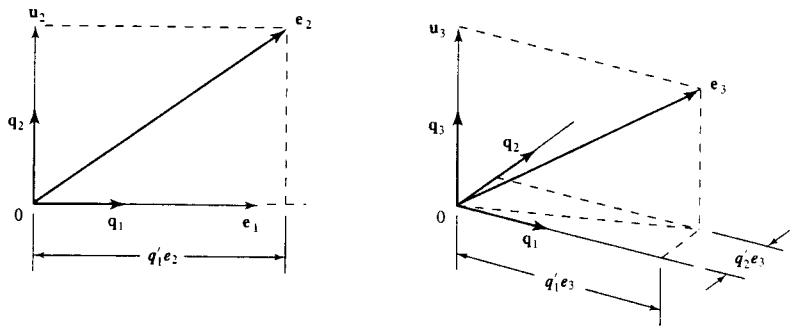


Figure 3.2 Schmidt orthonormization procedure.

defined as the dimension of the range space or, equivalently, the number of linearly independent columns in $\mathbf{A}$. A vector $\mathbf{x}$ is called a *null vector* of $\mathbf{A}$ if $\mathbf{A}\mathbf{x} = \mathbf{0}$. The *null space* of $\mathbf{A}$ consists of all its null vectors. The *nullity* is defined as the maximum number of linearly independent null vectors of $\mathbf{A}$ and is related to the rank by

$$\text{Nullity } (\mathbf{A}) = \text{ number of columns of } \mathbf{A} - \text{ rank } (\mathbf{A}) \tag{3.10}$$

EXAMPLE 3.2 Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 2 & 3 & 4 \\ 2 & 0 & 2 & 0 \end{bmatrix} =: [\mathbf{a}_1 \ \mathbf{a}_2 \ \mathbf{a}_3 \ \mathbf{a}_4] \tag{3.11}$$

where $\mathbf{a}_i$ denotes the ith column of $\mathbf{A}$. Clearly $\mathbf{a}_1$ and $\mathbf{a}_2$ are linearly independent. The third column is the sum of the first two columns or $\mathbf{a}_1 + \mathbf{a}_2 - \mathbf{a}_3 = \mathbf{0}$. The last column is twice the second column, or $2\mathbf{a}_2 - \mathbf{a}_4 = \mathbf{0}$. Thus $\mathbf{A}$ has two linearly independent columns and has rank 2. The set $\{\mathbf{a}_1, \mathbf{a}_2\}$ can be used as a basis of the range space of $\mathbf{A}$.

Equation (3.10) implies that the nullity of $\mathbf{A}$ is 2. It can readily be verified that the two vectors

$$\mathbf{n}_1 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0 \end{bmatrix} \qquad \mathbf{n}_2 = \begin{bmatrix} 0 \\ 2 \\ 0 \\ -1 \end{bmatrix} \tag{3.12}$$

meet the condition $\mathbf{A}\mathbf{n}_i = \mathbf{0}$. Because the two vectors are linearly independent, they form a basis of the null space.

The rank of $\mathbf{A}$ is defined as the number of linearly independent columns. It also equals the number of linearly independent rows. Because of this fact, if $\mathbf{A}$ is $m \times n$, then

$$\text{rank}(\mathbf{A}) \leq \min(m, n)$$

In MATLAB, the range space, null space, and rank can be obtained by calling the functions `orth`, `null`, and `rank`. For example, for the matrix in (3.11), we type

```
a=[0  1 1 2;1 2 3 4;2 0 2 0];
rank(a)
```

which yields 2. Note that MATLAB computes ranks by using singular-value decomposition (svd), which will be introduced later. The svd algorithm also yields the range and null spaces of the matrix. The MATLAB function `R=orth(a)` yields[1]

```
Ans     R=
```

$$\begin{matrix} 0.3782 & -0.3084 \\ 0.8877 & -0.1468 \\ 0.2627 & 0.9399 \end{matrix} \tag{3.13}$$

1. This is obtained using MATLAB Version 5. Earlier versions may yield different results.

The two columns of R form an orthonormal basis of the range space. To check the orthonormality, we type R'*R, which yields the unity matrix of order 2. The two columns in R are not obtained from the basis $\{\mathbf{a}_1, \mathbf{a}_2\}$ in (3.11) by using the Schmidt orthonormalization procedure; they are a by-product of svd. However, the two bases should span the same range space. This can be verified by typing

```
rank([a1 a2 R])
```

which yields 2. This confirms that $\{\mathbf{a}_1, \mathbf{a}_2\}$ span the same space as the two vectors of R. We mention that the rank of a matrix can be very sensitive to roundoff errors and imprecise data. For example, if we use the five-digit display of R in (3.13), the rank of [a1 a2 R] is 3. The rank is 2 if we use the R stored in MATLAB, which uses 16 digits plus exponent.

The null space of (3.11) can be obtained by typing null(a), which yields

```
Ans     N=
```

$$\begin{bmatrix} 0.3434 & -0.5802 \\ 0.8384 & 0.3395 \\ -0.3434 & 0.5802 \\ -0.2475 & -0.4598 \end{bmatrix} \tag{3.14}$$

The two columns are an orthonormal basis of the null space spanned by the two vectors $\{\mathbf{n}_1, \mathbf{n}_2\}$ in (3.12). All discussion for the range space applies here. That is, rank([n1 n2 N]) yields 3 if we use the five-digit display in (3.14). The rank is 2 if we use the N stored in MATLAB.

With this background, we are ready to discuss solutions of (3.9). We use ρ to denote the rank of a matrix.

Theorem 3.1

1. Given an $m \times n$ matrix $\mathbf{A}$ and an $m \times 1$ vector $\mathbf{y}$, an $n \times 1$ solution $\mathbf{x}$ exists in $\mathbf{Ax} = \mathbf{y}$ if and only if $\mathbf{y}$ lies in the range space of $\mathbf{A}$ or, equivalently,

$$\rho(\mathbf{A}) = \rho([\mathbf{A}\ \mathbf{y}])$$

where $[\mathbf{A}\ \mathbf{y}]$ is an $m \times (n+1)$ matrix with $\mathbf{y}$ appended to $\mathbf{A}$ as an additional column.

2. Given $\mathbf{A}$, a solution $\mathbf{x}$ exists in $\mathbf{Ax} = \mathbf{y}$ for every $\mathbf{y}$, if and only if $\mathbf{A}$ has rank m (full row rank).

The first statement follows directly from the definition of the range space. If $\mathbf{A}$ has full row rank, then the rank condition in (1) is always satisfied for every $\mathbf{y}$. This establishes the second statement.

Theorem 3.2 (Parameterization of all solutions)

Given an $m \times n$ matrix $\mathbf{A}$ and an $m \times 1$ vector $\mathbf{y}$, let $\mathbf{x}_p$ be a solution of $\mathbf{Ax} = \mathbf{y}$ and let $k := n - \rho(\mathbf{A})$ be the nullity of $\mathbf{A}$. If $\mathbf{A}$ has rank n (full column rank) or $k = 0$, then the solution $\mathbf{x}_p$ is unique. If $k > 0$, then for every real α_i, $i = 1, 2, \ldots, k$, the vector

$$\mathbf{x} = \mathbf{x}_p + \alpha_1 \mathbf{n}_1 + \cdots + \alpha_k \mathbf{n}_k \tag{3.15}$$

is a solution of $\mathbf{Ax} = \mathbf{y}$, where $\{\mathbf{n}_1, \ldots, \mathbf{n}_k\}$ is a basis of the null space of $\mathbf{A}$.

Substituting (3.15) into $\mathbf{Ax} = \mathbf{y}$ yields

$$\mathbf{Ax}_p + \sum_{i=1}^{k} \alpha_i \mathbf{An}_i = \mathbf{Ax}_p + \mathbf{0} = \mathbf{y}$$

Thus, for every α_i, (3.15) is a solution. Let $\bar{\mathbf{x}}$ be a solution or $\mathbf{A}\bar{\mathbf{x}} = \mathbf{y}$. Subtracting this from $\mathbf{Ax}_p = \mathbf{y}$ yields

$$\mathbf{A}(\bar{\mathbf{x}} - \mathbf{x}_p) = \mathbf{0}$$

which implies that $\bar{\mathbf{x}} - \mathbf{x}_p$ is in the null space. Thus $\bar{\mathbf{x}}$ can be expressed as in (3.15). This establishes Theorem 3.2.

EXAMPLE 3.3 Consider the equation

$$\mathbf{Ax} = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 2 & 3 & 4 \\ 2 & 0 & 2 & 0 \end{bmatrix} \mathbf{x} =: [\mathbf{a}_1\ \mathbf{a}_2\ \mathbf{a}_3\ \mathbf{a}_4]\mathbf{x} = \begin{bmatrix} -4 \\ -8 \\ 0 \end{bmatrix} = \mathbf{y} \tag{3.16}$$

This $\mathbf{y}$ clearly lies in the range space of $\mathbf{A}$ and $\mathbf{x}_p = [0\ \ -4\ \ 0\ \ 0]'$ is a solution. A basis of the null space of $\mathbf{A}$ was shown in (3.12). Thus the general solution of (3.16) can be expressed as

$$\mathbf{x} = \mathbf{x}_p + \alpha_1 \mathbf{n}_1 + \alpha_2 \mathbf{n}_2 = \begin{bmatrix} 0 \\ -4 \\ 0 \\ 0 \end{bmatrix} + \alpha_1 \begin{bmatrix} 1 \\ 1 \\ -1 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 2 \\ 0 \\ -1 \end{bmatrix} \tag{3.17}$$

for any real α_1 and α_2.

In application, we will also encounter $\mathbf{xA} = \mathbf{y}$, where the $m \times n$ matrix $\mathbf{A}$ and the $1 \times n$ vector $\mathbf{y}$ are given, and the $1 \times m$ vector $\mathbf{x}$ is to be solved. Applying Theorems 3.1 and 3.2 to the transpose of the equation, we can readily obtain the following result.

Corollary 3.2

1. Given an $m \times n$ matrix $\mathbf{A}$, a solution $\mathbf{x}$ exists in $\mathbf{xA} = \mathbf{y}$, for any $\mathbf{y}$, if and only if $\mathbf{A}$ has full column rank.

2. Given an $m \times n$ matrix $\mathbf{A}$ and an $1 \times n$ vector $\mathbf{y}$, let $\mathbf{x}_p$ be a solution of $\mathbf{xA} = \mathbf{y}$ and let $k = m - \rho(\mathbf{A})$. If $k = 0$, the solution $\mathbf{x}_p$ is unique. If $k > 0$, then for any α_i, $i = 1, 2, \ldots, k$, the vector

$$\mathbf{x} = \mathbf{x}_p + \alpha_1 \mathbf{n}_1 + \cdots + \alpha_k \mathbf{n}_k$$

is a solution of $\mathbf{xA} = \mathbf{y}$, where $\mathbf{n}_i \mathbf{A} = \mathbf{0}$ and the set $\{\mathbf{n}_1, \ldots, \mathbf{n}_k\}$ is linearly independent.

In MATLAB, the solution of $\mathbf{Ax} = \mathbf{y}$ can be obtained by typing A\y. Note the use of backslash, which denotes matrix left division. For example, for the equation in (3.16), typing

```
a=[0 1 1 2;1 2 3 4;2 0 2 0];y=[-4;-8;0];
a\y
```

yields `[0 -4 0 0]'`. The solution of $\mathbf{xA} = \mathbf{y}$ can be obtained by typing $\mathbf{y}/\mathbf{A}$. Here we use slash, which denotes matrix right division.

Determinant and inverse of square matrices The rank of a matrix is defined as the number of linearly independent columns or rows. It can also be defined using the determinant. The determinant of a 1×1 matrix is defined as itself. For $n = 2, 3, \ldots$, the determinant of $n \times n$ square matrix $\mathbf{A} = [a_{ij}]$ is defined recursively as, for any chosen j,

$$\det_{i} \mathbf{A} = \sum_{i}^{n} a_{ij} c_{ij} \tag{3.18}$$

where a_{ij} denotes the entry at the ith row and jth column of $\mathbf{A}$. Equation (3.18) is called the *Laplace expansion*. The number c_{ij} is the *cofactor* corresponding to a_{ij} and equals $(-1)^{i+j} \det M_{ij}$, where M_{ij} is the $(n-1) \times (n-1)$ submatrix of $\mathbf{A}$ by deleting its ith row and jth column. If $\mathbf{A}$ is diagonal or triangular, then $\det \mathbf{A}$ equals the product of all diagonal entries.

The determinant of any $r \times r$ submatrix of $\mathbf{A}$ is called a *minor* of order r. Then the rank can be defined as the largest order of all nonzero minors of $\mathbf{A}$. In other words, if $\mathbf{A}$ has rank r, then there is at least one nonzero minor of order r, and every minor of order larger than r is zero. A square matrix is said to be *nonsingular* if its determinant is nonzero. Thus a nonsingular square matrix has full rank and all its columns (rows) are linearly independent.

The *inverse* of a nonsingular square matrix $\mathbf{A} = [a_{ij}]$ is denoted by $\mathbf{A}^{-1}$. The inverse has the property $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ and can be computed as

$$\mathbf{A}^{-1} = \frac{\text{Adj } \mathbf{A}}{\det \mathbf{A}} = \frac{1}{\det \mathbf{A}} [c_{ij}]' \tag{3.19}$$

where c_{ij} is the cofactor. If a matrix is singular, its inverse does not exist. If $\mathbf{A}$ is 2×2, then we have

$$\mathbf{A}^{-1} := \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{3.20}$$

Thus the inverse of a 2×2 matrix is very simple: interchanging diagonal entries and changing the sign of off-diagonal entries (without changing position) and dividing the resulting matrix by the determinant of $\mathbf{A}$. In general, using (3.19) to compute the inverse is complicated. If $\mathbf{A}$ is triangular, it is simpler to compute its inverse by solving $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Note that the inverse of a triangular matrix is again triangular. The MATLAB function `inv` computes the inverse of $\mathbf{A}$.

Theorem 3.3

Consider $\mathbf{Ax} = \mathbf{y}$ with $\mathbf{A}$ square.

1. If $\mathbf{A}$ is nonsingular, then the equation has a unique solution for every $\mathbf{y}$ and the solution equals $\mathbf{A}^{-1}\mathbf{y}$. In particular, the only solution of $\mathbf{Ax} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$.

2. The homogeneous equation $\mathbf{Ax} = \mathbf{0}$ has nonzero solutions if and only if $\mathbf{A}$ is singular. The number of linearly independent solutions equals the nullity of $\mathbf{A}$.

3.4 Similarity Transformation

Consider an $n \times n$ matrix $\mathbf{A}$. It maps $\mathcal{R}^n$ into itself. If we associate with $\mathcal{R}^n$ the orthonormal basis $\{\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_n\}$ in (3.8), then the ith column of $\mathbf{A}$ is the representation of $\mathbf{A}\mathbf{i}_i$ with respect to the orthonormal basis. Now if we select a different set of basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$, then the matrix $\mathbf{A}$ has a different representation $\bar{\mathbf{A}}$. It turns out that *the ith column of $\bar{\mathbf{A}}$ is the representation of $\mathbf{A}\mathbf{q}_i$ with respect to the basis* $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$. This is illustrated by the example that follows.

EXAMPLE 3.4 Consider the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & -1 \\ -2 & 1 & 0 \\ 4 & 3 & 1 \end{bmatrix} \tag{3.21}$$

Let $\mathbf{b} = [0 \ 0 \ 1]'$. Then we have

$$\mathbf{Ab} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{A}^2\mathbf{b} = \mathbf{A}(\mathbf{Ab}) = \begin{bmatrix} -4 \\ 2 \\ -3 \end{bmatrix}, \quad \mathbf{A}^3\mathbf{b} = \mathbf{A}(\mathbf{A}^2\mathbf{b}) = \begin{bmatrix} -5 \\ 10 \\ -13 \end{bmatrix}$$

It can be verified that the following relation holds:

$$\mathbf{A}^3\mathbf{b} = 17\mathbf{b} - 15\mathbf{Ab} + 5\mathbf{A}^2\mathbf{b} \tag{3.22}$$

Because the three vectors $\mathbf{b}$, $\mathbf{Ab}$, and $\mathbf{A}^2\mathbf{b}$ are linearly independent, they can be used as a basis. We now compute the representation of $\mathbf{A}$ with respect to the basis. It is clear that

$$\mathbf{A}(\mathbf{b}) = [\mathbf{b} \ \mathbf{Ab} \ \mathbf{A}^2\mathbf{b}] \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\mathbf{A}(\mathbf{Ab}) = [\mathbf{b} \ \mathbf{Ab} \ \mathbf{A}^2\mathbf{b}] \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\mathbf{A}(\mathbf{A}^2\mathbf{b}) = [\mathbf{b} \ \mathbf{Ab} \ \mathbf{A}^2\mathbf{b}] \begin{bmatrix} 17 \\ -15 \\ 5 \end{bmatrix}$$

where the last equation is obtained from (3.22). Thus the representation of $\mathbf{A}$ with respect to the basis $\{\mathbf{b}, \mathbf{Ab}, \mathbf{A}^2\mathbf{b}\}$ is

$$\bar{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 17 \\ 1 & 0 & -15 \\ 0 & 1 & 5 \end{bmatrix} \tag{3.23}$$

The preceding discussion can be extended to the general case. Let $\mathbf{A}$ be an $n \times n$ matrix. If there exists an $n \times 1$ vector $\mathbf{b}$ such that the n vectors $\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{n-1}\mathbf{b}$ are linearly independent and if

$$\mathbf{A}^n = \beta_1\mathbf{b} + \beta_2\mathbf{Ab} + \cdots + \beta_n\mathbf{A}^{n-1}\mathbf{b}$$

then the representation of $\mathbf{A}$ with respect to the basis $\{\mathbf{b}, \mathbf{Ab}, \ldots, \mathbf{A}^{n-1}\mathbf{b}\}$ is

$$\bar{\mathbf{A}} = \begin{bmatrix} 0 & 0 & \cdots & 0 & \beta_1 \\ 1 & 0 & \cdots & 0 & \beta_2 \\ 0 & 1 & \cdots & 0 & \beta_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \beta_{n-1} \\ 0 & 0 & \cdots & 1 & \beta_n \end{bmatrix} \tag{3.24}$$

This matrix is said to be in a *companion* form.

Consider the equation

$$\mathbf{Ax} = \mathbf{y} \tag{3.25}$$

The square matrix $\mathbf{A}$ maps $\mathbf{x}$ in $\mathcal{R}^n$ into $\mathbf{y}$ in $\mathcal{R}^n$. With respect to the basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$, the equation becomes

$$\bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{y}} \tag{3.26}$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the representations of $\mathbf{x}$ and $\mathbf{y}$ with respect to the basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$. As discussed in (3.7), they are related by

$$\mathbf{x} = \mathbf{Q}\bar{\mathbf{x}} \qquad \mathbf{y} = \mathbf{Q}\bar{\mathbf{y}}$$

with

$$\mathbf{Q} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n] \tag{3.27}$$

an $n \times n$ nonsingular matrix. Substituting these into (3.25) yields

$$\mathbf{AQ}\bar{\mathbf{x}} = \mathbf{Q}\bar{\mathbf{y}} \quad \text{or} \quad \mathbf{Q}^{-1}\mathbf{AQ}\bar{\mathbf{x}} = \bar{\mathbf{y}} \tag{3.28}$$

Comparing this with (3.26) yields

$$\bar{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{AQ} \quad \text{or} \quad \mathbf{A} = \mathbf{Q}\bar{\mathbf{A}}\mathbf{Q}^{-1} \tag{3.29}$$

This is called the *similarity transformation* and $\mathbf{A}$ and $\bar{\mathbf{A}}$ are said to be *similar*. We write (3.29) as

$$\mathbf{AQ} = \mathbf{Q}\bar{\mathbf{A}}$$

or

$$\mathbf{A}[\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n] = [\mathbf{Aq}_1 \ \mathbf{Aq}_2 \ \cdots \ \mathbf{Aq}_n] = [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n]\bar{\mathbf{A}}$$

This shows that the ith column of $\bar{\mathbf{A}}$ is indeed the representation of $\mathbf{Aq}_i$ with respect to the basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$.

3.5 Diagonal Form and Jordan Form

A square matrix $\mathbf{A}$ has different representations with respect to different sets of basis. In this section, we introduce a set of basis so that the representation will be diagonal or block diagonal.

A real or complex number λ is called an *eigenvalue* of the $n \times n$ real matrix $\mathbf{A}$ if there exists a nonzero vector $\mathbf{x}$ such that $\mathbf{Ax} = \lambda\mathbf{x}$. Any nonzero vector $\mathbf{x}$ satisfying $\mathbf{Ax} = \lambda\mathbf{x}$ is called a (right) *eigenvector* of $\mathbf{A}$ associated with eigenvalue λ. In order to find the eigenvalue of $\mathbf{A}$, we write $\mathbf{Ax} = \lambda\mathbf{x} = \lambda\mathbf{Ix}$ as

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0} \tag{3.30}$$

where $\mathbf{I}$ is the unit matrix of order n. This is a homogeneous equation. If the matrix $(\mathbf{A} - \lambda\mathbf{I})$ is nonsingular, then the only solution of (3.30) is $\mathbf{x} = \mathbf{0}$ (Theorem 3.3). Thus in order for (3.30) to have a nonzero solution $\mathbf{x}$, the matrix $(\mathbf{A} - \lambda\mathbf{I})$ must be singular or have determinant 0. We define

$$\Delta(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A})$$

It is a monic polynomial of degree n with real coefficients and is called the *characteristic polynomial* of $\mathbf{A}$. A polynomial is called monic if its leading coefficient is 1. If λ is a root of the characteristic polynomial, then the determinant of $(\mathbf{A} - \lambda\mathbf{I})$ is 0 and (3.30) has at least one nonzero solution. Thus every root of $\Delta(\lambda)$ is an eigenvalue of $\mathbf{A}$. Because $\Delta(\lambda)$ has degree n, the $n \times n$ matrix $\mathbf{A}$ has n eigenvalues (not necessarily all distinct).

We mention that the matrices

$$\begin{bmatrix} 0 & 0 & 0 & -\alpha_4 \\ 1 & 0 & 0 & -\alpha_3 \\ 0 & 1 & 0 & -\alpha_2 \\ 0 & 0 & 1 & -\alpha_1 \end{bmatrix} \qquad \begin{bmatrix} -\alpha_1 & -\alpha_2 & -\alpha_3 & -\alpha_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

and their transposes

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\alpha_4 & -\alpha_3 & -\alpha_2 & -\alpha_1 \end{bmatrix} \qquad \begin{bmatrix} -\alpha_1 & 1 & 0 & 0 \\ -\alpha_2 & 0 & 1 & 0 \\ -\alpha_3 & 0 & 0 & 1 \\ -\alpha_4 & 0 & 0 & 0 \end{bmatrix}$$

all have the following characteristic polynomial:

$$\Delta(\lambda) = \lambda^4 + \alpha_1\lambda^3 + \alpha_2\lambda^2 + \alpha_3\lambda + \alpha_4$$

These matrices can easily be formed from the coefficients of $\Delta(\lambda)$ and are called *companion-form* matrices. The companion-form matrices will arise repeatedly later. The matrix in (3.24) is in such a form.

Eigenvalues of A are all distinct Let λ_i, $i = 1, 2, \ldots, n$, be the eigenvalues of $\mathbf{A}$ and be all distinct. Let $\mathbf{q}_i$ be an eigenvector of $\mathbf{A}$ associated with λ_i; that is, $\mathbf{A}\mathbf{q}_i = \lambda_i\mathbf{q}_i$. Then the set of eigenvectors $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$ is linearly independent and can be used as a basis. Let $\hat{\mathbf{A}}$ be the representation of $\mathbf{A}$ with respect to this basis. Then the first column of $\hat{\mathbf{A}}$ is the representation of $\mathbf{A}\mathbf{q}_1 = \lambda_1\mathbf{q}_1$ with respect to $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$. From

$$\mathbf{A}\mathbf{q}_1 = \lambda_1\mathbf{q}_1 = [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n]\begin{bmatrix}\lambda_1\\0\\0\\\vdots\\0\end{bmatrix}$$

we conclude that the first column of $\hat{\mathbf{A}}$ is $[\lambda_1 \ 0 \ \cdots \ 0]'$. The second column of $\hat{\mathbf{A}}$ is the representation of $\mathbf{A}\mathbf{q}_2 = \lambda_2\mathbf{q}_2$ with respect to $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n\}$, that is, $[0 \ \lambda_1 \ 0 \ \cdots \ 0]'$. Proceeding forward, we can establish

$$\hat{\mathbf{A}} = \begin{bmatrix}\lambda_1 & 0 & 0 & \cdots & 0\\0 & \lambda_2 & 0 & \cdots & 0\\0 & 0 & \lambda_3 & \cdots & 0\\\vdots & \vdots & \vdots & & \vdots\\0 & 0 & 0 & \cdots & \lambda_n\end{bmatrix} \tag{3.31}$$

This is a diagonal matrix. Thus we conclude that every matrix with distinct eigenvalues has a diagonal matrix representation by using its eigenvectors as a basis. Different orderings of eigenvectors will yield different diagonal matrices for the same $\mathbf{A}$.

If we define

$$\mathbf{Q} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n] \tag{3.32}$$

then the matrix $\hat{\mathbf{A}}$ equals

$$\hat{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} \tag{3.33}$$

as derived in (3.29). Computing (3.33) by hand is not simple because of the need to compute the inverse of $\mathbf{Q}$. However, if we know $\hat{\mathbf{A}}$, then we can verify (3.33) by checking $\mathbf{Q}\hat{\mathbf{A}} = \mathbf{A}\mathbf{Q}$.

EXAMPLE 3.5 Consider the matrix

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

Its characteristic polynomial is

$$\Delta(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = \det\begin{bmatrix}\lambda & 0 & 0\\-1 & \lambda & -2\\0 & -1 & \lambda - 1\end{bmatrix}$$
$$= \lambda[\lambda(\lambda - 1) - 2] = (\lambda - 2)(\lambda + 1)\lambda$$

Thus $\mathbf{A}$ has eigenvalues 2, -1, and 0. The eignevector associated with $\lambda = 2$ is any nonzero solution of

$$(\mathbf{A} - 2\mathbf{I})\mathbf{q}_1 = \begin{bmatrix}-2 & 0 & 0\\1 & -2 & 2\\0 & 1 & -1\end{bmatrix}\mathbf{q}_1 = \mathbf{0}$$

Thus $\mathbf{q}_1 = [0 \ 1 \ 1]'$ is an eigenvector associated with $\lambda = 2$. Note that the eigenvector is not unique. $[0 \ \alpha \ \alpha]'$ for any nonzero real α can also be chosen as an eigenvector. The eigenvector associated with $\lambda = -1$ is any nonzero solution of

$$(\mathbf{A} - (-1)\mathbf{I})\mathbf{q}_2 = \begin{bmatrix}1 & 0 & 0\\1 & 1 & 2\\0 & 1 & 2\end{bmatrix}\mathbf{q}_2 = \mathbf{0}$$

which yields $\mathbf{q}_2 = [0 \ -2 \ 1]'$. Similarly, the eigenvector associated with $\lambda = 0$ can be computed as $\mathbf{q}_3 = [2 \ 1 \ -1]'$. Thus the representation of $\mathbf{A}$ with respect to $\{\mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3\}$ is

$$\hat{\mathbf{A}} = \begin{bmatrix}2 & 0 & 0\\0 & -1 & 0\\0 & 0 & 0\end{bmatrix} \tag{3.34}$$

It is a diagonal matrix with eigenvalues on the diagonal. This matrix can also be obtained by computing

$$\hat{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$$

with

$$\mathbf{Q} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \mathbf{q}_3] = \begin{bmatrix}0 & 0 & 2\\1 & -2 & 1\\1 & 1 & -1\end{bmatrix} \tag{3.35}$$

However, it is simpler to verify $\mathbf{Q}\hat{\mathbf{A}} = \mathbf{A}\mathbf{Q}$ or

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

The result in this example can easily be obtained using MATLAB. Typing

```
a=[0 0 0;1 0 2;0 1 1]; [q,d]=eig(a)
```

yields

$$q = \begin{bmatrix}0 & 0 & 0.8186\\0.7071 & 0.8944 & 0.4082\\0.7071 & -0.4472 & -0.4082\end{bmatrix} \qquad d = \begin{bmatrix}2 & 0 & 0\\0 & -1 & 0\\0 & 0 & 0\end{bmatrix}$$

where `d` is the diagonal matrix in (3.34). The matrix `q` is different from the $\mathbf{Q}$ in (3.35); but their corresponding columns differ only by a constant. This is due to nonuniqueness of eigenvectors and every column of `q` is normalized to have norm 1 in MATLAB. If we type `eig(a)` without the left-hand-side argument, then MATLAB generates only the three eigenvalues `2, -1, 0`.

We mention that eigenvalues in MATLAB are *not* computed from the characteristic polynomial. Computing the characteristic polynomial using the Laplace expansion and then computing its roots are not numerically reliable, especially when there are repeated roots. Eigenvalues are computed in MATLAB directly from the matrix by using similarity transformations. Once all eigenvalues are computed, the characteristic polynomial equals $\prod(\lambda - \lambda_i)$. In MATLAB, typing `r=eig(a);poly(r)` yields the characteristic polynomial.

EXAMPLE 3.6 Consider the matrix

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 1 \\ 0 & 4 & -13 \\ 0 & 1 & 0 \end{bmatrix}$$

Its characteristic polynomial is $(\lambda+1)(\lambda^2 - 4\lambda + 13)$. Thus $\mathbf{A}$ has eigenvalues -1, $2 \pm 3j$. Note that complex conjugate eigenvalues must appear in pairs because $\mathbf{A}$ has only real coefficients. The eigenvectors associated with -1 and $2+3j$ are, respectively, $[1 \ 0 \ 0]'$ and $[j \ -3+2j \ j]'$. The eigenvector associated with $\lambda = 2 - 3j$ is $[-j \ -3-2j \ -j]'$, the complex conjugate of the eigenvector associated with $\lambda = 2 + 3j$. Thus we have

$$\mathbf{Q} = \begin{bmatrix} 1 & j & -j \\ 0 & -3+2j & -3-2j \\ 0 & j & j \end{bmatrix} \quad \text{and} \quad \hat{\mathbf{A}} = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2+3j & 0 \\ 0 & 0 & 2-3j \end{bmatrix} \tag{3.36}$$

The MATLAB function `[q,d]=eig(a)` yields

$$q = \begin{bmatrix} 1 & 0.2582j & -0.2582j \\ 0 & -0.7746 + 0.5164j & -0.7746 - 0.5164j \\ 0 & 0.2582j & -0.2582j \end{bmatrix}$$

and

$$d = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 2+3j & 0 \\ 0 & 0 & 2-3j \end{bmatrix}$$

All discussion in the preceding example applies here.

Complex eigenvalues Even though the data we encounter in practice are all real numbers, complex numbers may arise when we compute eigenvalues and eigenvectors. To deal with this problem, we must extend real linear spaces into complex linear spaces and permit all scalars such as α_i in (3.4) to assume complex numbers. To see the reason, we consider

$$\mathbf{A}\mathbf{v} = \begin{bmatrix} 1 & 1+j \\ 1-j & 2 \end{bmatrix} \mathbf{v} = \mathbf{0} \tag{3.37}$$

If we restrict $\mathbf{v}$ to real vectors, then (3.37) has no nonzero solution and the two columns of $\mathbf{A}$ are linearly independent. However, if $\mathbf{v}$ is permitted to assume complex numbers, then $\mathbf{v} = [-2 \ \ 1-j]'$ is a nonzero solution of (3.37). Thus the two columns of $\mathbf{A}$ are linearly dependent and $\mathbf{A}$ has rank 1. This is the rank obtained in MATLAB. Therefore, whenever complex eigenvalues arise, we consider complex linear spaces and complex scalars and

transpose is replaced by complex-conjugate transpose. By so doing, all concepts and results developed for real vectors and matrices can be applied to complex vectors and matrices. Incidentally, the diagonal matrix with complex eigenvalues in (3.36) can be transformed into a very useful real matrix as we will discuss in Section 4.3.1.

Eigenvalues of A are not all distinct An eigenvalue with multiplicity 2 or higher is called a *repeated* eigenvalue. In contrast, an eigenvalue with multiplicity 1 is called a *simple* eigenvalue. If $\mathbf{A}$ has only simple eigenvalues, it always has a diagonal-form representation. If $\mathbf{A}$ has repeated eigenvalues, then it may not have a diagonal form representation. However, it has a block-diagonal and triangular-form representation as we will discuss next.

Consider an $n \times n$ matrix $\mathbf{A}$ with eigenvalue λ and multiplicity n. In other words, $\mathbf{A}$ has only one distinct eigenvalue. To simplify the discussion, we assume $n = 4$. Suppose the matrix $(\mathbf{A} - \lambda\mathbf{I})$ has rank $n - 1 = 3$ or, equivalently, nullity 1; then the equation

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{q} = \mathbf{0}$$

has only one independent solution. Thus $\mathbf{A}$ has only one eigenvector associated with λ. We need $n - 1 = 3$ more linearly independent vectors to form a basis for $\mathcal{R}^4$. The three vectors $\mathbf{q}_2$, $\mathbf{q}_3$, $\mathbf{q}_4$ will be chosen to have the properties $(\mathbf{A} - \lambda\mathbf{I})^2\mathbf{q}_2 = \mathbf{0}$, $(\mathbf{A} - \lambda\mathbf{I})^3\mathbf{q}_3 = \mathbf{0}$, and $(\mathbf{A} - \lambda\mathbf{I})^4\mathbf{q}_4 = \mathbf{0}$.

A vector $\mathbf{v}$ is called a *generalized eigenvector* of grade n if

$$(\mathbf{A} - \lambda\mathbf{I})^n\mathbf{v} = \mathbf{0}$$

and

$$(\mathbf{A} - \lambda\mathbf{I})^{n-1}\mathbf{v} \neq \mathbf{0}$$

If $n = 1$, they reduce to $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ and $\mathbf{v} \neq \mathbf{0}$ and $\mathbf{v}$ is an ordinary eigenvector. For $n = 4$, we define

$$\begin{aligned} \mathbf{v}_4 &:= \mathbf{v} \\ \mathbf{v}_3 &:= (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_4 = (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} \\ \mathbf{v}_2 &:= (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_3 = (\mathbf{A} - \lambda\mathbf{I})^2\mathbf{v} \\ \mathbf{v}_1 &:= (\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_2 = (\mathbf{A} - \lambda\mathbf{I})^3\mathbf{v} \end{aligned}$$

They are called a chain of generalized eigenvectors of length $n = 4$ and have the properties $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v}_1 = \mathbf{0}$, $(\mathbf{A} - \lambda\mathbf{I})^2\mathbf{v}_2 = \mathbf{0}$, $(\mathbf{A} - \lambda\mathbf{I})^3\mathbf{v}_3 = \mathbf{0}$, and $(\mathbf{A} - \lambda\mathbf{I})^4\mathbf{v}_4 = \mathbf{0}$. These vectors, as generated, are automatically linearly independent and can be used as a basis. From these equations, we can readily obtain

$$\begin{aligned} \mathbf{A}\mathbf{v}_1 &= \lambda\mathbf{v}_1 \\ \mathbf{A}\mathbf{v}_2 &= \mathbf{v}_1 + \lambda\mathbf{v}_2 \\ \mathbf{A}\mathbf{v}_3 &= \mathbf{v}_2 + \lambda\mathbf{v}_3 \\ \mathbf{A}\mathbf{v}_4 &= \mathbf{v}_3 + \lambda\mathbf{v}_4 \end{aligned}$$

Then the representation of $\mathbf{A}$ with respect to the basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$ is

$$\mathbf{J} := \begin{bmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{bmatrix} \tag{3.38}$$

We verify this for the first and last columns. The first column of $\mathbf{J}$ is the representation of $\mathbf{A}\mathbf{v}_1 = \lambda\mathbf{v}_1$ with respect to $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$, which is $[\lambda \ 0 \ 0 \ 0]'$. The last column of $\mathbf{J}$ is the representation of $\mathbf{A}\mathbf{v}_4 = \mathbf{v}_3 + \lambda\mathbf{v}_4$ with respect to $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}$, which is $[0 \ 0 \ 1 \ \lambda]'$. This verifies the representation in (3.38). The matrix $\mathbf{J}$ has eigenvalues on the diagonal and 1 on the superdiagonal. If we reverse the order of the basis, then the 1's will appear on the subdiagonal. The matrix is called a *Jordan* block of order $n = 4$.

If $(\mathbf{A} - \lambda\mathbf{I})$ has rank $n - 2$ or, equivalently, nullity 2, then the equation

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{q} = \mathbf{0}$$

has two linearly independent solutions. Thus $\mathbf{A}$ has two linearly independent eigenvectors and we need $(n - 2)$ generalized eigenvectors. In this case, there exist two chains of generalized eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$ and $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_l\}$ with $k + l = n$. If $\mathbf{v}_1$ and $\mathbf{u}_1$ are linearly independent, then the set of n vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{u}_1, \ldots, \mathbf{u}_l\}$ is linearly independent and can be used as a basis. With respect to this basis, the representation of $\mathbf{A}$ is a block diagonal matrix of form

$$\hat{\mathbf{A}} = \text{diag}\{\mathbf{J}_1, \mathbf{J}_2\}$$

where $\mathbf{J}_1$ and $\mathbf{J}_2$ are, respectively, Jordan blocks of order k and l.

Now we discuss a specific example. Consider a 5×5 matrix $\mathbf{A}$ with repeated eigenvalue λ_1 with multiplicity 4 and simple eigenvalue λ_2. Then there exists a nonsingular matrix $\mathbf{Q}$ such that

$$\hat{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$$

assumes one of the following forms

$$\hat{\mathbf{A}}_1 = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix} \quad \hat{\mathbf{A}}_2 = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix}$$

$$\hat{\mathbf{A}}_3 = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix} \quad \hat{\mathbf{A}}_4 = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix}$$

$$\hat{\mathbf{A}}_5 = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix} \tag{3.39}$$

The first matrix occurs when the nullity of $(\mathbf{A} - \lambda_1\mathbf{I})$ is 1. If the nullity is 2, then $\hat{\mathbf{A}}$ has two Jordan blocks associated with λ_1; it may assume the form in $\hat{\mathbf{A}}_2$ or in $\hat{\mathbf{A}}_3$. If $(\mathbf{A} - \lambda_1\mathbf{I})$ has nullity 3, then $\hat{\mathbf{A}}$ has three Jordan blocks associated with λ_1 as shown in $\hat{\mathbf{A}}_4$. Certainly, the positions of the Jordan blocks can be changed by changing the order of the basis. If the nullity is 4, then $\hat{\mathbf{A}}$ is a diagonal matrix as shown in $\hat{\mathbf{A}}_5$. All these matrices are triangular and block diagonal with Jordan blocks on the diagonal; they are said to be in Jordan form. A diagonal matrix is a degenerated Jordan form: its Jordan blocks all have order 1. If $\mathbf{A}$ can be diagonalized, we can use `[q,d]=eig(a)` to generate $\mathbf{Q}$ and $\hat{\mathbf{A}}$ as shown in Examples 3.5 and 3.6. If $\mathbf{A}$ cannot be diagonized, $\mathbf{A}$ is said to be *defective* and `[q,d]=eig(a)` will yield an incorrect solution. In this case, we may use the MATLAB function `[q,d]=jordan(a)`. However, `jordan` will yield a correct result only if $\mathbf{A}$ has integers or ratios of small integers as its entries.

Jordan-form matrices are triangular and block diagonal and can be used to establish many general properties of matrices. For example, because $\det(\mathbf{CD}) = \det\mathbf{C}\det\mathbf{D}$ and $\det\mathbf{Q}\det\mathbf{Q}^{-1} = \det\mathbf{I} = 1$, from $\mathbf{A} = \mathbf{Q}\hat{\mathbf{A}}\mathbf{Q}^{-1}$, we have

$$\det\mathbf{A} = \det\mathbf{Q}\det\hat{\mathbf{A}}\det\mathbf{Q}^{-1} = \det\hat{\mathbf{A}}$$

The determinant of $\hat{\mathbf{A}}$ is the product of all diagonal entries or, equivalently, all eigenvalues of $\mathbf{A}$. Thus we have

$$\det\mathbf{A} = \text{ product of all eigenvalues of } \mathbf{A}$$

which implies that $\mathbf{A}$ *is nonsingular if and only if it has no zero eigenvalue.*

We discuss a useful property of Jordan blocks to conclude this section. Consider the Jordan block in (3.38) with order 4. Then we have

$$(\mathbf{J} - \lambda\mathbf{I}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (\mathbf{J} - \lambda\mathbf{I})^2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$(\mathbf{J} - \lambda\mathbf{I})^3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.40}$$

and $(\mathbf{J} - \lambda\mathbf{I})^k = \mathbf{0}$ for $k \geq 4$. This is called *nilpotent*.

3.6 Functions of a Square Matrix

This section studies functions of a square matrix. We use Jordan form extensively because many properties of functions can almost be visualized in terms of Jordan form. We study first polynomials and then general functions of a square matrix.

Polynomials of a square matrix Let $\mathbf{A}$ be a square matrix. If k is a positive integer, we define

$$\mathbf{A}^k := \mathbf{A}\mathbf{A}\cdots\mathbf{A} \quad (k \text{ terms})$$

and $\mathbf{A}^0 = \mathbf{I}$. Let $f(\lambda)$ be a polynomial such as $f(\lambda) = \lambda^3 + 2\lambda^2 - 6$ or $(\lambda + 2)(4\lambda - 3)$. Then $f(\mathbf{A})$ is defined as

$$f(\mathbf{A}) = \mathbf{A}^3 + 2\mathbf{A}^2 - 6\mathbf{I} \quad \text{or} \quad f(\mathbf{A}) = (\mathbf{A} + 2\mathbf{I})(4\mathbf{A} - 3\mathbf{I})$$

If $\mathbf{A}$ is block diagonal, such as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are square matrices of any order, then it is straightforward to verify

$$\mathbf{A}^k = \begin{bmatrix} \mathbf{A}_1^k & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2^k \end{bmatrix} \quad \text{and} \quad f(\mathbf{A}) = \begin{bmatrix} f(\mathbf{A}_1) & \mathbf{0} \\ \mathbf{0} & f(\mathbf{A}_2) \end{bmatrix} \tag{3.41}$$

Consider the similarity transformation $\hat{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$ or $\mathbf{A} = \mathbf{Q}\hat{\mathbf{A}}\mathbf{Q}^{-1}$. Because

$$\mathbf{A}^k = (\mathbf{Q}\hat{\mathbf{A}}\mathbf{Q}^{-1})(\mathbf{Q}\hat{\mathbf{A}}\mathbf{Q}^{-1})\cdots(\mathbf{Q}\hat{\mathbf{A}}\mathbf{Q}^{-1}) = \mathbf{Q}\hat{\mathbf{A}}^k\mathbf{Q}^{-1}$$

we have

$$f(\mathbf{A}) = \mathbf{Q}f(\hat{\mathbf{A}})\mathbf{Q}^{-1} \quad \text{or} \quad f(\hat{\mathbf{A}}) = \mathbf{Q}^{-1}f(\mathbf{A})\mathbf{Q} \tag{3.42}$$

A *monic* polynomial is a polynomial with 1 as its leading coefficient. The *minimal polynomial* of $\mathbf{A}$ is defined as the monic polynomial $\psi(\lambda)$ of least degree such that $\psi(\mathbf{A}) = \mathbf{0}$. Note that the $\mathbf{0}$ is a zero matrix of the same order as $\mathbf{A}$. A direct consequence of (3.42) is that $f(\mathbf{A}) = \mathbf{0}$ if and only if $f(\hat{\mathbf{A}}) = \mathbf{0}$. Thus $\mathbf{A}$ and $\hat{\mathbf{A}}$ or, more general, all similar matrices have the same minimal polynomial. Computing the minimal polynomial directly from $\mathbf{A}$ is not simple (see Problem 3.25); however, if the Jordan-form representation of $\mathbf{A}$ is available, the minimal polynomial can be read out by inspection.

Let λ_i be an eigenvalue of $\mathbf{A}$ with multiplicity n_i. That is, the characteristic polynomial of $\mathbf{A}$ is

$$\Delta(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = \prod_i (\lambda - \lambda_i)^{n_i}$$

Suppose the Jordan form of $\mathbf{A}$ is known. Associated with each eigenvalue, there may be one or more Jordan blocks. The *index* of λ_i, denoted by $\bar{n}_i$, is defined as the largest order of all Jordan blocks associated with λ_i. Clearly we have $\bar{n}_i \le n_i$. For example, the multiplicities of λ_1 in all five matrices in (3.39) are 4; their indices are, respectively, 4, 3, 2, 2, and 1. The multiplicities and indices of λ_2 in all five matrices in (3.39) are all 1. Using the indices of all eigenvalues, the minimal polynomial can be expressed as

$$\psi(\lambda) = \prod_i (\lambda - \lambda_i)^{\bar{n}_i}$$

with degree $\bar{n} = \sum \bar{n}_i \le \sum n_i = n =$ dimension of $\mathbf{A}$. For example, the minimal polynomials of the five matrices in (3.39) are

$$\psi_1 = (\lambda - \lambda_1)^4(\lambda - \lambda_2) \qquad \psi_2 = (\lambda - \lambda_1)^3(\lambda - \lambda_2)$$
$$\psi_3 = (\lambda - \lambda_1)^2(\lambda - \lambda_2) \qquad \psi_4 = (\lambda - \lambda_1)^2(\lambda - \lambda_2)$$
$$\psi_5 = (\lambda - \lambda_1)(\lambda - \lambda_2)$$

Their characteristic polynomials, however, all equal

$$\Delta(\lambda) = (\lambda - \lambda_1)^4(\lambda - \lambda_2)$$

We see that the minimal polynomial is a factor of the characteristic polynomial and has a degree less than or equal to the degree of the characteristic polynomial. Clearly, if all eigenvalues of $\mathbf{A}$ are distinct, then the minimal polynomial equals the characteristic polynomial.

Using the nilpotent property in (3.40), we can show that

$$\psi(\mathbf{A}) = \mathbf{0}$$

and that no polynomial of lesser degree meets the condition. Thus $\psi(\lambda)$ as defined is the minimal polynomial.

Theorem 3.4 (Cayley-Hamilton theorem)

Let

$$\Delta(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}) = \lambda^n + \alpha_1\lambda^{n-1} + \cdots + \alpha_{n-1}\lambda + \alpha_n$$

be the characteristic polynomial of $\mathbf{A}$. Then

$$\Delta(\mathbf{A}) = \mathbf{A}^n + \alpha_1\mathbf{A}^{n-1} + \cdots + \alpha_{n-1}\mathbf{A} + \alpha_n\mathbf{I} = \mathbf{0} \tag{3.43}$$

In words, a matrix satisfies its own characteristic polynomial. Because $n_i \ge \bar{n}_i$, the characteristic polynomial contains the minimal polynomial as a factor or $\Delta(\lambda) = \psi(\lambda)h(\lambda)$ for some polynomial $h(\lambda)$. Because $\psi(\mathbf{A}) = \mathbf{0}$, we have $\Delta(\mathbf{A}) = \psi(\mathbf{A})h(\mathbf{A}) = \mathbf{0} \cdot h(\mathbf{A}) = \mathbf{0}$. This establishes the theorem. The Cayley–Hamilton theorem implies that $\mathbf{A}^n$ can be written as a linear combination of $\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{n-1}\}$. Multiplying (3.43) by $\mathbf{A}$ yields

$$\mathbf{A}^{n+1} + \alpha_1\mathbf{A}^n + \cdots + \alpha_{n-1}\mathbf{A}2 + \alpha_n\mathbf{A} = \mathbf{0} \cdot \mathbf{A} = \mathbf{0}$$

which implies that $\mathbf{A}^{n+1}$ can be written as a linear combination of $\{\mathbf{A}, \mathbf{A}^2, \ldots, \mathbf{A}^n\}$, which, in turn, can be written as a linear combination of $\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{n-1}\}$. Proceeding forward, we conclude that, for any polynomial $f(\lambda)$, no matter how large its degree is, $f(\mathbf{A})$ can always be expressed as

$$f(\mathbf{A}) = \beta_0\mathbf{I} + \beta_1\mathbf{A} + \cdots + \beta_{n-1}\mathbf{A}^{n-1} \tag{3.44}$$

for some β_i. In other words, every polynomial of an $n \times n$ matrix $\mathbf{A}$ can be expressed as a linear combination of $\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{n-1}\}$. If the minimal polynomial of $\mathbf{A}$ with degree $\bar{n}$ is available, then every polynomial of $\mathbf{A}$ can be expressed as a linear combination of $\{\mathbf{I}, \mathbf{A}, \ldots, \mathbf{A}^{\bar{n}-1}\}$. This is a better result. However, because $\bar{n}$ may not be available, we discuss in the following only (3.44) with the understanding that all discussion applies to $\bar{n}$.

One way to compute (3.44) is to use long division to express $f(\lambda)$ as

$$f(\lambda) = q(\lambda)\Delta(\lambda) + h(\lambda) \tag{3.45}$$

where $q(\lambda)$ is the quotient and $h(\lambda)$ is the remainder with degree less than n. Then we have

$$f(\mathbf{A}) = q(\mathbf{A})\Delta(\mathbf{A}) + h(\mathbf{A}) = q(\mathbf{A})\mathbf{0} + h(\mathbf{A}) = h(\mathbf{A})$$

Long division is not convenient to carry out if the degree of $f(\lambda)$ is much larger than the degree of $\Delta(\lambda)$. In this case, we may solve $h(\lambda)$ directly from (3.45). Let

$$h(\lambda) := \beta_0 + \beta_1\lambda + \cdots + \beta_{n-1}\lambda^{n-1}$$

where the n unknowns β_i are to be solved. If all n eigenvalues of $\mathbf{A}$ are distinct, these β_i can be solved from the n equations

$$f(\lambda_i) = q(\lambda_i)\Delta(\lambda_i) + h(\lambda_i) = h(\lambda_i)$$

for $i = 1, 2, \ldots, n$. If $\mathbf{A}$ has repeated eigenvalues, then (3.45) must be differentiated to yield additional equations. This is stated as a theorem.

Theorem 3.5

We are given $f(\lambda)$ and an $n \times n$ matrix $\mathbf{A}$ with characteristic polynomial

$$\Delta(\lambda) = \prod_{i=1}^{m}(\lambda - \lambda_i)^{n_i}$$

where $n = \sum_{i=1}^{m} n_i$. Define

$$h(\lambda) := \beta_0 + \beta_1\lambda + \cdots + \beta_{n-1}\lambda^{n-1}$$

It is a polynomial of degree $n - 1$ with n unknown coefficients. These n unknowns are to be solved from the following set of n equations:

$$f^{(l)}(\lambda_i) = h^{(l)}(\lambda_i) \quad \text{for } l = 0, 1, \ldots, n_i - 1 \quad \text{and} \quad i = 1, 2, \ldots, m$$

where

$$f^{(l)}(\lambda_i) := \left.\frac{d^l f(\lambda)}{d\lambda^l}\right|_{\lambda=\lambda_i}$$

and $h^{(l)}(\lambda_i)$ is similarly defined. Then we have

$$f(\mathbf{A}) = h(\mathbf{A})$$

and $h(\lambda)$ is said to equal $f(\lambda)$ on the spectrum of $\mathbf{A}$.

Example 3.7 Compute $\mathbf{A}^{100}$ with

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

In other words, given $f(\lambda) = \lambda^{100}$, compute $f(\mathbf{A})$. The characteristic polynomial of $\mathbf{A}$ is $\Delta(\lambda) = \lambda^2 + 2\lambda + 1 = (\lambda + 1)^2$. Let $h(\lambda) = \beta_0 + \beta_1\lambda$. On the spectrum of $\mathbf{A}$, we have

$$f(-1) = h(-1): \qquad (-1)^{100} = \beta_0 - \beta_1$$
$$f'(-1) = h'(-1): \qquad 100 \cdot (-1)^{99} = \beta_1$$

Thus we have $\beta_1 = -100$, $\beta_0 = 1 + \beta_1 = -99$, $h(\lambda) = -99 - 100\lambda$, and

$$\mathbf{A}^{100} = \beta_0\mathbf{I} + \beta_1\mathbf{A} = -99\mathbf{I} - 100\mathbf{A}$$
$$= -99\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 100\begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} = \begin{bmatrix} -199 & -100 \\ 100 & 101 \end{bmatrix}$$

Clearly $\mathbf{A}^{100}$ can also be obtained by multiplying $\mathbf{A}$ 100 times. However, it is simpler to use Theorem 3.5.

Functions of a square matrix Let $f(\lambda)$ be any function, not necessarily a polynomial. One way to define $f(\mathbf{A})$ is to use Theorem 3.5. Let $h(\lambda)$ be a polynomial of degree $n - 1$, where n is the order of $\mathbf{A}$. We solve the coefficients of $h(\lambda)$ by equating $f(\lambda) = h(\lambda)$ on the spectrum of $\mathbf{A}$. Then $f(\mathbf{A})$ is defined as $h(\mathbf{A})$.

Example 3.8 Let

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

Compute $e^{\mathbf{A}_1 t}$. Or, equivalently, if $f(\lambda) = e^{\lambda t}$, what is $f(\mathbf{A}_1)$?

The characteristic polynomial of $\mathbf{A}_1$ is $(\lambda - 1)^2(\lambda - 2)$. Let $h(\lambda) = \beta_0 + \beta_1\lambda + \beta_2\lambda^2$. Then

$$f(1) = h(1): \qquad e^t = \beta_0 + \beta_1 + \beta_2$$
$$f'(1) = h'(1): \qquad te^t = \beta_1 + 2\beta_2$$
$$f(2) = h(2): \qquad e^{2t} = \beta_0 + 2\beta_1 + 4\beta_2$$

Note that, in the second equation, the differentiation is with respect to λ, not t. Solving these equations yields $\beta_0 = -2te^t + e^{2t}$, $\beta_1 = 3te^t + 2e^t - 2e^{2t}$, and $\beta_2 = e^{2t} - e^t - te^t$. Thus we have

$$e^{\mathbf{A}_1 t} = h(\mathbf{A}_1) = (-2te^t + e^{2t})\mathbf{I} + (3te^t + 2e^t - 2e^{2t})\mathbf{A}_1$$
$$+ (e^{2t} - e^t - te^t)\mathbf{A}_1^2 = \begin{bmatrix} 2e^t - e^{2t} & 0 & 2e^t - 2e^{2t} \\ 0 & e^t & 0 \\ e^{2t} - e^t & 0 & 2e^{2t} - e^t \end{bmatrix}$$

Example 3.9 Let

$$\mathbf{A}_2 = \begin{bmatrix} 0 & 2 & -2 \\ 0 & 1 & 0 \\ 1 & -1 & 3 \end{bmatrix}$$

Compute $e^{\mathbf{A}_2 t}$. The characteristic polynomial of $\mathbf{A}_2$ is $(\lambda - 1)^2(\lambda - 2)$, which is the same as for $\mathbf{A}_1$. Hence we have the same $h(\lambda)$ as in Example 3.8. Consequently, we have

$$e^{\mathbf{A}_2 t} = h(\mathbf{A}_2) = \begin{bmatrix} 2e^t - e^{2t} & 2te^t & 2e^t - 2e^{2t} \\ 0 & e^t & 0 \\ e^{2t} - e^t & -te^t & 2e^{2t} - e^t \end{bmatrix}$$

EXAMPLE 3.10 Consider the Jordan block of order 4:

$$\hat{\mathbf{A}} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 1 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix} \tag{3.46}$$

Its characteristic polynomial is $(\lambda-\lambda_1)^4$. Although we can select $h(\lambda)$ as $\beta_0+\beta_1\lambda+\beta_2\lambda+\beta_3\lambda^3$, it is computationally simpler to select $h(\lambda)$ as

$$h(\lambda) = \beta_0 + \beta_1(\lambda - \lambda_1) + \beta_2(\lambda - \lambda_1) + \beta_3(\lambda - \lambda_1)^3$$

This selection is permitted because $h(\lambda)$ has degree $(n - 1) = 3$ and $n = 4$ independent unknowns. The condition $f(\lambda) = h(\lambda)$ on the spectrum of $\hat{\mathbf{A}}$ yields immediately

$$\beta_0 = f(\lambda_1), \quad \beta_1 = f'(\lambda_1), \quad \beta_2 = \frac{f''(\lambda_1)}{2!}, \quad \beta_3 = \frac{f^{(3)}(\lambda_1)}{3!}$$

Thus we have

$$f(\hat{\mathbf{A}}) = f(\lambda_1)\mathbf{I} + \frac{f'(\lambda_1)}{1!}(\hat{\mathbf{A}} - \lambda_1\mathbf{I}) + \frac{f''(\lambda_1)}{2!}(\hat{\mathbf{A}} - \lambda_1\mathbf{I})^2 + \frac{f^{(3)}(\lambda_1)}{3!}(\hat{\mathbf{A}} - \lambda_1\mathbf{I})^3$$

Using the special forms of $(\hat{\mathbf{A}} - \lambda_1\mathbf{I})^k$ as discussed in (3.40), we can readily obtain

$$f(\hat{\mathbf{A}}) = \begin{bmatrix} f(\lambda_1) & f'(\lambda_1)/1! & f''(\lambda_1)/2! & f^{(3)}(\lambda_1)/3! \\ 0 & f(\lambda_1) & f'(\lambda_1)/1! & f''(\lambda_1)/2! \\ 0 & 0 & f(\lambda_1) & f'(\lambda_1)/1! \\ 0 & 0 & 0 & f(\lambda_1) \end{bmatrix} \tag{3.47}$$

If $f(\lambda) = e^{\lambda t}$, then

$$e^{\hat{\mathbf{A}}t} = \begin{bmatrix} e^{\lambda_1 t} & te^{\lambda_1 t} & t^2e^{\lambda_1 t}/2! & t^3e^{\lambda_1 t}/3! \\ 0 & e^{\lambda_1 t} & te^{\lambda_1 t} & t^2e^{\lambda_1 t}/2! \\ 0 & 0 & e^{\lambda_1 t} & te^{\lambda_1 t} \\ 0 & 0 & 0 & e^{\lambda_1 t} \end{bmatrix} \tag{3.48}$$

Because functions of $\mathbf{A}$ are defined through polynomials of $\mathbf{A}$, Equations (3.41) and (3.42) are applicable to functions.

EXAMPLE 3.11 Consider

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix}$$

It is block diagonal and contains two Jordan blocks. If $f(\lambda) = e^{\mathbf{A}t}$, then (3.41) and (3.48) imply

$$e^{\mathbf{A}t} = \begin{bmatrix} e^{\lambda_1 t} & te^{\lambda_1 t} & t^2e^{\lambda_1 t}/2! & 0 & 0 \\ 0 & e^{\lambda_1 t} & te^{\lambda_1 t} & 0 & 0 \\ 0 & 0 & e^{\lambda_1 t} & 0 & 0 \\ 0 & 0 & 0 & e^{\lambda_2 t} & te^{\lambda_2 t} \\ 0 & 0 & 0 & 0 & e^{\lambda_2 t} \end{bmatrix}$$

If $f(\lambda) = (s - \lambda)^{-1}$, then (3.41) and (3.47) imply

$$(s\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} \frac{1}{(s-\lambda_1)} & \frac{1}{(s-\lambda_1)^2} & \frac{1}{(s-\lambda_1)^3} & 0 & 0 \\ 0 & \frac{1}{(s-\lambda_1)} & \frac{1}{(s-\lambda_1)^2} & 0 & 0 \\ 0 & 0 & \frac{1}{(s-\lambda_1)} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{(s-\lambda_2)} & \frac{1}{(s-\lambda_2)^2} \\ 0 & 0 & 0 & 0 & \frac{1}{(s-\lambda_2)} \end{bmatrix} \tag{3.49}$$

Using power series The function of $\mathbf{A}$ was defined using a polynomial of finite degree. We now give an alternative definition by using an infinite power series. Suppose $f(\lambda)$ can be expressed as the power series

$$f(\lambda) = \sum_{i=0}^{\infty} \beta_i \lambda^i$$

with the radius of convergence ρ. If all eigenvalues of $\mathbf{A}$ have magnitudes less than ρ, then $f(\mathbf{A})$ can be defined as

$$f(\mathbf{A}) = \sum_{i=0}^{\infty} \beta_i \mathbf{A}^i \tag{3.50}$$

Instead of proving the equivalence of this definition and the definition based on Theorem 3.5, we use (3.50) to derive (3.47).

EXAMPLE 3.12 Consider the Jordan-form matrix $\hat{\mathbf{A}}$ in (3.46). Let

$$f(\lambda) = f(\lambda_1) + f'(\lambda_1)(\lambda - \lambda_1) + \frac{f''(\lambda_1)}{2!}(\lambda - \lambda_1)^2 + \cdots$$

then

$$f(\hat{\mathbf{A}}) = f(\lambda_1)\mathbf{I} + f'(\lambda_1)(\hat{\mathbf{A}} - \lambda_1\mathbf{I}) + \cdots + \frac{f^{(n-1)}(\lambda_1)}{(n-1)!}(\hat{\mathbf{A}} - \lambda_1\mathbf{I})^{n-1} + \cdots$$

Because $(\hat{\mathbf{A}} - \lambda_1\mathbf{I})^k = \mathbf{0}$ for $k \geq n = 4$ as discussed in (3.40), the infinite series reduces immediately to (3.47). Thus the two definitions lead to the same function of a matrix.

The most important function of $\mathbf{A}$ is the exponential function $e^{\mathbf{A}t}$. Because the Taylor series

$$e^{\lambda t} = 1 + \lambda t + \frac{\lambda^2 t^2}{2!} + \cdots + \frac{\lambda^n t^n}{n!} + \cdots$$

converges for all finite λ and t, we have

$$e^{\mathbf{A}t} = \mathbf{I} + t\mathbf{A} + \frac{t^2}{2!}\mathbf{A}^2 + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!} t^k \mathbf{A}^k \tag{3.51}$$

This series involves only multiplications and additions and may converge rapidly; therefore it is suitable for computer computation. We list in the following the program in MATLAB that computes (3.51) for $t = 1$:

```
Function E=expm2(A)
E=zeros(size(A));
F=eye(size(A));
k=1;
while norm(E+F-E,1)>0
      E=E+F;
      F=A*F/k;
      k=k+1;
end
```

In the program, E denotes the partial sum and F is the next term to be added to E. The first line defines the function. The next two lines initialize E and F. Let c_k denote the kth term of (3.51) with $t = 1$. Then we have $c_{k+1} = (\mathbf{A}/k)c_k$ for $k = 1, 2, \ldots$. Thus we have F = A * F/k. The computation stops if the 1-norm of E + F − E, denoted by norm(E + F − E, 1), is rounded to 0 in computers. Because the algorithm compares F and E, not F and 0, the algorithm uses norm(E + F − E, 1) instead of norm(F,1). Note that norm(a,1) is the 1-norm discussed in Section 3.2 and will be discussed again in Section 3.9. We see that the series can indeed be programmed easily. To improve the computed result, the techniques of scaling and squaring can be used. In MATLAB, the function `expm2` uses (3.51). The function `expm` or `expm1`, however, uses the so-called Padé approximation. It yields comparable results as `expm2` but requires only about half the computing time. Thus `expm` is preferred to `expm2`. The function `expm3` uses Jordan form, but it will yield an incorrect solution if a matrix is not diagonalizable. If a closed-form solution of $e^{\mathbf{A}t}$ is needed, we must use Theorem 3.5 or Jordan form to compute $e^{\mathbf{A}t}$.

We derive some important properties of $e^{\mathbf{A}t}$ to conclude this section. Using (3.51), we can readily verify the next two equalities

$$e^{\mathbf{0}} = \mathbf{I} \tag{3.52}$$

$$e^{\mathbf{A}(t_1+t_2)} = e^{\mathbf{A}t_1}e^{\mathbf{A}t_2} \tag{3.53}$$

$$[e^{\mathbf{A}t}]^{-1} = e^{-\mathbf{A}t} \tag{3.54}$$

To show (3.54), we set $t_2 = -t_1$. Then (3.53) and (3.52) imply

$$e^{\mathbf{A}t_1}e^{-\mathbf{A}t_1} = e^{\mathbf{A}\cdot 0} = e^{\mathbf{0}} = \mathbf{I}$$

which implies (3.54). Thus the inverse of $e^{\mathbf{A}t}$ can be obtained by simply changing the sign of t. Differentiating term by term of (3.51) yields

$$\frac{d}{dt}e^{\mathbf{A}t} = \sum_{k=1}^{\infty} \frac{1}{(k-1)!} t^{k-1}\mathbf{A}^k$$

$$= \mathbf{A}\left(\sum_{k=0}^{\infty} \frac{1}{k!} t^k \mathbf{A}^k\right) = \left(\sum_{k=0}^{\infty} \frac{1}{k!} t^k \mathbf{A}^k\right)\mathbf{A}$$

Thus we have

$$\frac{d}{dt}e^{\mathbf{A}t} = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A} \tag{3.55}$$

This is an important equation. We mention that

$$e^{(\mathbf{A}+\mathbf{B})t} \neq e^{\mathbf{A}t}e^{\mathbf{B}t} \tag{3.56}$$

The equality holds only if $\mathbf{A}$ and $\mathbf{B}$ commute or $\mathbf{AB} = \mathbf{BA}$. This can be verified by direct substitution of (3.51).

The Laplace transform of a function $f(t)$ is defined as

$$\hat{f}(s) := \mathcal{L}[f(t)] = \int_0^{\infty} f(t)e^{-st}dt$$

It can be shown that

$$\mathcal{L}\left[\frac{t^k}{k!}\right] = s^{-(k+1)}$$

Taking the Laplace transform of (3.51) yields

$$\mathcal{L}[e^{\mathbf{A}t}] = \sum_{k=0}^{\infty} s^{-(k+1)}\mathbf{A}^k = s^{-1}\sum_{k=0}^{\infty}(s^{-1}\mathbf{A})^k$$

Because the infinite series

$$\sum_{k=0}^{\infty}(s^{-1}\lambda)^k = 1 + s^{-1}\lambda + s^{-2}\lambda^2 + \cdots = (1 - s^{-1}\lambda)^{-1}$$

converges for $|s^{-1}\lambda| < 1$, we have

$$s^{-1}\sum_{k=0}^{\infty}(s^{-1}\mathbf{A})^k = s^{-1}\mathbf{I} + s^{-2}\mathbf{A} + s^{-3}\mathbf{A}^2 + \cdots$$

$$= s^{-1}(\mathbf{I} - s^{-1}\mathbf{A})^{-1} = \left[s(\mathbf{I} - s^{-1}\mathbf{A})\right]^{-1} = (s\mathbf{I} - \mathbf{A})^{-1} \tag{3.57}$$

and

$$\mathcal{L}[e^{\mathbf{A}t}] = (s\mathbf{I} - \mathbf{A})^{-1} \tag{3.58}$$

Although in the derivation of (3.57) we require s to be sufficiently large so that all eigenvalues of $s^{-1}\mathbf{A}$ have magnitudes less than 1, Equation (3.58) actually holds for all s except at the eigenvalues of $\mathbf{A}$. Equation (3.58) can also be established from (3.55). Because $\mathcal{L}[df(t)/dt] = s\mathcal{L}[f(t)] - f(0)$, applying the Laplace transform to (3.55) yields

$$s\mathcal{L}[e^{\mathbf{A}t}] - e^{\mathbf{0}} = \mathbf{A}\mathcal{L}[e^{\mathbf{A}t}]$$

or

$$(s\mathbf{I} - \mathbf{A})\mathcal{L}[e^{\mathbf{A}t}] = e^{\mathbf{0}} = \mathbf{I}$$

which implies (3.58). ∎

3.7 Lyapunov Equation

Consider the equation

$$\mathbf{AM} + \mathbf{MB} = \mathbf{C} \tag{3.59}$$

where $\mathbf{A}$ and $\mathbf{B}$ are, respectively, $n \times n$ and $m \times m$ constant matrices. In order for the equation to be meaningful, the matrices $\mathbf{M}$ and $\mathbf{C}$ must be of order $n \times m$. The equation is called the *Lyapunov* equation.

The equation can be written as a set of standard linear algebraic equations. To see this, we assume $n = 3$ and $m = 2$ and write (3.59) explicitly as

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \\ m_{31} & m_{32} \end{bmatrix} + \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \\ m_{31} & m_{32} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$
$$= \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \end{bmatrix}$$

Multiplying them out and then equating the corresponding entries on both sides of the equality, we obtain

$$\begin{bmatrix} a_{11}+b_{11} & a_{12} & a_{13} & b_{21} & 0 & 0 \\ a_{21} & a_{22}+b_{11} & a_{23} & 0 & b_{21} & 0 \\ a_{31} & a_{32} & a_{33}+b_{11} & 0 & 0 & b_{21} \\ b_{12} & 0 & 0 & a_{11}+b_{22} & a_{12} & a_{13} \\ 0 & b_{12} & 0 & a_{21} & a_{22}+b_{22} & a_{23} \\ 0 & 0 & b_{12} & a_{31} & a_{32} & a_{33}+b_{22} \end{bmatrix}$$

$$\times \begin{bmatrix} m_{11} \\ m_{21} \\ m_{31} \\ m_{12} \\ m_{22} \\ m_{32} \end{bmatrix} = \begin{bmatrix} c_{11} \\ c_{21} \\ c_{31} \\ c_{12} \\ c_{22} \\ c_{32} \end{bmatrix} \tag{3.60}$$

This is indeed a standard linear algebraic equation. The matrix on the preceding page is a square matrix of order $n \times m = 3 \times 2 = 6$.

Let us define $\mathcal{A}(\mathbf{M}) := \mathbf{AM} + \mathbf{MB}$. Then the Lyapunov equation can be written as $\mathcal{A}(\mathbf{M}) = \mathbf{C}$. It maps an nm-dimensional linear space into itself. A scalar η is called an eigenvalue of $\mathcal{A}$ if there exists a nonzero $\mathbf{M}$ such that

$$\mathcal{A}(\mathbf{M}) = \eta\mathbf{M}$$

Because $\mathcal{A}$ can be considered as a square matrix of order nm, it has nm eigenvalues η_k, for $k = 1, 2, \ldots, nm$. It turns out

$$\eta_k = \lambda_i + \mu_j \qquad \text{for } i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m$$

where λ_i, $i = 1, 2, \ldots, n$, and μ_j, $j = 1, 2, \ldots, m$, are, respectively, the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$. In other words, the eigenvalues of $\mathcal{A}$ are all possible sums of the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$.

We show intuitively why this is the case. Let $\mathbf{u}$ be an $n \times 1$ right eigenvector of $\mathbf{A}$ associated with λ_i; that is, $\mathbf{Au} = \lambda_i\mathbf{u}$. Let $\mathbf{v}$ be a $1 \times m$ left eigenvector of $\mathbf{B}$ associated with μ_j; that is, $\mathbf{vB} = \mathbf{v}\mu_j$. Applying $\mathcal{A}$ to the $n \times m$ matrix $\mathbf{uv}$ yields

$$\mathcal{A}(\mathbf{uv}) = \mathbf{Auv} + \mathbf{uvB} = \lambda_i\mathbf{uv} + \mathbf{uv}\mu_j = (\lambda_i + \mu_j)\mathbf{uv}$$

Because both $\mathbf{u}$ and $\mathbf{v}$ are nonzero, so is the matrix $\mathbf{uv}$. Thus $(\lambda_i + \mu_j)$ is an eigenvalue of $\mathcal{A}$.

The determinant of a square matrix is the product of all its eigenvalues. Thus a matrix is nonsingular if and only if it has no zero eigenvalue. If there are no i and j such that $\lambda_i + \mu_j = 0$, then the square matrix in (3.60) is nonsingular and, for every $\mathbf{C}$, there exists a unique $\mathbf{M}$ satisfying the equation. In this case, the Lyapunov equation is said to be nonsingular. If $\lambda_i + \mu_j = 0$ for some i and j, then for a given $\mathbf{C}$, solutions may or may not exist. If $\mathbf{C}$ lies in the range space of $\mathcal{A}$, then solutions exist and are not unique. See Problem 3.32.

The MATLAB function `m=lyap(a,b,-c)` computes the solution of the Lyapunov equation in (3.59).

3.8 Some Useful Formulas

This section discusses some formulas that will be needed later. Let $\mathbf{A}$ and $\mathbf{B}$ be $m \times n$ and $n \times p$ constant matrices. Then we have

$$\rho(\mathbf{AB}) \leq \min(\rho(\mathbf{A}), \rho(\mathbf{B})) \tag{3.61}$$

where ρ denotes the rank. This can be argued as follows. Let $\rho(\mathbf{B}) = \alpha$. Then $\mathbf{B}$ has α linearly independent rows. In $\mathbf{AB}$, $\mathbf{A}$ operates on the rows of $\mathbf{B}$. Thus the rows of $\mathbf{AB}$ are

linear combinations of the rows of $\mathbf{B}$. Thus $\mathbf{AB}$ has at most α linearly independent rows. In $\mathbf{AB}$, $\mathbf{B}$ operates on the columns of $\mathbf{A}$. Thus if $\mathbf{A}$ has β linearly independent columns, then $\mathbf{AB}$ has at most β linearly independent columns. This establishes (3.61). Consequently, if $\mathbf{A} = \mathbf{B}_1\mathbf{B}_2\mathbf{B}_3\cdots$, then the rank of $\mathbf{A}$ is equal to or smaller than the smallest rank of $\mathbf{B}_i$.

Let $\mathbf{A}$ be $m \times n$ and let $\mathbf{C}$ and $\mathbf{D}$ be any $n \times n$ and $m \times m$ nonsingular matrices. Then we have

$$\rho(\mathbf{AC}) = \rho(\mathbf{A}) = \rho(\mathbf{DA}) \tag{3.62}$$

In words, the rank of a matrix will not change after pre- or postmultiplying by a nonsingular matrix. To show (3.62), we define

$$\mathbf{P} := \mathbf{AC} \tag{3.63}$$

Because $\rho(\mathbf{A}) \leq \min(m, n)$ and $\rho(\mathbf{C}) = n$, we have $\rho(\mathbf{A}) \leq \rho(\mathbf{C})$. Thus (3.61) implies

$$\rho(\mathbf{P}) \leq \min(\rho(\mathbf{A}), \rho(\mathbf{C})) \leq \rho(\mathbf{A})$$

Next we write (3.63) as $\mathbf{A} = \mathbf{PC}^{-1}$. Using the same argument, we have $\rho(\mathbf{A}) \leq \rho(\mathbf{P})$. Thus we conclude $\rho(\mathbf{P}) = \rho(\mathbf{A})$. A consequence of (3.62) is that the rank of a matrix will not change by elementary operations. Elementary operations are (1) multiplying a row or a column by a nonzero number, (2) interchanging two rows or two columns, and (3) adding the product of one row (column) and a number to another row (column). These operations are the same as multiplying nonsingular matrices. See Reference [6, p. 542].

Let $\mathbf{A}$ be $m \times n$ and $\mathbf{B}$ be $n \times m$. Then we have

$$\det(\mathbf{I}_m + \mathbf{AB}) = \det(\mathbf{I}_n + \mathbf{BA}) \tag{3.64}$$

where $\mathbf{I}_m$ is the unit matrix of order m. To show (3.64), let us define

$$\mathbf{N} = \begin{bmatrix} \mathbf{I}_m & \mathbf{A} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} \mathbf{I}_m & \mathbf{0} \\ -\mathbf{B} & \mathbf{I}_n \end{bmatrix} \qquad \mathbf{P} = \begin{bmatrix} \mathbf{I}_m & -\mathbf{A} \\ \mathbf{B} & \mathbf{I}_n \end{bmatrix}$$

We compute

$$\mathbf{NP} = \begin{bmatrix} \mathbf{I}_m + \mathbf{AB} & \mathbf{0} \\ \mathbf{B} & \mathbf{I}_n \end{bmatrix}$$

and

$$\mathbf{QP} = \begin{bmatrix} \mathbf{I}_m & -\mathbf{A} \\ \mathbf{0} & \mathbf{I}_n + \mathbf{BA} \end{bmatrix}$$

Because $\mathbf{N}$ and $\mathbf{Q}$ are block triangular, their determinants equal the products of the determinant of their block-diagonal matrices or

$$\det \mathbf{N} = \det \mathbf{I}_m \cdot \det \mathbf{I}_n = 1 = \det \mathbf{Q}$$

Likewise, we have

$$\det(\mathbf{NP}) = \det(\mathbf{I}_m + \mathbf{AB}) \qquad \det(\mathbf{QP}) = \det(\mathbf{I}_n + \mathbf{BA})$$

Because

$$\det(\mathbf{NP}) = \det \mathbf{N} \det \mathbf{P} = \det \mathbf{P}$$

and

$$\det(\mathbf{QP}) = \det \mathbf{Q} \det \mathbf{P} = \det \mathbf{P}$$

we conclude $\det(\mathbf{I}_m + \mathbf{AB}) = \det(\mathbf{I}_n + \mathbf{BA})$.

In $\mathbf{N}$, $\mathbf{Q}$, and $\mathbf{P}$, if $\mathbf{I}_n$, $\mathbf{I}_m$, and $\mathbf{B}$ are replaced, respectively, by $\sqrt{s}\mathbf{I}_n$, $\sqrt{s}\mathbf{I}_m$, and $-\mathbf{B}$, then we can readily obtain

$$s^n \det(s\mathbf{I}_m - \mathbf{AB}) = s^m \det(s\mathbf{I}_n - \mathbf{BA}) \tag{3.65}$$

which implies, for $n = m$ or for $n \times n$ square matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\det(s\mathbf{I}_n - \mathbf{AB}) = \det(s\mathbf{I}_n - \mathbf{BA}) \tag{3.66}$$

They are useful formulas.

3.9 Quadratic Form and Positive Definiteness

An $n \times n$ real matrix $\mathbf{M}$ is said to be *symmetric* if its transpose equals itself. The scalar function $\mathbf{x}'\mathbf{Mx}$, where $\mathbf{x}$ is an $n \times 1$ real vector and $\mathbf{M}' = \mathbf{M}$, is called a *quadratic form*. We show that all eigenvalues of symmetric $\mathbf{M}$ are real.

The eigenvalues and eigenvectors of real matrices can be complex as shown in Example 3.6. Therefore we must allow $\mathbf{x}$ to assume complex numbers for the time being and consider the scalar function $\mathbf{x}^*\mathbf{Mx}$, where $\mathbf{x}^*$ is the complex conjugate transpose of $\mathbf{x}$. Taking the complex conjugate transpose of $\mathbf{x}^*\mathbf{Mx}$ yields

$$(\mathbf{x}^*\mathbf{Mx})^* = \mathbf{x}^*\mathbf{M}^*\mathbf{x} = \mathbf{x}^*\mathbf{M}'\mathbf{x} = \mathbf{x}^*\mathbf{Mx}$$

where we have used the fact that the complex conjugate transpose of a real $\mathbf{M}$ reduces to simply the transpose. Thus $\mathbf{x}^*\mathbf{Mx}$ is real for any complex $\mathbf{x}$. This assertion is not true if $\mathbf{M}$ is not symmetric. Let λ be an eigenvalue of $\mathbf{M}$ and $\mathbf{v}$ be its eigenvector; that is, $\mathbf{Mv} = \lambda\mathbf{v}$. Because

$$\mathbf{v}^*\mathbf{Mv} = \mathbf{v}^*\lambda\mathbf{v} = \lambda(\mathbf{v}^*\mathbf{v})$$

and because both $\mathbf{v}^*\mathbf{Mv}$ and $\mathbf{v}^*\mathbf{v}$ are real, the eigenvalue λ must be real. This shows that all eigenvalues of symmetric $\mathbf{M}$ are real. After establishing this fact, we can return our study to exclusively real vector $\mathbf{x}$.

We claim that every symmetric matrix can be diagonalized using a similarity transformation even it has repeated eigenvalue λ. To show this, we show that there is no generalized eigenvector of grade 2 or higher. Suppose $\mathbf{x}$ is a generalized eigenvector of grade 2 or

$$(\mathbf{M} - \lambda\mathbf{I})^2\mathbf{x} = \mathbf{0} \tag{3.67}$$

$$(\mathbf{M} - \lambda\mathbf{I})\mathbf{x} \neq \mathbf{0} \tag{3.68}$$

Consider

$$[(\mathbf{M}-\lambda\mathbf{I})\mathbf{x}]'(\mathbf{M}-\lambda\mathbf{I})\mathbf{x} = \mathbf{x}'(\mathbf{M}'-\lambda\mathbf{I}')(\mathbf{M}-\lambda\mathbf{I})\mathbf{x} = \mathbf{x}'(\mathbf{M}-\lambda\mathbf{I})^2\mathbf{x}$$

which is nonzero according to (3.68) but is zero according to (3.67). This is a contradiction. Therefore the Jordan form of $\mathbf{M}$ has no Jordan block of order 2. Similarly, we can show that the Jordan form of $\mathbf{M}$ has no Jordan block of order 3 or higher. Thus we conclude that there exists a nonsingular $\mathbf{Q}$ such that

$$\mathbf{M} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} \tag{3.69}$$

where $\mathbf{D}$ is a diagonal matrix with real eigenvalues of $\mathbf{M}$ on the diagonal.

A square matrix $\mathbf{A}$ is called an *orthogonal matrix* if all columns of $\mathbf{A}$ are orthonormal. Clearly $\mathbf{A}$ is nonsingular and we have

$$\mathbf{A}'\mathbf{A} = \mathbf{I} \quad \text{and} \quad \mathbf{A}^{-1} = \mathbf{A}'$$

which imply $\mathbf{A}\mathbf{A}' = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}'\mathbf{A}$. Thus the inverse of an orthogonal matrix equals its transpose. Consider (3.69). Because $\mathbf{D}' = \mathbf{D}$ and $\mathbf{M}' = \mathbf{M}$, (3.69) equals its own transpose or

$$\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1} = [\mathbf{Q}\mathbf{D}\mathbf{Q}^{-1}]' = [\mathbf{Q}^{-1}]'\mathbf{D}\mathbf{Q}'$$

which implies $\mathbf{Q}^{-1} = \mathbf{Q}'$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}$. Thus $\mathbf{Q}$ is an orthogonal matrix; its columns are orthonormalized eigenvectors of $\mathbf{M}$. This is summarized as a theorem.

Theorem 3.6

For every real symmetric matrix $\mathbf{M}$, there exists an orthogonal matrix $\mathbf{Q}$ such that

$$\mathbf{M} = \mathbf{Q}\mathbf{D}\mathbf{Q}' \quad \text{or} \quad \mathbf{D} = \mathbf{Q}'\mathbf{M}\mathbf{Q}$$

where $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{M}$, which are all real, on the diagonal.

A symmetric matrix $\mathbf{M}$ is said to be *positive definite*, denoted by $\mathbf{M} > 0$, if $\mathbf{x}'\mathbf{M}\mathbf{x} > 0$ for every nonzero $\mathbf{x}$. It is *positive semidefinite*, denoted by $\mathbf{M} \geq 0$, if $\mathbf{x}'\mathbf{M}\mathbf{x} \geq 0$ for every nonzero $\mathbf{x}$. If $\mathbf{M} > 0$, then $\mathbf{x}'\mathbf{M}\mathbf{x} = 0$ if and only if $\mathbf{x} = \mathbf{0}$. If $\mathbf{M}$ is positive semidefinite, then there exists a nonzero $\mathbf{x}$ such that $\mathbf{x}'\mathbf{M}\mathbf{x} = 0$. This property will be used repeatedly later.

Theorem 3.7

A symmetric $n \times n$ matrix $\mathbf{M}$ is positive definite (positive semidefinite) if and only if any one of the following conditions holds.

1. Every eigenvalue of $\mathbf{M}$ is positive (zero or positive).
2. All the *leading* principal minors of $\mathbf{M}$ are positive (all the principal minors of $\mathbf{M}$ are zero or positive).
3. There exists an $n \times n$ nonsingular matrix $\mathbf{N}$ (an $n \times n$ singular matrix $\mathbf{N}$ or an $m \times n$ matrix $\mathbf{N}$ with $m < n$) such that $\mathbf{M} = \mathbf{N}'\mathbf{N}$.

Condition (1) can readily be proved by using Theorem 3.6. Next we consider Conditon (3). If $\mathbf{M} = \mathbf{N}'\mathbf{N}$, then

$$\mathbf{x}'\mathbf{M}\mathbf{x} = \mathbf{x}'\mathbf{N}'\mathbf{N}\mathbf{x} = (\mathbf{N}\mathbf{x})'(\mathbf{N}\mathbf{x}) = ||\mathbf{N}\mathbf{x}||_2^2 \geq 0$$

for any $\mathbf{x}$. If $\mathbf{N}$ is nonsingular, the only $\mathbf{x}$ to make $\mathbf{N}\mathbf{x} = \mathbf{0}$ is $\mathbf{x} = \mathbf{0}$. Thus $\mathbf{M}$ is positive definite. If $\mathbf{N}$ is singular, there exists a nonzero $\mathbf{x}$ to make $\mathbf{N}\mathbf{x} = \mathbf{0}$. Thus $\mathbf{M}$ is positive semidefinite. For a proof of Condition (2), see Reference [10].

We use an example to illustrate the principal minors and leading principal minors. Consider

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}$$

Its principal minors are m_{11}, m_{22}, m_{33},

$$\det\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}, \quad \det\begin{bmatrix} m_{11} & m_{13} \\ m_{31} & m_{33} \end{bmatrix}, \quad \det\begin{bmatrix} m_{22} & m_{23} \\ m_{32} & m_{33} \end{bmatrix}$$

and $\det \mathbf{M}$. Thus the principal minors are the determinants of all submatrices of $\mathbf{M}$ whose diagonals coincide with the diagonal of $\mathbf{M}$. The leading principal minors of $\mathbf{M}$ are

$$m_{11}, \quad \det\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}, \quad \text{and} \quad \det \mathbf{M}$$

Thus the leading principal minors of $\mathbf{M}$ are the determinants of the submatrices of $\mathbf{M}$ obtained by deleting the last k columns and last k rows for $k = 2, 1, 0$.

Theorem 3.8

1. An $m \times n$ matrix $\mathbf{H}$, with $m \geq n$, has rank n, if and only if the $n \times n$ matrix $\mathbf{H}'\mathbf{H}$ has rank n or $\det(\mathbf{H}'\mathbf{H}) \neq 0$.

2. An $m \times n$ matrix $\mathbf{H}$, with $m \leq n$, has rank m, if and only if the $m \times m$ matrix $\mathbf{H}\mathbf{H}'$ has rank m or $\det(\mathbf{H}\mathbf{H}') \neq 0$.

The symmetric matrix $\mathbf{H}'\mathbf{H}$ is always positive semidefinite. It becomes positive definite if $\mathbf{H}'\mathbf{H}$ is nonsingular. We give a proof of this theorem. The argument in the proof will be used to establish the main results in Chapter 6; therefore the proof is spelled out in detail.

Proof: *Necessity*: The condition $\rho(\mathbf{H}'\mathbf{H}) = n$ implies $\rho(\mathbf{H}) = n$. We show this by contradiction. Suppose $\rho(\mathbf{H}'\mathbf{H}) = n$ but $\rho(\mathbf{H}) < n$. Then there exists a nonzero vector $\mathbf{v}$ such that $\mathbf{H}\mathbf{v} = \mathbf{0}$, which implies $\mathbf{H}'\mathbf{H}\mathbf{v} = \mathbf{0}$. This contradicts $\rho(\mathbf{H}'\mathbf{H}) = n$. Thus $\rho(\mathbf{H}'\mathbf{H}) = n$ implies $\rho(\mathbf{H}) = n$.

Sufficiency: The condition $\rho(\mathbf{H}) = n$ implies $\rho(\mathbf{H}'\mathbf{H}) = n$. Suppose not, or $\rho(\mathbf{H}'\mathbf{H}) < n$; then there exists a nonzero vector $\mathbf{v}$ such that $\mathbf{H}'\mathbf{H}\mathbf{v} = \mathbf{0}$, which implies $\mathbf{v}'\mathbf{H}'\mathbf{H}\mathbf{v} = 0$ or

$$0 = \mathbf{v}'\mathbf{H}'\mathbf{H}\mathbf{v} = (\mathbf{H}\mathbf{v})'(\mathbf{H}\mathbf{v}) = ||\mathbf{H}\mathbf{v}||_2^2$$

Thus we have $\mathbf{H}\mathbf{v} = \mathbf{0}$. This contradicts the hypotheses that $\mathbf{v} \neq 0$ and $\rho(\mathbf{H}) = n$. Thus $\rho(\mathbf{H}) =$ implies $\rho(\mathbf{H}'\mathbf{H}) = n$. This establishes the first part of Theorem 3.8. The second part can be established similarly. Q.E.D.

We discuss the relationship between the eigenvalues of $\mathbf{H}'\mathbf{H}$ and those of $\mathbf{H}\mathbf{H}'$. Because both $\mathbf{H}'\mathbf{H}$ and $\mathbf{H}\mathbf{H}'$ are symmetric and positive semidefinite, their eigenvalues are real and nonnegative (zero or

positive). If $\mathbf{H}$ is $m \times n$, then $\mathbf{H}'\mathbf{H}$ has n eigenvalues and $\mathbf{H}\mathbf{H}'$ has m eigenvalues. Let $\mathbf{A} = \mathbf{H}$ and $\mathbf{B} = \mathbf{H}'$. Then (3.65) becomes

$$\det(s\mathbf{I}_m - \mathbf{H}\mathbf{H}') = s^{m-n} \det(s\mathbf{I}_n - \mathbf{H}'\mathbf{H}) \tag{3.70}$$

This implies that the characteristic polynomials of $\mathbf{H}\mathbf{H}'$ and $\mathbf{H}'\mathbf{H}$ differ only by s^{m-n}. Thus we conclude that $\mathbf{H}\mathbf{H}'$ and $\mathbf{H}'\mathbf{H}$ have the same nonzero eigenvalues but may have different numbers of zero eigenvalues. Furthermore, they have at most $\bar{n} := \min(m, n)$ number of nonzero eigenvalues.

3.10 Singular-Value Decomposition

Let $\mathbf{H}$ be an $m \times n$ real matrix. Define $\mathbf{M} := \mathbf{H}'\mathbf{H}$. Clearly $\mathbf{M}$ is $n \times n$, symmetric, and semidefinite. Thus all eigenvalues of $\mathbf{M}$ are real and nonnegative (zero or positive). Let r be the number of its positive eigenvalues. Then the eigenvalues of $\mathbf{M} = \mathbf{H}'\mathbf{H}$ can be arranged as

$$\lambda_1^2 \geq \lambda_2^2 \geq \cdots \lambda_r^2 > 0 = \lambda_{r+1} = \cdots = \lambda_n$$

Let $\bar{n} := \min(m, n)$. Then the set

$$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_{\bar{n}}$$

is called the *singular values* of $\mathbf{H}$. The singular values are usually arranged in descending order in magnitude.

EXAMPLE 3.13 Consider the 2×3 matrix

$$\mathbf{H} = \begin{bmatrix} -4 & -1 & 2 \\ 2 & 0.5 & -1 \end{bmatrix}$$

We compute

$$\mathbf{M} = \mathbf{H}'\mathbf{H} = \begin{bmatrix} 20 & 5 & -10 \\ 5 & 1.25 & -2.5 \\ -10 & -2.5 & 5 \end{bmatrix}$$

and compute its characteristic polynomial as

$$\det(\lambda\mathbf{I} - \mathbf{M}) = \lambda^3 - 26.25\lambda^2 = \lambda^2(\lambda - 26.25)$$

Thus the eigenvalues of $\mathbf{H}'\mathbf{H}$ are 26.25, 0, and 0, and the singular values of $\mathbf{H}$ are $\sqrt{26.25} = 5.1235$ and 0. Note that the number of singular values equals $\min(n, m)$.

In view of (3.70), we can also compute the singular values of $\mathbf{H}$ from the eigenvalues of $\mathbf{H}\mathbf{H}'$. Indeed, we have

$$\bar{\mathbf{M}} := \mathbf{H}\mathbf{H}' = \begin{bmatrix} 21 & -10.5 \\ -10.5 & 5.25 \end{bmatrix}$$

and

$$\det(\lambda\mathbf{I} - \bar{\mathbf{M}}) = \lambda^2 - 26.25\lambda = \lambda(\lambda - 26.25)$$

Thus the eigenvalues of $\mathbf{H}\mathbf{H}'$ are 26.25 and 0 and the singular values of $\mathbf{H}'$ are 5.1235 and 0. We see that the eigenvalues of $\mathbf{H}'\mathbf{H}$ differ from those of $\mathbf{H}\mathbf{H}'$ only in the number of zero eigenvalues and the singular values of $\mathbf{H}$ equal the singular values of $\mathbf{H}'$.

For $\mathbf{M} = \mathbf{H}'\mathbf{H}$, there exists, following Theorem 3.6, an orthogonal matrix $\mathbf{Q}$ such that

$$\mathbf{Q}'\mathbf{H}'\mathbf{H}\mathbf{Q} = \mathbf{D} =: \mathbf{S}'\mathbf{S} \tag{3.71}$$

where $\mathbf{D}$ is an $n \times n$ diagonal matrix with λ_i^2 on the diagonal. The matrix $\mathbf{S}$ is $m \times n$ with the singular values λ_i on the diagonal. Manipulation on (3.71) will lead eventially to the theorem that follows.

Theorem 3.9 (Singular-value decomposition)

Every $m \times n$ matrix $\mathbf{H}$ can be transformed into the form

$$\mathbf{H} = \mathbf{R}\mathbf{S}\mathbf{Q}'$$

with $\mathbf{R}'\mathbf{R} = \mathbf{R}\mathbf{R}' = \mathbf{I}_m$, $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}_n$, and $\mathbf{S}$ being $m \times n$ with the singular values of $\mathbf{H}$ on the diagonal.

The columns of $\mathbf{Q}$ are orthonormalized eigenvectors of $\mathbf{H}'\mathbf{H}$ and the columns of $\mathbf{R}$ are orthonormalized eigenvectors of $\mathbf{H}\mathbf{H}'$. Once $\mathbf{R}$, $\mathbf{S}$, and $\mathbf{Q}$ are computed, the rank of $\mathbf{H}$ equals the number of nonzero singular values. If the rank of $\mathbf{H}$ is r, the first r columns of $\mathbf{R}$ are an orthonormal basis of the range space of $\mathbf{H}$. The last $(n - r)$ columns of $\mathbf{Q}$ are an orthonormal basis of the null space of $\mathbf{H}$. Although computing singular-value decomposition is time consuming, it is very reliable and gives a quantitative measure of the rank. Thus it is used in MATLAB to compute the rank, range space, and null space. In MATLAB, the singular values of $\mathbf{H}$ can be obtained by typing `s=svd(H)`. Typing `[R,S,Q]=svd(H)` yields the three matrices in the theorem. Typing `orth(H)` and `null(H)` yields, respectively, orthonormal bases of the range space and null space of $\mathbf{H}$. The function `null` will be used repeatedly in Chapter 7.

EXAMPLE 3.14 Consider the matrix in (3.11). We type

```
a=[0 1 1 2;1 2 3 4;2 0 2 0];
[r,s,q]=svd(a)
```

which yield

$$r = \begin{bmatrix} 0.3782 & -0.3084 & 0.8729 \\ 0.8877 & -0.1468 & -0.4364 \\ 0.2627 & 0.9399 & 0.2182 \end{bmatrix} \quad s = \begin{bmatrix} 6.1568 & 0 & 0 & 0 \\ 0 & 2.4686 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$q = \begin{bmatrix} 0.2295 & 0.7020 & 0.3434 & -0.5802 \\ 0.3498 & -0.2439 & 0.8384 & 0.3395 \\ 0.5793 & 0.4581 & -0.3434 & 0.5802 \\ 0.6996 & -0.4877 & -0.2475 & -0.4598 \end{bmatrix}$$

Thus the singular values of the matrix $\mathbf{A}$ in (3.11) are 6.1568, 2.4686, and 0. The matrix has two nonzero singular values, thus its rank is 2 and, consequently, its nullity is $4 - \rho(\mathbf{A}) = 2$. The first two columns of r are the orthonormal basis in (3.13) and the last two columns of q are the orthonormal basis in (3.14).

3.11 Norms of Matrices

The concept of norms for vectors can be extended to matrices. This concept is needed in Chapter 5. Let $\mathbf{A}$ be an $m \times n$ matrix. The norm of $\mathbf{A}$ can be defined as

$$||\mathbf{A}|| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{||\mathbf{Ax}||}{||\mathbf{x}||} = \sup_{||\mathbf{x}||=1} ||\mathbf{Ax}|| \tag{3.72}$$

where sup stands for supremum or the least upper bound. This norm is defined through the norm of $\mathbf{x}$ and is therefore called an *induced norm*. For different $||\mathbf{x}||$, we have different $||\mathbf{A}||$. For example, if the 1-norm $||\mathbf{x}||_1$ is used, then

$$||\mathbf{A}||_1 = \max_j \left(\sum_{i=1}^{m} |a_{ij}| \right) = \text{largest column absolute sum}$$

where a_{ij} is the ijth element of $\mathbf{A}$. If the Euclidean norm $||\mathbf{x}||_2$ is used, then

$$\begin{aligned} ||\mathbf{A}||_2 &= \text{largest singular value of } \mathbf{A} \\ &= (\text{largest eigenvalue of } \mathbf{A}'\mathbf{A})^{1/2} \end{aligned}$$

If the infinite-norm $||\mathbf{x}||_\infty$ is used, then

$$||\mathbf{A}||_\infty = \max_i \left(\sum_{j=1}^{n} |a_{ij}| \right) = \text{largest row absolute sum}$$

These norms are all different for the same $\mathbf{A}$. For example, if

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

then $||\mathbf{A}||_1 = 3 + |-1| = 4$, $||\mathbf{A}||_2 = 3.7$, and $||\mathbf{A}||_\infty = 3 + 2 = 5$, as shown in Fig. 3.3. The MATLAB functions `norm(a,1)`, `norm(a,2)` =`norm(a)`, and `norm(a,inf)` compute the three norms.

The norm of matrices has the following properties:

$$||\mathbf{Ax}|| \leq ||\mathbf{A}||||\mathbf{x}||$$
$$||\mathbf{A} + \mathbf{B}|| \leq ||\mathbf{A}|| + ||\mathbf{B}||$$
$$||\mathbf{AB}|| \leq ||\mathbf{A}||||\mathbf{B}||$$

PROBLEMS

The reader should try first to solve all problems involving numerical numbers by hand and then verify the results using MATLAB or any software.

3.1 Consider Fig. 3.1. What is the representation of the vector $\mathbf{x}$ with respect to the basis $\{\mathbf{q}_1, \mathbf{i}_2\}$? What is the representation of $\mathbf{q}_1$ with respect to $\{\mathbf{i}_2, \mathbf{q}_2\}$?

3.2 What are the 1-norm, 2-norm, and infinite-norm of the vectors

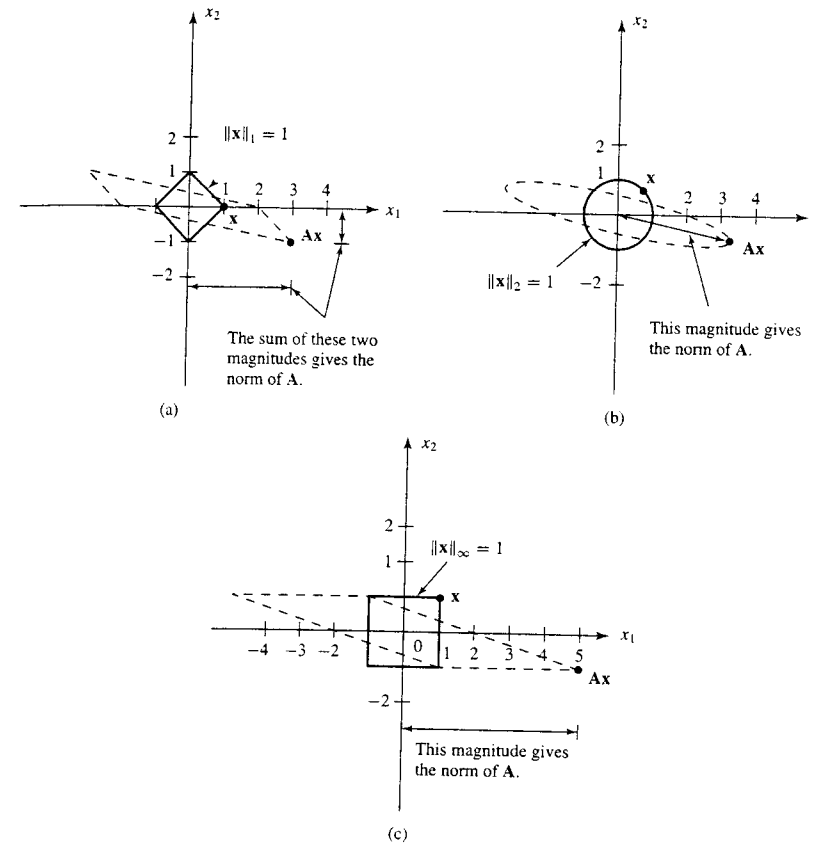


Figure 3.3 Different norms of $\mathbf{A}$.

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix} \qquad \mathbf{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

3.3 Find two orthonormal vectors that span the same space as the two vectors in Problem 3.2.

3.4 Consider an $n \times m$ matrix $\mathbf{A}$ with $n \geq m$. If all columns of $\mathbf{A}$ are orthonormal, then $\mathbf{A}'\mathbf{A} = \mathbf{I}_m$. What can you say about $\mathbf{AA}'$?

3.5 Find the ranks and nullities of the following matrices:

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{A}_2 = \begin{bmatrix} 4 & 1 & -1 \\ 3 & 2 & 0 \\ 1 & 1 & 0 \end{bmatrix} \qquad \mathbf{A}_3 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -1 & -2 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

3.6 Find bases of the range spaces and null spaces of the matrices in Problem 3.5.

3.7 Consider the linear algebraic equation

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

It has three equations and two unknowns. Does a solution $\mathbf{x}$ exist in the equation? Is the solution unique? Does a solution exist if $\mathbf{y} = [1 \ \ 1 \ \ 1]'$?

3.8 Find the general solution of

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

How many parameters do you have?

3.9 Find the solution in Example 3.3 that has the smallest Euclidean norm.

3.10 Find the solution in Problem 3.8 that has the smallest Euclidean norm.

3.11 Consider the equation

$$\mathbf{x}[n] = \mathbf{A}^n x[0] + \mathbf{A}^{n-1}\mathbf{b}u[0] + \mathbf{A}^{n-2}\mathbf{b}u[1] + \cdots + \mathbf{A}\mathbf{b}u[n-2] + \mathbf{b}u[n-1]$$

where $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{b}$ is an $n \times 1$ column vector. Under what conditions on $\mathbf{A}$ and $\mathbf{b}$ will there exist $u[0], u[1], \ldots, u[n-1]$ to meet the equation for any $\mathbf{x}[n]$ and $\mathbf{x}[0]$? [*Hint:* Write the equation in the form

$$\mathbf{x}[n] - \mathbf{A}^n\mathbf{x}[0] = [\mathbf{b} \ \ \mathbf{A}\mathbf{b} \ \ \cdots \ \ \mathbf{A}^{n-1}\mathbf{b}] \begin{bmatrix} u[n-1] \\ u[n-2] \\ \vdots \\ u[0] \end{bmatrix}]$$

3.12 Given

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \qquad \bar{\mathbf{b}} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 1 \end{bmatrix}$$

what are the representations of $\mathbf{A}$ with respect to the basis $\{\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \mathbf{A}^3\mathbf{b}\}$ and the basis $\{\bar{\mathbf{b}}, \mathbf{A}\bar{\mathbf{b}}, \mathbf{A}^2\bar{\mathbf{b}}, \mathbf{A}^3\bar{\mathbf{b}}\}$, respectively? (Note that the representations are the same!)

3.13 Find Jordan-form representations of the following matrices:

$$\mathbf{A}_1 = \begin{bmatrix} 1 & 4 & 10 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad \mathbf{A}_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -2 & -4 & -3 \end{bmatrix}$$

$$\mathbf{A}_3 = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \qquad \mathbf{A}_4 = \begin{bmatrix} 0 & 4 & 3 \\ 0 & 20 & 16 \\ 0 & -25 & -20 \end{bmatrix}$$

Note that all except $\mathbf{A}_4$ can be diagonalized.

3.14 Consider the companion-form matrix

$$\mathbf{A} = \begin{bmatrix} -\alpha_1 & -\alpha_2 & -\alpha_3 & -\alpha_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

Show that its characteristic polynomial is given by

$$\Delta(\lambda) = \lambda^4 + \alpha_1\lambda^3 + \alpha_2\lambda^2 + \alpha_3\lambda + \alpha_4$$

Show also that if λ_i is an eigenvalue of $\mathbf{A}$ or a solution of $\Delta(\lambda) = 0$, then $[\lambda_i^3 \ \ \lambda_i^2 \ \ \lambda_i \ \ 1]'$ is an eigenvector of $\mathbf{A}$ associated with λ_i.

3.15 Show that the *Vandermonde* determinant

$$\begin{bmatrix} \lambda_1^3 & \lambda_2^3 & \lambda_3^3 & \lambda_4^3 \\ \lambda_1^2 & \lambda_2^2 & \lambda_3^2 & \lambda_4^2 \\ \lambda_1 & \lambda_2 & \lambda_3 & \lambda_4 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

equals $\prod_{1 \le i < j \le 4}(\lambda_j - \lambda_i)$. Thus we conclude that the matrix is nonsingular or, equivalently, the eigenvectors are linearly independent if all eigenvalues are distinct.

3.16 Show that the companion-form matrix in Problem 3.14 is nonsingular if and only if $\alpha_4 \neq 0$. Under this assumption, show that its inverse equals

$$\mathbf{A}^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -1/\alpha_4 & -\alpha_1/\alpha_4 & -\alpha_2/\alpha_4 & -\alpha_3/\alpha_4 \end{bmatrix}$$

3.17 Consider

$$\mathbf{A} = \begin{bmatrix} \lambda & \lambda T & \lambda T^2/2 \\ 0 & \lambda & \lambda T \\ 0 & 0 & \lambda \end{bmatrix}$$

with $\lambda \neq 0$ and $T > 0$. Show that $[0 \ \ 0 \ \ 1]'$ is a generalized eigenvector of grade 3 and the three columns of

$$\mathbf{Q} = \begin{bmatrix} \lambda^2 T^2 & \lambda T^2 & 0 \\ 0 & \lambda T & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

constitute a chain of generalized eigenvectors of length 3. Verify

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}$$

3.18 Find the characteristic polynomials and the minimal polynomials of the following matrices:

$$\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_2 \end{bmatrix} \quad \begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}$$

$$\begin{bmatrix} \lambda_1 & 1 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix} \quad \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & \lambda_1 \end{bmatrix}$$

3.19 Show that if λ is an eigenvalue of $\mathbf{A}$ with eigenvector $\mathbf{x}$, then $f(\lambda)$ is an eigenvalue of $f(\mathbf{A})$ with the same eigenvector $\mathbf{x}$.

3.20 Show that an $n \times n$ matrix has the property $\mathbf{A}^k = \mathbf{0}$ for $k \geq m$ if and only if $\mathbf{A}$ has eigenvalues 0 with multiplicity n and index m or less. Such a matrix is called a *nilpotent* matrix.

3.21 Given

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

find $\mathbf{A}^{10}$, $\mathbf{A}^{103}$, and $e^{\mathbf{A}t}$.

3.22 Use two different methods to compute $e^{\mathbf{A}t}$ for $\mathbf{A}_1$ and $\mathbf{A}_4$ in Problem 3.13.

3.23 Show that functions of the same matrix commute; that is,

$$f(\mathbf{A})g(\mathbf{A}) = g(\mathbf{A})f(\mathbf{A})$$

Consequently we have $\mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}$.

3.24 Let

$$\mathbf{C} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

Find a matrix $\mathbf{B}$ such that $e^{\mathbf{B}} = \mathbf{C}$. Show that if $\lambda_i = 0$ for some i, then $\mathbf{B}$ does not exist. Let

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

Find a $\mathbf{B}$ such that $e^{\mathbf{B}} = \mathbf{C}$. Is it true that, for any nonsingular $\mathbf{C}$, there exists a matrix $\mathbf{B}$ such that $e^{\mathbf{B}} = \mathbf{C}$?

3.25 Let

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{\Delta(s)}\text{Adj}\,(s\mathbf{I} - \mathbf{A})$$

and let $m(s)$ be the monic greatest common divisor of all entries of Adj $(s\mathbf{I} - \mathbf{A})$. Verify for the matrix $\mathbf{A}_3$ in Problem 3.13 that the minimal polynominal of $\mathbf{A}$ equals $\Delta(s)/m(s)$.

3.26 Define

$$(s\mathbf{I} - \mathbf{A})^{-1} := \frac{1}{\Delta(s)}\left[\mathbf{R}_0 s^{n-1} + \mathbf{R}_1 s^{n-2} + \cdots + \mathbf{R}_{n-2}s + \mathbf{R}_{n-1}\right]$$

where

$$\Delta(s) := \det(s\mathbf{I} - \mathbf{A}) := s^n + \alpha_1 s^{n-1} + \alpha_2 s^{n-2} + \cdots + \alpha_n$$

and $\mathbf{R}_i$ are constant matrices. This definition is valid because the degree in s of the adjoint of $(s\mathbf{I} - \mathbf{A})$ is at most $n - 1$. Verify

$$\alpha_1 = -\frac{\text{tr}(\mathbf{A}\mathbf{R}_0)}{1} \qquad \mathbf{R}_0 = \mathbf{I}$$

$$\alpha_2 = -\frac{\text{tr}(\mathbf{A}\mathbf{R}_1)}{2} \qquad \mathbf{R}_1 = \mathbf{A}\mathbf{R}_0 + \alpha_1\mathbf{I} = \mathbf{A} + \alpha_1\mathbf{I}$$

$$\alpha_3 = -\frac{\text{tr}(\mathbf{A}\mathbf{R}_2)}{3} \qquad \mathbf{R}_2 = \mathbf{A}\mathbf{R}_1 + \alpha_2\mathbf{I} = \mathbf{A}^2 + \alpha_1\mathbf{A} + \alpha_2\mathbf{I}$$

$$\vdots$$

$$\alpha_{n-1} = -\frac{\text{tr}(\mathbf{A}\mathbf{R}_{n-2})}{n-1} \qquad \mathbf{R}_{n-1} = \mathbf{A}\mathbf{R}_{n-2} + \alpha_{n-1}\mathbf{I} = \mathbf{A}^{n-1} + \alpha_1\mathbf{A}^{n-2} + \cdots + \alpha_{n-2}\mathbf{A} + \alpha_{n-1}\mathbf{I}$$

$$\alpha_n = -\frac{\text{tr}(\mathbf{A}\mathbf{R}_{n-1})}{n} \qquad \mathbf{0} = \mathbf{A}\mathbf{R}_{n-1} + \alpha_n\mathbf{I}$$

where tr stands for the *trace* of a matrix and is defined as the sum of all its diagonal entries. This process of computing α_i and $\mathbf{R}_i$ is called the *Leverrier algorithm*.

3.27 Use Problem 3.26 to prove the Cayley–Hamilton theorem.

3.28 Use Problem 3.26 to show

$$(s\mathbf{I} - \mathbf{A})^{-1} = \frac{1}{\Delta(s)}\left[\mathbf{A}^{n-1} + (s + \alpha_1)\mathbf{A}^{n-2} + (s^2 + \alpha_1 s + \alpha_2)\mathbf{A}^{n-3}\right.$$

$$+\cdots+(s^{n-1}+\alpha_1 s^{n-2}+\cdots+\alpha_{n-1})\mathbf{I}]$$

3.29 Let all eigenvalues of $\mathbf{A}$ be distinct and let $\mathbf{q}_i$ be a right eigenvector of $\mathbf{A}$ associated with λ_i; that is, $\mathbf{A}\mathbf{q}_i = \lambda_i \mathbf{q}_i$. Define $\mathbf{Q} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n]$ and define

$$\mathbf{P} := \mathbf{Q}^{-1} =: \begin{bmatrix} \mathbf{p}_1 \\ \mathbf{p}_2 \\ \vdots \\ \mathbf{p}_n \end{bmatrix}$$

where $\mathbf{p}_i$ is the ith row of $\mathbf{P}$. Show that $\mathbf{p}_i$ is a left eigenvector of $\mathbf{A}$ associated with λ_i; that is, $\mathbf{p}_i\mathbf{A} = \lambda_i \mathbf{p}_i$.

3.30 Show that if all eigenvalues of $\mathbf{A}$ are distinct, then $(s\mathbf{I} - \mathbf{A})^{-1}$ can be expressed as

$$(s\mathbf{I} - \mathbf{A})^{-1} = \sum \frac{1}{s - \lambda_i} \mathbf{q}_i \mathbf{p}_i$$

where $\mathbf{q}_i$ and $\mathbf{p}_i$ are right and left eigenvectors of $\mathbf{A}$ associated with λ_i.

3.31 Find the $\mathbf{M}$ to meet the Lyapunov equation in (3.59) with

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

What are the eigenvalues of the Lyapunov equation? Is the Lyapunov equation singular? Is the solution unique?

3.32 Repeat Problem 3.31 for

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix} \qquad \mathbf{B} = 1 \qquad \mathbf{C}_1 = \begin{bmatrix} 3 \\ 3 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 3 \\ -3 \end{bmatrix}$$

with two different $\mathbf{C}$.

3.33 Check to see if the following matrices are positive definite or semidefinite:

$$\begin{bmatrix} 2 & 3 & 2 \\ 3 & 1 & 0 \\ 2 & 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 2 \end{bmatrix} \qquad \begin{bmatrix} a_1a_1 & a_1a_2 & a_1a_3 \\ a_2a_1 & a_2a_2 & a_2a_3 \\ a_3a_1 & a_3a_2 & a_3a_3 \end{bmatrix}$$

3.34 Compute the singular values of the following matrices:

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

3.35 If $\mathbf{A}$ is symmetric, what is the relationship between its eigenvalues and singular values?

3.36 Show

$$\det\left(\mathbf{I}_n + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} [b_1 \ b_2 \ \cdots \ b_n]\right) = 1 + \sum_{m=1}^{n} a_m b_m$$

3.37 Show (3.65).

3.38 Consider $\mathbf{A}\mathbf{x} = \mathbf{y}$, where $\mathbf{A}$ is $m \times n$ and has rank m. Is $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}$ a solution? If not, under what condition will it be a solution? Is $\mathbf{A}'(\mathbf{A}\mathbf{A}')^{-1}\mathbf{y}$ a solution?

Chapter 4

State-Space Solutions and Realizations

4.1 Introduction

We showed in Chapter 2 that linear systems can be described by convolutions and, if lumped, by state-space equations. This chapter discusses how to find their solutions. First we discuss briefly how to compute solutions of the input-output description. There is no simple analytical way of computing the convolution

$$y(t) = \int_{\tau=t_0}^{t} g(t, \tau)u(\tau)\, d\tau$$

The easiest way is to compute it numerically on a digital computer. Before doing so, the equation must be discretized. One way is to discretize it as

$$y(k\Delta) = \sum_{m=k_0}^{k} g(k\Delta, m\Delta)u(m\Delta)\Delta \tag{4.1}$$

where Δ is called the integration step size. This is basically the discrete convolution discussed in (2.34). This discretization is the easiest but yields the least accurate result for the same integration step size. For other integration methods, see, for example, Reference [17].

For the linear time-invariant (LTI) case, we can also use $\hat{y}(s) = \hat{g}(s)\hat{u}(s)$ to compute the solution. If a system is distributed, $\hat{g}(s)$ will not be a rational function of s. Except for some special cases, it is simpler to compute the solution directly in the time domain as in (4.1). If the system is lumped, $\hat{g}(s)$ will be a rational function of s. In this case, if the Laplace transform of $u(t)$ is also a rational function of s, then the solution can be obtained by taking the inverse Laplace transform of $\hat{g}(s)\hat{u}(s)$. This method requires computing poles, carrying out

partial fraction expansion, and then using a Laplace transform table. These can be carried out using the MATLAB functions `roots` and `residue`. However, when there are repeated poles, the computation may become very sensitive to small changes in the data, including roundoff errors; therefore computing solutions using the Laplace transform is not a viable method on digital computers. A better method is to transform transfer functions into state-space equations and then compute the solutions. This chapter discusses solutions of state equations, how to transform transfer functions into state equations, and other related topics. We discuss first the time-invariant case and then the time-varying case.

4.2 Solution of LTI State Equations

Consider the linear time-invariant (LTI) state-space equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{4.2}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{4.3}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are, respectively, $n \times n$, $n \times p$, $q \times n$, and $q \times p$ constant matrices. The problem is to find the solution excited by the initial state $\mathbf{x}(0)$ and the input $\mathbf{u}(t)$. The solution hinges on the exponential function of $\mathbf{A}$ studied in Section 3.6. In particular, we need the property in (3.55) or

$$\frac{d}{dt}e^{\mathbf{A}t} = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}$$

to develop the solution. Premultiplying $e^{-\mathbf{A}t}$ on both sides of (4.2) yields

$$e^{-\mathbf{A}t}\dot{\mathbf{x}}(t) - e^{-\mathbf{A}t}\mathbf{A}\mathbf{x}(t) = e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}(t)$$

which implies

$$\frac{d}{dt}\left(e^{-\mathbf{A}t}\mathbf{x}(t)\right) = e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}(t)$$

Its integration from 0 to t yields

$$e^{-\mathbf{A}\tau}\mathbf{x}(\tau)\Big|_{\tau=0}^{t} = \int_0^t e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)\, d\tau$$

Thus we have

$$e^{-\mathbf{A}t}\mathbf{x}(t) - e^{\mathbf{0}}\mathbf{x}(0) = \int_0^t e^{-\mathbf{A}\tau}\mathbf{B}\mathbf{u}(\tau)\, d\tau \tag{4.4}$$

Because the inverse of $e^{-\mathbf{A}t}$ is $e^{\mathbf{A}t}$ and $e^{\mathbf{0}} = \mathbf{I}$ as discussed in (3.54) and (3.52), (4.4) implies

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\, d\tau \tag{4.5}$$

This is the solution of (4.2).

It is instructive to verify that (4.5) is the solution of (4.2). To verify this, we must show that (4.5) satisfies (4.2) and the initial condition $\mathbf{x}(t) = \mathbf{x}(0)$ at $t = 0$. Indeed, at $t = 0$, (4.5) reduces to

$$\mathbf{x}(0) = e^{\mathbf{A}\cdot 0}\mathbf{x}(0) = e^{0}\mathbf{x}(0) = \mathbf{I}\mathbf{x}(0) = \mathbf{x}(0)$$

Thus (4.5) satisfies the initial condition. We need the equation

$$\frac{\partial}{\partial t}\int_{t_0}^{t} f(t,\tau)\,d\tau = \int_{t_0}^{t}\left(\frac{\partial}{\partial t}f(t,\tau)\right)d\tau + f(t,\tau)|_{\tau=t} \tag{4.6}$$

to show that (4.5) satisfies (4.2). Differentiating (4.5) and using (4.6), we obtain

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \frac{d}{dt}\left[e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau\right]\\ &= \mathbf{A}e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t \mathbf{A}e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau + e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\big|_{\tau=t}\\ &= \mathbf{A}\left(e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau\right) + e^{\mathbf{A}\cdot 0}\mathbf{B}\mathbf{u}(t)\end{aligned}$$

which becomes, after substituting (4.5),

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$

Thus (4.5) meets (4.2) and the initial condition $\mathbf{x}(0)$ and is the solution of (4.2).

Substituting (4.5) into (4.3) yields the solution of (4.3) as

$$\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}(0) + \mathbf{C}\int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau + \mathbf{D}\mathbf{u}(t) \tag{4.7}$$

This solution and (4.5) are computed directly in the time domain. We can also compute the solutions by using the Laplace transform. Applying the Laplace transform to (4.2) and (4.3) yields, as derived in (2.14) and (2.15),

$$\begin{aligned}\hat{\mathbf{x}}(s) &= (s\mathbf{I} - \mathbf{A})^{-1}[\mathbf{x}(0) + \mathbf{B}\hat{\mathbf{u}}(s)]\\ \hat{\mathbf{y}}(s) &= \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}[\mathbf{x}(0) + \mathbf{B}\hat{\mathbf{u}}(s)] + \mathbf{D}\hat{\mathbf{u}}(s)\end{aligned}$$

Once $\hat{\mathbf{x}}(s)$ and $\hat{\mathbf{y}}(s)$ are computed algebraically, their inverse Laplace transforms yield the time-domain solutions.

We now give some remarks concerning the computation of $e^{\mathbf{A}t}$. We discussed in Section 3.6 three methods of computing functions of a matrix. They can all be used to compute $e^{\mathbf{A}t}$:

1. Using Theorem 3.5: First, compute the eigenvalues of $\mathbf{A}$; next, find a polynomial $h(\lambda)$ of degree $n-1$ that equals $e^{\lambda t}$ on the spectrum of $\mathbf{A}$; then $e^{\mathbf{A}t} = h(\mathbf{A})$.
2. Using Jordan form of $\mathbf{A}$: Let $\mathbf{A} = \mathbf{Q}\hat{\mathbf{A}}\mathbf{Q}^{-1}$; then $e^{\mathbf{A}t} = \mathbf{Q}e^{\hat{\mathbf{A}}t}\mathbf{Q}^{-1}$, where $\hat{\mathbf{A}}$ is in Jordan form and $e^{\hat{\mathbf{A}}t}$ can readily be obtained by using (3.48).
3. Using the infinite power series in (3.51): Although the series will not, in general, yield a closed-form solution, it is suitable for computer computation, as discussed following (3.51).

In addition, we can use (3.58) to compute $e^{\mathbf{A}t}$, that is,

$$e^{\mathbf{A}t} = \mathcal{L}^{-1}(s\mathbf{I} - \mathbf{A})^{-1} \tag{4.8}$$

The inverse of $(s\mathbf{I} - \mathbf{A})$ is a function of $\mathbf{A}$; therefore, again, we have many methods to compute it:

1. Taking the inverse of $(s\mathbf{I} - \mathbf{A})$.
2. Using Theorem 3.5.
3. Using $(s\mathbf{I} - \mathbf{A})^{-1} = \mathbf{Q}(s\mathbf{I} - \hat{\mathbf{A}})^{-1}\mathbf{Q}^{-1}$ and (3.49).
4. Using the infinite power series in (3.57).
5. Using the Leverrier algorithm discussed in Problem 3.26.

EXAMPLE 4.1 We use Methods 1 and 2 to compute $(s\mathbf{I} - \mathbf{A})^{-1}$, where

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

Method 1: We use (3.20) to compute

$$\begin{aligned}(s\mathbf{I} - \mathbf{A})^{-1} &= \begin{bmatrix} s & 1 \\ -1 & s+2 \end{bmatrix}^{-1} = \frac{1}{s^2+2s+1}\begin{bmatrix} s+2 & -1 \\ 1 & s \end{bmatrix}\\ &= \begin{bmatrix} (s+2)/(s+1)^2 & -1/(s+1)^2 \\ 1/(s+1)^2 & s/(s+)^2 \end{bmatrix}\end{aligned}$$

Method 2: The eigenvalues of $\mathbf{A}$ are $-1, -1$. Let $h(\lambda) = \beta_0 + \beta_1\lambda$. If $h(\lambda)$ equals $f(\lambda) := (s-\lambda)^{-1}$ on the spectrum of $\mathbf{A}$, then

$$\begin{aligned}f(-1) = h(-1): &\qquad (s+1)^{-1} = \beta_0 - \beta_1\\ f'(-1) = h'(-1): &\qquad (s+1)^{-2} = \beta_1\end{aligned}$$

Thus we have

$$h(\lambda) = [(s+1)^{-1} + (s+1)^{-2}] + (s+1)^{-2}\lambda$$

and

$$\begin{aligned}(s\mathbf{I} - \mathbf{A})^{-1} = h(\mathbf{A}) &= [(s+1)^{-1} + (s+1)^{-2}]\mathbf{I} + (s+1)^{-2}\mathbf{A}\\ &= \begin{bmatrix} (s+2)/(s+1)^2 & -1/(s+1)^2 \\ 1/(s+1)^2 & s/(s+)^2 \end{bmatrix}\end{aligned}$$

EXAMPLE 4.2 Consider the equation

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & -1 \\ 1 & -2 \end{bmatrix}\mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}u(t)$$

Its solution is

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0) + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}u(\tau)\,d\tau$$

The matrix function $e^{\mathbf{A}t}$ is the inverse Laplace transform of $(s\mathbf{I} - \mathbf{A})^{-1}$, which was computed in the preceding example. Thus we have

$$e^{\mathbf{A}t} = \mathcal{L}^{-1}\begin{bmatrix} \dfrac{s+2}{(s+1)^2} & \dfrac{-1}{(s+1)^2} \\ \dfrac{1}{(s+1)^2} & \dfrac{s}{(s+)^2} \end{bmatrix} = \begin{bmatrix} (1+t)e^{-t} & -te^{-t} \\ te^{-t} & (1-t)e^{-t} \end{bmatrix}$$

and

$$\mathbf{x}(t) = \begin{bmatrix} (1+t)e^{-t} & -te^{-t} \\ te^{-t} & (1-t)e^{-t} \end{bmatrix}\mathbf{x}(0) + \begin{bmatrix} -\int_0^t (t-\tau)e^{-(t-\tau)}u(\tau)\,d\tau \\ \int_0^t [1-(t-\tau)]e^{-(t-\tau)}u(\tau)\,d\tau \end{bmatrix}$$

We discuss a general property of the zero-input response $e^{\mathbf{A}t}\mathbf{x}(0)$. Consider the second matrix in (3.39). Then we have

$$e^{\mathbf{A}t} = \mathbf{Q}\begin{bmatrix} e^{\lambda_1 t} & te^{\lambda_1 t} & t^2e^{\lambda_1 t}/2 & 0 & 0 \\ 0 & e^{\lambda_1 t} & te^{\lambda_1 t} & 0 & 0 \\ 0 & 0 & e^{\lambda_1 t} & 0 & 0 \\ 0 & 0 & 0 & e^{\lambda_1 t} & 0 \\ 0 & 0 & 0 & 0 & e^{\lambda_2 t} \end{bmatrix}\mathbf{Q}^{-1}$$

Every entry of $e^{\mathbf{A}t}$ and, consequently, of the zero-input response is a linear combination of terms $\{e^{\lambda_1 t},\ te^{\lambda_1 t},\ t^2e^{\lambda_1 t},\ e^{\lambda_2 t}\}$. These terms are dictated by the eigenvalues and their indices. In general, if $\mathbf{A}$ has eigenvalue λ_1 with index $\bar{n}_1$, then every entry of $e^{\mathbf{A}t}$ is a linear combination of

$$e^{\lambda_1 t}\quad te^{\lambda_1 t}\quad \cdots\quad t^{\bar{n}_1-1}e^{\lambda_1 t}$$

Every such term is *analytic* in the sense that it is infinitely differentiable and can be expanded in a Taylor series at every t. This is a nice property and will be used in Chapter 6.

If every eigenvalue, simple or repeated, of $\mathbf{A}$ has a negative real part, then every zero-input response will approach zero as $t \to \infty$. If $\mathbf{A}$ has an eigenvalue, simple or repeated, with a positive real part, then most zero-input responses will grow unbounded as $t \to \infty$. If $\mathbf{A}$ has some eigenvalues with zero real part and all with index 1 and the remaining eigenvalues all have negative real parts, then no zero-input response will grow unbounded. However, if the index is 2 or higher, then some zero-input response may become unbounded. For example, if $\mathbf{A}$ has eigenvalue 0 with index 2, then $e^{\mathbf{A}t}$ contains the terms $\{1,\ t\}$. If a zero-input response contains the term t, then it will grow unbounded.

4.2.1 Discretization

Consider the continuous-time state equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{4.9}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{4.10}$$

If the set of equations is to be computed on a digital computer, it must be discretized. Because

$$\dot{\mathbf{x}}(t) = \lim_{T\to 0}\frac{\mathbf{x}(t+T)-\mathbf{x}(t)}{T}$$

we can approximate (4.9) as

$$\mathbf{x}(t+T) = \mathbf{x}(t) + \mathbf{A}\mathbf{x}(t)T + \mathbf{B}\mathbf{u}(t)T \tag{4.11}$$

If we compute $\mathbf{x}(t)$ and $\mathbf{y}(t)$ only at $t = kT$ for $k = 0,\ 1,\ \ldots$, then (4.11) and (4.10) become

$$\mathbf{x}((k+1)T) = (\mathbf{I} + T\mathbf{A})\mathbf{x}(kT) + T\mathbf{B}\mathbf{u}(kT)$$

$$\mathbf{y}(kT) = \mathbf{C}\mathbf{x}(kT) + \mathbf{D}\mathbf{u}(kT)$$

This is a discrete-time state-space equation and can easily be computed on a digital computer. This discretization is the easiest to carry out but yields the least accurate results for the same T. We discuss next a different discretization.

If an input $\mathbf{u}(t)$ is generated by a digital computer followed by a digital-to-analog converter, then $\mathbf{u}(t)$ will be piecewise constant. This situation often arises in computer control of control systems. Let

$$\mathbf{u}(t) = \mathbf{u}(kT) =: \mathbf{u}[k] \qquad \text{for } kT \le t < (k+1)T \tag{4.12}$$

for $k = 0,\ 1,\ 2,\ \ldots$. This input changes values only at discrete-time instants. For this input, the solution of (4.9) still equals (4.5). Computing (4.5) at $t = kT$ and $t = (k+1)T$ yields

$$\mathbf{x}[k] := \mathbf{x}(kT) = e^{\mathbf{A}kT}\mathbf{x}(0) + \int_0^{kT} e^{\mathbf{A}(kT-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau \tag{4.13}$$

and

$$\mathbf{x}[k+1] := \mathbf{x}((k+1)T) = e^{\mathbf{A}(k+1)T}\mathbf{x}(0) + \int_0^{(k+1)T} e^{\mathbf{A}((k+1)T-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau \tag{4.14}$$

Equation (4.14) can be written as

$$\mathbf{x}[k+1] = e^{\mathbf{A}T}\left[e^{\mathbf{A}kT}\mathbf{x}(0) + \int_0^{kT} e^{\mathbf{A}(kT-\tau)}\mathbf{B}\mathbf{u}(\tau)d\tau\right] + \int_{kT}^{(k+1)T} e^{\mathbf{A}(kT+T-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau$$

which becomes, after substituting (4.12) and (4.13) and introducing the new variable $\alpha := kT + T - \tau$,

$$\mathbf{x}[k+1] = e^{\mathbf{A}T}\mathbf{x}[k] + \left(\int_0^T e^{\mathbf{A}\alpha}d\alpha\right)\mathbf{B}\mathbf{u}[k]$$

Thus, if an input changes value only at discrete-time instants kT and if we compute only the responses at $t = kT$, then (4.9) and (4.10) become

$$\mathbf{x}[k+1] = \mathbf{A}_d\mathbf{x}[k] + \mathbf{B}_d\mathbf{u}[k] \tag{4.15}$$

$$\mathbf{y}[k] = \mathbf{C}_d\mathbf{x}[k] + \mathbf{D}_d\mathbf{u}[k] \tag{4.16}$$

with

$$\mathbf{A}_d = e^{\mathbf{A}T} \qquad \mathbf{B}_d = \left(\int_0^T e^{\mathbf{A}\tau} d\tau\right)\mathbf{B} \qquad \mathbf{C}_d = \mathbf{C} \qquad \mathbf{D}_d = \mathbf{D} \tag{4.17}$$

This is a discrete-time state-space equation. Note that there is no approximation involved in this derivation and (4.15) yields the exact solution of (4.9) at $t = kT$ if the input is piecewise constant.

We discuss the computation of $\mathbf{B}_d$. Using (3.51), we have

$$\int_0^T \left(\mathbf{I} + \mathbf{A}\tau + \mathbf{A}^2\frac{\tau^2}{2!} + \cdots\right) d\tau$$
$$= T\mathbf{I} + \frac{T^2}{2!}\mathbf{A} + \frac{T^3}{3!}\mathbf{A}^2 + \frac{T^4}{4!}\mathbf{A}^3 + \cdots$$

This power series can be computed recursively as in computing (3.51). If $\mathbf{A}$ is nonsingular, then the series can be written as, using (3.51),

$$\mathbf{A}^{-1}\left(T\mathbf{A} + \frac{T^2}{2!}\mathbf{A}^2 + \frac{T^3}{3!}\mathbf{A}^3 + \cdots + \mathbf{I} - \mathbf{I}\right) = \mathbf{A}^{-1}(e^{\mathbf{A}T} - \mathbf{I})$$

Thus we have

$$\mathbf{B}_d = \mathbf{A}^{-1}(\mathbf{A}_d - \mathbf{I})\mathbf{B} \qquad \text{(if } \mathbf{A} \text{ is nonsingular)} \tag{4.18}$$

Using this formula, we can avoid computing an infinite series.

The MATLAB function `[ad,bd]=c2d(a,b,T)` transforms the continuous-time state equation in (4.9) into the discrete-time state equation in (4.15).

4.2.2 Solution of Discrete-Time Equations

Consider the discrete-time state-space equation

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k]$$
$$\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k] \tag{4.19}$$

where the subscript d has been dropped. It is understood that if the equation is obtained from a continuous-time equation, then the four matrices must be computed from (4.17). The two equations in (4.19) are algebraic equations. Once $\mathbf{x}[0]$ and $\mathbf{u}[k]$, $k = 0, 1, \ldots$, are given, the response can be computed recursively from the equations.

The MATLAB function `dstep` computes unit-step responses of discrete-time state-space equations. It also computes unit-step responses of discrete transfer functions; internally, it first transforms the transfer function into a discrete-time state-space equation by calling `tf2ss`, which will be discussed later, and then uses `dstep`. The function `dlsim`, an acronym for discrete linear simulation, computes responses excited by any input. The function `step` computes unit-step responses of continuous-time state-space equations. Internally, it first uses the function `c2d` to transform a continuous-time state equation into a discrete-time equation and then carries out the computation. If the function `step` is applied to a continuous-time transfer function, then it first uses `tf2ss` to transform the transfer function into a continuous-time state equation and then discretizes it by using `c2d` and then uses `dstep` to compute the response. Similar remarks apply to `lsim`, which computes responses of continuous-time state equations or transfer functions excited by any input.

In order to discuss the general behavior of discrete-time state equations, we will develop a general form of solutions. We compute

$$\mathbf{x}[1] = \mathbf{A}\mathbf{x}[0] + \mathbf{B}\mathbf{u}[0]$$
$$\mathbf{x}[2] = \mathbf{A}\mathbf{x}[1] + \mathbf{B}\mathbf{u}[1] = \mathbf{A}^2\mathbf{x}[0] + \mathbf{A}\mathbf{B}\mathbf{u}[0] + \mathbf{B}\mathbf{u}[1]$$

Proceeding forward, we can readily obtain, for $k > 0$,

$$\mathbf{x}[k] = \mathbf{A}^k\mathbf{x}[0] + \sum_{m=0}^{k-1}\mathbf{A}^{k-1-m}\mathbf{B}\mathbf{u}[m] \tag{4.20}$$

$$\mathbf{y}[k] = \mathbf{C}\mathbf{A}^k\mathbf{x}[0] + \sum_{m=0}^{k-1}\mathbf{C}\mathbf{A}^{k-1-m}\mathbf{B}\mathbf{u}[m] + \mathbf{D}\mathbf{u}[k] \tag{4.21}$$

They are the discrete counterparts of (4.5) and (4.7). Their derivations are considerably simpler than the continuous-time case.

We discuss a general property of the zero-input response $\mathbf{A}^k\mathbf{x}[0]$. Suppose $\mathbf{A}$ has eigenvalue λ_1 with multiplicity 4 and eigenvalue λ_2 with multiplicity 1 and suppose its Jordan form is as shown in the second matrix in (3.39). In other words, λ_1 has index 3 and λ_2 has index 1. Then we have

$$\mathbf{A}^k = \mathbf{Q}\begin{bmatrix} \lambda_1^k & k\lambda_1^{k-1} & k(k-1)\lambda_1^{k-2}/2 & 0 & 0 \\ 0 & \lambda_1^k & k\lambda_1^{k-1} & 0 & 0 \\ 0 & 0 & \lambda_1^k & 0 & 0 \\ 0 & 0 & 0 & \lambda_1^k & 0 \\ 0 & 0 & 0 & 0 & \lambda_2^k \end{bmatrix}\mathbf{Q}^{-1}$$

which implies that every entry of the zero-input response is a linear combination of $\{\lambda_1^k,\ k\lambda_1^{k-1},\ k^2\lambda_1^{k-2},\ \lambda_2^k\}$. These terms are dictated by the eigenvalues and their indices.

If every eigenvalue, simple or repeated, of $\mathbf{A}$ has magnitude less than 1, then every zero-input response will approach zero as $k \to \infty$. If $\mathbf{A}$ has an eigenvalue, simple or repeated, with magnitude larger than 1, then most zero-input responses will grow unbounded as $k \to \infty$. If $\mathbf{A}$ has some eigenvalues with magnitude 1 and all with index 1 and the remaining eigenvalues all have magnitudes less than 1, then no zero-input response will grow unbounded. However, if the index is 2 or higher, then some zero-state response may become unbounded. For example, if $\mathbf{A}$ has eigenvalue 1 with index 2, then $\mathbf{A}^k$ contains the terms $\{1,\ k\}$. If a zero-input response contains the term k, then it will grow unbounded as $k \to \infty$.

4.3 Equivalent State Equations

The example that follows provides a motivation for studying equivalent state equations.

EXAMPLE 4.3 Consider the network shown in Fig. 4.1. It consists of one capacitor, one inductor, one resistor, and one voltage source. First we select the inductor current x_1 and

capacitor voltage x_2 as state variables as shown. The voltage across the inductor is $\dot{x}_1$ and the current through the capacitor is $\dot{x}_2$. The voltage across the resistor is x_2; thus its current is $x_2/1 = x_2$. Clearly we have $x_1 = x_2 + \dot{x}_2$ and $\dot{x}_1 + x_2 - u = 0$. Thus the network is described by the following state equation:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u$$
$$y = [0 \quad 1]\mathbf{x} \tag{4.22}$$

If, instead, the loop currents $\bar{x}_1$ and $\bar{x}_2$ are chosen as state variables as shown, then the voltage across the inductor is $\dot{\bar{x}}_1$ and the voltage across the resistor is $(\bar{x}_1 - \bar{x}_2) \cdot 1$. From the left-hand-side loop, we have

$$u = \dot{\bar{x}}_1 + \bar{x}_1 - \bar{x}_2 \quad \text{or} \quad \dot{\bar{x}}_1 = -\bar{x}_1 + \bar{x}_2 + u$$

The voltage across the capacitor is the same as the one across the resistor, which is $\bar{x}_1 - \bar{x}_2$. Thus the current through the capacitor is $\dot{\bar{x}}_1 - \dot{\bar{x}}_2$, which equals $\bar{x}_2$ or

$$\dot{\bar{x}}_2 = \dot{\bar{x}}_1 - \bar{x}_2 = -\bar{x}_1 + u$$

Thus the network is also described by the state equation

$$\begin{bmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u$$
$$y = [1 \quad -1]\bar{\mathbf{x}} \tag{4.23}$$

The state equations in (4.22) and (4.23) describe the same network; therefore they must be closely related. In fact, they are equivalent as will be established shortly.

Consider the n-dimensional state equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$$
$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{4.24}$$

where $\mathbf{A}$ is an $n \times n$ constant matrix mapping an n-dimensional real space $\mathcal{R}^n$ into itself. The state $\mathbf{x}$ is a vector in $\mathcal{R}^n$ for all t; thus the real space is also called the state space. The state equation in (4.24) can be considered to be associated with the orthonormal basis in (3.8). Now we study the effect on the equation by choosing a different basis.

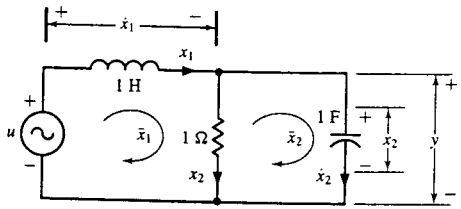


Figure 4.1 Network with two different sets of state variables.

Definition 4.1 *Let* $\mathbf{P}$ *be an* $n \times n$ *real nonsingular matrix and let* $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$. *Then the state equation,*

$$\dot{\bar{\mathbf{x}}}(t) = \bar{\mathbf{A}}\bar{\mathbf{x}}(t) + \bar{\mathbf{B}}\mathbf{u}(t)$$
$$\mathbf{y}(t) = \bar{\mathbf{C}}\bar{\mathbf{x}}(t) + \bar{\mathbf{D}}\mathbf{u}(t) \tag{4.25}$$

where

$$\bar{\mathbf{A}} = \mathbf{P}\mathbf{A}\mathbf{P}^{-1} \qquad \bar{\mathbf{B}} = \mathbf{P}\mathbf{B} \qquad \bar{\mathbf{C}} = \mathbf{C}\mathbf{P}^{-1} \qquad \bar{\mathbf{D}} = \mathbf{D} \tag{4.26}$$

is said to be (algebraically) equivalent *to (4.24) and* $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$ *is called an* equivalence transformation.

Equation (4.26) is obtained from (4.24) by substituting $\mathbf{x}(t) = \mathbf{P}^{-1}\bar{\mathbf{x}}(t)$ and $\dot{\mathbf{x}}(t) = \mathbf{P}^{-1}\dot{\bar{\mathbf{x}}}(t)$. In this substitution, we have changed, as in Equation (3.7), the basis vectors of the state space from the orthonormal basis to the columns of $\mathbf{P}^{-1} =: \mathbf{Q}$. Clearly $\mathbf{A}$ and $\bar{\mathbf{A}}$ are similar and $\bar{\mathbf{A}}$ is simply a different representation of $\mathbf{A}$. To be precise, let $\mathbf{Q} = \mathbf{P}^{-1} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n]$. Then the ith column of $\bar{\mathbf{A}}$ is, as discussed in Section 3.4, the representation of $\mathbf{A}\mathbf{q}_i$ with respect to the basis $\{\mathbf{q}_1, \ \mathbf{q}_2, \ \cdots, \ \mathbf{q}_n\}$. From the equation $\bar{\mathbf{B}} = \mathbf{P}\mathbf{B}$ or $\mathbf{B} = \mathbf{P}^{-1}\bar{\mathbf{B}} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \cdots \ \mathbf{q}_n]\bar{\mathbf{B}}$, we see that the ith column of $\bar{\mathbf{B}}$ is the representation of the ith column of $\mathbf{B}$ with respect to the basis $\{\mathbf{q}_1, \ \mathbf{q}_2, \ \cdots, \ \mathbf{q}_n\}$. The matrix $\bar{\mathbf{C}}$ is to be computed from $\mathbf{C}\mathbf{P}^{-1}$. The matrix $\mathbf{D}$, called the *direct transmission part* between the input and output, has nothing to do with the state space and is not affected by the equivalence transformation.

We show that (4.24) and (4.25) have the same set of eigenvalues and the same transfer matrix. Indeed, we have, using $\det(\mathbf{P})\det(\mathbf{P}^{-1}) = 1$,

$$\bar{\Delta}(\lambda) = \det(\lambda\mathbf{I} - \bar{\mathbf{A}}) = \det(\lambda\mathbf{P}\mathbf{P}^{-1} - \mathbf{P}\mathbf{A}\mathbf{P}^{-1}) = \det[\mathbf{P}(\lambda\mathbf{I} - \mathbf{A})\mathbf{P}^{-1}]$$
$$= \det(\mathbf{P})\det(\lambda\mathbf{I} - \mathbf{A})\det(\mathbf{P}^{-1}) = \det(\lambda\mathbf{I} - \mathbf{A}) = \Delta(\lambda)$$

and

$$\hat{\bar{\mathbf{G}}}(s) = \bar{\mathbf{C}}(s\mathbf{I} - \bar{\mathbf{A}})^{-1}\bar{\mathbf{B}} + \bar{\mathbf{D}} = \mathbf{C}\mathbf{P}^{-1}[\mathbf{P}(s\mathbf{I} - \mathbf{A})\mathbf{P}^{-1}]^{-1}\mathbf{P}\mathbf{B} + \mathbf{D}$$
$$= \mathbf{C}\mathbf{P}^{-1}\mathbf{P}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{P}^{-1}\mathbf{P}\mathbf{B} + \mathbf{D} = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} = \hat{\mathbf{G}}(s)$$

Thus equivalent state equations have the same characteristic polynomial and, consequently, the same set of eigenvalues and same transfer matrix. In fact, all properties of (4.24) are preserved or invariant under any equivalence transformation.

Consider again the network shown in Fig. 4.1, which can be described by (4.22) and (4.23). We show that the two equations are equivalent. From Fig. 4.1, we have $x_1 = \bar{x}_1$. Because the voltage across the resistor is x_2, its current is $x_2/1$ and equals $\bar{x}_1 - \bar{x}_2$. Thus we have

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

or

$$\begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix}^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{4.27}$$

Note that, for this $\mathbf{P}$, its inverse happens to equal itself. It is straightforward to verify that (4.22) and (4.23) are related by the equivalence transformation in (4.26).

The MATLAB function `[ab,bb,cb,db]=ss2ss(a,b,c,d,p)` carries out equivalence transformations.

Two state equations are said to be *zero-state equivalent* if they have the same transfer matrix or

$$\mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \bar{\mathbf{D}} + \bar{\mathbf{C}}(s\mathbf{I} - \bar{\mathbf{A}})^{-1}\bar{\mathbf{B}}$$

This becomes, after substituting (3.57),

$$\mathbf{D} + \mathbf{C}\mathbf{B}s^{-1} + \mathbf{C}\mathbf{A}\mathbf{B}s^{-2} + \mathbf{C}\mathbf{A}^2\mathbf{B}s^{-3} + \cdots = \bar{\mathbf{D}} + \bar{\mathbf{C}}\bar{\mathbf{B}}s^{-1} + \bar{\mathbf{C}}\bar{\mathbf{A}}\bar{\mathbf{B}}s^{-2} + \bar{\mathbf{C}}\bar{\mathbf{A}}^2\bar{\mathbf{B}}s^{-3} + \cdots$$

Thus we have the theorem that follows.

Theorem 4.1

Two linear time-invariant state equations $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ and $\{\bar{\mathbf{A}}, \bar{\mathbf{B}}, \bar{\mathbf{C}}, \bar{\mathbf{D}}\}$ are zero-state equivalent or have the same transfer matrix if and only if $\mathbf{D} = \bar{\mathbf{D}}$ and

$$\mathbf{C}\mathbf{A}^m\mathbf{B} = \bar{\mathbf{C}}\bar{\mathbf{A}}^m\bar{\mathbf{B}} \qquad m = 0, 1, 2, \ldots$$

It is clear that (algebraic) equivalence implies zero-state equivalence. In order for two state equations to be equivalent, they must have the same dimension. This is, however, not the case for zero-state equivalence, as the next example shows.

EXAMPLE 4.4 Consider the two networks shown in Fig. 4.2. The capacitor is assumed to have capacitance -1 F. Such a negative capacitance can be realized using an op-amp circuit. For the circuit in Fig. 4.2(a), we have $y(t) = 0.5 \cdot u(t)$ or $\hat{y}(s) = 0.5\hat{u}(s)$. Thus its transfer function is 0.5. To compute the transfer function of the network in Fig. 4.2(b), we may assume the initial voltage across the capacitor to be zero. Because of the symmetry of the four resistors, half of the current will go through each resistor or $i(t) = 0.5u(t)$, where $i(t)$ denotes the right upper resistor's current. Consequently, $y(t) = i(t) \cdot 1 = 0.5u(t)$ and the transfer function also equals 0.5. Thus the two networks, or more precisely their state equations, are zero-state equivalent. This fact can also be verified by using Theorem 4.1. The network in Fig. 4.2(a) is described by the zero-dimensional state equation $y(t) = 0.5u(t)$ or $\mathbf{A} = \mathbf{B} = \mathbf{C} = 0$ and $\mathbf{D} = 0.5$. To

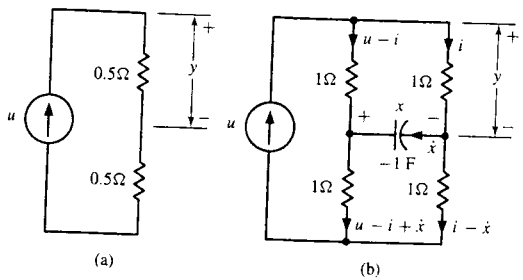


Figure 4.2 Two zero-state equivalent networks.

develop a state equation for the network in Fig. 4.2(b), we assign the capacitor voltage as state variable x with polarity shown. Its current is $\dot{x}$ flowing from the negative to positive polarity because of the negative capacitance. If we assign the right upper resistor's current as $i(t)$, then the right lower resistor's current is $i - \dot{x}$, the left upper resistor's current is $u - i$, and the left lower resistor's current is $u - i + \dot{x}$. The total voltage around the upper right-hand loop is 0:

$$i - x - (u - i) = 0 \quad \text{or} \quad i = 0.5(x + u)$$

which implies

$$y = 1 \cdot i = i = 0.5(x + u)$$

The total voltage around the lower right-hand loop is 0:

$$x + (i - \dot{x}) - (u - i + \dot{x}) = 0$$

which implies

$$2\dot{x} = 2i + x - u = x + u + x - u = 2x$$

Thus the network in Fig. 4.2(b) is described by the one-dimensional state equation

$$\dot{x}(t) = x(t)$$
$$y(t) = 0.5x(t) + 0.5u(t)$$

with $\bar{\mathbf{A}} = 1$, $\bar{\mathbf{B}} = 0$, $\bar{\mathbf{C}} = 0.5$, and $\bar{\mathbf{D}} = 0.5$. We see that $\mathbf{D} = \bar{\mathbf{D}} = 0.5$ and $\mathbf{C}\mathbf{A}^m\mathbf{B} = \bar{\mathbf{C}}\bar{\mathbf{A}}^m\bar{\mathbf{B}} = 0$ for $m = 0, 1, \ldots$. Thus the two equations are zero-state equivalent.

4.3.1 Canonical Forms

MATLAB contains the function `[ab,bb,cb,db,P]=canon(a,b,c,d,'type')`. If `type=companion`, the function will generate an equivalent state equation with $\bar{\mathbf{A}}$ in the companion form in (3.24). This function works only if $\mathbf{Q} := [\mathbf{b}_1 \ \mathbf{A}\mathbf{b}_1 \ \cdots \ \mathbf{A}^{n-1}\mathbf{b}_1]$ is nonsingular, where $\mathbf{b}_1$ is the first column of $\mathbf{B}$. This condition is the same as $\{\mathbf{A}, \mathbf{b}_1\}$ controllable, as we will discuss in Chapter 6. The $\mathbf{P}$ that the function `canon` generates equals $\mathbf{Q}^{-1}$. See the discussion in Section 3.4.

We discuss a different canonical form. Suppose $\mathbf{A}$ has two real eigenvalues and two complex eigenvalues. Because $\mathbf{A}$ has only real coefficients, the two complex eigenvalues must be complex conjugate. Let λ_1, λ_2, $\alpha + j\beta$, and $\alpha - j\beta$ be the eigenvalues and $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, and $\mathbf{q}_4$ be the corresponding eigenvectors, where λ_1, λ_2, α, β, $\mathbf{q}_1$, and $\mathbf{q}_2$ are all real and $\mathbf{q}_4$ equals the complex conjugate of $\mathbf{q}_3$. Define $\mathbf{Q} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \mathbf{q}_3 \ \mathbf{q}_4]$. Then we have

$$\mathbf{J} := \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \alpha + j\beta & 0 \\ 0 & 0 & 0 & \alpha - j\beta \end{bmatrix} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{Q}$$

Note that $\mathbf{Q}$ and $\mathbf{J}$ can be obtained from `[q,j]=eig(a)` in MATLAB as shown in Examples 3.5 and 3.6. This form is useless in practice but can be transformed into a real matrix by the following similarity transformation

$$\bar{\mathbf{Q}}^{-1}\mathbf{J}\bar{\mathbf{Q}} := \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & j & -j \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \alpha+j\beta & 0 \\ 0 & 0 & 0 & \alpha-j\beta \end{bmatrix}$$
$$\cdot \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & -0.5j \\ 0 & 0 & 0.5 & 0.5j \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \alpha & \beta \\ 0 & 0 & -\beta & \alpha \end{bmatrix} =: \bar{\mathbf{A}}$$

We see that this transformation transforms the complex eigenvalues on the diagonal into a block with the real part of the eigenvalues on the diagonal and the imaginary part on the off-diagonal. This new A-matrix is said to be in *modal* form. The MATLAB function `[ab,bb,cb,db,P]=canon(a,b,c,d,'modal')` or `canon(a,b,c,d)` with no type specified will yield an equivalent state equation with $\bar{\mathbf{A}}$ in modal form. Note that there is no need to transform $\mathbf{A}$ into a diagonal form and then to a modal form. The two transformations can be combined into one as

$$\mathbf{P}^{-1} = \mathbf{Q}\bar{\mathbf{Q}} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \mathbf{q}_3 \ \mathbf{q}_4] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.5 & -0.5j \\ 0 & 0 & 0.5 & 0.5j \end{bmatrix}$$
$$= [\mathbf{q}_1 \ \mathbf{q}_2 \ \mathrm{Re}(\mathbf{q}_3) \ \mathrm{Im}(\mathbf{q}_3)]$$

where Re and Im stand, respectively, for the real part and imaginary part and we have used in the last equality the fact that $\mathbf{q}_4$ is the complex conjugate of $\mathbf{q}_3$. We give one more example. The modal form of a matrix with real eigenvalue λ_1 and two pairs of distinct complex conjugate eigenvalues $\alpha_i \pm j\beta_i$, for $i = 1, 2$, is

$$\bar{\mathbf{A}} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \alpha_1 & \beta_1 & 0 & 0 \\ 0 & -\beta_1 & \alpha_1 & 0 & 0 \\ 0 & 0 & 0 & \alpha_2 & \beta_2 \\ 0 & 0 & 0 & -\beta_2 & \alpha_2 \end{bmatrix} \tag{4.28}$$

It is block diagonal and can be obtained by the similarity transformation

$$\mathbf{P}^{-1} = [\mathbf{q}_1 \ \mathrm{Re}(\mathbf{q}_2) \ \mathrm{Im}(\mathbf{q}_2) \ \mathrm{Re}(\mathbf{q}_4) \ \mathrm{Im}(\mathbf{q}_4)]$$

where $\mathbf{q}_1$, $\mathbf{q}_2$, and $\mathbf{q}_4$ are, respectively, eigenvectors associated with λ_1, $\alpha_1 + j\beta_1$, and $\alpha_2 + j\beta_2$. This form is useful in state-space design.

4.3.2 Magnitude Scaling in Op-Amp Circuits

As discussed in Section 2.3.1, every LTI state equation can be implemented using an op-amp circuit.[1] In actual op-amp circuits, all signals are limited by power supplies. If we use ±15-volt power supplies, then all signals are roughly limited to ±13 volts. If any signal goes outside the range, the circuit will saturate and will not behave as the state equation dictates. Therefore saturation is an important issue in actual op-amp circuit implementation.

1. This subsection may be skipped without loss of continuity.

Consider an LTI state equation and suppose all signals must be limited to $\pm M$. For linear systems, if the input magnitude increases by α, so do the magnitudes of all state variables and the output. Thus there must be a limit on input magnitude. Clearly it is desirable to have the admissible input magnitude as large as possible. One way to achieve this is to use an equivalence transformation so that

$$|x_i(t)| \le |y(t)| \le M$$

for all i and for all t. The equivalence transformation, however, will not alter the relationship between the input and output; therefore we can use the original state equation to find the input range to achieve $|y(t)| \le M$. In addition, we can use the same transformation to amplify some state variables to increase visibility or accuracy. This is illustrated in the next example.

EXAMPLE 4.5 Consider the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} -0.1 & 2 \\ 0 & -1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 10 \\ 0.1 \end{bmatrix} u$$
$$y = [0.1 \quad -1]\mathbf{x}$$

Suppose the input is a step function of various magnitude and the equation is to be implemented using an op-amp circuit in which all signals must be limited to ±10. First we use MATLAB to find its unit-step response. We type

```
a=[-0.1 2;0 -1];b=[10;0.1];c=[0.2 -1];d=0;
[y,x,t]=step(a,b,c,d);
plot(t,y,t,x)
```

which yields the plot in Fig. 4.3(a). We see that $|x_1|_{max} = 100 > |y|_{max} = 20$ and $|x_2| << |y|_{max}$. The state variable x_2 is hardly visible and its largest magnitude is found to be 0.1 by plotting it separately (not shown). From the plot, we see that if $|u(t)| \le 0.5$, then the output will not saturate but $x_1(t)$ will.

Let us introduce new state variables as

$$\bar{x}_1 = \frac{20}{100}x_1 = 0.2x_1 \qquad \bar{x}_2 = \frac{20}{0.1}x_2 = 200x_2$$

With this transformation, the maximum magnitudes of $\bar{x}_1(t)$ and $\bar{x}_2(t)$ will equal $|y|_{max}$. Thus if $y(t)$ does not saturate, neither will all the state variables $\bar{x}_i$. The transformation can be expressed as $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$ with

$$\mathbf{P} = \begin{bmatrix} 0.2 & 0 \\ 0 & 200 \end{bmatrix} \qquad \mathbf{P}^{-1} = \begin{bmatrix} 5 & 0 \\ 0 & 0.005 \end{bmatrix}$$

Then its equivalent state equation can readily be computed from (4.26) as

$$\dot{\bar{\mathbf{x}}} = \begin{bmatrix} -0.1 & 0.002 \\ 0 & -1 \end{bmatrix}\bar{\mathbf{x}} + \begin{bmatrix} 2 \\ 20 \end{bmatrix} u$$
$$y = [1 \quad -0.005]\bar{\mathbf{x}}$$

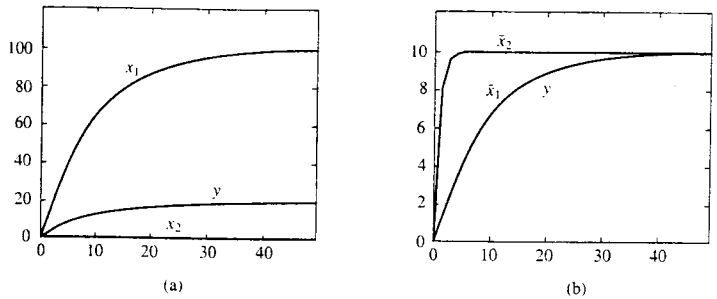


Figure 4.3 Time responses.

Its step responses due to $u(t) = 0.5$ are plotted in Fig. 4.3(b). We see that all signals lie inside the range ± 10 and occupy the full scale. Thus the equivalence state equation is better for op-amp circuit implementation or simulation.

The magnitude scaling is important in using op-amp circuits to implement or simulate continuous-time systems. Although we discuss only step inputs, the idea is applicable to any input. We mention that analog computers are essentially op-amp circuits. Before the advent of digital computers, magnitude scaling in analog computer simulation was carried out by trial and error. With the help of digital computer simulation, the magnitude scaling can now be carried out easily.

4.4 Realizations

Every linear time-invariant (LTI) system can be described by the input–output description

$$\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}(s)\hat{\mathbf{u}}(s)$$

and, if the system is lumped as well, by the state-space equation description

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{aligned} \tag{4.29}$$

If the state equation is known, the transfer matrix can be computed as $\hat{\mathbf{G}}(s) = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$. The computed transfer matrix is unique. Now we study the converse problem, that is, to find a state-space equation from a given transfer matrix. This is called the *realization* problem. This terminology is justified by the fact that, by using the state equation, we can build an op-amp circuit for the transfer matrix.

A transfer matrix $\hat{\mathbf{G}}(s)$ is said to be *realizable* if there exists a finite-dimensional state equation (4.29) or, simply, $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ such that

$$\hat{\mathbf{G}}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

and $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$ is called a *realization* of $\hat{\mathbf{G}}(s)$. An LTI distributed system can be described by a transfer matrix, but not by a finite-dimensional state equation. Thus not every $\hat{\mathbf{G}}(s)$ is realizable. If $\hat{\mathbf{G}}(s)$ is realizable, then it has infinitely many realizations, not necessarily of the same dimension. Thus the realization problem is fairly complex. We study here only the realizability condition. The other issues will be studied in later chapters.

Theorem 4.2

A transfer matrix $\hat{\mathbf{G}}(s)$ is realizable if and only if $\hat{\mathbf{G}}(s)$ is a proper rational matrix.

We use (3.19) to write

$$\hat{\mathbf{G}}_{sp}(s) := \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \frac{1}{\det(s\mathbf{I} - \mathbf{A})}\mathbf{C}[\text{Adj } (s\mathbf{I} - \mathbf{A})]\mathbf{B} \tag{4.30}$$

If $\mathbf{A}$ is $n \times n$, then $\det(s\mathbf{I} - \mathbf{A})$ has degree n. Every entry of $\text{Adj}\ (s\mathbf{I} - \mathbf{A})$ is the determinant of an $(n-1) \times (n-1)$ submatrix of $(s\mathbf{I} - \mathbf{A})$; thus it has at most degree $(n-1)$. Their linear combinations again have at most degree $(n-1)$. Thus we conclude that $\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}$ is a strictly proper rational matrix. If $\mathbf{D}$ is a nonzero matrix, then $\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}+\mathbf{D}$ is proper. This shows that if $\hat{\mathbf{G}}(s)$ is realizable, then it is a proper rational matrix. Note that we have

$$\hat{\mathbf{G}}(\infty) = \mathbf{D}$$

Next we show the converse; that is, if $\hat{\mathbf{G}}(s)$ is a $q \times p$ proper rational matrix, then there exists a realization. First we decompose $\hat{\mathbf{G}}(s)$ as

$$\hat{\mathbf{G}}(s) = \hat{\mathbf{G}}(\infty) + \hat{\mathbf{G}}_{sp}(s) \tag{4.31}$$

where $\hat{\mathbf{G}}_{sp}$ is the strictly proper part of $\hat{\mathbf{G}}(s)$. Let

$$d(s) = s^r + \alpha_1 s^{r-1} + \cdots + \alpha_{r-1}s + \alpha_r \tag{4.32}$$

be the least common denominator of all entries of $\hat{\mathbf{G}}_{sp}(s)$. Here we require $d(s)$ to be monic; that is, its leading coefficient is 1. Then $\hat{\mathbf{G}}_{sp}(s)$ can be expressed as

$$\hat{\mathbf{G}}_{sp}(s) = \frac{1}{d(s)}[\mathbf{N}(s)] = \frac{1}{d(s)}\left[\mathbf{N}_1 s^{r-1} + \mathbf{N}_2 s^{r-2} + \cdots + \mathbf{N}_{r-1}s + \mathbf{N}_r\right] \tag{4.33}$$

where $\mathbf{N}_i$ are $q \times p$ constant matrices. Now we claim that the set of equations

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} -\alpha_1\mathbf{I}_p & -\alpha_2\mathbf{I}_p & \cdots & -\alpha_{r-1}\mathbf{I}_p & -\alpha_r\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_p & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_p & \mathbf{0} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \mathbf{I}_p \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}\mathbf{u} \\ \mathbf{y} &= \begin{bmatrix} \mathbf{N}_1 & \mathbf{N}_2 & \cdots & \mathbf{N}_{r-1} & \mathbf{N}_r \end{bmatrix}\mathbf{x} + \hat{\mathbf{G}}(\infty)\mathbf{u} \end{aligned} \tag{4.34}$$

is a realization of $\hat{\mathbf{G}}(s)$. The matrix $\mathbf{I}_p$ is the $p \times p$ unit matrix and every $\mathbf{0}$ is a $p \times p$ zero matrix. The A-matrix is said to be in block companion form; it consists of r rows and r columns of $p \times p$ matrices; thus the A-matrix has order $rp \times rp$. The B-matrix has order $rp \times p$. Because the C-matrix consists of r number of $\mathbf{N}_i$, each of order $q \times p$, the C-matrix has order $q \times rp$. The realization has dimension rp and is said to be in *controllable canonical form*.

We show that (4.34) is a realization of $\hat{\mathbf{G}}(s)$ in (4.31) and (4.33). Let us define

$$\mathbf{Z} := \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \vdots \\ \mathbf{Z}_r \end{bmatrix} := (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} \tag{4.35}$$

where $\mathbf{Z}_i$ is $p \times p$ and $\mathbf{Z}$ is $rp \times p$. Then the transfer matrix of (4.34) equals

$$\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \hat{\mathbf{G}}(\infty) = \mathbf{N}_1\mathbf{Z}_1 + \mathbf{N}_2\mathbf{Z}_2 + \cdots + \mathbf{N}_r\mathbf{Z}_r + \hat{\mathbf{G}}(\infty) \tag{4.36}$$

We write (4.35) as $(s\mathbf{I} - \mathbf{A})\mathbf{Z} = \mathbf{B}$ or

$$s\mathbf{Z} = \mathbf{A}\mathbf{Z} + \mathbf{B} \tag{4.37}$$

Using the shifting property of the companion form of $\mathbf{A}$, from the second to the last block of equations in (4.37), we can readily obtain

$$s\mathbf{Z}_2 = \mathbf{Z}_1, \quad s\mathbf{Z}_3 = \mathbf{Z}_2, \quad \cdots, \quad s\mathbf{Z}_r = \mathbf{Z}_{r-1}$$

which implies

$$\mathbf{Z}_2 = \frac{1}{s}\mathbf{Z}_1, \quad \mathbf{Z}_3 = \frac{1}{s^2}\mathbf{Z}_1, \quad \cdots, \quad \mathbf{Z}_r = \frac{1}{s^{r-1}}\mathbf{Z}_1$$

Substituting these into the first block of equations in (4.37) yields

$$\begin{aligned} s\mathbf{Z}_1 &= -\alpha_1\mathbf{Z}_1 - \alpha_2\mathbf{Z}_2 - \cdots - \alpha_r\mathbf{Z}_r + \mathbf{I}_p \\ &= -\left(\alpha_1 + \frac{\alpha_2}{s} + \cdots + \frac{\alpha_r}{s^{r-1}}\right)\mathbf{Z}_1 + \mathbf{I}_p \end{aligned}$$

or, using (4.32),

$$\left(s + \alpha_1 + \frac{\alpha_2}{s} + \cdots + \frac{\alpha_r}{s^{r-1}}\right)\mathbf{Z}_1 = \frac{d(s)}{s^{r-1}}\mathbf{Z}_1 = \mathbf{I}_p$$

Thus we have

$$\mathbf{Z}_1 = \frac{s^{r-1}}{d(s)}\mathbf{I}_p, \quad \mathbf{Z}_2 = \frac{s^{r-2}}{d(s)}\mathbf{I}_p, \quad \cdots, \quad \mathbf{Z}_r = \frac{1}{d(s)}\mathbf{I}_p$$

Substituting these into (4.36) yields

$$\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \hat{\mathbf{G}}(\infty) = \frac{1}{d(s)}[\mathbf{N}_1 s^{r-1} + \mathbf{N}_2 s^{r-2} + \cdots + \mathbf{N}_r] + \hat{\mathbf{G}}(\infty)$$

This equals $\hat{\mathbf{G}}(s)$ in (4.31) and (4.33). This shows that (4.34) is a realization of $\hat{\mathbf{G}}(s)$.

EXAMPLE 4.6 Consider the proper rational matrix

$$\begin{aligned} \hat{\mathbf{G}}(s) &= \begin{bmatrix} \dfrac{4s-10}{2s+1} & \dfrac{3}{s+2} \\ \dfrac{1}{(2s+1)(s+2)} & \dfrac{s+1}{(s+2)^2} \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \dfrac{-12}{2s+1} & \dfrac{3}{s+2} \\ \dfrac{1}{(2s+1)(s+2)} & \dfrac{s+1}{(s+2)^2} \end{bmatrix} \end{aligned} \tag{4.38}$$

where we have decomposed $\hat{\mathbf{G}}(s)$ into the sum of a constant matrix and a strictly proper rational matrix $\hat{\mathbf{G}}_{sp}(s)$. The monic least common denominator of $\hat{\mathbf{G}}_{sp}(s)$ is $d(s) = (s+0.5)(s+2)^2 = s^3 + 4.5s^2 + 6s + 2$. Thus we have

$$\begin{aligned} \hat{\mathbf{G}}_{sp}(s) &= \frac{1}{s^3 + 4.5s^2 + 6s + 2}\begin{bmatrix} -6(s+2)^2 & 3(s+2)(s+0.5) \\ 0.5(s+2) & (s+1)(s+0.5) \end{bmatrix} \\ &= \frac{1}{d(s)}\left(\begin{bmatrix} -6 & 3 \\ 0 & 1 \end{bmatrix}s^2 + \begin{bmatrix} -24 & 7.5 \\ 0.5 & 1.5 \end{bmatrix}s + \begin{bmatrix} -24 & 3 \\ 1 & 0.5 \end{bmatrix}\right) \end{aligned}$$

and a realization of (4.38) is

$$\dot{\mathbf{x}} = \left[\begin{array}{cc:cc:cc} -4.5 & 0 & -6 & 0 & -2 & 0 \\ 0 & -4.5 & 0 & -6 & 0 & -2 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{array}\right]\mathbf{x} + \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ \cdots & \cdots \\ 0 & 0 \\ 0 & 0 \\ \cdots & \cdots \\ 0 & 0 \\ 0 & 0 \end{array}\right]\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\mathbf{y} = \left[\begin{array}{cc:cc:cc} -6 & 3 & -24 & 7.5 & -24 & 3 \\ 0 & 1 & 0.5 & 1.5 & 1 & 0.5 \end{array}\right]\mathbf{x} + \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{4.39}$$

This is a six-dimensional realization.

We discuss a special case of (4.31) and (4.34) in which $p = 1$. To save space, we assume $r = 4$ and $q = 2$. However, the discussion applies to any positive integers r and q. Consider the 2×1 proper rational matrix

$$\begin{aligned} \hat{\mathbf{G}}(s) = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} &+ \frac{1}{s^4 + \alpha_1 s^3 + \alpha_2 s^2 + \alpha_3 s + \alpha_4} \\ &\cdot \begin{bmatrix} \beta_{11}s^3 + \beta_{12}s^2 + \beta_{13}s + \beta_{14} \\ \beta_{21}s^3 + \beta_{22}s^2 + \beta_{23}s + \beta_{24} \end{bmatrix} \end{aligned} \tag{4.40}$$

Then its realization can be obtained directly from (4.34) as

$$\dot{\mathbf{x}} = \begin{bmatrix} -\alpha_1 & -\alpha_2 & -\alpha_3 & -\alpha_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} u$$

$$\mathbf{y} = \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} \\ \beta_{21} & \beta_{22} & \beta_{23} & \beta_{24} \end{bmatrix} \mathbf{x} + \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} u \tag{4.41}$$

This controllable-canonical-form realization can be read out from the coefficients of $\hat{\mathbf{G}}(s)$ in (4.40).

There are many ways to realize a proper transfer matrix. For example, Problem 4.9 gives a different realization of (4.33) with dimension rq. Let $\hat{\mathbf{G}}_{ci}(s)$ be the ith column of $\hat{\mathbf{G}}(s)$ and let u_i be the ith component of the input vector $\mathbf{u}$. Then $\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}(s)\hat{\mathbf{u}}(s)$ can be expressed as

$$\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}_{c1}(s)\hat{u}_1(s) + \hat{\mathbf{G}}_{c2}(s)\hat{u}_2(s) + \cdots =: \hat{\mathbf{y}}_{c1}(s) + \hat{\mathbf{y}}_{c2}(s) + \cdots$$

as shown in Fig. 4.4(a). Thus we can realize each column of $\hat{\mathbf{G}}(s)$ and then combine them to yield a realization of $\hat{\mathbf{G}}(s)$. Let $\hat{\mathbf{G}}_{ri}(s)$ be the ith row of $\hat{\mathbf{G}}(s)$ and let y_i be the ith component of the output vector $\mathbf{y}$. Then $\hat{\mathbf{y}}(s) = \hat{\mathbf{G}}(s)\hat{\mathbf{u}}(s)$ can be expressed as

$$\hat{y}_i(s) = \hat{\mathbf{G}}_{ri}(s)\hat{\mathbf{u}}(s)$$

as shown in Fig. 4.4(b). Thus we can realize each row of $\hat{\mathbf{G}}(s)$ and then combine them to obtain a realization of $\hat{\mathbf{G}}(s)$. Clearly we can also realize each entry of $\hat{\mathbf{G}}(s)$ and then combine them to obtain a realization of $\hat{\mathbf{G}}(s)$. See Reference [6, pp. 158–160].

The MATLAB function `[a,b,c,d]=tf2ss(num,den)` generates the controllable-canonical-form realization shown in (4.41) for any single-input multiple-output transfer matrix $\hat{\mathbf{G}}(s)$. In its employment, there is no need to decompose $\hat{\mathbf{G}}(s)$ as in (4.31). But we must compute its least common denominator, not necessarily monic. The next example will apply `tf2ss` to each column of $\hat{\mathbf{G}}(s)$ in (4.38) and then combine them to form a realization of $\hat{\mathbf{G}}(s)$.

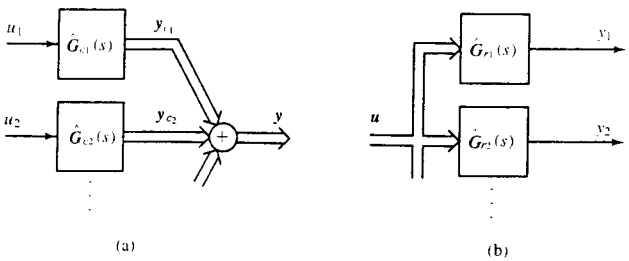


Figure 4.4 Realizations of $\hat{\mathbf{G}}(s)$ by columns and by rows.

EXAMPLE 4.7 Consider the proper rational matrix in (4.38). Its first column is

$$\hat{\mathbf{G}}_{c1}(s) = \begin{bmatrix} \dfrac{4s-10}{2s+1} \\ \dfrac{1}{(2s+1)(s+2)} \end{bmatrix} = \begin{bmatrix} \dfrac{(4s-10)(s+2)}{(2s+1)(s+2)} \\ \dfrac{1}{2s^2+5s+2} \end{bmatrix} = \begin{bmatrix} \dfrac{4s^2-2s-20}{2s^2+5s+2} \\ \dfrac{1}{2s^2+5s+2} \end{bmatrix}$$

Typing

```
n1=[4 -2 -20;0 0 1];d1=[2 5 2]; [a,b,c,d]=tf2ss(n1,d1)
```

yields the following realization for the first column of $\hat{\mathbf{G}}(s)$:

$$\dot{\mathbf{x}}_1 = \mathbf{A}_1\mathbf{x}_1 + \mathbf{b}_1u_1 = \begin{bmatrix} -2.5 & -1 \\ 1 & 0 \end{bmatrix}\mathbf{x}_1 + \begin{bmatrix} 1 \\ 0 \end{bmatrix}u_1$$

$$\mathbf{y}_{c1} = \mathbf{C}_1\mathbf{x}_1 + \mathbf{d}_1u_1 = \begin{bmatrix} -6 & -12 \\ 0 & 0.5 \end{bmatrix}\mathbf{x}_1 + \begin{bmatrix} 2 \\ 0 \end{bmatrix}u_1 \tag{4.42}$$

Similarly, the function `tf2ss` can generate the following realization for the second column of $\hat{\mathbf{G}}(s)$:

$$\dot{\mathbf{x}}_2 = \mathbf{A}_2\mathbf{x}_2 + \mathbf{b}_2u_2 = \begin{bmatrix} -4 & -4 \\ 1 & 0 \end{bmatrix}\mathbf{x}_2 + \begin{bmatrix} 1 \\ 0 \end{bmatrix}u_2$$

$$\mathbf{y}_{c2} = \mathbf{C}_2\mathbf{x}_2 + \mathbf{d}_2u_2 = \begin{bmatrix} 3 & 6 \\ 1 & 1 \end{bmatrix}\mathbf{x}_2 + \begin{bmatrix} 0 \\ 0 \end{bmatrix}u_2 \tag{4.43}$$

These two realizations can be combined as

$$\begin{bmatrix} \dot{\mathbf{x}}_1 \\ \dot{\mathbf{x}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{b}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_2 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$\mathbf{y} = \mathbf{y}_{c1} + \mathbf{y}_{c2} = [\mathbf{C}_1 \ \ \mathbf{C}_2]\mathbf{x} + [\mathbf{d}_1 \ \ \mathbf{d}_2]\mathbf{u}$$

or

$$\dot{\mathbf{x}} = \begin{bmatrix} -2.5 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & 1 & 0 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}\mathbf{u}$$

$$\mathbf{y} = \begin{bmatrix} -6 & -12 & 3 & 6 \\ 0 & 0.5 & 1 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}\mathbf{u} \tag{4.44}$$

This is a different realization of the $\hat{\mathbf{G}}(s)$ in (4.38). This realization has dimension 4, two less than the one in (4.39).

The two state equations in (4.39) and (4.44) are zero-state equivalent because they have the same transfer matrix. They are, however, not algebraically equivalent. More will be said

in Chapter 7 regarding realizations. We mention that all discussion in this section, including `tf2ss`, applies without any modification to the discrete-time case

4.5 Solution of Linear Time-Varying (LTV) Equations

Consider the linear time-varying (LTV) state equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{4.45}$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \tag{4.46}$$

It is assumed that, for every initial state $\mathbf{x}(t_0)$ and any input $\mathbf{u}(t)$, the state equation has a unique solution. A sufficient condition for such an assumption is that every entry of $\mathbf{A}(t)$ is a continuous function of t. Before considering the general case, we first discuss the solutions of $\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t)$ and the reasons why the approach taken in the time-invariant case cannot be used here.

The solution of the time-invariant equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ can be extended from the scalar equation $\dot{x} = ax$. The solution of $\dot{x} = ax$ is $x(t) = e^{at}x(0)$ with $d(e^{at})/dt = ae^{at} = e^{at}a$. Similarly, the solution of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is $\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0)$ with

$$\frac{d}{dt}e^{\mathbf{A}t} = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}$$

where the commutative property is crucial. Note that, in general, we have $\mathbf{AB} \neq \mathbf{BA}$ and $e^{(\mathbf{A}+\mathbf{B})t} \neq e^{\mathbf{A}t}e^{\mathbf{B}t}$.

The solution of the scalar time-varying equation $\dot{x} = a(t)x$ due to $x(0)$ is

$$x(t) = e^{\int_0^t a(\tau)d\tau}x(0)$$

with

$$\frac{d}{dt}e^{\int_0^t a(\tau)d\tau} = a(t)e^{\int_0^t a(\tau)d\tau} = e^{\int_0^t a(\tau)d\tau}a(t)$$

Extending this to the matrix case becomes

$$\mathbf{x}(t) = e^{\int_0^t \mathbf{A}(\tau)d\tau}\mathbf{x}(0) \tag{4.47}$$

with, using (3.51),

$$e^{\int_0^t \mathbf{A}(\tau)d\tau} = \mathbf{I} + \int_0^t \mathbf{A}(\tau)d\tau + \frac{1}{2}\left(\int_0^t \mathbf{A}(\tau)d\tau\right)\left(\int_0^t \mathbf{A}(s)ds\right) + \cdots$$

This extension, however, is not valid because

$$\frac{d}{dt}e^{\int_0^t \mathbf{A}(\tau)d\tau} = \mathbf{A}(t) + \frac{1}{2}\mathbf{A}(t)\left(\int_0^t \mathbf{A}(s)ds\right) + \frac{1}{2}\left(\int_0^t \mathbf{A}(\tau)d\tau\right)\mathbf{A}(t) + \cdots$$

$$\neq \mathbf{A}(t)e^{\int_0^t \mathbf{A}(\tau)d\tau} \tag{4.48}$$

Thus, in general, (4.47) is not a solution of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$. In conclusion, we cannot extend the

solution of scalar time-varying equations to the matrix case and must use a different approach to develop the solution.

Consider

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} \tag{4.49}$$

where $\mathbf{A}$ is $n \times n$ with continuous functions of t as its entries. Then for every initial state $\mathbf{x}_i(t_0)$, there exists a unique solution $\mathbf{x}_i(t)$, for $i = 1, 2, \ldots, n$. We can arrange these n solutions as $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n]$, a square matrix of order n. Because every $\mathbf{x}_i$ satisfies (4.49), we have

$$\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t) \tag{4.50}$$

If $\mathbf{X}(t_0)$ is nonsingular or the n initial states are linearly independent, then $\mathbf{X}(t)$ is called a *fundamental matrix* of (4.49). Because the initial states can arbitrarily be chosen, as long as they are linearly independent, the fundamental matrix is not unique.

EXAMPLE 4.8 Consider the homogeneous equation

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 0 \\ t & 0 \end{bmatrix}\mathbf{x}(t) \tag{4.51}$$

or

$$\dot{x}_1(t) = 0 \qquad \dot{x}_2(t) = tx_1(t)$$

The solution of $\dot{x}_1(t) = 0$ for $t_0 = 0$ is $x_1(t) = x_1(0)$; the solution of $\dot{x}_2(t) = tx_1(t) = tx_1(0)$ is

$$x_2(t) = \int_0^t \tau x_1(0)d\tau + x_2(0) = 0.5t^2x_1(0) + x_2(0)$$

Thus we have

$$\mathbf{x}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Rightarrow \mathbf{x}(t) = \begin{bmatrix} 1 \\ 0.5t^2 \end{bmatrix}$$

and

$$\mathbf{x}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \Rightarrow \mathbf{x}(t) = \begin{bmatrix} 1 \\ 0.5t^2 + 2 \end{bmatrix}$$

The two initial states are linearly independent; thus

$$\mathbf{X}(t) = \begin{bmatrix} 1 & 1 \\ 0.5t^2 & 0.5t^2 + 2 \end{bmatrix} \tag{4.52}$$

is a fundamental matrix.

A very important property of the fundamental matrix is that $\mathbf{X}(t)$ is nonsingular for all t. For example, $\mathbf{X}(t)$ in (4.52) has determinant $0.5t^2 + 2 - 0.5t^2 = 2$; thus it is nonsingular for all t. We argue intuitively why this is the case. If $\mathbf{X}(t)$ is singular at some t_1, then there exists a nonzero vector $\mathbf{v}$ such that $\mathbf{x}(t_1) := \mathbf{X}(t_1)\mathbf{v} = \mathbf{0}$, which, in turn, implies $\mathbf{x}(t) := \mathbf{X}(t)\mathbf{v} \equiv \mathbf{0}$ for all t, in particular, at $= t_0$. This is a contradiction. Thus $\mathbf{X}(t)$ is nonsingular for all t.

Definition 4.2 *Let* $\mathbf{X}(t)$ *be any fundamental matrix of* $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$. *Then*

$$\boldsymbol{\Phi}(t, t_0) := \mathbf{X}(t)\mathbf{X}^{-1}(t_0)$$

is called the state transition matrix *of* $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$. *The state transition matrix is also the unique solution of*

$$\frac{\partial}{\partial t}\boldsymbol{\Phi}(t, t_0) = \mathbf{A}(t)\boldsymbol{\Phi}(t, t_0) \tag{4.53}$$

with the initial condition $\boldsymbol{\Phi}(t_0, t_0) = \mathbf{I}$.

Because $\mathbf{X}(t)$ is nonsingular for all t, its inverse is well defined. Equation (4.53) follows directly from (4.50). From the definition, we have the following important properties of the state transition matrix:

$$\boldsymbol{\Phi}(t, t) = \mathbf{I} \tag{4.54}$$

$$\boldsymbol{\Phi}^{-1}(t, t_0) = [\mathbf{X}(t)\mathbf{X}^{-1}(t_0)]^{-1} = \mathbf{X}(t_0)\mathbf{X}^{-1}(t) = \boldsymbol{\Phi}(t_0, t) \tag{4.55}$$

$$\boldsymbol{\Phi}(t, t_0) = \boldsymbol{\Phi}(t, t_1)\boldsymbol{\Phi}(t_1, t_0) \tag{4.56}$$

for every t, t_0, and t_1.

EXAMPLE 4.9 Consider the homogeneous equation in Example 4.8. Its fundamental matrix was computed as

$$\mathbf{X}(t) = \begin{bmatrix} 1 & 1 \\ 0.5t^2 & 0.5t^2 + 2 \end{bmatrix}$$

Its inverse is, using (3.20),

$$\mathbf{X}^{-1}(t) = \begin{bmatrix} 0.25t^2 + 1 & -0.5 \\ -0.25t^2 & 0.5 \end{bmatrix}$$

Thus the state transition matrix is given by

$$\begin{aligned} \boldsymbol{\Phi}(t, t_0) &= \begin{bmatrix} 1 & 1 \\ 0.5t^2 & 0.5t^2 + 2 \end{bmatrix} \begin{bmatrix} 0.25t_0^2 + 1 & -0.5 \\ -0.25t_0^2 & 0.5 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0.5(t^2 - t_0^2) & 1 \end{bmatrix} \end{aligned}$$

It is straightforward to verify that this transition matrix satisfies (4.53) and has the three properties listed in (4.54) through (4.56).

Now we claim that the solution of (4.45) excited by the initial state $\mathbf{x}(t_0) = \mathbf{x}_0$ and the input $\mathbf{u}(t)$ is given by

$$\mathbf{x}(t) = \boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \tag{4.57}$$

$$= \boldsymbol{\Phi}(t, t_0)\left[\mathbf{x}_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(t_0, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau\right] \tag{4.58}$$

where $\boldsymbol{\Phi}(t, \tau)$ is the state transition matrix of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$. Equation (4.58) follows from (4.57) by using $\boldsymbol{\Phi}(t, \tau) = \boldsymbol{\Phi}(t, t_0)\boldsymbol{\Phi}(t_0, \tau)$. We show that (4.57) satisfies the initial condition and the state equaion. At $t = t_0$, we have

$$\mathbf{x}(t_0) = \boldsymbol{\Phi}(t_0, t_0)\mathbf{x}_0 + \int_{t_0}^{t_0} \boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau = \mathbf{I}\mathbf{x}_0 + \mathbf{0} = \mathbf{x}_0$$

Thus (4.57) satisfies the initial condition. Using (4.53) and (4.6), we have

$$\begin{aligned} \frac{d}{dt}\mathbf{x}(t) &= \frac{\partial}{\partial t}\boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \frac{\partial}{\partial t}\int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \\ &= \mathbf{A}(t)\boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \left(\frac{\partial}{\partial t}\boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\right) d\tau + \boldsymbol{\Phi}(t, t)\mathbf{B}(t)\mathbf{u}(t) \\ &= \mathbf{A}(t)\boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \mathbf{A}(t)\boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau + \mathbf{B}(t)\mathbf{u}(t) \\ &= \mathbf{A}(t)\left[\boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau\right] + \mathbf{B}(t)\mathbf{u}(t) \\ &= \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \end{aligned}$$

Thus (4.57) is the solution. Substituting (4.57) into (4.46) yields

$$\mathbf{y}(t) = \mathbf{C}(t)\boldsymbol{\Phi}(t, t_0)\mathbf{x}_0 + \mathbf{C}(t)\int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau + \mathbf{D}(t)\mathbf{u}(t) \tag{4.59}$$

If the input is identically zero, then Equation (4.57) reduces to

$$\mathbf{x}(t) = \boldsymbol{\Phi}(t, t_0)\mathbf{x}_0$$

This is the zero-input response. Thus the state transition matrix governs the unforced propagation of the state vector. If the initial state is zero, then (4.59) reduces to

$$\begin{aligned} \mathbf{y}(t) &= \mathbf{C}(t)\int_{t_0}^{t} \boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau + \mathbf{D}(t)\mathbf{u}(t) \\ &= \int_{t_0}^{t} [\mathbf{C}(t)\boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau) + \mathbf{D}\delta(t - \tau)]\,\mathbf{u}(\tau)\, d\tau \end{aligned} \tag{4.60}$$

This is the zero-state response. As discussed in (2.5), the zero-state response can be described by

$$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{G}(t, \tau)\mathbf{u}(\tau)\, d\tau \tag{4.61}$$

where $\mathbf{G}(t, \tau)$ is the impulse response matrix and is the output at time t excited by an impulse input applied at time τ. Comparing (4.60) and (4.61) yields

$$\begin{aligned} \mathbf{G}(t, \tau) &= \mathbf{C}(t)\boldsymbol{\Phi}(t, \tau)\mathbf{B}(\tau) + \mathbf{D}(t)\delta(t - \tau) \\ &= \mathbf{C}(t)\mathbf{X}(t)\mathbf{X}^{-1}(\tau)\mathbf{B}(\tau) + \mathbf{D}(t)\delta(t - \tau) \end{aligned} \tag{4.62}$$

This relates the input–output and state-space descriptions.

The solutions in (4.57) and (4.59) hinge on solving (4.49) or (4.53). If $\mathbf{A}(t)$ is triangular such as

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} a_{11}(t) & 0 \\ a_{21}(t) & a_{22}(t) \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

we can solve the scalar equation $\dot{x}_1(t) = a_{11}(t)x_1(t)$ and then substitute it into

$$\dot{x}_2(t) = a_{22}(t)x_2(t) + a_{21}(t)x_1(t)$$

Because $x_1(t)$ has been solved, the preceding scalar equation can be solved for $x_2(t)$. This is what we did in Example 4.8. If $\mathbf{A}(t)$, such as $\mathbf{A}(t)$ diagonal or constant, has the commutative property

$$\mathbf{A}(t)\left(\int_{t_0}^{t} \mathbf{A}(\tau)\,d\tau\right) = \left(\int_{t_0}^{t} \mathbf{A}(\tau)\,d\tau\right)\mathbf{A}(t)$$

for all t_0 and t, then the solution of (4.53) can be shown to be

$$\mathbf{\Phi}(t, t_0) = e^{\int_{t_0}^{t} \mathbf{A}(\tau)\,d\tau} = \sum_{k=0}^{\infty} \frac{1}{k!}\left(\int_{t_0}^{t} \mathbf{A}(\tau)\,d\tau\right)^k \tag{4.63}$$

For $\mathbf{A}(t)$ constant, (4.63) reduces to

$$\mathbf{\Phi}(t, \tau) = e^{\mathbf{A}(t-\tau)} = \mathbf{\Phi}(t - \tau)$$

and $\mathbf{X}(t) = e^{\mathbf{A}t}$. Other than the preceding special cases, computing state transition matrices is generally difficult.

4.5.1 Discrete-Time Case

Consider the discrete-time state equation

$$\mathbf{x}[k+1] = \mathbf{A}[k]\mathbf{x}[k] + \mathbf{B}[k]\mathbf{u}[k] \tag{4.64}$$

$$\mathbf{y}[k] = \mathbf{C}[k]\mathbf{x}[k] + \mathbf{D}[k]\mathbf{u}[k] \tag{4.65}$$

The set consists of algebraic equations and their solutions can be computed recursively once the initial state $\mathbf{x}[k_0]$ and the input $\mathbf{u}[k]$, for $k \geq k_0$, are given. The situation here is much simpler than the continuous-time case.

As in the continuous-time case, we can define the discrete state transition matrix as the solution of

$$\mathbf{\Phi}[k+1, k_0] = \mathbf{A}[k]\mathbf{\Phi}[k, k_0] \qquad \text{with } \mathbf{\Phi}[k_0, k_0] = \mathbf{I}$$

for $k = k_0,\ k_0 + 1,\ \ldots$. This is the discrete counterpart of (4.53) and its solution can be obtained directly as

$$\mathbf{\Phi}[k, k_0] = \mathbf{A}[k-1]\mathbf{A}[k-2]\cdots\mathbf{A}[k_0] \tag{4.66}$$

for $k > k_0$ and $\mathbf{\Phi}[k_0, k_0] = \mathbf{I}$. We discuss a significant difference between the continuous- and discrete-time cases. Because the fundamental matrix in the continuous-time case is nonsingular

for all t, the state transition matrix $\mathbf{\Phi}(t, t_0)$ is defined for $t \geq t_0$ and $t < t_0$ and can govern the propagation of the state vector in the positive-time and negative-time directions. In the discrete-time case, the A-matrix can be singular; thus the inverse of $\mathbf{\Phi}[k, k_0]$ may not be defined. Thus $\mathbf{\Phi}[k, k_0]$ is defined only for $k \geq k_0$ and governs the propagation of the state vector in only the positive-time direction. Therefore the discrete counterpart of (4.56) or

$$\mathbf{\Phi}[k, k_0] = \mathbf{\Phi}[k, k_1]\mathbf{\Phi}[k_1, k_0]$$

holds only for $k \geq k_1 \geq k_0$.

Using the discrete state transition matrix, we can express the solutions of (4.64) and (4.65) as, for $k > k_0$,

$$\begin{aligned} \mathbf{x}[k] &= \mathbf{\Phi}[k, k_0]\mathbf{x}_0 + \sum_{m=k_0}^{k-1} \mathbf{\Phi}[k, m+1]\mathbf{B}[m]\mathbf{u}[m] \\ \mathbf{y}[k] &= \mathbf{C}[k]\mathbf{\Phi}[k, k_0]\mathbf{x}_0 + \mathbf{C}[k]\sum_{m=k_0}^{k-1} \mathbf{\Phi}[k, m+1]\mathbf{B}[m]\mathbf{u}[m] + \mathbf{D}[k]\mathbf{u}[k] \end{aligned} \tag{4.67}$$

Their derivations are similar to those of (4.20) and (4.21) and will not be repeated.

If the initial state is zero, Equation (4.67) reduces to

$$\mathbf{y}[k] = \mathbf{C}[k]\sum_{m=k_0}^{k-1} \mathbf{\Phi}[k, m+1]\mathbf{B}[m]\mathbf{u}[m] + \mathbf{D}[k]\mathbf{u}[k] \tag{4.68}$$

for $k > k_0$. This describes the zero-state response of (4.65). If we define $\mathbf{\Phi}[k, m] = \mathbf{0}$ for $k < m$, then (4.68) can be written as

$$\mathbf{y}[k] = \sum_{m=k_0}^{k} \left(\mathbf{C}[k]\mathbf{\Phi}[k, m+1]\mathbf{B}[m] + \mathbf{D}[m]\delta[k-m]\right)\mathbf{u}[m]$$

where the impulse sequence $\delta[k-m]$ equals 1 if $k = m$ and 0 if $k \neq m$. Comparing this with the multivariable version of (2.34), we have

$$\mathbf{G}[k, m] = \mathbf{C}[k]\mathbf{\Phi}[k, m+1]\mathbf{B}[m] + \mathbf{D}[m]\delta[k-m]$$

for $k \geq m$. This relates the impulse response sequence and the state equation and is the discrete counterpart of (4.62).

4.6 Equivalent Time-Varying Equations

This section extends the equivalent state equations discussed in Section 4.3 to the time-varying case. Consider the n-dimensional linear time-varying state equation

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \\ \mathbf{y} &= \mathbf{C}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u} \end{aligned} \tag{4.69}$$

Let $\mathbf{P}(t)$ be an $n \times n$ matrix. It is assumed that $\mathbf{P}(t)$ is nonsingular and both $\mathbf{P}(t)$ and $\dot{\mathbf{P}}(t)$ are continuous for all t. Let $\bar{\mathbf{x}} = \mathbf{P}(t)\mathbf{x}$. Then the state equation

$$\begin{aligned} \dot{\bar{\mathbf{x}}} &= \bar{\mathbf{A}}(t)\bar{\mathbf{x}} + \bar{\mathbf{B}}(t)\mathbf{u} \\ \mathbf{y} &= \bar{\mathbf{C}}(t)\bar{\mathbf{x}} + \bar{\mathbf{D}}(t)\mathbf{u} \end{aligned} \tag{4.70}$$

where

$$\begin{aligned} \bar{\mathbf{A}}(t) &= [\mathbf{P}(t)\mathbf{A}(t) + \dot{\mathbf{P}}(t)]\mathbf{P}^{-1}(t) \\ \bar{\mathbf{B}}(t) &= \mathbf{P}(t)\mathbf{B}(t) \\ \bar{\mathbf{C}}(t) &= \mathbf{C}(t)\mathbf{P}^{-1}(t) \\ \bar{\mathbf{D}}(t) &= \mathbf{D}(t) \end{aligned}$$

is said to be (algebraically) equivalent to (4.69) and $\mathbf{P}(t)$ is called an (*algebraic*) *equivalence* transformation.

Equation (4.70) is obtained from (4.69) by substituting $\bar{\mathbf{x}} = \mathbf{P}(t)\mathbf{x}$ and $\dot{\bar{\mathbf{x}}} = \dot{\mathbf{P}}(t)\mathbf{x} + \mathbf{P}(t)\dot{\mathbf{x}}$. Let $\mathbf{X}$ be a fundamental matrix of (4.69). Then we claim that

$$\bar{\mathbf{X}}(t) := \mathbf{P}(t)\mathbf{X}(t) \tag{4.71}$$

is a fundamental matrix of (4.70). By definition, $\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t)$ and $\mathbf{X}(t)$ is nonsingular for all t. Because the rank of a matrix will not change by multiplying a nonsingular matrix, the matrix $\mathbf{P}(t)\mathbf{X}(t)$ is also nonsingular for all t. Now we show that $\mathbf{P}(t)\mathbf{X}(t)$ satisfies the equation $\dot{\bar{\mathbf{x}}} = \bar{\mathbf{A}}(t)\bar{\mathbf{x}}$. Indeed, we have

$$\begin{aligned} \frac{d}{dt}[\mathbf{P}(t)\mathbf{X}(t)] &= \dot{\mathbf{P}}(t)\mathbf{X}(t) + \mathbf{P}(t)\dot{\mathbf{X}}(t) = \dot{\mathbf{P}}(t)\mathbf{X}(t) + \mathbf{P}(t)\mathbf{A}(t)\mathbf{X}(t) \\ &= [\dot{\mathbf{P}}(t) + \mathbf{P}(t)\mathbf{A}(t)][\mathbf{P}^{-1}(t)\mathbf{P}(t)]\mathbf{X}(t) = \bar{\mathbf{A}}(t)[\mathbf{P}(t)\mathbf{X}(t)] \end{aligned}$$

Thus $\mathbf{P}(t)\mathbf{X}(t)$ is a fundamental matrix of $\dot{\bar{\mathbf{x}}}(t) = \bar{\mathbf{A}}(t)\bar{\mathbf{x}}(t)$.

Theorem 4.3

Let $\mathbf{A}_o$ be an arbitrary constant matrix. Then there exists an equivalence transformation that transforms (4.69) into (4.70) with $\bar{\mathbf{A}}(t) = \mathbf{A}_o$.

Proof: Let $\mathbf{X}(t)$ be a fundamental matrix of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$. The differentiation of $\mathbf{X}^{-1}(t)\mathbf{X}(t) = \mathbf{I}$ yields

$$\dot{\mathbf{X}}^{-1}(t)\mathbf{X}(t) + \mathbf{X}^{-1}\dot{\mathbf{X}}(t) = \mathbf{0}$$

which implies

$$\dot{\mathbf{X}}^{-1}(t) = -\mathbf{X}^{-1}(t)\mathbf{A}(t)\mathbf{X}(t)\mathbf{X}^{-1}(t) = -\mathbf{X}^{-1}(t)\mathbf{A}(t) \tag{4.72}$$

Because $\bar{\mathbf{A}}(t) = \mathbf{A}_o$ is a constant matrix, $\bar{\mathbf{X}}(t) = e^{\mathbf{A}_o t}$ is a fundamental matrix of $\dot{\bar{\mathbf{x}}} = \bar{\mathbf{A}}(t)\bar{\mathbf{x}} = \mathbf{A}_o\bar{\mathbf{x}}$. Following (4.71), we define

$$\mathbf{P}(t) := \bar{\mathbf{X}}(t)\mathbf{X}^{-1}(t) = e^{\mathbf{A}_o t}\mathbf{X}^{-1}(t) \tag{4.73}$$

and compute

$$\begin{aligned} \bar{\mathbf{A}}(t) &= [\mathbf{P}(t)\mathbf{A}(t) + \dot{\mathbf{P}}(t)]\mathbf{P}^{-1}(t) \\ &= [e^{\mathbf{A}_o t}\mathbf{X}^{-1}(t)\mathbf{A}(t) + \mathbf{A}_o e^{\mathbf{A}_o t}\mathbf{X}^{-1}(t) + e^{\mathbf{A}_o t}\dot{\mathbf{X}}^{-1}(t)]\mathbf{X}(t)e^{-\mathbf{A}_o t} \end{aligned}$$

which becomes, after substituting (4.72),

$$\bar{\mathbf{A}}(t) = \mathbf{A}_o e^{\mathbf{A}_o t}\mathbf{X}^{-1}(t)\mathbf{X}(t)e^{-\mathbf{A}_o t} = \mathbf{A}_o$$

This establishes the theorem. Q.E.D.

If $\mathbf{A}_o$ is chosen as a zero matrix, then $\mathbf{P}(t) = \mathbf{X}^{-1}(t)$ and (4.70) reduces to

$$\bar{\mathbf{A}}(t) = \mathbf{0} \qquad \bar{\mathbf{B}}(t) = \mathbf{X}^{-1}(t)\mathbf{B}(t) \qquad \bar{\mathbf{C}}(t) = \mathbf{C}(t)\mathbf{X}(t) \qquad \bar{\mathbf{D}}(t) = \mathbf{D}(t) \tag{4.74}$$

The block diagrams of (4.69) with $\mathbf{A}(t) \neq \mathbf{0}$ and $\mathbf{A}(t) = \mathbf{0}$ are plotted in Fig. 4.5. The block diagram with $\mathbf{A}(t) = \mathbf{0}$ has no feedback and is considerably simpler. Every time-varying state equation can be transformed into such a block diagram. However, in order to do so, we must know its fundamental matrix.

The impulse response matrix of (4.69) is given in (4.62). The impulse response matrix of (4.70) is, using (4.71) and (4.72),

$$\bar{\mathbf{G}}(t, \tau) = \bar{\mathbf{C}}(t)\bar{\mathbf{X}}(t)\bar{\mathbf{X}}^{-1}(\tau)\bar{\mathbf{B}}(\tau) + \bar{\mathbf{D}}(t)\delta(t - \tau)$$

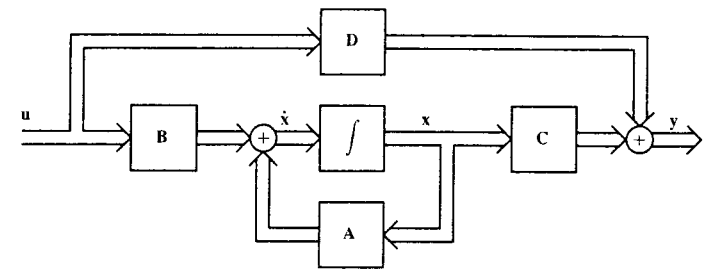


Figure 4.5 Block daigrams with feedback and without feedback.

$$= \mathbf{C}(t)\mathbf{P}^{-1}(t)\mathbf{P}(t)\mathbf{X}(t)\mathbf{X}^{-1}(\tau)\mathbf{P}^{-1}(\tau)\mathbf{P}(\tau)\mathbf{B}(\tau) + \mathbf{D}(t)\delta(t-\tau)$$
$$= \mathbf{C}(t)\mathbf{X}(t)\mathbf{X}^{-1}(\tau)\mathbf{B}(\tau) + \mathbf{D}(t)\delta(t-\tau) = \mathbf{G}(t,\tau)$$

Thus the impulse response matrix is invariant under any equivalence transformation. The property of the A-matrix, however, may not be preserved in equivalence transformations. For example, every A-matrix can be transformed, as shown in Theorem 4.3, into a constant or a zero matrix. Clearly the zero matrix does not have any property of $\mathbf{A}(t)$. In the time-invariant case, equivalence transformations will preserve all properties of the original state equation. Thus the equivalence transformation in the time-invariant case is not a special case of the time-varying case.

Definition 4.3 *A matrix* $\mathbf{P}(t)$ *is called a* Lyapunov transformation *if* $\mathbf{P}(t)$ *is nonsingular,* $\mathbf{P}(t)$ *and* $\dot{\mathbf{P}}(t)$ *are continuous, and* $\mathbf{P}(t)$ *and* $\mathbf{P}^{-1}(t)$ *are bounded for all* t. *Equations (4.69) and (4.70) are said to be* Lyapunov equivalent *if* $\mathbf{P}(t)$ *is a Lyapunov transformation.*

It is clear that if $\mathbf{P}(t) = \mathbf{P}$ is a constant matrix, then it is a Lyapunov transformation. Thus the (algebraic) transformation in the time-invariant case is a special case of the Lyapunov transformation. If $\mathbf{P}(t)$ is required to be a Lyapunov transformation, then Theorem 4.3 does not hold in general. In other words, not every time-varying state equation can be Lyapunov equivalent to a state equation with a constant A-matrix. However, this is true if $\mathbf{A}(t)$ is periodic.

Periodic state equations Consider the linear time-varying state equation in (4.69). It is assumed that

$$\mathbf{A}(t+T) = \mathbf{A}(t)$$

for all t and for some positive constant T. That is, $\mathbf{A}(t)$ is periodic with period T. Let $\mathbf{X}(t)$ be a fundamental matrix of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ or $\dot{\mathbf{X}}(t) = \mathbf{A}(t)\mathbf{X}(t)$ with $\mathbf{X}(0)$ nonsingular. Then we have

$$\dot{\mathbf{X}}(t+T) = \mathbf{A}(t+T)\mathbf{X}(t+T) = \mathbf{A}(t)\mathbf{X}(t+T)$$

Thus $\mathbf{X}(t+T)$ is also a fundamental matrix. Furthermore, it can be expressed as

$$\mathbf{X}(t+T) = \mathbf{X}(t)\mathbf{X}^{-1}(0)\mathbf{X}(T) \tag{4.75}$$

This can be verified by direct substitution. Let us define $\mathbf{Q} = \mathbf{X}^{-1}(0)\mathbf{X}(T)$. It is a constant nonsingular matrix. For this $\mathbf{Q}$ there exists a constant matrix $\bar{\mathbf{A}}$ such that $e^{\bar{\mathbf{A}}T} = \mathbf{Q}$ (Problem 3.24). Thus (4.75) can be written as

$$\mathbf{X}(t+T) = \mathbf{X}(t)e^{\bar{\mathbf{A}}T} \tag{4.76}$$

Define

$$\mathbf{P}(t) := e^{\bar{\mathbf{A}}t}\mathbf{X}^{-1}(t) \tag{4.77}$$

We show that $\mathbf{P}(t)$ is periodic with period T:

$$\mathbf{P}(t+T) = e^{\bar{\mathbf{A}}(t+T)}\mathbf{X}^{-1}(t+T) = e^{\bar{\mathbf{A}}t}e^{\bar{\mathbf{A}}T}[e^{-\bar{\mathbf{A}}T}\mathbf{X}^{-1}(t)]$$
$$= e^{\bar{\mathbf{A}}t}\mathbf{X}^{-1}(t) = \mathbf{P}(t)$$

▶ **Theorem 4.4**

Consider (4.69) with $\mathbf{A}(t) = \mathbf{A}(t+T)$ for all t and some $T > 0$. Let $\mathbf{X}(t)$ be a fundamental matrix of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$. Let $\bar{\mathbf{A}}$ be the constant matrix computed from $e^{\bar{\mathbf{A}}T} = \mathbf{X}^{-1}(0)\mathbf{X}(T)$. Then (4.69) is Lyapunov equivalent to

$$\dot{\bar{\mathbf{x}}}(t) = \bar{\mathbf{A}}\bar{\mathbf{x}}(t) + \mathbf{P}(t)\mathbf{B}(t)\mathbf{u}(t)$$
$$\bar{\mathbf{y}}(t) = \mathbf{C}(t)\mathbf{P}^{-1}(t)\bar{\mathbf{x}}(t) + \mathbf{D}(t)\mathbf{u}(t)$$

where $\mathbf{P}(t) = e^{\bar{\mathbf{A}}t}\mathbf{X}^{-1}(t)$.

The matrix $\mathbf{P}(t)$ in (4.77) satisfies all conditions in Definition 4.3; thus it is a Lyapunov transformation. The rest of the theorem follows directly from Theorem 4.3. The homogeneous part of Theorem 4.4 is the *theory of Floquet*. It states that if $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ and if $\mathbf{A}(t+T) = \mathbf{A}(t)$ for all t, then its fundamental matrix is of the form $\mathbf{P}^{-1}(t)e^{\bar{\mathbf{A}}t}$, where $\mathbf{P}^{-1}(t)$ is a periodic function. Furthermore, $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ is Lyapunov equivalent to $\dot{\bar{\mathbf{x}}} = \bar{\mathbf{A}}\bar{\mathbf{x}}$.

4.7 Time-Varying Realizations

We studied in Section 4.4 the realization problem for linear time-invariant systems. In this section, we study the corresponding problem for linear time-varying systems. The Laplace transform cannot be used here; therefore we study the problem directly in the time domain.

Every linear time-varying system can be described by the input–output description

$$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{G}(t,\tau)\mathbf{u}(\tau)\,d\tau$$

and, if the system is lumped as well, by the state equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t)$$
$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \tag{4.78}$$

If the state equation is available, the impulse response matrix can be computed from

$$\mathbf{G}(t,\tau) = \mathbf{C}(t)\mathbf{X}(t)\mathbf{X}^{-1}(\tau)\mathbf{B}(\tau) + \mathbf{D}(t)\delta(t-\tau) \qquad \text{for } t \geq \tau \tag{4.79}$$

where $\mathbf{X}(t)$ is a fundamental matrix of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$. The converse problem is to find a state equation from a given impulse response matrix. An impulse response matrix $\mathbf{G}(t,\tau)$ is said to be *realizable* if there exists $\{\mathbf{A}(t), \mathbf{B}(t), \mathbf{C}(t), \mathbf{D}(t)\}$ to meet (4.79).

▶ **Theorem 4.5**

A $q \times p$ impulse response matrix $\mathbf{G}(t,\tau)$ is realizable if and only if $\mathbf{G}(t,\tau)$ can be decomposed as

$$\mathbf{G}(t,\tau) = \mathbf{M}(t)\mathbf{N}(\tau) + \mathbf{D}(t)\delta(t-\tau) \tag{4.80}$$

for all $t \geq \tau$, where $\mathbf{M}$, $\mathbf{N}$, and $\mathbf{D}$ are, respectively, $q \times n$, $n \times p$, and $q \times p$ matrices for some integer n.

Proof: If $\mathbf{G}(t, \tau)$ is realizable, there exists a realization that meets (4.79). Identifying $\mathbf{M}(t) = \mathbf{C}(t)\mathbf{X}(t)$ and $\mathbf{N}(\tau) = \mathbf{X}^{-1}(\tau)\mathbf{B}(\tau)$ establishes the necessary part of the theorem.

If $\mathbf{G}(t, \tau)$ can be decomposed as in (4.80), then the n-dimensional state equation

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{N}(t)\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{M}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{u}(t) \end{aligned} \tag{4.81}$$

is a realization. Indeed, a fundamental matrix of $\dot{\mathbf{x}} = \mathbf{0} \cdot \mathbf{x}$ is $\mathbf{X}(t) = \mathbf{I}$. Thus the impulse response matrix of (4.81) is

$$\mathbf{M}(t)\mathbf{I} \cdot \mathbf{I}^{-1}\mathbf{N}(\tau) + \mathbf{D}(t)\delta(t - \tau)$$

which equals $\mathbf{G}(t, \tau)$. This shows the sufficiency of the theorem. Q.E.D.

Although Theorem 4.5 can also be applied to time-invariant systems, the result is not useful in practical implementation, as the next example illustrates.

EXAMPLE 4.10 Consider $g(t) = te^{\lambda t}$ or

$$g(t, \tau) = g(t - \tau) = (t - \tau)e^{\lambda(t-\tau)}$$

It is straightforward to verify

$$g(t - \tau) = [e^{\lambda t} \quad te^{\lambda t}] \begin{bmatrix} -\tau e^{-\lambda \tau} \\ e^{-\lambda \tau} \end{bmatrix}$$

Thus the two-dimensional time-varying state equation

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} -te^{-\lambda t} \\ e^{-\lambda t} \end{bmatrix} u(t) \\ y(t) &= [e^{\lambda t} \quad te^{\lambda t}]\mathbf{x}(t) \end{aligned} \tag{4.82}$$

is a realization of the impulse response $g(t) = te^{\lambda t}$.

The Laplace transform of the impulse response is

$$\hat{g}(s) = \mathcal{L}[te^{\lambda t}] = \frac{1}{(s - \lambda)^2} = \frac{1}{s^2 - 2\lambda s + \lambda^2}$$

Using (4.41), we can readily obtain

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \begin{bmatrix} 2\lambda & -\lambda^2 \\ 1 & 0 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t) \\ y(t) &= [0 \quad 1]\mathbf{x}(t) \end{aligned} \tag{4.83}$$

This LTI state equation is a different realization of the same impulse response. This realization is clearly more desirable because it can readily be implemented using an op-amp circuit. The implementation of (4.82) is much more difficult in practice.

PROBLEMS

4.1 An oscillation can be generated by

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \mathbf{x}$$

Show that its solution is

$$\mathbf{x}(t) = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \mathbf{x}(0)$$

4.2 Use two different methods to find the unit-step response of

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 0 & 1 \\ -2 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u \\ y &= [2 \quad 3]\mathbf{x} \end{aligned}$$

4.3 Discretize the state equation in Problem 4.2 for $T = 1$ and $T = \pi$.

4.4 Find the companion-form and modal-form equivalent equations of

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} -2 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & -2 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} u \\ y &= [1 \quad -1 \quad 0]\mathbf{x} \end{aligned}$$

4.5 Find an equivalent state equation of the equation in Problem 4.4 so that all state variables have their largest magnitudes roughly equal to the largest magnitude of the output. If all signals are required to lie inside ± 10 volts and if the input is a step function with magnitude a, what is the permissible largest a?

4.6 Consider

$$\dot{\mathbf{x}} = \begin{bmatrix} \lambda & 0 \\ 0 & \bar{\lambda} \end{bmatrix} \mathbf{x} + \begin{bmatrix} b_1 \\ \bar{b}_1 \end{bmatrix} u \qquad y = [c_1 \quad \bar{c}_1]\mathbf{x}$$

where the overbar denotes complex conjugate. Verify that the equation can be transformed into

$$\dot{\bar{\mathbf{x}}} = \bar{\mathbf{A}}\bar{\mathbf{x}} + \bar{\mathbf{b}}u \qquad y = \bar{\mathbf{c}}\bar{\mathbf{x}}$$

with

$$\bar{\mathbf{A}} = \begin{bmatrix} 0 & 1 \\ -\lambda\bar{\lambda} & \lambda + \bar{\lambda} \end{bmatrix} \qquad \bar{\mathbf{b}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \bar{\mathbf{c}}_1 = [-2\mathrm{Re}(\bar{\lambda}b_1c_1) \quad 2\mathrm{Re}(b_1c_1)]$$

by using the transformation $\mathbf{x} = \mathbf{Q}\bar{\mathbf{x}}$ with

$$\mathbf{Q}_1 = \begin{bmatrix} -\bar{\lambda}b_1 & b_1 \\ -\lambda\bar{b}_1 & \bar{b}_1 \end{bmatrix}$$

4.7 Verify that the Jordan-form equation

$$\dot{\mathbf{x}} = \begin{bmatrix} \lambda & 1 & 0 & 0 & 0 & 0 \\ 0 & \lambda & 1 & 0 & 0 & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 \\ 0 & 0 & 0 & \bar{\lambda} & 1 & 0 \\ 0 & 0 & 0 & 0 & \bar{\lambda} & 1 \\ 0 & 0 & 0 & 0 & 0 & \bar{\lambda} \end{bmatrix} \mathbf{x} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \bar{b}_1 \\ \bar{b}_2 \\ \bar{b}_3 \end{bmatrix} u$$

$$y = [c_1 \ c_2 \ c_3 \ \bar{c}_1 \ \bar{c}_2 \ \bar{c}_3]\mathbf{x}$$

can be transformed into

$$\dot{\bar{\mathbf{x}}} = \begin{bmatrix} \bar{\mathbf{A}} & \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{A}} & \mathbf{I}_2 \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{A}} \end{bmatrix} \bar{\mathbf{x}} + \begin{bmatrix} \bar{\mathbf{b}} \\ \bar{\mathbf{b}} \\ \bar{\mathbf{b}} \end{bmatrix} u \qquad y = [\bar{\mathbf{c}}_1 \ \bar{\mathbf{c}}_2 \ \bar{\mathbf{c}}_3]\bar{\mathbf{x}}$$

where $\bar{\mathbf{A}}$, $\bar{\mathbf{b}}$, and $\bar{\mathbf{c}}_i$ are defined in Problem 4.6 and $\mathbf{I}_2$ is the unit matrix of order 2. [*Hint*: Change the order of the state variables from $[x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6]'$ to $[x_1 \ x_4 \ x_2 \ x_5 \ x_3 \ x_6]'$ and then apply the equivalence transformation $\mathbf{x} = \mathbf{Q}\bar{\mathbf{x}}$ with $\mathbf{Q} = \text{diag}(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3)$.]

4.8 Are the two sets of state equations

$$\dot{\mathbf{x}} = \begin{bmatrix} 2 & 1 & 2 \\ 0 & 2 & 2 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u \qquad y = [1 \ -1 \ 0]\mathbf{x}$$

and

$$\dot{\mathbf{x}} = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{x} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} u \qquad y = [1 \ -1 \ 0]\mathbf{x}$$

equivalent? Are they zero-state equivalent?

4.9 Verify that the transfer matrix in (4.33) has the following realization:

$$\dot{\mathbf{x}} = \begin{bmatrix} -\alpha_1\mathbf{I}_q & \mathbf{I}_q & \mathbf{0} & \cdots & \mathbf{0} \\ -\alpha_2\mathbf{I}_q & \mathbf{0} & \mathbf{I}_q & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ -\alpha_{r-1}\mathbf{I}_q & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_q \\ -\alpha_r\mathbf{I}_q & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{N}_1 \\ \mathbf{N}_2 \\ \vdots \\ \mathbf{N}_{r-1} \\ \mathbf{N}_r \end{bmatrix} \mathbf{u}$$

$$\mathbf{y} = [\mathbf{I}_q \ \mathbf{0} \ \mathbf{0} \ \cdots \ \mathbf{0}]\mathbf{x}$$

This is called the *observable canonical form realization* and has dimension rq. It is dual to (4.34).

4.10 Consider the 1×2 proper rational matrix

$$\hat{\mathbf{G}}(s) = [d_1 \ \ d_2] + \frac{1}{s^4 + \alpha_1 s^3 + \alpha_2 s^2 + \alpha_3 s + \alpha_4}$$

$$\times [\beta_{11}s^3 + \beta_{21}s^2 + \beta_{31}s + \beta_{41} \quad \beta_{12}s^3 + \beta_{22}s^2 + \beta_{32}s + \beta_{42}]$$

Show that its observable canonical form realization can be reduced from Problem 4.9 as

$$\dot{\mathbf{x}} = \begin{bmatrix} -\alpha_1 & 1 & 0 & 0 \\ -\alpha_2 & 0 & 1 & 0 \\ -\alpha_3 & 0 & 0 & 1 \\ -\alpha_4 & 0 & 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \\ \beta_{41} & \beta_{42} \end{bmatrix} \mathbf{u}$$

$$y = [1 \ 0 \ 0 \ 0]\mathbf{x} + [d_1 \ d_2]\mathbf{u}$$

4.11 Find a realization for the proper rational matrix

$$\hat{\mathbf{G}}(s) = \begin{bmatrix} \dfrac{2}{s+1} & \dfrac{2s-3}{(s+1)(s+2)} \\ \dfrac{s-2}{s+1} & \dfrac{s}{s+2} \end{bmatrix}$$

4.12 Find a realization for each column of $\hat{\mathbf{G}}(s)$ in Problem 4.11 and then connect them, as shown in Fig. 4.4(a), to obtain a realization of $\hat{\mathbf{G}}(s)$. What is the dimension of this realization? Compare this dimension with the one in Problem 4.11.

4.13 Find a realization for each row of $\hat{\mathbf{G}}(s)$ in Problem 4.11 and then connect them, as shown in Fig. 4.4(b), to obtain a realization of $\hat{\mathbf{G}}(s)$. What is the dimension of this realization? Compare this dimension with the ones in Problems 4.11 and 4.12.

4.14 Find a realization for

$$\hat{\mathbf{G}}(s) = \begin{bmatrix} \dfrac{-(12s+6)}{3s+34} & \dfrac{22s+23}{3s+34} \end{bmatrix}$$

4.15 Consider the n-dimensional state equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{b}u \qquad y = \mathbf{c}\mathbf{x}$$

Let $\hat{g}(s)$ be its transfer function. Show that $\hat{g}(s)$ has m zeros or, equivalently, the numerator of $\hat{g}(s)$ has degree m if and only if

$$\mathbf{c}\mathbf{A}^i\mathbf{b} = 0 \qquad \text{for } i = 0, 1, 2, \ldots, n-m-2$$

and $\mathbf{c}\mathbf{A}^{n-m-1}\mathbf{b} \neq 0$. Or, equivalently, the difference between the degrees of the denominator and numerator of $\hat{g}(s)$ is $\alpha = n - m$ if and only if

$$\mathbf{c}\mathbf{A}^{\alpha-1}\mathbf{b} \neq 0 \quad \text{and} \quad \mathbf{c}\mathbf{A}^i\mathbf{b} = 0$$

for $i = 0, 1, 2, \ldots, \alpha - 2$.

4.16 Find fundamental matrices and state transition matrices for

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & t \end{bmatrix} \mathbf{x}$$

and

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & e^{2t} \\ 0 & -1 \end{bmatrix} \mathbf{x}$$

4.17 Show $\partial \mathbf{\Phi}(t_0, t)/\partial t = -\mathbf{\Phi}(t_0, t)\mathbf{A}(t)$.

4.18 Given

$$\mathbf{A}(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) \\ a_{21}(t) & a_{22}(t) \end{bmatrix}$$

show

$$\det \mathbf{\Phi}(t, t_0) = \exp\left[\int_{t_0}^{t} (a_{11}(\tau) + a_{22}(\tau))\, d\tau\right]$$

4.19 Let

$$\mathbf{\Phi}(t, t_0) = \begin{bmatrix} \mathbf{\Phi}_{11}(t, t_0) & \mathbf{\Phi}_{12}(t, t_0) \\ \mathbf{\Phi}_{21}(t, t_0) & \mathbf{\Phi}_{22}(t, t_0) \end{bmatrix}$$

be the state transition matrix of

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \mathbf{A}_{11}(t) & \mathbf{A}_{12}(t) \\ \mathbf{0} & \mathbf{A}_{22}(t) \end{bmatrix} \mathbf{x}(t)$$

Show that $\mathbf{\Phi}_{21}(t, t_0) = \mathbf{0}$ for all t and t_0 and that $(\partial/\partial t)\mathbf{\Phi}_{ii}(t, t_0) = \mathbf{A}_{ii}\mathbf{\Phi}_{ii}(t, t_0)$, for $i = 1, 2$.

4.20 Find the state transition matrix of

$$\dot{\mathbf{x}} = \begin{bmatrix} -\sin t & 0 \\ 0 & -\cos t \end{bmatrix} \mathbf{x}$$

4.21 Verify that $\mathbf{X}(t) = e^{\mathbf{A}t}\mathbf{C}e^{\mathbf{B}t}$ is the solution of

$$\dot{\mathbf{X}} = \mathbf{AX} + \mathbf{XB} \qquad \mathbf{X}(0) = \mathbf{C}$$

4.22 Show that if $\dot{\mathbf{A}}(t) = \mathbf{A}_1\mathbf{A}(t) - \mathbf{A}(t)\mathbf{A}_1$, then

$$\mathbf{A}(t) = e^{\mathbf{A}_1 t}\mathbf{A}(0)e^{-\mathbf{A}_1 t}$$

Show also that the eigenvalues of $\mathbf{A}(t)$ are independent of t.

4.23 Find an equivalent time-invariant state equation of the equation in Problem 4.20.

4.24 Transform a time-invariant $(\mathbf{A}, \mathbf{B}, \mathbf{C})$ into $(\mathbf{0}, \bar{\mathbf{B}}(t), \bar{\mathbf{C}}(t))$ by a time-varying equivalence transformation.

4.25 Find a time-varying realization and a time-invariant realization of the impulse response $g(t) = t^2 e^{\lambda t}$.

4.26 Find a realization of $g(t, \tau) = \sin t(e^{-(t-\tau)}) \cos \tau$. Is it possible to find a time-invariant state equation realization?

Chapter 5

Stability

5.1 Introduction

Systems are designed to perform some tasks or to process signals. If a system is not stable, the system may burn out, disintegrate, or saturate when a signal, no matter how small, is applied. Therefore an unstable system is useless in practice and stability is a basic requirement for all systems. In addition to stability, systems must meet other requirements, such as to track desired signals and to suppress noise, to be really useful in practice.

The response of linear systems can always be decomposed as the zero-state response and the zero-input response. It is customary to study the stabilities of these two responses separately. We will introduce the BIBO (bounded-input bounded-output) stability for the zero-state response and marginal and asymptotic stabilities for the zero-input response. We study first the time-invariant case and then the time-varying case.

5.2 Input–Output Stability of LTI Systems

Consider a SISO linear time-invariant (LTI) system described by

$$y(t) = \int_0^t g(t - \tau)u(\tau)\, d\tau = \int_0^t g(\tau)u(t - \tau)\, d\tau \tag{5.1}$$

where $g(t)$ is the impulse response or the output excited by an impulse input applied at $t = 0$. Recall that in order to be describable by (5.1), the system must be linear, time-invariant, and causal. In addition, the system must be initially relaxed at $t = 0$.

An input $u(t)$ is said to be *bounded* if $u(t)$ does not grow to positive or negative infinity or, equivalently, there exists a constant u_m such that

$$|u(t)| \leq u_m < \infty \qquad \text{for all } t \geq 0$$

A system is said to be *BIBO stable* (bounded-input bounded-output stable) if every bounded input excites a bounded output. This stability is defined for the zero-state response and is applicable only if the system is initially relaxed.

Theorem 5.1

A SISO system described by (5.1) is BIBO stable if and only if $g(t)$ is absolutely integrable in $[0, \infty)$, or

$$\int_0^\infty |g(t)|dt \leq M < \infty$$

for some constant M.

Proof: First we show that if $g(t)$ is absolutely integrable, then every bounded input excites a bounded output. Let $u(t)$ be an arbitrary input with $|u(t)| \leq u_m < \infty$ for all $t \geq 0$. Then

$$|y(t)| = \left| \int_0^t g(\tau)u(t-\tau)\,d\tau \right| \leq \int_0^t |g(\tau)||u(t-\tau)|\,d\tau$$

$$\leq u_m \int_0^\infty |g(\tau)|d\tau \leq u_m M$$

Thus the output is bounded. Next we show intuitively that if $g(t)$ is not absolutely integrable, then the system is not BIBO stable. If $g(t)$ is not absolutely integrable, then there exists a t_1 such that

$$\int_0^{t_1} |g(\tau)|\,d\tau = \infty$$

Let us choose

$$u(t_1 - \tau) = \begin{cases} 1 & \text{if } g(\tau) \geq 0 \\ -1 & \text{if } g(\tau) < 0 \end{cases}$$

Clearly u is bounded. However, the output excited by this input equals

$$y(t_1) = \int_0^{t_1} g(\tau)u(t_1 - \tau)\,d\tau = \int_0^{t_1} |g(\tau)|\,d\tau = \infty$$

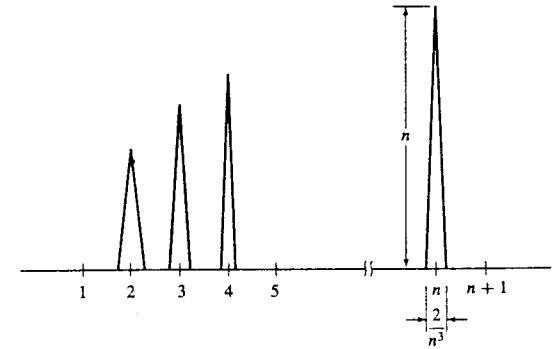
which is not bounded. Thus the system is not BIBO stable. This completes the proof of Theorem 5.1. Q.E.D.

A function that is absolutely integrable may not be bounded or may not approach zero as $t \to \infty$. Indeed, consider the function defined by

$$f(t-n) = \begin{cases} n + (t-n)n^4 & \text{for } n - 1/n^3 \leq t \leq n \\ n - (t-n)n^4 & \text{for } n < t \leq n + 1/n^3 \end{cases}$$

for $n = 2, 3, \ldots$ and plotted in Fig. 5.1. The area under each triangle is $1/n^2$. Thus the absolute

Figure 5.1 Function.



integration of the function equals $\sum_{n=2}^{\infty}(1/n^2) < \infty$. This function is absolutely integrable but is not bounded and does not approach zero as $t \to \infty$.

Theorem 5.2

If a system with impulse response $g(t)$ is BIBO stable, then, as $t \to \infty$:

1. The output excited by $u(t) = a$, for $t \geq 0$, approaches $\hat{g}(0) \cdot a$.
2. The output excited by $u(t) = \sin \omega_o t$, for $t \geq 0$, approaches

$$|\hat{g}(j\omega_o)| \sin(\omega_o t + \measuredangle \hat{g}(j\omega_o))$$

where $\hat{g}(s)$ is the Laplace transform of $g(t)$ or

$$\hat{g}(s) = \int_0^\infty g(\tau)e^{-s\tau}\,d\tau \tag{5.2}$$

Proof: If $u(t) = a$ for all $t \geq 0$, then (5.1) becomes

$$y(t) = \int_0^t g(\tau)u(t-\tau)\,d\tau = a\int_0^t g(\tau)\,d\tau$$

which implies

$$y(t) \to a\int_0^\infty g(\tau)\,d\tau = a\hat{g}(0) \qquad \text{as } t \to \infty$$

where we have used (5.2) with $s = 0$. This establishes the first part of Theorem 5.2. If $u(t) = \sin \omega_o t$, then (5.1) becomes

$$y(t) = \int_0^t g(\tau)\sin\omega_o(t-\tau)\,d\tau$$

$$= \int_0^t g(\tau)\,[\sin\omega_o t\cos\omega_o\tau - \cos\omega_o t\sin\omega_o\tau]\,d\tau$$

$$= \sin\omega_o t\int_0^t g(\tau)\cos\omega_o\tau\,d\tau - \cos\omega_o t\int_0^t g(\tau)\sin\omega_o\tau\,d\tau$$

Thus we have, as $t \to \infty$,

$$y(t) \to \sin \omega_o t \int_0^\infty g(\tau) \cos \omega_o \tau \, d\tau - \cos \omega_o t \int_0^\infty g(\tau) \sin \omega_o \tau \, d\tau \tag{5.3}$$

If $g(t)$ is absolutely integrable, we can replace s by $j\omega$ in (5.2) to yield

$$\hat{g}(j\omega) = \int_0^\infty g(\tau)[\cos \omega\tau - j \sin \omega\tau] \, d\tau$$

The impulse response $g(t)$ is assumed implicitly to be real; thus we have

$$\text{Re}[\hat{g}(j\omega)] = \int_0^\infty g(\tau) \cos \omega\tau \, d\tau$$

and

$$\text{Im}[\hat{g}(j\omega)] = -\int_0^\infty g(\tau) \sin \omega\tau \, d\tau$$

where Re and Im denote, respectively, the real part and imaginary part. Substituting these into (5.3) yields

$$\begin{aligned} y(t) &\to \sin \omega_o t (\text{Re}[\hat{g}(j\omega_o)]) + \cos \omega_o t (\text{Im}[\hat{g}(j\omega_o)]) \\ &= |\hat{g}(j\omega_o)| \sin(\omega_o t + \measuredangle \hat{g}(j\omega_o)) \end{aligned}$$

This completes the proof of Theorem 5.2. Q.E.D.

Theorem 5.2 is a basic result; filtering of signals is based essentially on this theorem. Next we state the BIBO stability condition in terms of proper rational transfer functions.

Theorem 5.3

A SISO system with proper rational transfer function $\hat{g}(s)$ is BIBO stable if and only if every pole of $\hat{g}(s)$ has a negative real part or, equivalently, lies inside the left-half s-plane.

If $\hat{g}(s)$ has pole p_i with multiplicity m_i, then its partial fraction expansion contains the factors

$$\frac{1}{s - p_i}, \quad \frac{1}{(s - p_i)^2}, \quad \cdots, \quad \frac{1}{(s - p_i)^{m_i}}$$

Thus the inverse Laplace transform of $\hat{g}(s)$ or the impulse response contains the factors

$$e^{p_i t}, \quad te^{p_i t}, \quad \cdots, \quad t^{m_i - 1} e^{p_i t}$$

It is straightforward to verify that every such term is absolutely integrable if and only if p_i has a negative real part. Using this fact, we can establish Theorem 5.3.

EXAMPLE 5.1 Consider the positive feedback system shown in Fig. 2.5(a). Its impulse response was computed in (2.9) as

$$g(t) = \sum_{i=1}^{\infty} a^i \delta(t - i)$$

where the gain a can be positive or negative. The impulse is defined as the limit of the pulse in Fig. 2.3 and can be considered to be positive. Thus we have

$$|g(t)| = \sum_{i=1}^{\infty} |a|^i \delta(t - i)$$

and

$$\int_0^\infty |g(t)| \, dt = \sum_{i=1}^{\infty} |a|^i = \begin{cases} \infty & \text{if } |a| \geq 1 \\ |a|/(1 - |a|) < \infty & \text{if } |a| < 1 \end{cases}$$

Thus we conclude that the positive feedback system in Fig. 2.5(a) is BIBO stable if and only if the gain a has a magnitude less than 1.

The transfer function of the system was computed in (2.12) as

$$\hat{g}(s) = \frac{ae^{-s}}{1 - ae^{-s}}$$

It is an irrational function of s and Theorem 5.3 is not applicable. In this case, it is simpler to use Theorem 5.1 to check its stability.

For multivariable systems, we have the following results.

Theorem 5.M1

A multivariable system with impulse response matrix $\mathbf{G}(t) = [g_{ij}(t)]$ is BIBO stable if and only if every $g_{ij}(t)$ is absolutely integrable in $[0, \infty)$.

Theorem 5.M3

A multivariable system with proper rational transfer matrix $\hat{\mathbf{G}}(s) = [\hat{g}_{ij}(s)]$ is BIBO stable if and only if every pole of every $\hat{g}_{ij}(s)$ has a negative real part.

We now discuss the BIBO stability of state equations. Consider

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) &= \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \end{aligned} \tag{5.4}$$

Its transfer matrix is

$$\hat{\mathbf{G}}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

Thus Equation (5.4) or, to be more precise, the zero-state response of (5.4) is BIBO stable if and only if every pole of $\hat{\mathbf{G}}(s)$ has a negative real part. Recall that every pole of every entry of $\hat{\mathbf{G}}(s)$ is called a pole of $\hat{\mathbf{G}}(s)$.

We discuss the relationship between the poles of $\hat{\mathbf{G}}(s)$ and the eigenvalues of $\mathbf{A}$. Because of

$$\hat{\mathbf{G}}(s) = \frac{1}{\det(s\mathbf{I} - \mathbf{A})} \mathbf{C}[\text{Adj}(s\mathbf{I} - \mathbf{A})]\mathbf{B} + \mathbf{D} \tag{5.5}$$

every pole of $\hat{G}(s)$ is an eigenvalue of **A**. Thus if every eigenvalue of **A** has a negative real part, then every pole has a negative real part and (5.4) is BIBO stable. On the other hand, because of possible cancellation in (5.5), not every eigenvalue is a pole. Thus, even if **A** has some eigenvalues with zero or positive real part, (5.4) may still be BIBO stable, as the next example shows.

EXAMPLE 5.2 Consider the network shown in Fig. 4.2(b). Its state equation was derived in Example 4.4 as

$$\dot{x}(t) = x(t) + 0 \cdot u(t) \qquad y(t) = 0.5x(t) + 0.5u(t) \tag{5.6}$$

The A-matrix is 1 and its eigenvalue is 1. It has a positive real part. The transfer function of the equation is

$$\hat{g}(s) = 0.5(s - 1)^{-1} \cdot 0 + 0.5 = 0.5$$

The transfer function equals 0.5. It has no pole and no condition to meet. Thus (5.6) is BIBO stable even though it has an eigenvalue with a positive real part. We mention that BIBO stability does not say anything about the zero-input response, which will be discussed later.

5.2.1 Discrete-Time Case

Consider a discrete-time SISO system described by

$$y[k] = \sum_{m=0}^{k} g[k-m]u[m] = \sum_{m=0}^{k} g[m]u[k-m] \tag{5.7}$$

where $g[k]$ is the impulse response sequence or the output sequence excited by an impulse sequence applied at $k = 0$. Recall that in order to be describable by (5.7), the discrete-time system must be linear, time-invariant, and causal. In addition, the system must be initially relaxed at $k = 0$.

An input sequence $u[k]$ is said to be *bounded* if $u[k]$ does not grow to positive or negative infinity or there exists a constant u_m such that

$$|u[k]| \le u_m < \infty \qquad \text{for } k = 0, 1, 2, \ldots$$

A system is said to be *BIBO stable* (bounded-input bounded-output stable) if every bounded-input sequence excites a bounded-output sequence. This stability is defined for the zero-state response and is applicable only if the system is initially relaxed.

▶ **Theorem 5.D1**

A discrete-time SISO system described by (5.7) is BIBO stable if and only if $g[k]$ is absolutely summable in $[0, \infty)$ or

$$\sum_{k=0}^{\infty} |g[k]| \le M < \infty$$

for some constant M.

Its proof is similar to the proof of Theorem 5.1 and will not be repeated. We give a discrete counterpart of Theorem 5.2 in the following.

Theorem 5.D2

If a discrete-time system with impulse response sequence $g[k]$ is BIBO stable, then, as $k \to \infty$:

1. The output excited by $u[k] = a$, for $k \ge 0$, approaches $\hat{g}(1) \cdot a$.
2. The output excited by $u[k] = \sin \omega_o k$, for $k \ge 0$, approaches

$$|\hat{g}(e^{j\omega_o})| \sin(\omega_o k + \measuredangle \hat{g}(e^{j\omega_o}))$$

where $\hat{g}(z)$ is the z-transform of $g[k]$ or

$$\hat{g}(z) = \sum_{m=0}^{\infty} g[m]z^{-m} \tag{5.8}$$

⇒ ***Proof:*** If $u[k] = a$ for all $k \ge 0$, then (5.7) becomes

$$y[k] = \sum_{0}^{k} g[m]u[k-m] = a \sum_{k=0}^{k} g[m]$$

which implies

$$y[k] \to a \sum_{m=0}^{\infty} g[m] = a\hat{g}(1) \qquad \text{as } k \to \infty$$

where we have used (5.8) with $z = 1$. This establishes the first part of Theorem 5.D2. If $u[k] = \sin \omega_o k$, then (5.7) becomes

$$\begin{aligned} y[k] &= \sum_{m=0}^{k} g[m] \sin \omega_o [k-m] \\ &= \sum_{m=0}^{k} g[m](\sin \omega_o k \cos \omega_o m - \cos \omega_o k \sin \omega_o m) \\ &= \sin \omega_o k \sum_{m=0}^{k} g[m] \cos \omega_o m - \cos \omega_o k \sum_{m=0}^{k} g[m] \sin \omega_o m \end{aligned}$$

Thus we have, as $k \to \infty$,

$$y[k] \to \sin \omega_o k \sum_{m=0}^{\infty} g[m] \cos \omega_o m - \cos \omega_o k \sum_{m=0}^{\infty} g[m] \sin \omega_o m \tag{5.9}$$

If $g[k]$ is absolutely summable, we can replace z by $e^{j\omega}$ in (5.8) to yield

$$\hat{g}(e^{j\omega}) = \sum_{m=0}^{\infty} g[m]e^{-j\omega m} = \sum_{m=0}^{\infty} g[m][\cos \omega m - j \sin \omega m]$$

Thus (5.9) becomes

$$y[k] \to \sin \omega_o k(\mathrm{Re}[\hat{g}(e^{j\omega_o})]) + \cos \omega_o k(\mathrm{Im}[\hat{g}(e^{j\omega_o})])$$
$$= |\hat{g}(e^{j\omega_o})| \sin(\omega_o k + \measuredangle \hat{g}(e^{j\omega_o}))$$

This completes the proof of Theorem 5.D2. Q.E.D.

Theorem 5.D2 is a basic result in digital signal processing. Next we state the BIBO stability in terms of discrete proper rational transfer functions.

Theorem 5.D3

A discrete-time SISO system with proper rational transfer function $\hat{g}(z)$ is BIBO stable if and only if every pole of $\hat{g}(z)$ has a magnitude less than 1 or, equivalently, lies inside the unit circle on the z-plane.

If $\hat{g}(z)$ has pole p_i with multiplicity m_i, then its partial fraction expansion contains the factors

$$\frac{1}{z - p_i}, \quad \frac{1}{(z - p_i)^2}, \quad \cdots, \quad \frac{1}{(z - p_i)^{m_i}}$$

Thus the inverse z-transform of $\hat{g}(z)$ or the impulse response sequence contains the factors

$$p_i^k, \quad kp_i^k, \quad \cdots, \quad k^{m_i - 1} p_i^k$$

It is straightforward to verify that every such term is absolutely summable if and only if p_i has a magnitude less than 1. Using this fact, we can establish Theorem 5.D3.

In the continuous-time case, an absolutely integrable function $f(t)$, as shown in Fig. 5.1, may not be bounded and may not approach zero as $t \to \infty$. In the discrete-time case, if $g[k]$ is absolutely summable, then it must be bounded and approach zero as $k \to \infty$. However, the converse is not true as the next example shows.

EXAMPLE 5.3 Consider a discrete-time LTI system with impulse response sequence $g[k] = 1/k$, for $k = 1, 2, \ldots$, and $g[0] = 0$. We compute

$$S := \sum_{k=0}^{\infty} |g[k]| = \sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$
$$= 1 + \frac{1}{2} + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \cdots + \frac{1}{8}\right) + \left(\frac{1}{9} + \cdots + \frac{1}{16}\right) + \cdots$$

There are two terms, each is $\frac{1}{4}$ or larger, in the first pair of parentheses; therefore their sum is larger than $\frac{1}{2}$. There are four terms, each is $\frac{1}{8}$ or larger, in the second pair of parentheses; therefore their sum is larger than $\frac{1}{2}$. Proceeding forward we conclude

$$S > 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty$$

This impulse response sequence is bounded and approaches 0 as $k \to \infty$ but is not absolutely summable. Thus the discrete-time system is not BIBO stable according to Theorem 5.D1. The transfer function of the system can be shown to equal

$$\hat{g}(z) = -\ln(1 + z^{-1})$$

It is not a rational function of z and Theorem 5.D3 is not applicable.

For multivariable discrete-time systems, we have the following results.

Theorem 5.MD1

A MIMO discrete-time system with impulse response sequence matrix $\mathbf{G}[k] = [g_{ij}[k]]$ is BIBO stable if and only if every $g_{ij}[k]$ is absolutely summable.

Theorem 5.MD3

A MIMO discrete-time system with discrete proper rational transfer matrix $\hat{\mathbf{G}}(z) = [\hat{g}_{ij}(z)]$ is BIBO stable if and only if every pole of every $\hat{g}_{ij}(z)$ has a magnitude less than 1.

We now discuss the BIBO stability of discrete-time state equations. Consider

$$\mathbf{x}[k + 1] = \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k]$$
$$\mathbf{y}[k] = \mathbf{C}\mathbf{x}[k] + \mathbf{D}\mathbf{u}[k] \tag{5.10}$$

Its discrete transfer matrix is

$$\hat{\mathbf{G}}(z) = \mathbf{C}(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

Thus Equation (5.10) or, to be more precise, the zero-state response of (5.10) is BIBO stable if and only if every pole of $\hat{\mathbf{G}}(z)$ has a magnitude less than 1.

We discuss the relationship between the poles of $\hat{\mathbf{G}}(z)$ and the eigenvalues of $\mathbf{A}$. Because of

$$\hat{\mathbf{G}}(z) = \frac{1}{\det(z\mathbf{I} - \mathbf{A})} \mathbf{C}[\mathrm{Adj}(z\mathbf{I} - \mathbf{A})]\mathbf{B} + \mathbf{D}$$

every pole of $\hat{\mathbf{G}}(z)$ is an eigenvalue of $\mathbf{A}$. Thus if every eigenvalue of $\mathbf{A}$ has a negative real part, then (5.10) is BIBO stable. On the other hand, even if $\mathbf{A}$ has some eigenvalues with zero or positive real part, (5.10) may, as in the continuous-time case, still be BIBO stable.

5.3 Internal Stability

The BIBO stability is defined for the zero-state response. Now we study the stability of the zero-input response or the response of

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) \tag{5.11}$$

excited by nonzero initial state $\mathbf{x}_o$. Clearly, the solution of (5.11) is

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}_o \tag{5.12}$$

Definition 5.1 *The zero-input response of (5.4) or the equation* $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ *is* marginally stable *or* stable in the sense of Lyapunov *if every finite initial state* $\mathbf{x}_o$ *excites a bounded response. It is* asymptotically stable *if every finite initial state excites a bounded response, which, in addition, approaches* $\mathbf{0}$ *as* $t \to \infty$.

We mention that this definition is applicable only to linear systems. The definition that is applicable to both linear and nonlinear systems must be defined using the concept of equivalence states and can be found, for example, in Reference [6, pp. 401–403]. This text studies only linear systems; therefore we use the simplified Definition 5.1.

Theorem 5.4

1. The equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is marginally stable if and only if all eigenvalues of $\mathbf{A}$ have zero or negative real parts and those with zero real parts are simple roots of the minimal polynomial of $\mathbf{A}$.

2. The equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is asymptotically stable if and only if all eigenvalues of $\mathbf{A}$ have negative real parts.

We first mention that any (algebraic) equivalence transformation will not alter the stability of a state equation. Consider $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$, where $\mathbf{P}$ is a nonsingular matrix. Then $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is equivalent to $\dot{\bar{\mathbf{x}}} = \bar{\mathbf{A}}\bar{\mathbf{x}} = \mathbf{P}\mathbf{A}\mathbf{P}^{-1}\bar{\mathbf{x}}$. Because $\mathbf{P}$ is nonsingular, if $\mathbf{x}$ is bounded, so is $\bar{\mathbf{x}}$; if $\mathbf{x}$ approaches $\mathbf{0}$ as $t \to \infty$, so does $\bar{\mathbf{x}}$. Thus we may study the stability of $\mathbf{A}$ by using $\bar{\mathbf{A}}$. Note that the eigenvalues of $\mathbf{A}$ and of $\bar{\mathbf{A}}$ are the same as discussed in Section 4.3.

The response of $\dot{\bar{\mathbf{x}}} = \bar{\mathbf{A}}\bar{\mathbf{x}}$ excited by $\bar{\mathbf{x}}(0)$ equals $\bar{\mathbf{x}}(t) = e^{\bar{\mathbf{A}}t}\bar{\mathbf{x}}(0)$. It is clear that the response is bounded if and only if every entry of $e^{\bar{\mathbf{A}}t}$ is bounded for all $t \geq 0$. If $\bar{\mathbf{A}}$ is in Jordan form, then $e^{\bar{\mathbf{A}}t}$ is of the form shown in (3.48). Using (3.48), we can show that if an eigenvalue has a negative real part, then every entry of (3.48) is bounded and approaches 0 as $t \to \infty$. If an eigenvalue has zero real part and has no Jordan block of order 2 or higher, then the corresponding entry in (3.48) is a constant or is sinusoidal for all t and is, therefore, bounded. This establishes the sufficiency of the first part of Theorem 5.4. If $\bar{\mathbf{A}}$ has an eigenvalue with a positive real part, then every entry in (3.48) will grow without bound. If $\bar{\mathbf{A}}$ has an eigenvalue with zero real part and its Jordan block has order 2 or higher, then (3.48) has at least one entry that grows unbounded. This completes the proof of the first part. To be asymptotically stable, every entry of (3.48) must approach zero as $t \to \infty$. Thus no eigenvalue with zero real part is permitted. This establishes the second part of the theroem.

EXAMPLE 5.4 Consider

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{x}$$

Its characteristic polynomial is $\Delta(\lambda) = \lambda^2(\lambda + 1)$ and its mimimal polynomial is $\psi(\lambda) = \lambda(\lambda + 1)$. The matrix has eigenvalues 0, 0, and -1. The eigenvlaue 0 is a simple root of the minimal polynomial. Thus the equation is marginally stable. The equation

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \mathbf{x}$$

is not marginally stable, however, because its minimal polynomial is $\lambda^2(\lambda + 1)$ and $\lambda = 0$ is not a simple root of the minimal polynomial.

As discussed earlier, every pole of the transfer matrix

$$\hat{\mathbf{G}}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$$

is an eigenvalue of $\mathbf{A}$. Thus asymptotic stability implies BIBO stability. Note that asymptotic stability is defined for the zero-input response, whereas BIBO stability is defined for the zero-state response. The system in Example 5.2 has eigenvalue 1 and is not asymptotically stable; however, it is BIBO stable. Thus BIBO stability, in general, does not imply asymptotic stability. We mention that marginal stability is useful only in the design of oscillators. Other than oscillators, every physical system is designed to be asymptotically stable or BIBO stable with some additional conditions, as we will discuss in Chapter 7.

5.3.1 Discrete-Time Case

This subsection studies the internal stability of discrete-time systems or the stability of

$$\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k] \tag{5.13}$$

excited by nonzero initial state $\mathbf{x}_o$. The solution of (5.13) is, as derived in (4.20),

$$\mathbf{x}[k] = \mathbf{A}^k\mathbf{x}_o \tag{5.14}$$

Equation (5.13) is said to be *marginally stable* or *stable in the sense of Lyapunov* if every finite initial state $\mathbf{x}_o$ excites a bounded response. It is *asymptotically stable* if every finite initial state excites a bounded response, which, in addition, approaches $\mathbf{0}$ as $k \to \infty$. These definitions are identical to the continuous-time case.

Theorem 5.D4

1. The equation $\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k]$ is marginally stable if and only if all eigenvalues of $\mathbf{A}$ have magnitudes less than or equal to 1 and those equal to 1 are simple roots of the minimal polynomial of $\mathbf{A}$.

2. The equation $\mathbf{x}[k+1] = \mathbf{A}\mathbf{x}[k]$ is asymptotically stable if and only if all eigenvalues of $\mathbf{A}$ have magnitudes less than 1.

As in the continuous-time case, any (algebraic) equivalence transformation will not alter the stability of a state equation. Thus we can use Jordan form to establish the theorem. The proof is similar to the continuous-time case and will not be repeated. Asymptotic stability

implies BIBO stability but not the converse. We mention that marginal stability is useful only in the design of discrete-time oscillators. Other than oscillators, every discrete-time physical system is designed to be asymptotically stable or BIBO stable with some additional conditions, as we will discuss in Chapter 7.

5.4 Lyapunov Theorem

This section introduces a different method of checking asymptotic stability of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$. For convenience, we call $\mathbf{A}$ stable if every eigenvalue of $\mathbf{A}$ has a negative real part.

Theorem 5.5

All eigenvalues of $\mathbf{A}$ have negative real parts if and only if for any given positive definite symmetric matrix $\mathbf{N}$, the *Lyapunov* equation

$$\mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} = -\mathbf{N} \tag{5.15}$$

has a unique symmetric solution $\mathbf{M}$ and $\mathbf{M}$ is positive definite.

Corollary 5.5

All eigenvalues of an $n \times n$ matrix $\mathbf{A}$ have negative real parts if and only if for any given $m \times n$ matrix $\bar{\mathbf{N}}$ with $m < n$ and with the property

$$\text{rank } O := \text{rank} \begin{bmatrix} \bar{\mathbf{N}} \\ \bar{\mathbf{N}}\mathbf{A} \\ \vdots \\ \bar{\mathbf{N}}\mathbf{A}^{n-1} \end{bmatrix} = n \quad \text{(full column rank)} \tag{5.16}$$

where O is an $nm \times n$ matrix, the Lyapunov equation

$$\mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} = -\bar{\mathbf{N}}'\bar{\mathbf{N}} =: -\mathbf{N} \tag{5.17}$$

has a unique symmetric solution $\mathbf{M}$ and $\mathbf{M}$ is positive definite.

For any $\bar{\mathbf{N}}$, the matrix $\mathbf{N}$ in (5.17) is positive semidefinite (Theorem 3.7). Theorem 5.5 and its corollary are valid for any given $\mathbf{N}$; therefore we shall use the simplest possible $\mathbf{N}$. Even so, using them to check stability of $\mathbf{A}$ is not simple. It is much simpler to compute, using MATLAB, the eigenvalues of $\mathbf{A}$ and then check their real parts. Thus the importance of Theorem 5.5 and its corollary is not in checking the stability of $\mathbf{A}$ but rather in studying the stability of nonlinear systems. They are essential in using the so-called second method of Lyapunov. We mention that Corollary 5.5 can be used to prove the Routh–Hurwitz test. See Reference [6, pp. 417–419].

⇒ ***Proof of Theorem 5.5*** *Necessity:* Equation (5.15) is a special case of (3.59) with $\mathbf{A} = \mathbf{A}'$ and $\mathbf{B} = \mathbf{A}$. Because $\mathbf{A}$ and $\mathbf{A}'$ have the same set of eigenvalues, if $\mathbf{A}$ is stable, $\mathbf{A}$ has no two eigenvalues such that $\lambda_i + \lambda_j = 0$. Thus the Lyapunov equation is nonsingular and has a unique solution $\mathbf{M}$ for any $\mathbf{N}$. We claim that the solution can be expressed as

$$\mathbf{M} = \int_0^\infty e^{\mathbf{A}'t}\mathbf{N}e^{\mathbf{A}t}\,dt \tag{5.18}$$

Indeed, substituting (5.18) into (5.15) yields

$$\begin{aligned} \mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} &= \int_0^\infty \mathbf{A}'e^{\mathbf{A}'t}\mathbf{N}e^{\mathbf{A}t}\,dt + \int_0^\infty e^{\mathbf{A}'t}\mathbf{N}e^{\mathbf{A}t}\mathbf{A}\,dt \\ &= \int_0^\infty \frac{d}{dt}\left(e^{\mathbf{A}'t}\mathbf{N}e^{\mathbf{A}t}\right)dt = e^{\mathbf{A}'t}\mathbf{N}e^{\mathbf{A}t}\Big|_{t=0}^{\infty} \\ &= \mathbf{0} - \mathbf{N} = -\mathbf{N} \end{aligned} \tag{5.19}$$

where we have used the fact $e^{\mathbf{A}t} = \mathbf{0}$ at $t = \infty$ for stable $\mathbf{A}$. This shows that the $\mathbf{M}$ in (5.18) is the solution. It is clear that if $\mathbf{N}$ is symmetric, so is $\mathbf{M}$. Let us decompose $\mathbf{N}$ as $\mathbf{N} = \bar{\mathbf{N}}'\bar{\mathbf{N}}$, where $\bar{\mathbf{N}}$ is nonsingular (Theorem 3.7) and consider

$$\mathbf{x}'\mathbf{M}\mathbf{x} = \int_0^\infty \mathbf{x}'e^{\mathbf{A}'t}\bar{\mathbf{N}}'\bar{\mathbf{N}}e^{\mathbf{A}t}\mathbf{x}\,dt = \int_0^\infty \|\bar{\mathbf{N}}e^{\mathbf{A}t}\mathbf{x}\|_2^2\,dt \tag{5.20}$$

Because both $\bar{\mathbf{N}}$ and $e^{\mathbf{A}t}$ are nonsingular, for any nonzero $\mathbf{x}$, the integrand of (5.20) is positive for every t. Thus $\mathbf{x}'\mathbf{M}\mathbf{x}$ is positive for any $\mathbf{x} \neq \mathbf{0}$. This shows the positive definiteness of $\mathbf{M}$.

Sufficiency: We show that if $\mathbf{N}$ and $\mathbf{M}$ are positive definite, then $\mathbf{A}$ is stable. Let λ be an eigenvalue of $\mathbf{A}$ and $\mathbf{v} \neq \mathbf{0}$ be a corresponding eigenvector; that is, $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Even though $\mathbf{A}$ is a real matrix, its eigenvalue and eigenvector can be complex, as shown in Example 3.6. Taking the complex-conjugate transpose of $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ yields $\mathbf{v}^*\mathbf{A}^* = \mathbf{v}^*\mathbf{A}' = \lambda^*\mathbf{v}^*$, where the asterisk denotes complex-conjugate transpose. Premultiplying $\mathbf{v}^*$ and postmultiplying $\mathbf{v}$ to (5.15) yields

$$\begin{aligned} -\mathbf{v}^*\mathbf{N}\mathbf{v} &= \mathbf{v}^*\mathbf{A}'\mathbf{M}\mathbf{v} + \mathbf{v}^*\mathbf{M}\mathbf{A}\mathbf{v} \\ &= (\lambda^* + \lambda)\mathbf{v}^*\mathbf{M}\mathbf{v} = 2\text{Re}(\lambda)\mathbf{v}^*\mathbf{M}\mathbf{v} \end{aligned} \tag{5.21}$$

Because $\mathbf{v}^*\mathbf{M}\mathbf{v}$ and $\mathbf{v}^*\mathbf{N}\mathbf{v}$ are, as discussed in Section 3.9, both real and positive, (5.21) implies $\text{Re}(\lambda) < 0$. This shows that every eigenvalue of $\mathbf{A}$ has a negative real part. Q.E.D.

The proof of Corollary 5.5 follows the proof of Theorem 5.5 with some modification. We discuss only where the proof of Theorem 5.5 is not applicable. Consider (5.20). Now $\bar{\mathbf{N}}$ is $m \times n$ with $m < n$ and $\mathbf{N} = \bar{\mathbf{N}}'\bar{\mathbf{N}}$ is positive semidefinite. Even so, $\mathbf{M}$ in (5.18) can still be positive definite if the integrand of (5.20) is not identically zero for all t. Suppose the integrand of (5.20) is identically zero or $\bar{\mathbf{N}}e^{\mathbf{A}t}\mathbf{x} \equiv \mathbf{0}$. Then its derivative with respect to t yields $\bar{\mathbf{N}}\mathbf{A}e^{\mathbf{A}t}\mathbf{x} = \mathbf{0}$. Proceeding forward, we can obtain

$$\begin{bmatrix} \bar{\mathbf{N}} \\ \bar{\mathbf{N}}\mathbf{A} \\ \vdots \\ \bar{\mathbf{N}}\mathbf{A}^{n-1} \end{bmatrix} e^{\mathbf{A}t}\mathbf{x} = \mathbf{0} \tag{5.22}$$

This equation implies that, because of (5.16) and the nonsingularity of $e^{\mathbf{A}t}$, the only $\mathbf{x}$ meeting

(5.22) is $\mathbf{0}$. Thus the integrand of (5.20) cannot be identically zero for any $\mathbf{x} \neq \mathbf{0}$. Thus $\mathbf{M}$ is positive definite under the condition in (5.16). This shows the necessity of Corollary 5.5. Next we consider (5.21) with $\mathbf{N} = \bar{\mathbf{N}}'\bar{\mathbf{N}}$ or[1]

$$2\text{Re}(\lambda)\mathbf{v}^*\mathbf{M}\mathbf{v} = -\mathbf{v}^*\bar{\mathbf{N}}'\bar{\mathbf{N}}\mathbf{v} = -||\bar{\mathbf{N}}\mathbf{v}||_2^2 \tag{5.23}$$

We show that $\bar{\mathbf{N}}\mathbf{v}$ is nonzero under (5.16). Because of $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, we have $\mathbf{A}^2\mathbf{v} = \lambda\mathbf{A}\mathbf{v} = \lambda^2\mathbf{v}$, $\ldots$, $\mathbf{A}^{n-1}\mathbf{v} = \lambda^{n-1}\mathbf{v}$. Consider

$$\begin{bmatrix} \bar{\mathbf{N}} \\ \bar{\mathbf{N}}\mathbf{A} \\ \vdots \\ \bar{\mathbf{N}}\mathbf{A}^{n-1} \end{bmatrix} \mathbf{v} = \begin{bmatrix} \bar{\mathbf{N}}\mathbf{v} \\ \bar{\mathbf{N}}\mathbf{A}\mathbf{v} \\ \vdots \\ \bar{\mathbf{N}}\mathbf{A}^{n-1}\mathbf{v} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{N}}\mathbf{v} \\ \lambda\bar{\mathbf{N}}\mathbf{v} \\ \vdots \\ \lambda^{n-1}\bar{\mathbf{N}}\mathbf{v} \end{bmatrix}$$

If $\bar{\mathbf{N}}\mathbf{v} = \mathbf{0}$, the rightmost matrix is zero; the leftmost matrix, however, is nonzero under the conditions of (5.16) and $\mathbf{v} \neq \mathbf{0}$. This is a contradiction. Thus $\bar{\mathbf{N}}\mathbf{v}$ is nonzero and (5.23) implies $\text{Re}(\lambda) < 0$. This completes the proof of Corollary 5.5.

In the proof of Theorem of 5.5, we have established the following result. For easy reference, we state it as a theorem.

Theorem 5.6

If all eigenvalues of $\mathbf{A}$ have negative real parts, then the Lyapunov equation

$$\mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} = -\mathbf{N}$$

has a unique solution for every $\mathbf{N}$, and the solution can be expressed as

$$\mathbf{M} = \int_0^\infty e^{\mathbf{A}'t}\mathbf{N}e^{\mathbf{A}t}dt \tag{5.24}$$

Because of the importance of this theorem, we give a different proof of the uniqueness of the solution. Suppose there are two solutions $\mathbf{M}_1$ and $\mathbf{M}_2$. Then we have

$$\mathbf{A}'(\mathbf{M}_1 - \mathbf{M}_2) + (\mathbf{M}_1 - \mathbf{M}_2)\mathbf{A} = \mathbf{0}$$

which implies

$$e^{\mathbf{A}'t}[\mathbf{A}'(\mathbf{M}_1 - \mathbf{M}_2) + (\mathbf{M}_1 - \mathbf{M}_2)\mathbf{A}]e^{\mathbf{A}t} = \frac{d}{dt}[e^{\mathbf{A}'t}(\mathbf{M}_1 - \mathbf{M}_2)e^{\mathbf{A}t}] = \mathbf{0}$$

Its integration from 0 to ∞ yields

$$[e^{\mathbf{A}'t}(\mathbf{M}_1 - \mathbf{M}_2)e^{\mathbf{A}t}]\Big|_0^\infty = \mathbf{0}$$

or, using $e^{\mathbf{A}t} \to \mathbf{0}$ as $t \to \infty$,

$$\mathbf{0} - (\mathbf{M}_1 - \mathbf{M}_2) = \mathbf{0}$$

1. Note that if $\mathbf{x}$ is a complex vector, then the Euclidean norm defined in Section 3.2 must be modified as $||\mathbf{x}||_2^2 = \mathbf{x}^*\mathbf{x}$, where $\mathbf{x}^*$ is the complex conjugate transpose of $\mathbf{x}$.

This shows the uniqueness of $\mathbf{M}$. Although the solution can be expressed as in (5.24), the integration is not used in computing the solution. It is simpler to arrange the Lyapunov equation, after some transformations, into a standard linear algebraic equation as in (3.60) and then solve the equation. Note that even if $\mathbf{A}$ is not stable, a unique solution still exists if $\mathbf{A}$ has no two eigenvalues such that $\lambda_i + \lambda_j = 0$. The solution, however, cannot be expressed as in (5.24); the integration will diverge and is meaningless. If $\mathbf{A}$ is singular or, equivalently, has at least one zero eigenvalue, then the Lyapunov equation is always singular and solutions may or may not exist depending on whether or not $\mathbf{N}$ lies in the range space of the equation.

5.4.1 Discrete-Time Case

Before discussing the discrete counterpart of Theorems 5.5 and 5.6, we discuss the discrete counterpart of the Lyapunov equation in (3.59). Consider

$$\mathbf{M} - \mathbf{A}\mathbf{M}\mathbf{B} = \mathbf{C} \tag{5.25}$$

where $\mathbf{A}$ and $\mathbf{B}$ are, respectively, $n \times n$ and $m \times m$ matrices, and $\mathbf{M}$ and $\mathbf{C}$ are $n \times m$ matrices. As (3.60), Equation (5.25) can be expressed as $\mathbf{Y}\mathbf{m} = \mathbf{c}$, where $\mathbf{Y}$ is an $nm \times nm$ matrix; $\mathbf{m}$ and $\mathbf{c}$ are $nm \times 1$ column vectors with the m columns of $\mathbf{M}$ and $\mathbf{C}$ stacked in order. Thus (5.25) is essentially a set of linear algebraic equations. Let η_k be an eigenvalue of $\mathbf{Y}$ or of (5.25). Then we have

$$\eta_k = 1 - \lambda_i\mu_j \qquad \text{for } i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m$$

where λ_i and μ_j are, respectively, the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$. This can be established intuitively as follows. Let us define $\mathcal{A}(\mathbf{M}) := \mathbf{M} - \mathbf{A}\mathbf{M}\mathbf{B}$. Then (5.25) can be written as $\mathcal{A}(\mathbf{M}) = \mathbf{C}$. A scalar η is an eigenvalue of $\mathcal{A}$ if there exists a nonzero $\mathbf{M}$ such that $\mathcal{A}(\mathbf{M}) = \eta\mathbf{M}$. Let $\mathbf{u}$ be an $n \times 1$ right eigenvector of $\mathbf{A}$ associated with λ_i; that is, $\mathbf{A}\mathbf{u} = \lambda_i\mathbf{u}$. Let $\mathbf{v}$ be a $1 \times m$ left eigenvector of $\mathbf{B}$ associated with μ_j; that is, $\mathbf{v}\mathbf{B} = \mathbf{v}\mu_j$. Applying $\mathcal{A}$ to the $n \times m$ nonzero matrix $\mathbf{u}\mathbf{v}$ yields

$$\mathcal{A}(\mathbf{u}\mathbf{v}) = \mathbf{u}\mathbf{v} - \mathbf{A}\mathbf{u}\mathbf{v}\mathbf{B} = (1 - \lambda_i\mu_j)\mathbf{u}\mathbf{v}$$

Thus the eigenvalues of (5.25) are $1 - \lambda_i\mu_j$, for all i and j. If there are no i and j such that $\lambda_i\mu_j = 1$, then (5.25) is nonsingular and, for any $\mathbf{C}$, a unique solution $\mathbf{M}$ exists in (5.25). If $\lambda_i\mu_j = 1$ for some i and j, then (5.25) is singular and, for a given $\mathbf{C}$, solutions may or may not exist. The situation here is similar to what was discussed in Section 3.7.

Theorem 5.D5

All eigenvalues of an $n \times n$ matrix $\mathbf{A}$ have magnitudes less than 1 if and only if for any given positive definite symmetric matrix $\mathbf{N}$ or for $\mathbf{N} = \bar{\mathbf{N}}'\bar{\mathbf{N}}$, where $\bar{\mathbf{N}}$ is any given $m \times n$ matrix with $m < n$ and with the property in (5.16), the discrete Lyapunov equation

$$\mathbf{M} - \mathbf{A}'\mathbf{M}\mathbf{A} = \mathbf{N} \tag{5.26}$$

has a unique symmetric solution $\mathbf{M}$ and $\mathbf{M}$ is positive definite.

We sketch briefly its proof for $\mathbf{N} > 0$. If all eigenvalues of $\mathbf{A}$ and, consequently, of $\mathbf{A}'$ have magnitudes less than 1, then we have $|\lambda_i \lambda_j| < 1$ for all i and j. Thus $\lambda_i \lambda_j \neq 1$ and (5.26) is nonsingular. Therefore, for any $\mathbf{N}$, a unique solution exists in (5.26). We claim that the solution can be expressed as

$$\mathbf{M} = \sum_{m=0}^{\infty} (\mathbf{A}')^m \mathbf{N} \mathbf{A}^m \tag{5.27}$$

Because $|\lambda_i| < 1$ for all i, this infinite series converges and is well defined. Substituting (5.27) into (5.26) yields

$$\sum_{m=0}^{\infty} (\mathbf{A}')^m \mathbf{N} \mathbf{A}^m - \mathbf{A}' \left(\sum_{m=0}^{\infty} (\mathbf{A}')^m \mathbf{N} \mathbf{A}^m \right) \mathbf{A}$$

$$= \mathbf{N} + \sum_{m=1}^{\infty} (\mathbf{A}')^m \mathbf{N} \mathbf{A}^m - \sum_{m=1}^{\infty} (\mathbf{A}')^m \mathbf{N} \mathbf{A}^m = \mathbf{N}$$

Thus (5.27) is the solution. If $\mathbf{N}$ is symmetric, so is $\mathbf{M}$. If $\mathbf{N}$ is positive definite, so is $\mathbf{M}$. This establishes the necessity. To show sufficiency, let λ be an eigenvalue of $\mathbf{A}$ and $\mathbf{v} \neq \mathbf{0}$ be a corresponding eigenvector; that is, $\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$. Then we have

$$\mathbf{v}^* \mathbf{N} \mathbf{v} = \mathbf{v}^* \mathbf{M} \mathbf{v} - \mathbf{v}^* \mathbf{A}' \mathbf{M} \mathbf{A} \mathbf{v}$$

$$= \mathbf{v}^* \mathbf{M} \mathbf{v} - \lambda^* \mathbf{v}^* \mathbf{M} \mathbf{v} \lambda = (1 - |\lambda|^2) \mathbf{v}^* \mathbf{M} \mathbf{v}$$

Because both $\mathbf{v}^*\mathbf{N}\mathbf{v}$ and $\mathbf{v}^*\mathbf{M}\mathbf{v}$ are real and positive, we conclude $(1 - |\lambda|^2) > 0$ or $|\lambda|^2 < 1$. This establishes the theorem for $\mathbf{N} > 0$. The case $\mathbf{N} \geq 0$ can similarly be established.

Theorem 5.D6

If all eigenvalues of $\mathbf{A}$ have magnitudes less than 1, then the discrete Lyapunov equation

$$\mathbf{M} - \mathbf{A}' \mathbf{M} \mathbf{A} = \mathbf{N}$$

has a unique solution for every $\mathbf{N}$, and the solution can be expressed as

$$\mathbf{M} = \sum_{m=0}^{\infty} (\mathbf{A}')^m \mathbf{N} \mathbf{A}^m$$

It is important to mention that even if $\mathbf{A}$ has one or more eigenvalues with magnitudes larger than 1, a unique solution still exists in the discrete Lyapunov equation if $\lambda_i \lambda_j \neq 1$ for all i and j. In this case, the solution cannot be expressed as in (5.27) but can be computed from a set of linear algebraic equations.

Let us discuss the relationships between the continuous-time and discrete-time Lyapunov equations. The stability condition for continuous-time systems is that all eigenvalues lie inside the open left-half s-plane. The stability condition for discrete-time systems is that all eigenvalues lie inside the unit circle on the z-plane. These conditions can be related by the bilinear transformation

$$s = \frac{z - 1}{z + 1} \qquad z = \frac{1 + s}{1 - s} \tag{5.28}$$

which maps the left-half s-plane into the interior of the unit circle on the z-plane and vice versa. To differentiate the continuous-time and discrete-time cases, we write

$$\mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} = -\mathbf{N} \tag{5.29}$$

and

$$\mathbf{M}_d - \mathbf{A}_d' \mathbf{M}_d \mathbf{A}_d = \mathbf{N}_d \tag{5.30}$$

Following (5.28), these two equations can be related by

$$\mathbf{A} = (\mathbf{A}_d + \mathbf{I})^{-1}(\mathbf{A}_d - \mathbf{I}) \qquad \mathbf{A}_d = (\mathbf{I} + \mathbf{A})(\mathbf{I} - \mathbf{A})^{-1}$$

Substituting the right-hand-side equation into (5.30) and performing a simple manipulation, we find

$$\mathbf{A}'\mathbf{M}_d + \mathbf{M}_d \mathbf{A} = -0.5(\mathbf{I} - \mathbf{A}')\mathbf{N}_d(\mathbf{I} - \mathbf{A})$$

Comparing this with (5.29) yields

$$\mathbf{A} = (\mathbf{A}_d + \mathbf{I})^{-1}(\mathbf{A}_d - \mathbf{I}) \qquad \mathbf{M} = \mathbf{M}_d \qquad \mathbf{N} = 0.5(\mathbf{I} - \mathbf{A}')\mathbf{N}_d(\mathbf{I} - \mathbf{A}) \tag{5.31}$$

These relate (5.29) and (5.30).

The MATLAB function `lyap` computes the Lyapunov equation in (5.29) and `dlyap` computes the discrete Lyapunov equation in (5.30). The function `dlyap` transforms (5.30) into (5.29) by using (5.31) and then calls `lyap`. The result yields $\mathbf{M} = \mathbf{M}_d$.

5.5 Stability of LTV Systems

Consider a SISO linear time-varying (LTV) system described by

$$y(t) = \int_{t_0}^{t} g(t, \tau) u(\tau) \, d\tau \tag{5.32}$$

The system is said to be BIBO stable if every bounded input excites a bounded output. The condition for (5.32) to be BIBO stable is that there exists a finite constant M such that

$$\int_{t_0}^{t} |g(t, \tau)| \, d\tau \leq M < \infty \tag{5.33}$$

for all t and t_0 with $t \geq t_0$. The proof in the time-invariant case applies here with only minor modification.

For the multivariable case, (5.32) becomes

$$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{G}(t, \tau) \mathbf{u}(\tau) \, d\tau \tag{5.34}$$

The condition for (5.34) to be BIBO stable is that every entry of $\mathbf{G}(t, \tau)$ meets the condition in (5.33). For multivariable systems, we can also express the condition in terms of norms. Any norm discussed in Section 3.11 can be used. However, the infinite-norm

$$\|\mathbf{u}\|_\infty = \max_i |u_i| \qquad \|\mathbf{G}\|_\infty = \text{largest row absolute sum}$$

is probably the most convenient to use in stability study. For convenience, no subscript will be attached to any norm. The necessary and sufficient condition for (5.34) to be BIBO stable is that there exists a finite constant M such that

$$\int_{t_0}^{t} ||\mathbf{G}(t, \tau)||\, d\tau \leq M < \infty$$

for all t and t_0 with $t \geq t_0$.

The impulse response matrix of

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u} \\ \mathbf{y} &= \mathbf{C}(t)\mathbf{x} + \mathbf{D}(t)\mathbf{u} \end{aligned} \tag{5.35}$$

is

$$\mathbf{G}(t, \tau) = \mathbf{C}(t)\mathbf{\Phi}(t, \tau)\mathbf{B}(\tau) + \mathbf{D}(t)\delta(t - \tau)$$

and the zero-state response is

$$\mathbf{y}(t) = \int_{t_0}^{t} \mathbf{C}(t)\mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau + \mathbf{D}(t)\mathbf{u}(t)$$

Thus (5.35) or, more precisely, the zero-state response of (5.35) is BIBO stable if and only if there exist constants M_1 and M_2 such that

$$||\mathbf{D}(t)|| \leq M_1 < \infty$$

and

$$\int_{t_0}^{t} ||\mathbf{G}(t, \tau)||\, d\tau \leq M_2 < \infty$$

for all t and t_0 with $t \geq t_0$.

Next we study the stability of the zero-input response of (5.35). As in the time-invariant case, we define the zero-input response of (5.35) or the equation $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ to be marginally stable if every finite initial state excites a bounded response. Because the response is governed by

$$\mathbf{x}(t) = \mathbf{\Phi}(t, t_0)\mathbf{x}(t_0) \tag{5.36}$$

we conclude that the response is marginally stable if and only if there exists a finite constant M such that

$$||\mathbf{\Phi}(t, t_0)|| \leq M < \infty \tag{5.37}$$

for all t_0 and for all $t \geq t_0$. The equation $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ is asymptotically stable if the response excited by every finite initial state is bounded and approaches zero as $t \to \infty$. The asymptotic stability conditions are the boundedness condition in (5.37) and

$$||\mathbf{\Phi}(t, t_0)|| \to 0 \qquad \text{as } t \to \infty \tag{5.38}$$

A great deal can be said regarding these definitions and conditions. Does the constant M in (5.37) depend on t_0? What is the rate for the state transition matrix to approach 0 in (5.38)? The interested reader is referred to References [4, 15].

A time-invariant equation $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is asymptotically stable if all eigenvalues of $\mathbf{A}$ have negative real parts. Is this also true for the time-varying case? The answer is negative as the next example shows.

EXAMPLE 5.5 Consider the linear time-varying equation

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} = \begin{bmatrix} -1 & e^{2t} \\ 0 & -1 \end{bmatrix}\mathbf{x} \tag{5.39}$$

The characteristic polynomial of $\mathbf{A}(t)$ is

$$\det(\lambda\mathbf{I} - \mathbf{A}(t)) = \det\begin{bmatrix} \lambda + 1 & -e^{2t} \\ 0 & \lambda + 1 \end{bmatrix} = (\lambda + 1)^2$$

Thus $\mathbf{A}(t)$ has eigenvalues -1 and -1 for all t. It can be verified directly that

$$\mathbf{\Phi}(t, 0) = \begin{bmatrix} e^{-t} & 0.5(e^{t} - e^{-t}) \\ 0 & e^{-t} \end{bmatrix}$$

meets (4.53) and is therefore the state transition matrix of (5.39). See also Problem 4.16. Because the (1,2)th entry of $\mathbf{\Phi}$ grows without bound, the equation is neither asymptotically stable nor marginally stable. This example shows that even though the eigenvalues can be defined for $\mathbf{A}(t)$ at every t, the concept of eigenvalues is not useful in the time-varying case.

All stability properties in the time-invariant case are invariant under any equivalence transformation. In the time-varying case, this is so only for BIBO stability, because the impulse response matrix is preserved. An equivalence transformation can transform, as shown in Theorem 4.3, any $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ into $\dot{\bar{\mathbf{x}}} = \mathbf{A}_o\bar{\mathbf{x}}$, where $\mathbf{A}_o$ is any constant matrix; therefore, in the time-varying case, marginal and asymptotic stabilities are not invariant under any equivalence transformation.

▶ **Theorem 5.7**

Marginal and asymptotic stabilities of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ are invariant under any Lyapunov transformation.

As discussed in Section 4.6, if $\mathbf{P}(t)$ and $\dot{\mathbf{P}}(t)$ are continuous, and $\mathbf{P}(t)$ is nonsingular for all t, then $\bar{\mathbf{x}} = \mathbf{P}(t)\mathbf{x}$ is an algebraic transformation. If, in addition, $\mathbf{P}(t)$ and $\mathbf{P}^{-1}(t)$ are bounded for all t, then $\bar{\mathbf{x}} = \mathbf{P}(t)\mathbf{x}$ is a Lyapunov transformation. The fundamental matrix $\mathbf{X}(t)$ of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$ and the fundamental matrix $\bar{\mathbf{X}}(t)$ of $\dot{\bar{\mathbf{x}}} = \bar{\mathbf{A}}(t)\bar{\mathbf{x}}$ are related by, as derived in (4.71),

$$\bar{\mathbf{X}}(t) = \mathbf{P}(t)\mathbf{X}(t)$$

which implies

$$\begin{aligned} \bar{\mathbf{\Phi}}(t, \tau) &= \bar{\mathbf{X}}(t)\bar{\mathbf{X}}^{-1}(\tau) = \mathbf{P}(t)\mathbf{X}(t)\mathbf{X}^{-1}(\tau)\mathbf{P}^{-1}(\tau) \\ &= \mathbf{P}(t)\mathbf{\Phi}(t, \tau)\mathbf{P}^{-1}(\tau) \end{aligned} \tag{5.40}$$

Because both $\mathbf{P}(t)$ and $\mathbf{P}^{-1}(t)$ are bounded, if $||\mathbf{\Phi}(t, \tau)||$ is bounded, so is $||\bar{\mathbf{\Phi}}(t, \tau)||$; if $||\mathbf{\Phi}(t, \tau)|| \to 0$ as $t \to \infty$, so is $||\bar{\mathbf{\Phi}}(t, \tau)||$. This establishes Theorem 5.7.

In the time-invariant case, asymptotic stability of zero-input responses always implies BIBO stability of zero-state responses. This is not necessarily so in the time-varying case. A time-varying equation is asymptotically stable if

$$||\mathbf{\Phi}(t, t_0)|| \to 0 \qquad \text{as } t \to \infty \tag{5.41}$$

for all t, t_0 with $t \geq t_0$. It is BIBO stable if

$$\int_{t_0}^{t} ||\mathbf{C}(t)\mathbf{\Phi}(t, \tau)\mathbf{B}(\tau)||\, d\tau < \infty \tag{5.42}$$

for all t, t_0 with $t \geq t_0$. A function that approaches 0, as $t \to \infty$, may not be absolutely integrable. Thus asymptotic stability may not imply BIBO stability in the time-varying case. However, if $||\mathbf{\Phi}(t, \tau)||$ decreases to zero rapidly, in particular, exponentially, and if $\mathbf{C}(t)$ and $\mathbf{B}(t)$ are bounded for all t, then asymptotic stability does imply BIBO stability. See References [4, 6, 15].

PROBLEMS

5.1 Is the network shown in Fig. 5.2 BIBO stable? If not, find a bounded input that will excite an unbounded output.

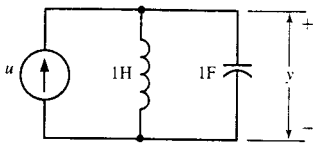


Figure 5.2

5.2 Consider a system with an irrational transfer function $\hat{g}(s)$. Show that a necessary condition for the system to be BIBO stable is that $|\hat{g}(s)|$ is finite for all Re $s \geq 0$.

5.3 Is a system with impulse response $g(t) = 1/(t+1)$ BIBO stable? How about $g(t) = te^{-t}$ for $t \geq 0$?

5.4 Is a system with transfer function $\hat{g}(s) = e^{-2s}/(s+1)$ BIBO stable?

5.5 Show that the negative-feedback system shown in Fig. 2.5(b) is BIBO stable if and only if the gain a has a magnitude less than 1. For $a = 1$, find a bounded input $r(t)$ that will excite an unbounded output.

5.6 Consider a system with transfer function $\hat{g}(s) = (s-2)/(s+1)$. What are the steady-state responses excited by $u(t) = 3$, for $t \geq 0$, and by $u(t) = \sin 2t$, for $t \geq 0$?

5.7 Consider

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 10 \\ 0 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} -2 \\ 0 \end{bmatrix}u$$

$$y = [-2 \quad 3]\mathbf{x} - 2u$$

Is it BIBO stable?

5.8 Consider a discrete-time system with impulse response sequence

$$g[k] = k(0.8)^k \qquad \text{for } k \geq 0$$

Is the system BIBO stable?

5.9 Is the state equation in Problem 5.7 marginally stable? Asymptotically stable?

5.10 Is the homogeneous state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\mathbf{x}$$

marginally stable? Asymptotically stable?

5.11 Is the homogeneous state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}\mathbf{x}$$

marginally stable? Asymptotically stable?

5.12 Is the discrete-time homogeneous state equation

$$\mathbf{x}[k+1] = \begin{bmatrix} 0.9 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\mathbf{x}[k]$$

marginally stable? Asymptotically stable?

5.13 Is the discrete-time homogeneous state equation

$$\mathbf{x}[k+1] = \begin{bmatrix} 0.9 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}\mathbf{x}[k]$$

marginally stable? Asymptotically stable?

5.14 Use Theorem 5.5 to show that all eigenvalues of

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

have negative real parts.

5.15 Use Theorem 5.D5 to show that all eigenvalues of the $\mathbf{A}$ in Problem 5.14 have magnitudes less than 1.

5.16 For any distinct negative real λ_i and any nonzero real a_i, show that the matrix

$$\mathbf{M} = \begin{bmatrix} -\frac{a_1^2}{2\lambda_1} & -\frac{a_1 a_2}{\lambda_1+\lambda_2} & -\frac{a_1 a_3}{\lambda_1+\lambda_3} \\ -\frac{a_2 a_1}{\lambda_2+\lambda_1} & -\frac{a_2^2}{2\lambda_2} & -\frac{a_2 a_3}{\lambda_2+\lambda_3} \\ -\frac{a_3 a_1}{\lambda_3+\lambda_1} & -\frac{a_3 a_2}{\lambda_3+\lambda_2} & -\frac{a_3^2}{2\lambda_3} \end{bmatrix}$$

is positive definite. [*Hint*: Use Corollary 5.5 and $\mathbf{A}$=diag$(\lambda_1, \lambda_2, \lambda_3)$.]

5.17 A real matrix $\mathbf{M}$ (not necessarily symmetric) is defined to be positive definite if $\mathbf{x}'\mathbf{M}\mathbf{x} > 0$ for any nonzero $\mathbf{x}$. Is it true that the matrix $\mathbf{M}$ is positive definite if all eigenvalues of $\mathbf{M}$ are real and positive or if all its leading principal minors are positive? If not, how do you check its positive definiteness? [*Hint*: Try

$$\begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix} \quad \begin{bmatrix} 2 & 1 \\ 1.9 & 1 \end{bmatrix}]$$

5.18 Show that all eigenvalues of $\mathbf{A}$ have real parts less than $-\mu < 0$ if and only if, for any given positive definite symmetric matrix $\mathbf{N}$, the equation

$$\mathbf{A}'\mathbf{M} + \mathbf{M}\mathbf{A} + 2\mu\mathbf{M} = -\mathbf{N}$$

has a unique symmetric solution $\mathbf{M}$ and $\mathbf{M}$ is positive definite.

5.19 Show that all eigenvalues of $\mathbf{A}$ have magnitudes less than ρ if and only if, for any given positive definite symmetric matrix $\mathbf{N}$, the equation

$$\rho^2\mathbf{M} - \mathbf{A}'\mathbf{M}\mathbf{A} = \rho^2\mathbf{N}$$

has a unique symmetric solution $\mathbf{M}$ and $\mathbf{M}$ is positive definite.

5.20 Is a system with impulse response $g(t, \tau) = e^{-2|t|-|\tau|}$, for $t \geq \tau$, BIBO stable? How about $g(t, \tau) = \sin t(e^{-(t-\tau)}) \cos \tau$?

5.21 Consider the time-varying equation

$$\dot{x} = 2tx + u \qquad y = e^{-t^2}x$$

Is the equation BIBO stable? Marginally stable? Asymptotically stable?

5.22 Show that the equation in Problem 5.21 can be transformed by using $\bar{x} = P(t)x$, with $P(t) = e^{-t^2}$, into

$$\dot{\bar{x}} = 0 \cdot \bar{x} + e^{-t^2}u \qquad y = \bar{x}$$

Is the equation BIBO stable? Marginally stable? Asymptotically stable? Is the transformation a Lyapunov transformation?

5.23 Is the homogeneous equation

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 0 \\ -e^{-3t} & 0 \end{bmatrix}\mathbf{x}$$

for $t_0 \geq 0$, marginally stable? Asymptotically stable?

Chapter 6

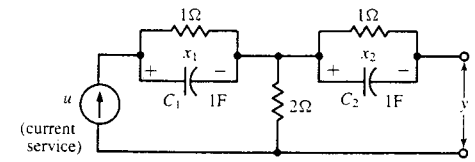
Controllability and Observability

6.1 Introduction

This chapter introduces the concepts of controllability and observability. Controllability deals with whether or not the state of a state-space equation can be controlled from the input, and observability deals with whether or not the initial state can be observed from the output. These concepts can be illustrated using the network shown in Fig. 6.1. The network has two state variables. Let x_i be the voltage across the capacitor with capacitance C_i, for $i = 1, 2$. The input u is a current source and the output y is the voltage shown. From the network, we see that, because of the open circuit across y, the input has no effect on x_2 or cannot control x_2. The current passing through the 2-Ω resistor always equals the current source u; therefore the response excited by the initial state x_1 will not appear in y. Thus the initial state x_1 cannot be observed from the output. Thus the equation describing the network cannot be controllable and observable.

These concepts are essential in discussing the internal structure of linear systems. They are also needed in studying control and filtering problems. We study first continuous-time

Figure 6.1 Network.



linear time-invariant (LTI) state equations and then discrete-time LTI state equations. Finally, we study the time-varying case.

6.2 Controllability

Consider the n-dimensional p-input state equation

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \tag{6.1}$$

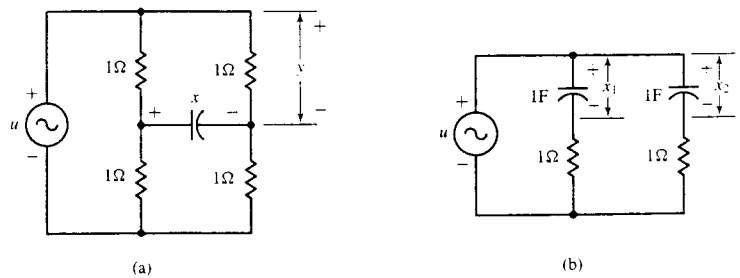
where $\mathbf{A}$ and $\mathbf{B}$ are, respectively, $n \times n$ and $n \times p$ real constant matrices. Because the output does not play any role in controllability, we will disregard the output equation in this study.

> ***Definition 6.1*** *The state equation (6.1) or the pair* $(\mathbf{A}, \mathbf{B})$ *is said to be* controllable *if for any initial state* $\mathbf{x}(0) = \mathbf{x}_0$ *and any final state* $\mathbf{x}_1$, *there exists an input that transfers* $\mathbf{x}_0$ *to* $\mathbf{x}_1$ *in a finite time. Otherwise (6.1) or* $(\mathbf{A}, \mathbf{B})$ *is said to be* uncontrollable.

This definition requires only that the input be capable of moving any state in the state space to any other state in a finite time; what trajectory the state should take is not specified. Furthermore, there is no constraint imposed on the input; its magnitude can be as large as desired. We give an example to illustrate the concept.

EXAMPLE 6.1 Consider the network shown in Fig. 6.2(a). Its state variable x is the voltage across the capacitor. If $x(0) = 0$, then $x(t) = 0$ for all $t \geq 0$ no matter what input is applied. This is due to the symmetry of the network, and the input has no effect on the voltage across the capacitor. Thus the system or, more precisely, the state equation that describes the system is not controllable.

Next we consider the network shown in Fig. 6.2(b). It has two state variables x_1 and x_2 as shown. The input can transfer x_1 *or* x_2 to any value; but it cannot transfer x_1 *and* x_2 to any values. For example, if $x_1(0) = x_2(0) = 0$, then no matter what input is applied, $x_1(t)$ always equals $x_2(t)$ for all $t \geq 0$. Thus the equation that describes the network is not controllable.



(a) (b)

Figure 6.2 Uncontrollable networks.

Theorem 6.1

The following statements are equivalent.

1. The n-dimensional pair $(\mathbf{A}, \mathbf{B})$ is controllable.
2. The $n \times n$ matrix

$$\mathbf{W}_c(t) = \int_0^t e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'\tau}\,d\tau = \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'(t-\tau)}\,d\tau \tag{6.2}$$

   is nonsingular for any $t > 0$.
3. The $n \times np$ *controllability matrix*

$$C = [\mathbf{B}\ \mathbf{A}\mathbf{B}\ \mathbf{A}^2\mathbf{B}\ \cdots\ \mathbf{A}^{n-1}\mathbf{B}] \tag{6.3}$$

   has rank n (full row rank).
4. The $n \times (n + p)$ matrix $[\mathbf{A} - \lambda\mathbf{I}\ \ \mathbf{B}]$ has full row rank at every eigenvalue, λ, of $\mathbf{A}$.[1]
5. If, in addition, all eigenvalues of $\mathbf{A}$ have negative real parts, then the unique solution of

$$\mathbf{A}\mathbf{W}_c + \mathbf{W}_c\mathbf{A}' = -\mathbf{B}\mathbf{B}' \tag{6.4}$$

   is positive definite. The solution is called the *controllability Gramian* and can be expressed as

$$\mathbf{W}_c = \int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'\tau}d\tau \tag{6.5}$$

⇒ ***Proof:*** (1) ↔ (2): First we show the equivalence of the two forms in (6.2). Define $\bar{\tau} := t - \tau$. Then we have

$$\int_{\tau=0}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'(t-\tau)}\,d\tau = \int_{\bar{\tau}=t}^{0} e^{\mathbf{A}\bar{\tau}}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'\bar{\tau}}(-d\bar{\tau})$$

$$= \int_{\bar{\tau}=0}^{t} e^{\mathbf{A}\bar{\tau}}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'\bar{\tau}}d\bar{\tau}$$

It becomes the first form of (6.2) after replacing $\bar{\tau}$ by τ. Because of the form of the integrand, $\mathbf{W}_c(t)$ is always positive semidefinite; it is positive definite if and only if it is nonsingular. See Section 3.9.

First we show that if $\mathbf{W}_c(t)$ is nonsingular, then (6.1) is controllable. The response of (6.1) at time t_1 was derived in (4.5) as

$$\mathbf{x}(t_1) = e^{\mathbf{A}t_1}\mathbf{x}(0) + \int_0^{t_1} e^{\mathbf{A}(t_1-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau \tag{6.6}$$

We claim that for any $\mathbf{x}(0) = \mathbf{x}_0$ and any $\mathbf{x}(t_1) = \mathbf{x}_1$, the input

$$\mathbf{u}(t) = -\mathbf{B}'e^{\mathbf{A}'(t_1-t)}\mathbf{W}_c^{-1}(t_1)[e^{\mathbf{A}t_1}\mathbf{x}_0 - \mathbf{x}_1] \tag{6.7}$$

will transfer $\mathbf{x}_0$ to $\mathbf{x}_1$ at time t_1. Indeed, substituting (6.7) into (6.6) yields

1. If λ is complex, then we must use complex numbers as scalars in checking the rank. See the discussion regarding (3.37).

$$\mathbf{x}(t_1) = e^{\mathbf{A}t_1}\mathbf{x}_0 - \left(\int_0^{t_1} e^{\mathbf{A}(t_1-\tau)}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'(t_1-\tau)}d\tau\right)\mathbf{W}_c^{-1}(t_1)[e^{\mathbf{A}t_1}\mathbf{x}_0 - \mathbf{x}_1]$$

$$= e^{\mathbf{A}t_1}\mathbf{x}_0 - \mathbf{W}_c(t_1)\mathbf{W}_c^{-1}(t_1)[e^{\mathbf{A}t_1}\mathbf{x}_0 - \mathbf{x}_1] = \mathbf{x}_1$$

This shows that if $\mathbf{W}_c$ is nonsingular, then the pair $(\mathbf{A}, \mathbf{B})$ is controllable. We show the converse by contradiction. Suppose the pair is controllable but $\mathbf{W}_c(t_1)$ is not positive definite for some t_1. Then there exists an $n \times 1$ nonzero vector $\mathbf{v}$ such that

$$\mathbf{v}'\mathbf{W}_c(t_1)\mathbf{v} = \int_0^{t_1} \mathbf{v}'e^{\mathbf{A}(t_1-\tau)}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'(t_1-\tau)}\mathbf{v}d\tau$$

$$= \int_0^{t_1} \|\mathbf{B}'e^{\mathbf{A}'(t_1-\tau)}\mathbf{v}\|^2 d\tau = 0$$

which implies

$$\mathbf{B}'e^{\mathbf{A}'(t_1-\tau)}\mathbf{v} \equiv \mathbf{0} \quad \text{or} \quad \mathbf{v}'e^{\mathbf{A}(t_1-\tau)}\mathbf{B} \equiv \mathbf{0} \tag{6.8}$$

for all τ in $[0, t_1]$. If (6.1) is controllable, there exists an input that transfers the initial state $\mathbf{x}(0) = e^{-\mathbf{A}t_1}\mathbf{v}$ to $\mathbf{x}(t_1) = \mathbf{0}$ and (6.6) becomes

$$\mathbf{0} = \mathbf{v} + \int_0^{t_1} e^{\mathbf{A}(t_1-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau$$

Its premultiplication by $\mathbf{v}'$ yields

$$0 = \mathbf{v}'\mathbf{v} + \int_0^{t_1} \mathbf{v}'e^{\mathbf{A}(t_1-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau = \|\mathbf{v}\|^2 + 0$$

which contradicts $\mathbf{v} \neq \mathbf{0}$. This establishes the equivalence of (1) and (2).

(2) $\leftrightarrow$ (3): Because every entry of $e^{\mathbf{A}t}\mathbf{B}$ is, as discussed at the end of Section 4.2, an analytical function of t, if $\mathbf{W}_c(t)$ is nonsingular for some t, then it is nonsingular for all t in $(-\infty, \infty)$. See Reference [6, p. 554]. Because of the equivalence of the two forms in (6.2), (6.8) implies that $\mathbf{W}_c(t)$ is nonsingular if and only if there exists no $n \times 1$ nonzero vector $\mathbf{v}$ such that

$$\mathbf{v}'e^{\mathbf{A}t}\mathbf{B} = \mathbf{0} \qquad \text{for all } t \tag{6.9}$$

Now we show that if $\mathbf{W}_c(t)$ is nonsingular, then the controllability matrix C has full row rank. Suppose C does not have full row rank, then there exists an $n \times 1$ nonzero vector $\mathbf{v}$ such that $\mathbf{v}'C = \mathbf{0}$ or

$$\mathbf{v}'\mathbf{A}^k\mathbf{B} = \mathbf{0} \qquad \text{for } k = 0, 1, 2, \ldots, n-1$$

Because $e^{\mathbf{A}t}\mathbf{B}$ can be expressed as a linear combination of $\{\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{n-1}\mathbf{B}\}$ (Theorem 3.5), we conclude $\mathbf{v}'e^{\mathbf{A}t}\mathbf{B} = \mathbf{0}$. This contradicts the nonsingularity assumption of $\mathbf{W}_c(t)$. Thus Condition (2) implies Condition (3). To show the converse, suppose C has full row rank but $\mathbf{W}_c(t)$ is singular. Then there exists a nonzero $\mathbf{v}$ such that (6.9) holds. Setting $t = 0$, we have $\mathbf{v}'\mathbf{B} = \mathbf{0}$. Differentiating (6.9) and then setting $t = 0$, we have $\mathbf{v}'\mathbf{AB} = \mathbf{0}$. Proceeding forward yields $\mathbf{v}'\mathbf{A}^k\mathbf{B} = \mathbf{0}$ for $k = 0, 1, 2, \ldots$. They can be arranged as

$$\mathbf{v}'[\mathbf{B}\ \mathbf{AB}\ \cdots\ \mathbf{A}^{n-1}\mathbf{B}] = \mathbf{v}'C = \mathbf{0}$$

This contradicts the hypothesis that C has full row rank. This shows the equivalence of (2) and (3).

(3) $\leftrightarrow$ (4): If C has full row rank, then the matrix $[\mathbf{A} - \lambda\mathbf{I}\ \ \mathbf{B}]$ has full row rank at every eigenvalue of $\mathbf{A}$. If not, there exists an eigenvalue λ_1 and a $1 \times n$ vector $\mathbf{q} \neq \mathbf{0}$ such that

$$\mathbf{q}[\mathbf{A} - \lambda_1\mathbf{I}\ \ \mathbf{B}] = \mathbf{0}$$

which implies $\mathbf{qA} = \lambda_1\mathbf{q}$ and $\mathbf{qB} = \mathbf{0}$. Thus $\mathbf{q}$ is a left eigenvector of $\mathbf{A}$. We compute

$$\mathbf{qA}^2 = (\mathbf{qA})\mathbf{A} = (\lambda_1\mathbf{q})\mathbf{A} = \lambda_1^2\mathbf{q}$$

Proceeding forward, we have $\mathbf{qA}^k = \lambda_1^k\mathbf{q}$. Thus we have

$$\mathbf{q}[\mathbf{B}\ \mathbf{AB}\ \cdots\ \mathbf{A}^{n-1}\mathbf{B}] = [\mathbf{qB}\ \lambda_1\mathbf{qB}\ \cdots\ \lambda_1^{n-1}\mathbf{qB}] = \mathbf{0}$$

This contradicts the hypothesis that C has full row rank.

In order to show that $\rho(C) < n$ implies $\rho([\mathbf{A} - \lambda\mathbf{I}\ \ \mathbf{B}]) < n$ at some eigenvalue λ_1 of $\mathbf{A}$, we need Theorems 6.2 and 6.6, which will be established later. Theorem 6.2 states that controllability is invaraint under any equivalence transformation. Therefore we may show $\rho([\bar{\mathbf{A}} - \lambda\mathbf{I}\ \bar{\mathbf{B}}]) < n$ at some eigenvalue of $\bar{\mathbf{A}}$, where $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is equivalent to $(\mathbf{A}, \mathbf{B})$. Theorem 6.6 states that if the rank of C is less than n or $\rho(C) = n - m$, for some integer $m \geq 1$, then there exists a nonsingular matrix $\mathbf{P}$ such that

$$\bar{\mathbf{A}} = \mathbf{PAP}^{-1} = \begin{bmatrix} \bar{\mathbf{A}}_c & \bar{\mathbf{A}}_{12} \\ \mathbf{0} & \bar{\mathbf{A}}_{\bar{c}} \end{bmatrix} \qquad \bar{\mathbf{B}} = \mathbf{PB} = \begin{bmatrix} \bar{\mathbf{B}}_c \\ \mathbf{0} \end{bmatrix}$$

where $\bar{\mathbf{A}}_{\bar{c}}$ is $m \times m$. Let λ_1 be an eigenvalue of $\bar{\mathbf{A}}_{\bar{c}}$ and $\mathbf{q}_1$ be a corresponding $1 \times m$ nonzero left eigenvector or $\mathbf{q}_1\bar{\mathbf{A}}_{\bar{c}} = \lambda_1\mathbf{q}_1$. Then we have $\mathbf{q}_1(\bar{\mathbf{A}}_{\bar{c}} - \lambda_1\mathbf{I}) = \mathbf{0}$. Now we form the $1 \times n$ vector $\mathbf{q} := [\mathbf{0}\ \ \mathbf{q}_1]$. We compute

$$\mathbf{q}[\bar{\mathbf{A}} - \lambda_1\mathbf{I}\ \ \bar{\mathbf{B}}] = [\mathbf{0}\ \ \mathbf{q}_1]\begin{bmatrix} \bar{\mathbf{A}}_c - \lambda_1\mathbf{I} & \bar{\mathbf{A}}_{12} & \bar{\mathbf{B}}_c \\ \mathbf{0} & \bar{\mathbf{A}}_{\bar{c}} - \lambda_1\mathbf{I} & \mathbf{0} \end{bmatrix} = \mathbf{0} \tag{6.10}$$

which implies $\rho([\bar{\mathbf{A}} - \lambda\mathbf{I}\ \bar{\mathbf{B}}]) < n$ and, consequently, $\rho([\mathbf{A} - \lambda\mathbf{I}\ \ \mathbf{B}]) < n$ at some eigenvalue of $\mathbf{A}$. This establishes the equivalence of (3) and (4).

(2) $\leftrightarrow$ (5): If $\mathbf{A}$ is stable, then the unique solution of (6.4) can be expressed as in (6.5) (Theorem 5.6). The Gramian $\mathbf{W}_c$ is always positive semidefinite. It is positive definite if and only if $\mathbf{W}_c$ is nonsingular. This establishes the equivalence of (2) and (5). Q.E.D.

EXAMPLE 6.2 Consider the inverted pendulum studied in Example 2.8. Its state equation was developed in (2.27). Suppose for a given pendulum, the equation becomes

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 5 & 0 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ -2 \end{bmatrix}u \tag{6.11}$$

$$y = [1\ 0\ 0\ 0]\mathbf{x}$$

We compute

$$C = [\mathbf{B}\ \mathbf{AB}\ \mathbf{A}^2\mathbf{B}\ \mathbf{A}^3\mathbf{B}] = \begin{bmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 2 & 0 \\ 0 & -2 & 0 & -10 \\ -2 & 0 & -10 & 0 \end{bmatrix}$$

This matrix can be shown to have rank 4; thus the system is controllable. Therefore, if $x_3 = \theta$ deviates from zero slightly, we can find a control u to push it back to zero. In fact, a control exists to bring $x_1 = y$, x_3, and their derivatives back to zero. This is consistent with our experience of balancing a broom on our palm.

The MATLAB functions `ctrb` and `gram` will generate the controllability matrix and controllability Gramian. Note that the controllability Gramian is not computed from (6.5); it is obtained by solving a set of linear algebraic equations. Whether a state equation is controllable can then be determined by computing the rank of the controllability matrix or Gramian by using `rank` in MATLAB.

EXAMPLE 6.3 Consider the platform system shown in Fig. 6.3; it can be used to study suspension systems of automobiles. The system consists of one platform; both ends of the platform are supported on the ground by means of springs and dashpots, which provide viscous friction. The mass of the platform is assumed to be zero; thus the movements of the two spring systems are independent and half of the force is applied to each spring system. The spring constants of both springs are assumed to be 1 and the viscous friction coefficients are assumed to be 2 and 1 as shown. If the displacements of the two spring systems from equilibrium are chosen as state variables x_1 and x_2, then we have $x_1 + 2\dot{x}_1 = u$ and $x_2 + \dot{x}_2 = u$ or

$$\dot{\mathbf{x}} = \begin{bmatrix} -0.5 & 0 \\ 0 & -1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} u \tag{6.12}$$

This state equation describes the system.

Now if the initial displacements are different from zero, and if no force is applied, the platform will return to zero exponentially. In theory, it will take an infinite time for x_i to equal 0 exactly. Now we pose the problem. If $x_1(0) = 10$ and $x_2(0) = -1$, can we apply a force to bring the platform to equilibrium in 2 seconds? The answer does not seem to be obvious because the *same* force is applied to the two spring systems.

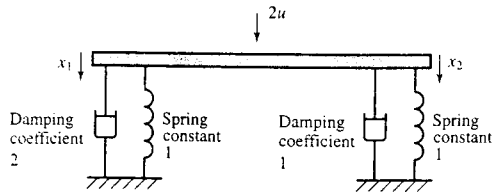


Figure 6.3 Platform system.

For Equation (6.12), we compute

$$\rho([\mathbf{B}\ \mathbf{AB}]) = \rho\begin{bmatrix} 0.5 & -0.25 \\ 1 & -1 \end{bmatrix} = 2$$

Thus the equation is controllable and, for any $\mathbf{x}(0)$, there exists an input that transfers $\mathbf{x}(0)$ to $\mathbf{0}$ in 2 seconds or in any finite time. We compute (6.2) and (6.7) for this system at $t_1 = 2$:

$$\mathbf{W}_c(2) = \int_0^2 \left(\begin{bmatrix} e^{-0.5\tau} & 0 \\ 0 & e^{-\tau} \end{bmatrix} \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} [0.5\ \ 1] \begin{bmatrix} e^{-0.5\tau} & 0 \\ 0 & e^{-\tau} \end{bmatrix} \right) d\tau$$
$$= \begin{bmatrix} 0.2162 & 0.3167 \\ 0.3167 & 0.4908 \end{bmatrix}$$

and

$$u_1(t) = -[0.5\ \ 1] \begin{bmatrix} e^{-0.5(2-t)} & 0 \\ 0 & e^{-(2-t)} \end{bmatrix} \mathbf{W}_c^{-1}(2) \begin{bmatrix} e^{-1} & 0 \\ 0 & e^{-2} \end{bmatrix} \begin{bmatrix} 10 \\ -1 \end{bmatrix}$$
$$= -58.82e^{0.5t} + 27.96e^{t}$$

for t in $[0,\ 2]$. This input force will transfer $\mathbf{x}(0) = [10\ \ -1]'$ to $[0\ \ 0]'$ in 2 seconds as shown in Fig. 6.4(a) in which the input is also plotted. It is obtained by using the MATLAB function `lsim`, an acronym for *linear simulation*. The largest magnitude of the input is about 45.

Figure 6.4(b) plots the input u_2 that transfers $\mathbf{x}(0) = [10\ \ -1]'$ to $\mathbf{0}$ in 4 seconds. We see that the smaller the time interval, the larger the input magnitude. If no restriction is imposed on the input, we can transfer $\mathbf{x}(0)$ to zero in an arbitrarily small time interval; however, the input magnitude may become very large. If some restriction is imposed on the input magnitude, then we cannot achieve the transfer as fast as desired. For example, if we require $|u(t)| < 9$, for all t, in Example 6.3, then we cannot transfer $\mathbf{x}(0)$ to $\mathbf{0}$ in less than 4 seconds. We remark that the input $\mathbf{u}(t)$ in (6.7) is called the *minimal energy control* in the sense that for any other input $\bar{\mathbf{u}}(t)$ that achieves the same transfer, we have

$$\int_{t_0}^{t_1} \bar{\mathbf{u}}'(t)\bar{\mathbf{u}}(t)\,dt \geq \int_{t_0}^{t_1} \mathbf{u}'(t)\mathbf{u}(t)\,dt$$

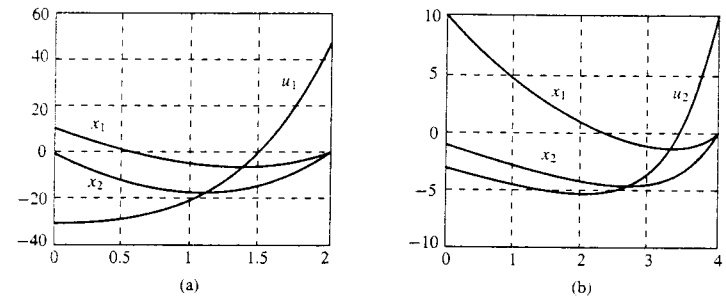


Figure 6.4 Transfer $\mathbf{x}(0) = [10\ \ -1]'$ to $[0\ 0]'$.

Its proof can be found in Reference [6, pp. 556–558].

EXAMPLE 6.4 Consider again the platform system shown in Fig. 6.3. We now assume that the viscous friction coefficients and spring constants of both spring systems all equal 1. Then the state equation that describes the system becomes

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u$$

Clearly we have

$$\rho(C) = \rho \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} = 1$$

and the state equation is not controllable. If $x_1(0) \neq x_2(0)$, no input can transfer $\mathbf{x}(0)$ to zero in a finite time.

6.2.1 Controllability Indices

Let $\mathbf{A}$ and $\mathbf{B}$ be $n \times n$ and $n \times p$ constant matrices. We assume that $\mathbf{B}$ has rank p or full column rank. If $\mathbf{B}$ does not have full column rank, there is a redundancy in inputs. For example, if the second column of $\mathbf{B}$ equals the first column of $\mathbf{B}$, then the effect of the second input on the system can be generated from the first input. Thus the second input is redundant. In conclusion, deleting linearly dependent columns of $\mathbf{B}$ and the corresponding inputs will not affect the control of the system. Thus it is reasonable to assume that $\mathbf{B}$ has full column rank.

If $(\mathbf{A}, \mathbf{B})$ is controllable, its controllability matrix C has rank n and, consequently, n linearly independent columns. Note that there are np columns in C; therefore it is possible to find many sets of n linearly independent columns in C. We discuss in the following the most important way of searching these columns; the search also happens to be most natural. Let $\mathbf{b}_i$ be the ith column of $\mathbf{B}$. Then C can be written explicitly as

$$C = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_p \vdots \mathbf{A}\mathbf{b}_1 \ \cdots \ \mathbf{A}\mathbf{b}_p \vdots \ \cdots \ \vdots \mathbf{A}^{n-1}\mathbf{b}_1 \ \cdots \ \mathbf{A}^{n-1}\mathbf{b}_p] \tag{6.13}$$

Let us search linearly independent columns of C from left to right. Because of the pattern of C, if $\mathbf{A}^i\mathbf{b}_m$ depends on its left-hand-side (LHS) columns, then $\mathbf{A}^{i+1}\mathbf{b}_m$ will also depend on its LHS columns. It means that once a column associated with $\mathbf{b}_m$ becomes linearly dependent, then all columns associated with $\mathbf{b}_m$ thereafter are linearly dependent. Let μ_m be the number of the linearly independent columns associated with $\mathbf{b}_m$ in C. That is, the columns

$$\mathbf{b}_m, \ \mathbf{A}\mathbf{b}_m, \ \ldots, \ \mathbf{A}^{\mu_m - 1}\mathbf{b}_m$$

are linearly independent in C and $\mathbf{A}^{\mu_m + i}\mathbf{b}_m$ are linearly dependent for $i = 0, 1, \ldots$. It is clear that if C has rank n, then

$$\mu_1 + \mu_2 + \cdots + \mu_p = n \tag{6.14}$$

The set $\{\mu_1, \mu_2, \ldots, \mu_p\}$ is called the *controllability indices* and

$$\mu = \max(\mu_1, \ \mu_2, \ \ldots, \ \mu_p)$$

is called the *controllability index* of $(\mathbf{A}, \mathbf{B})$. Or, equivalently, if $(\mathbf{A}, \mathbf{B})$ is controllable, the controllability index μ is the least integer such that

$$\rho(C_\mu) = \rho([\mathbf{B} \ \mathbf{A}\mathbf{B} \ \cdots \ \mathbf{A}^{\mu-1}\mathbf{B}]) = n \tag{6.15}$$

Now we give a range of μ. If $\mu_1 = \mu_2 = \cdots = \mu_p$, then $n/p \leq \mu$. If all μ_m, except one, equal 1, then $\mu = n - (p - 1)$; this is the largest possible μ. Let $\bar{n}$ be the degree of the minimal polynomial of $\mathbf{A}$. Then, by definition, there exist α_i such that

$$\mathbf{A}^{\bar{n}} = \alpha_1\mathbf{A}^{\bar{n}-1} + \alpha_2\mathbf{A}^{\bar{n}-2} + \cdots + \alpha_{\bar{n}}\mathbf{I}$$

which implies that $\mathbf{A}^{\bar{n}}\mathbf{B}$ can be written as a linear combination of $\{\mathbf{B}, \mathbf{AB}, \ldots, \mathbf{A}^{\bar{n}-1}\mathbf{B}\}$. Thus we conclude

$$n/p \leq \mu \leq \min(\bar{n}, n - p + 1) \tag{6.16}$$

where $\rho(\mathbf{B}) = p$. Because of (6.16), in checking controllability, it is unnecessary to check the $n \times np$ matrix C. It is sufficient to check a matrix of lesser columns. Because the degree of the minimal polynomial is generally not available—whereas the rank of $\mathbf{B}$ can readily be computed—we can use the following corollary to check controllability. The second part of the corollary follows Theorem 3.8.

Corollary 6.1

The n-dimensional pair $(\mathbf{A}, \mathbf{B})$ is controllable if and only if the matrix

$$C_{n-p+1} := [\mathbf{B} \ \mathbf{AB} \ \cdots \ \mathbf{A}^{n-p}\mathbf{B}] \tag{6.17}$$

where $\rho(\mathbf{B}) = p$, has rank n or the $n \times n$ matrix $C_{n-p+1}C'_{n-p+1}$ is nonsingular.

EXAMPLE 6.5 Consider the satellite system studied in Fig. 2.13. Its linearized state equation was developed in (2.29). From the equation, we can see that the control of the first four state variables by the first two inputs and the control of the last two state variables by the last input are decoupled; therefore we can consider only the following subequation of (2.29):

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & -2 & 0 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{u}$$

$$\mathbf{y} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \mathbf{x} \tag{6.18}$$

where we have assumed, for simplicity, $\omega_o = m = r_o = 1$. The controllability matrix of (6.18) is of order 4×8. If we use Corollary 6.1, then we can check its controllability by using the following 4×6 matrix:

$$[\mathbf{B} \ \mathbf{AB} \ \mathbf{A}^2\mathbf{B}] = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 0 \\ 0 & 1 & -2 & 0 & 0 & -4 \end{bmatrix} \tag{6.19}$$

It has rank 4. Thus (6.18) is controllable. From (6.19), we can readily verify that the controllability indices are 2 and 2, and the controllability index is 2.

Theorem 6.2

The controllability property is invariant under any equivalence transformation.

⇒ ***Proof:*** Consider the pair $(\mathbf{A}, \mathbf{B})$ with controllability matrix

$$C = [\mathbf{B}\ \mathbf{AB}\ \cdots\ \mathbf{A}^{n-1}\mathbf{B}]$$

and its equivalent pair $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ with $\bar{\mathbf{A}} = \mathbf{PAP}^{-1}$ and $\bar{\mathbf{B}} = \mathbf{PB}$, where $\mathbf{P}$ is a nonsingular matrix. The controllability matrix of $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$ is

$$\begin{aligned}\bar{C} &= [\bar{\mathbf{B}}\ \bar{\mathbf{A}}\bar{\mathbf{B}}\ \cdots\ \bar{\mathbf{A}}^{n-1}\bar{\mathbf{B}}] \\ &= [\mathbf{PB}\ \mathbf{PAP}^{-1}\mathbf{PB}\ \cdots\ \mathbf{PA}^{n-1}\mathbf{P}^{-1}\mathbf{PB}] \\ &= [\mathbf{PB}\ \mathbf{PAB}\ \cdots\ \mathbf{PA}^{n-1}\mathbf{B}] \\ &= \mathbf{P}[\mathbf{B}\ \mathbf{AB}\ \cdots\ \mathbf{A}^{n-1}\mathbf{B}] = \mathbf{P}C\end{aligned} \tag{6.20}$$

Because $\mathbf{P}$ is nonsingular, we have $\rho(C) = \rho(\bar{C})$ (see Equation (3.62)). This establishes Theorem 6.2. Q.E.D.

Theorem 6.3

The set of the controllability indices of $(\mathbf{A}, \mathbf{B})$ is invariant under any equivalence transformation and any reordering of the columns of $\mathbf{B}$.

⇒ ***Proof:*** Let us define

$$C_k = [\mathbf{B}\ \mathbf{AB}\ \cdots\ \mathbf{A}^{k-1}\mathbf{B}] \tag{6.21}$$

Then we have, following the proof of Theorem 6.2,

$$\rho(C_k) = \rho(\bar{C}_k)$$

for $k = 0, 1, 2, \ldots$. Thus the set of controllability indices is invariant under any equivalence transformation.

The rearrangement of the columns of $\mathbf{B}$ can be achieved by

$$\hat{\mathbf{B}} = \mathbf{BM}$$

where $\mathbf{M}$ is a $p \times p$ nonsingular permutation matrix. It is straightforward to verify

$$\hat{C}_k := [\hat{\mathbf{B}}\ \mathbf{A}\hat{\mathbf{B}}\ \cdots\ \mathbf{A}^{k-1}\hat{\mathbf{B}}] = C_k \text{diag}(\mathbf{M}, \mathbf{M}, \ldots, \mathbf{M})$$

Because diag $(\mathbf{M}, \mathbf{M}, \ldots, \mathbf{M})$ is nonsingular, we have $\rho(\hat{C}_k) = \rho(C_k)$ for $k = 0, 1, \ldots$. Thus the set of controllability indices is invariant under any reordering of the columns of $\mathbf{B}$. Q.E.D.

Because the set of the controllability indices is invariant under any equivalence transformation and any rearrangement of the inputs, it is an intrinsic property of the system that the state equation describes. The physical significance of the controllability index is not transparent here; but it becomes obvious in the discrete-time case. As we will discuss in later chapters, the controllability index can also be computed from transfer matrices and dictates the minimum degree required to achieve pole placement and model matching.

6.3 Observability

The concept of observability is dual to that of controllability. Roughly speaking, controllability studies the possibility of steering the state from the input; observability studies the possibility of estimating the state from the output. These two concepts are defined under the assumption that the state equation or, equivalently, all $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are known. Thus the problem of observability is different from the problem of realization or identification, which is to determine or estimate $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ from the information collected at the input and output terminals.

Consider the n-dimensional p-input q-output state equation

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx} + \mathbf{Du}\end{aligned} \tag{6.22}$$

where $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are, respectively, $n \times n$, $n \times p$, $q \times n$, and $q \times p$ constant matrices.

> ***Definition 6.01*** *The state equation (6.22) is said to be* observable *if for any unknown initial state* $\mathbf{x}(0)$, *there exists a finite* $t_1 > 0$ *such that the knowledge of the input* $\mathbf{u}$ *and the output* $\mathbf{y}$ *over* $[0, t_1]$ *suffices to determine uniquely the initial state* $\mathbf{x}(0)$. *Otherwise, the equation is said to be* unobservable.

EXAMPLE 6.6 Consider the network shown in Fig. 6.5. If the input is zero, no matter what the initial voltage across the capacitor is, the output is identically zero because of the symmetry of the four resistors. We know the input and output (both are identically zero), but we cannot determine uniquely the initial state. Thus the network or, more precisely, the state equation that describes the network is not observable.

EXAMPLE 6.7 Consider the network shown in Fig. 6.6(a). The network has two state variables: the current x_1 through the inductor and the voltage x_2 across the capacitor. The input u is a

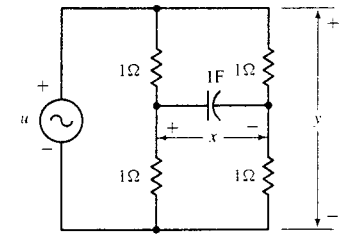


Figure 6.5 Unobservable network.

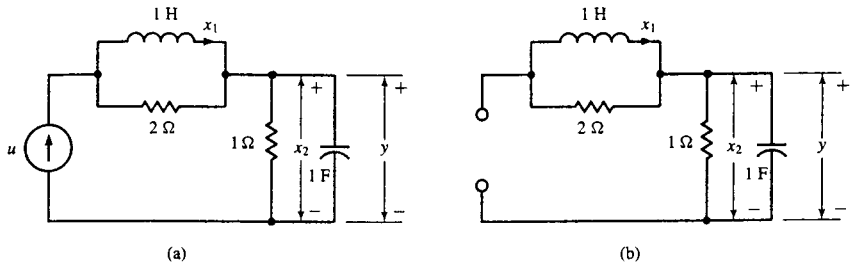


Figure 6.6 Unobservable network.

current source. If $u = 0$, the network reduces to the one shown in Fig. 6.6(b). If $x_1(0) = a \neq 0$ and $x_2 = 0$, then the output is identically zero. Any $\mathbf{x}(0) = [a \ \ 0]'$ and $u(t) \equiv 0$ yield the same output $y(t) \equiv 0$. Thus there is no way to determine the initial state $[a \ \ 0]'$ uniquely and the equation that describes the network is not observable.

The response of (6.22) excited by the initial state $\mathbf{x}(0)$ and the input $\mathbf{u}(t)$ was derived in (4.7) as

$$\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}(0) + \mathbf{C}\int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau + \mathbf{D}\mathbf{u}(t) \tag{6.23}$$

In the study of observability, the output $\mathbf{y}$ and the input $\mathbf{u}$ are assumed to be known; the inital state $\mathbf{x}(0)$ is the only unknown. Thus we can write (6.23) as

$$\mathbf{C}e^{\mathbf{A}t}\mathbf{x}(0) = \bar{\mathbf{y}}(t) \tag{6.24}$$

where

$$\bar{\mathbf{y}}(t) := \mathbf{y}(t) - \mathbf{C}\int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau - \mathbf{D}\mathbf{u}(t)$$

is a known function. Thus the observability problem reduces to solving $\mathbf{x}(0)$ from (6.24). If $\mathbf{u} \equiv \mathbf{0}$, then $\bar{\mathbf{y}}(t)$ reduces to the zero-input response $\mathbf{C}e^{\mathbf{A}t}\mathbf{x}(0)$. Thus Definition 6.O1 can be modified as follows: Equation (6.22) is observable if and only if the initial state $\mathbf{x}(0)$ can be determined uniquely from its zero-input response over a finite time interval.

Next we discuss how to solve $\mathbf{x}(0)$ from (6.24). For a fixed t, $\mathbf{C}e^{\mathbf{A}t}$ is a $q \times n$ constant matrix, and $\bar{\mathbf{y}}(t)$ is a $q \times 1$ constant vector. Thus (6.24) is a set of linear algebraic equations with n unknowns. Because of the way it is developed, for every fixed t, $\bar{\mathbf{y}}(t)$ is in the range space of $\mathbf{C}e^{\mathbf{A}t}$ and solutions always exist in (6.24). The only question is whether the solution is unique. If $q < n$, as is the case in general, the $q \times n$ matrix $\mathbf{C}e^{\mathbf{A}t}$ has rank at most q and, consequently, has nullity $n - q$ or larger. Thus solutions are not unique (Theorem 3.2). In conclusion, we cannot find a unique $\mathbf{x}(0)$ from (6.24) at an isolated t. In order to determine $\mathbf{x}(0)$ uniquely from (6.24), we must use the knowledge of $\mathbf{u}(t)$ and $\mathbf{y}(t)$ over a nonzero time interval as stated in the next theorem.

▶ **Theorem 6.4**

The state equation (6.22) is observable if and only if the $n \times n$ matrix

$$\mathbf{W}_o(t) = \int_0^t e^{\mathbf{A}'\tau}\mathbf{C}'\mathbf{C}e^{\mathbf{A}\tau}\,d\tau \tag{6.25}$$

is nonsingular for any $t > 0$.

■ ***Proof:*** We premultiply (6.24) by $e^{\mathbf{A}'t}\mathbf{C}'$ and then integrate it over $[0, \ t_1]$ to yield

$$\left(\int_0^{t_1} e^{\mathbf{A}'t}\mathbf{C}'\mathbf{C}e^{\mathbf{A}t}\,dt\right)\mathbf{x}(0) = \int_0^{t_1} e^{\mathbf{A}'t}\mathbf{C}'\bar{\mathbf{y}}(t)\,dt$$

If $\mathbf{W}_o(t_1)$ is nonsingular, then

$$\mathbf{x}(0) = \mathbf{W}_o^{-1}(t_1)\int_0^{t_1} e^{\mathbf{A}'t}\mathbf{C}'\bar{\mathbf{y}}(t)\,dt \tag{6.26}$$

This yields a unique $\mathbf{x}(0)$. This shows that if $\mathbf{W}_o(t)$, for any $t > 0$, is nonsingular, then (6.22) is observable. Next we show that if $\mathbf{W}_o(t_1)$ is singular or, equivalently, positive semidefinite for all t_1, then (6.22) is not observable. If $\mathbf{W}_o(t_1)$ is positive semidefinite, there exists an $n \times 1$ nonzero constant vector $\mathbf{v}$ such that

$$\begin{aligned}\mathbf{v}'\mathbf{W}_o(t_1)\mathbf{v} &= \int_0^{t_1} \mathbf{v}'e^{\mathbf{A}'\tau}\mathbf{C}'\mathbf{C}e^{\mathbf{A}\tau}\mathbf{v}\,d\tau \\ &= \int_0^{t_1} \|\mathbf{C}e^{\mathbf{A}\tau}\mathbf{v}\|^2 d\tau = 0\end{aligned}$$

which implies

$$\mathbf{C}e^{\mathbf{A}\tau}\mathbf{v} \equiv \mathbf{0} \tag{6.27}$$

for all t in $[0, \ t_1]$. If $\mathbf{u} \equiv \mathbf{0}$, then $\mathbf{x}_1(0) = \mathbf{v} \neq \mathbf{0}$ and $\mathbf{x}_2(0) = \mathbf{0}$ both yield the same

$$\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{x}_i(0) \equiv \mathbf{0}$$

Two different initial states yield the same zero-input response; therefore we cannot uniquely determine $\mathbf{x}(0)$. Thus (6.22) is not observable. This completes the proof of Theorem 6.4. Q.E.D.

We see from this theorem that observability depends only on $\mathbf{A}$ and $\mathbf{C}$. This can also be deduced from Definition 6.O1 by choosing $\mathbf{u}(t) \equiv \mathbf{0}$. Thus observability is a property of the pair $(\mathbf{A}, \mathbf{C})$ and is independent of $\mathbf{B}$ and $\mathbf{D}$. As in the controllability part, if $\mathbf{W}_o(t)$ is nonsingular for some t, then it is nonsingular for every t and the initial state can be computed from (6.26) by using any nonzero time interval.

▶ **Theorem 6.5 (Theorem of duality)**

The pair $(\mathbf{A}, \mathbf{B})$ is controllable if and only if the pair $(\mathbf{A}', \mathbf{B}')$ is observable.

⇒ *Proof:* The pair $(\mathbf{A}, \mathbf{B})$ is controllable if and only if

$$\mathbf{W}_c(t) = \int_0^t e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'\tau}d\tau$$

is nonsingular for any t. The pair $(\mathbf{A}', \mathbf{B}')$ is observable if and only if, by replacing $\mathbf{A}$ by $\mathbf{A}'$ and $\mathbf{C}$ by $\mathbf{B}'$ in (6.25),

$$\mathbf{W}_o(t) = \int_0^t e^{\mathbf{A}\tau}\mathbf{B}\mathbf{B}'e^{\mathbf{A}'\tau}\,d\tau$$

is nonsingular for any t. The two conditions are identical and the theorem follows. Q.E.D.

We list in the following the observability counterpart of Theorem 6.1. It can be proved either directly or by applying the theorem of duality.

Theorem 6.01

The following statements are equivalent.

1. The n-dimensional pair $(\mathbf{A}, \mathbf{C})$ is observable.
2. The $n \times n$ matrix

$$\mathbf{W}_o(t) = \int_0^t e^{\mathbf{A}'\tau}\mathbf{C}'\mathbf{C}e^{\mathbf{A}\tau}\,d\tau \tag{6.28}$$

is nonsingular for any $t > 0$.

3. The $nq \times n$ *observability matrix*

$$O = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix} \tag{6.29}$$

has rank n (full column rank). This matrix can be generated by calling `obsv` in MATLAB.

4. The $(n + q) \times n$ matrix

$$\begin{bmatrix} \mathbf{A} - \lambda\mathbf{I} \\ \mathbf{C} \end{bmatrix}$$

has full column rank at every eigenvalue, λ, of $\mathbf{A}$.

5. If, in addition, all eigenvalues of $\mathbf{A}$ have negative real parts, then the unique solution of

$$\mathbf{A}'\mathbf{W}_o + \mathbf{W}_o\mathbf{A} = -\mathbf{C}'\mathbf{C} \tag{6.30}$$

is positive definite. The solution is called the *observability Gramian* and can be expressed as

$$\mathbf{W}_o = \int_0^\infty e^{\mathbf{A}'\tau}\mathbf{C}'\mathbf{C}e^{\mathbf{A}\tau}d\tau \tag{6.31}$$

6.3.1 Observability Indices

Let $\mathbf{A}$ and $\mathbf{C}$ be $n \times n$ and $q \times n$ constant matrices. We assume that $\mathbf{C}$ has rank q (full row rank). If $\mathbf{C}$ does not have full row rank, then the output at some output terminal can be expressed as a linear combination of other outputs. Thus the output does not offer any new information regarding the system and the terminal can be eliminated. By deleting the corresponding row, the reduced $\mathbf{C}$ will then have full row rank.

If $(\mathbf{A}, \mathbf{C})$ is observable, its observability matrix O has rank n and, consequently, n linearly independent rows. Let $\mathbf{c}_i$ be the ith row of $\mathbf{C}$. Let us search linearly independent rows of O in order from top to bottom. Dual to the controllability part, if a row associated with $\mathbf{c}_m$ becomes linearly dependent on its upper rows, then all rows associated with $\mathbf{c}_m$ thereafter will also be dependent. Let ν_m be the number of the linearly independent rows associated with $\mathbf{c}_m$. It is clear that if O has rank n, then

$$\nu_1 + \nu_2 + \cdots + \nu_q = n \tag{6.32}$$

The set $\{\nu_1, \nu_2, \ldots, \nu_q\}$ is called the *observability indices* and

$$\nu = \max(\nu_1, \nu_2, \ldots, \nu_q) \tag{6.33}$$

is called the *observability index* of $(\mathbf{A}, \mathbf{C})$. If $(\mathbf{A}, \mathbf{C})$ is observable, it is the least integer such that

$$\rho(O_\nu) := \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \\ \vdots \\ \mathbf{CA}^{\nu-1} \end{bmatrix} = n$$

Dual to the controllability part, we have

$$n/q \le \nu \le \min(\bar{n}, n - q + 1) \tag{6.34}$$

where $\rho(\mathbf{C}) = q$ and $\bar{n}$ is the degree of the minimal polynomial of A.

Corollary 6.01

The n-dimensional pair $(\mathbf{A}, \mathbf{C})$ is observable if and only if the matrix

$$O_{n-q+1} = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-q} \end{bmatrix} \tag{6.35}$$

where $\rho(\mathbf{C}) = q$, has rank n or the $n \times n$ matrix $O'_{n-q+1}O_{n-q+1}$ is nonsingular.

Theorem 6.02

The observability property is invariant under any equivalence transformation.

Theorem 6.03

The set of the observability indices of $(\mathbf{A}, \mathbf{C})$ is invariant under any equivalence transformation and any reordering of the rows of $\mathbf{C}$.

Before concluding this section, we discuss a different way of solving (6.24). Differentiating (6.24) repeatedly and setting $t = 0$, we can obtain

$$\begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{\nu-1} \end{bmatrix} \mathbf{x}(0) = \begin{bmatrix} \bar{\mathbf{y}}(0) \\ \dot{\bar{\mathbf{y}}}(0) \\ \vdots \\ \bar{\mathbf{y}}^{(\nu-1)}(0) \end{bmatrix}$$

or

$$O_\nu \mathbf{x}(0) = \bar{\mathbf{y}}(0) \tag{6.36}$$

where $\bar{\mathbf{y}}^{(i)}(t)$ is the ith derivative of $\bar{\mathbf{y}}(t)$, and $\bar{\mathbf{y}}(0) := [\bar{\mathbf{y}}'(0)\ \dot{\bar{\mathbf{y}}}'(0)\ \cdots\ (\bar{\mathbf{y}}^{(\nu-1)})']'$. Equation (6.36) is a set of linear algebraic equations. Because of the way it is developed, $\bar{\mathbf{y}}(0)$ must lie in the range space of O_ν. Thus a solution $\mathbf{x}(0)$ exists in (6.36). If $(\mathbf{A}, \mathbf{C})$ is observable, then O_ν has full column rank and, following Theorem 3.2, the solution is unique. Premultiplying (6.36) by O_ν' and then using Theorem 3.8, we can obtain the solution as

$$\mathbf{x}(0) = \left[O_\nu' O_\nu\right]^{-1} O_\nu' \bar{\mathbf{y}}(0) \tag{6.37}$$

We mention that in order to obtain $\dot{\bar{\mathbf{y}}}(0), \ddot{\bar{\mathbf{y}}}(0), \ldots$, we need knowledge of $\bar{\mathbf{y}}(t)$ in the neighborhood of $t = 0$. This is consistent with the earlier assertion that we need knowledge of $\bar{\mathbf{y}}(t)$ over a nonzero time interval in order to determine $\mathbf{x}(0)$ uniquely from (6.24). In conclusion, the initial state can be computed using (6.26) or (6.37).

The output $y(t)$ measured in practice is often corrupted by high-frequency noise. Because

- differentiation will amplify high-frequency noise and
- integration will suppress or smooth high-frequency noise,

the result obtained from (6.36) or (6.37) may differ greatly from the actual initial state. Thus (6.26) is preferable to (6.36) in computing initial states.

The physical significance of the observability index can be seen from (6.36). It is the smallest integer in order to determine $\mathbf{x}(0)$ uniquely from (6.36) or (6.37). It also dictates the minimum degree required to achieve pole placement and model matching, as we will discuss in Chapter 9.

6.4 Canonical Decomposition

This section discusses canonical decomposition of state equations. This fundamental result will be used to establish the relationship between the state-space description and the transfer-matrix description. Consider

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{Ax} + \mathbf{Bu} \\ \mathbf{y} &= \mathbf{Cx} + \mathbf{Du} \end{aligned} \tag{6.38}$$

Let $\bar{\mathbf{x}} = \mathbf{Px}$, where $\mathbf{P}$ is a nonsingular matrix. Then the state equation

$$\begin{aligned} \dot{\bar{\mathbf{x}}} &= \bar{\mathbf{A}}\bar{\mathbf{x}} + \bar{\mathbf{B}}\mathbf{u} \\ \mathbf{y} &= \bar{\mathbf{C}}\bar{\mathbf{x}} + \bar{\mathbf{D}}\mathbf{u} \end{aligned} \tag{6.39}$$

with $\bar{\mathbf{A}} = \mathbf{PAP}^{-1}$, $\bar{\mathbf{B}} = \mathbf{PB}$, $\bar{\mathbf{C}} = \mathbf{CP}^{-1}$, and $\bar{\mathbf{D}} = \mathbf{D}$ is equivalent to (6.38). All properties of (6.38), including stability, controllability, and observability, are preserved in (6.39). We also have

$$\bar{C} = PC \qquad \bar{O} = OP^{-1}$$

Theorem 6.6

Consider the n-dimensional state equation in (6.38) with

$$\rho(C) = \rho([\mathbf{B}\ \mathbf{AB}\ \cdots\ \mathbf{A}^{n-1}\mathbf{B}]) = n_1 < n$$

We form the $n \times n$ matrix

$$\mathbf{P}^{-1} := [\mathbf{q}_1\ \cdots\ \mathbf{q}_{n_1} \cdots\ \mathbf{q}_n]$$

where the first n_1 columns are any n_1 linearly independent columns of C, and the remaining columns can arbitrarily be chosen as long as $\mathbf{P}$ is nonsingular. Then the equivalence transformation $\bar{\mathbf{x}} = \mathbf{Px}$ or $\mathbf{x} = \mathbf{P}^{-1}\bar{\mathbf{x}}$ will transform (6.38) into

$$\begin{aligned} \begin{bmatrix} \dot{\bar{\mathbf{x}}}_c \\ \dot{\bar{\mathbf{x}}}_{\bar{c}} \end{bmatrix} &= \begin{bmatrix} \bar{\mathbf{A}}_c & \bar{\mathbf{A}}_{12} \\ \mathbf{0} & \bar{\mathbf{A}}_{\bar{c}} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_c \\ \bar{\mathbf{x}}_{\bar{c}} \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{B}}_c \\ \mathbf{0} \end{bmatrix} \mathbf{u} \\ \mathbf{y} &= [\bar{\mathbf{C}}_c\ \bar{\mathbf{C}}_{\bar{c}}] \begin{bmatrix} \bar{\mathbf{x}}_c \\ \bar{\mathbf{x}}_{\bar{c}} \end{bmatrix} + \mathbf{Du} \end{aligned} \tag{6.40}$$

where $\bar{\mathbf{A}}_c$ is $n_1 \times n_1$ and $\bar{\mathbf{A}}_{\bar{c}}$ is $(n - n_1) \times (n - n_1)$, and the n_1-dimensional subequation of (6.40),

$$\begin{aligned} \dot{\bar{\mathbf{x}}}_c &= \bar{\mathbf{A}}_c\bar{\mathbf{x}}_c + \bar{\mathbf{B}}_c\mathbf{u} \\ \bar{\mathbf{y}} &= \bar{\mathbf{C}}_c\bar{\mathbf{x}}_c + \mathbf{Du} \end{aligned} \tag{6.41}$$

is controllable and has the same transfer matrix as (6.38).

Proof: As discussed in Section 4.3, the transformation $\mathbf{x} = \mathbf{P}^{-1}\bar{\mathbf{x}}$ changes the basis of the state space from the orthonormal basis in (3.8) to the columns of $\mathbf{Q} := \mathbf{P}^{-1}$ or $\{\mathbf{q}_1, \ldots, \mathbf{q}_{n_1}, \ldots, \mathbf{q}_n\}$. The ith column of $\bar{\mathbf{A}}$ is the representation of $\mathbf{Aq}_i$ with respect to $\{\mathbf{q}_1, \ldots, \mathbf{q}_{n_1}, \ldots, \mathbf{q}_n\}$. Now the vector $\mathbf{Aq}_i$, for $i = 1, 2, \ldots, n_1$, are linearly dependent on the set $\{\mathbf{q}_1, \ldots, \mathbf{q}_{n_1}\}$; they are linearly independent of $\{\mathbf{q}_{n_1+1}, \ldots, \mathbf{q}_n\}$. Thus the matrix $\bar{\mathbf{A}}$ has the form shown in (6.40). The columns of $\bar{\mathbf{B}}$ are the representation of the columns of $\mathbf{B}$ with respect to $\{\mathbf{q}_1, \ldots, \mathbf{q}_{n_1}, \ldots, \mathbf{q}_n\}$. Now the columns of $\mathbf{B}$ depend only on $\{\mathbf{q}_1, \ldots, \mathbf{q}_{n_1}\}$; thus $\bar{\mathbf{B}}$ has the form shown in (6.40). We mention that if the $n \times p$

matrix $\mathbf{B}$ has rank p and if its columns are chosen as the first p columns of $\mathbf{P}^{-1}$, then the upper part of $\bar{\mathbf{B}}$ is the unit matrix of order p.

Let $\bar{C}$ be the controllability matrix of (6.40). Then we have $\rho(C) = \rho(\bar{C}) = n_1$. It is straightforward to verify

$$\bar{C} = \begin{bmatrix} \bar{\mathbf{B}}_c & \bar{\mathbf{A}}_c\bar{\mathbf{B}}_c & \cdots & \bar{\mathbf{A}}_c^{n_1}\bar{\mathbf{B}}_c & \cdots & \bar{\mathbf{A}}_c^{n-1}\bar{\mathbf{B}}_c \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \bar{C}_c & \bar{\mathbf{A}}_c^{n_1}\bar{\mathbf{B}}_c & \cdots & \bar{\mathbf{A}}_c^{n-1}\bar{\mathbf{B}}_c \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$

where $\bar{C}_c$ is the controllability matrix of $(\bar{\mathbf{A}}_c, \bar{\mathbf{B}}_c)$. Because the columns of $\bar{\mathbf{A}}_c^k\bar{\mathbf{B}}_c$, for $k \geq n_1$, are linearly dependent on the columns of $\bar{C}_c$, the condition $\rho(C) = n_1$ implies $\rho(\bar{C}_c) = n_1$. Thus the n_1-dimensional state equation in (6.41) is controllable.

Next we show that (6.41) has the same transfer matrix as (6.38). Because (6.38) and (6.40) have the same transfer matrix, we need to show only that (6.40) and (6.41) have the same transfer matrix. By direct verification, we can show

$$\begin{bmatrix} s\mathbf{I} - \bar{\mathbf{A}}_c & -\bar{\mathbf{A}}_{12} \\ \mathbf{0} & s\mathbf{I} - \bar{\mathbf{A}}_{\bar{c}} \end{bmatrix}^{-1} = \begin{bmatrix} (s\mathbf{I} - \bar{\mathbf{A}}_c)^{-1} & \mathbf{M} \\ \mathbf{0} & (s\mathbf{I} - \bar{\mathbf{A}}_{\bar{c}})^{-1} \end{bmatrix} \tag{6.42}$$

where

$$\mathbf{M} = (s\mathbf{I} - \bar{\mathbf{A}}_c)^{-1}\bar{\mathbf{A}}_{12}(s\mathbf{I} - \bar{\mathbf{A}}_{\bar{c}})^{-1}$$

Thus the transfer matrix of (6.40) is

$$\begin{aligned} &\begin{bmatrix} \bar{\mathbf{C}}_c & \bar{\mathbf{C}}_{\bar{c}} \end{bmatrix} \begin{bmatrix} s\mathbf{I} - \bar{\mathbf{A}}_c & -\bar{\mathbf{A}}_{12} \\ \mathbf{0} & s\mathbf{I} - \bar{\mathbf{A}}_{\bar{c}} \end{bmatrix}^{-1} \begin{bmatrix} \bar{\mathbf{B}}_c \\ \mathbf{0} \end{bmatrix} + \mathbf{D} \\ &= \begin{bmatrix} \bar{\mathbf{C}}_c & \bar{\mathbf{C}}_{\bar{c}} \end{bmatrix} \begin{bmatrix} (s\mathbf{I} - \bar{\mathbf{A}}_c)^{-1} & \mathbf{M} \\ \mathbf{0} & (s\mathbf{I} - \bar{\mathbf{A}}_{\bar{c}})^{-1} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{B}}_c \\ \mathbf{0} \end{bmatrix} + \mathbf{D} \\ &= \bar{\mathbf{C}}_c(s\mathbf{I} - \bar{\mathbf{A}}_c)^{-1}\bar{\mathbf{B}}_c + \mathbf{D} \end{aligned}$$

which is the transfer matrix of (6.41). This completes the proof of Theorem 6.6. Q.E.D.

In the equivalence transformation $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$, the n-dimensional state space is divided into two subspaces. One is the n_1-dimensional subspace that consists of all vectors of the form $[\bar{\mathbf{x}}_c' \ \mathbf{0}']'$; the other is the $(n - n_1)$-dimensional subspace that consists of all vectors of the form $[\mathbf{0}' \ \bar{\mathbf{x}}_{\bar{c}}']'$. Because (6.41) is controllable, the input $\mathbf{u}$ can transfer $\bar{\mathbf{x}}_c$ from any state to any other state. However, the input $\mathbf{u}$ cannot control $\bar{\mathbf{x}}_{\bar{c}}$ because, as we can see from (6.40), $\mathbf{u}$ does not affect $\bar{\mathbf{x}}_{\bar{c}}$ directly, nor indirectly through the state $\bar{\mathbf{x}}_c$. By dropping the uncontrollable state vector, we obtain a controllable state equation of lesser dimension that is zero-state equivalent to the original equation.

EXAMPLE 6.8 Consider the three-dimensional state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}u \qquad y = [1 \ 1 \ 1]\mathbf{x} \tag{6.43}$$

The rank of $\mathbf{B}$ is 2; therefore we can use $C_2 = [\mathbf{B} \ \mathbf{AB}]$, instead of $C = [\mathbf{B} \ \mathbf{AB} \ \mathbf{A}^2\mathbf{B}]$, to check the controllability of (6.43) (Corollary 6.1). Because

$$\rho(C_2) = \rho([\mathbf{B} \ \mathbf{AB}]) = \rho\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} = 2 < 3$$

the state equation in (6.43) is not controllable. Let us choose

$$\mathbf{P}^{-1} = \mathbf{Q} := \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

The first two columns of $\mathbf{Q}$ are the first two linearly independent columns of C_2; the last column is chosen arbitrarily to make $\mathbf{Q}$ nonsingular. Let $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$. We compute

$$\bar{\mathbf{A}} = \mathbf{P}\mathbf{A}\mathbf{P}^{-1} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$= \left[\begin{array}{ccc:c} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 1 \end{array}\right]$$

$$\bar{\mathbf{B}} = \mathbf{P}\mathbf{B} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \cdots & \cdots \\ 0 & 0 \end{bmatrix}$$

$$\bar{\mathbf{C}} = \mathbf{C}\mathbf{P}^{-1} = [1 \ 1 \ 1]\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = [1 \ 2 \ \vdots \ 1]$$

Note that the 1×2 submatrix $\bar{\mathbf{A}}_{21}$ of $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}_{\bar{c}}$ are zero as expected. The 2×1 submatrix $\bar{\mathbf{A}}_{12}$ happens to be zero; it could be nonzero. The upper part of $\bar{\mathbf{B}}$ is a unit matrix because the columns of $\mathbf{B}$ are the first two columns of $\mathbf{Q}$. Thus (6.43) can be reduced to

$$\dot{\bar{\mathbf{x}}}_c = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}\bar{\mathbf{x}}_c + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\mathbf{u} \qquad y = [1 \ \ 2]\bar{\mathbf{x}}_c$$

This equation is controllable and has the same transfer matrix as (6.43).

The MATLAB function `ctrbf` transforms (6.38) into (6.40) except that the order of the columns in $\mathbf{P}^{-1}$ is reversed. Thus the resulting equation has the form

$$\begin{bmatrix} \bar{\mathbf{A}}_{\bar{c}} & \mathbf{0} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_c \end{bmatrix} \qquad \begin{bmatrix} \mathbf{0} \\ \bar{\mathbf{B}}_c \end{bmatrix}$$

Theorem 6.6 is established from the controllability matrix. In actual computation, it is unnecessary to form the controllability matrix. The result can be obtained by carrying out a sequence

of similarity transformations to transform [**B A**] into a Hessenberg form. See Reference [6, pp. 220–222]. This procedure is efficient and numerically stable and should be used in actual computation.

Dual to Theorem 6.6, we have the following theorem for unobservable state equations.

Theorem 6.O6

Consider the n-dimensional state equation in (6.38) with

$$\rho(O) = \rho \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix} = n_2 < n$$

We form the $n \times n$ matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_{n_2} \\ \vdots \\ \mathbf{p}_n \end{bmatrix}$$

where the first n_2 rows are any n_2 linearly independent rows of O, and the remaining rows can be chosen arbitrarily as long as $\mathbf{P}$ is nonsingular. Then the equivalence transformation $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$ will transform (6.38) into

$$\begin{bmatrix} \dot{\bar{\mathbf{x}}}_o \\ \dot{\bar{\mathbf{x}}}_{\bar{o}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{A}}_o & \mathbf{0} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{\bar{o}} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_o \\ \bar{\mathbf{x}}_{\bar{o}} \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{B}}_o \\ \bar{\mathbf{B}}_{\bar{o}} \end{bmatrix} \mathbf{u}$$

$$\mathbf{y} = [\bar{\mathbf{C}}_o \ \mathbf{0}] \begin{bmatrix} \bar{\mathbf{x}}_o \\ \bar{\mathbf{x}}_{\bar{o}} \end{bmatrix} + \mathbf{D}\mathbf{u} \tag{6.44}$$

where $\bar{\mathbf{A}}_o$ is $n_2 \times n_2$ and $\bar{\mathbf{A}}_{\bar{o}}$ is $(n - n_2) \times (n - n_2)$, and the n_2-dimensional subequation of (6.44),

$$\dot{\bar{\mathbf{x}}}_o = \bar{\mathbf{A}}_o \bar{\mathbf{x}}_o + \bar{\mathbf{B}}_o \mathbf{u}$$
$$\bar{\mathbf{y}} = \bar{\mathbf{C}}_o \bar{\mathbf{x}}_o + \mathbf{D}\mathbf{u}$$

is observable and has the same transfer matrix as (6.38).

In the equivalence transformation $\bar{\mathbf{x}} = \mathbf{P}\mathbf{x}$, the n-dimensional state space is divided into two subspaces. One is the n_2-dimensional subspace that consists of all vectors of the form $[\bar{\mathbf{x}}_o' \ \mathbf{0}']'$; the other is the $(n - n_2)$-dimensional subspace consisting of all vectors of the form $[\mathbf{0}' \ \bar{\mathbf{x}}_{\bar{o}}']'$. The state $\bar{\mathbf{x}}_o$ can be detected from the output. However, $\bar{\mathbf{x}}_{\bar{o}}$ cannot be detected from the output because, as we can see from (6.44), it is not connected to the output either directly, or indirectly through the state $\bar{\mathbf{x}}_o$. By dropping the unobservable state vector, we obtain an observable state equation of lesser dimension that is zero-state equivalent to the original equation. The MATLAB function `obsvf` is the counterpart of `ctrbf`. Combining Theorems 6.6 and 6.06, we have the following *Kalman decomposition theorem.*

Theorem 6.7

Every state-space equation can be transformed, by an equivalence transformation, into the following canonical form

$$\begin{bmatrix} \dot{\bar{\mathbf{x}}}_{co} \\ \dot{\bar{\mathbf{x}}}_{c\bar{o}} \\ \dot{\bar{\mathbf{x}}}_{\bar{c}o} \\ \dot{\bar{\mathbf{x}}}_{\bar{c}\bar{o}} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{A}}_{co} & \mathbf{0} & \bar{\mathbf{A}}_{13} & \mathbf{0} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{c\bar{o}} & \bar{\mathbf{A}}_{23} & \bar{\mathbf{A}}_{24} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{A}}_{\bar{c}o} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{A}}_{43} & \bar{\mathbf{A}}_{\bar{c}\bar{o}} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_{co} \\ \bar{\mathbf{x}}_{c\bar{o}} \\ \bar{\mathbf{x}}_{\bar{c}o} \\ \bar{\mathbf{x}}_{\bar{c}\bar{o}} \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{B}}_{co} \\ \bar{\mathbf{B}}_{c\bar{o}} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \mathbf{u} \tag{6.45}$$

$$\mathbf{y} = [\bar{\mathbf{C}}_{co} \ \mathbf{0} \ \bar{\mathbf{C}}_{\bar{c}o} \ \mathbf{0}]\bar{\mathbf{x}} + \mathbf{D}\mathbf{u}$$

where the vector $\bar{\mathbf{x}}_{co}$ is controllable and observable, $\bar{\mathbf{x}}_{c\bar{o}}$ is controllable but not observable, $\bar{\mathbf{x}}_{\bar{c}o}$ is observable but not controllable, and $\bar{\mathbf{x}}_{\bar{c}\bar{o}}$ is neither controllable nor observable. Furthermore, the state equation is zero-state equivalent to the controllable and observable state equation

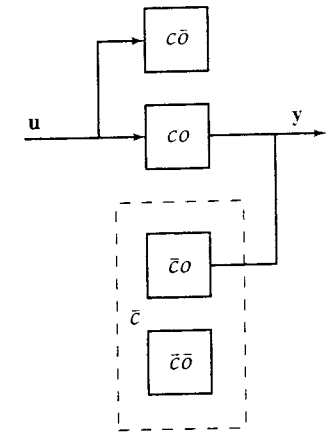
$$\dot{\bar{\mathbf{x}}}_{co} = \bar{\mathbf{A}}_{co}\bar{\mathbf{x}}_{co} + \bar{\mathbf{B}}_{co}\mathbf{u}$$
$$\mathbf{y} = \bar{\mathbf{C}}_{co}\bar{\mathbf{x}}_{co} + \mathbf{D}\mathbf{u} \tag{6.46}$$

and has the transfer matrix

$$\hat{\mathbf{G}}(s) = \bar{\mathbf{C}}_{co}(s\mathbf{I} - \bar{\mathbf{A}}_{co})^{-1}\bar{\mathbf{B}}_{co} + \mathbf{D}$$

This theorem can be illustrated symbolically as shown in Fig. 6.7. The equation is first decomposed, using Theorem 6.6, into controllable and uncontrollable subequations. We then decompose each subequation, using Theorem 6.O6, into observable and unobservable parts. From the figure, we see that only the controllable and observable part is connected to both the input and output terminals. Thus the transfer matrix describes only this part of the system. This is the reason that the transfer-function description and the state-space description are not necessarily equivalent. For example, if any A-matrix other than $\bar{\mathbf{A}}_{co}$ has an eigenvalue with a

Figure 6.7 Kalman decomposition.



positive real part, then some state variable may grow without bound and the system may burn out. This phenomenon, however, cannot be detected from the transfer matrix.

The MATLAB function `minreal`, an acronym for *minimal realization*, can reduce any state equation to (6.46). The reason for calling it minimal realization will be given in the next chapter.

EXAMPLE 6.9 Consider the network shown in Fig. 6.8(a). Because the input is a current source, responses due to the initial conditions in C_1 and L_1 will not appear at the output. Thus the state variables associated with C_1 and L_1 are not observable; whether or not they are controllable is immaterial in subsequent discussion. Similarly, the state variable associated with L_2 is not controllable. Because of the symmetry of the four 1-Ω resistors, the state variable associated with C_2 is neither controllable nor observable. By dropping the state variables that are either uncontrollable or unobservable, the network in Fig. 6.8(a) can be reduced to the one in Fig. 6.8(b). The current in each branch is $u/2$; thus the output y equals $2 \cdot (u/2)$ or $y = u$. Thus the transfer function of the network in Fig. 6.8(a) is $\hat{g}(s) = 1$.

If we assign state variables as shown, then the network can be described by

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & -0.5 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -0.5 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0.5 \\ 0 \\ 0 \\ 0 \end{bmatrix} u$$

$$y = [0 \ 0 \ 0 \ 1]\mathbf{x} + u$$

Because the equation is already of the form shown in (6.40), it can be reduced to the following controllable state equation

$$\dot{\mathbf{x}}_c = \begin{bmatrix} 0 & -0.5 \\ 1 & 0 \end{bmatrix} \mathbf{x}_c + \begin{bmatrix} 0.5 \\ 0 \end{bmatrix} u$$

$$y = [0 \ 0]\mathbf{x}_c + u$$

The output is independent of $\mathbf{x}_c$; thus the equation can be further reduced to $y = u$. This is what we will obtain by using the MATLAB function `minreal`.

6.5 Conditions in Jordan-Form Equations

Controllability and observability are invariant under any equivalence transformation. If a state equation is transformed into Jordan form, then the controllability and observability conditions become very simple and can often be checked by inspection. Consider the state equation

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{J}\mathbf{x} + \mathbf{B}\mathbf{x} \\ \mathbf{y} &= \mathbf{C}\mathbf{x} \end{aligned} \tag{6.47}$$

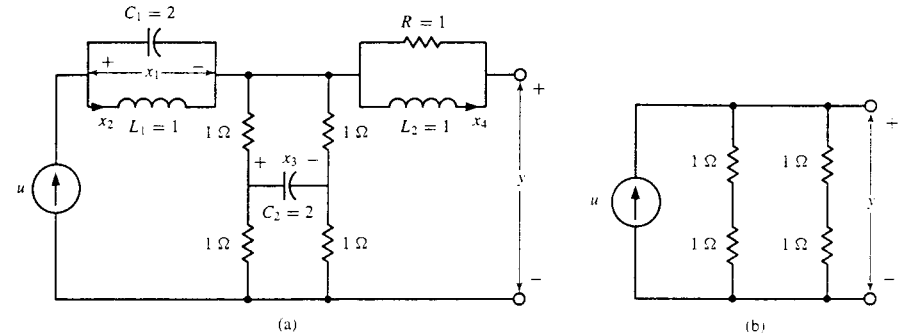


Figure 6.8 Networks.

where $\mathbf{J}$ is in Jordan form. To simplify discussion, we assume that $\mathbf{J}$ has only two distinct eigenvalues λ_1 and λ_2 and can be written as

$$\mathbf{J} = \text{diag}\,(\mathbf{J}_1, \mathbf{J}_2)$$

where $\mathbf{J}_1$ consists of all Jordan blocks associated with λ_1 and $\mathbf{J}_2$ consists of all Jordan blocks associated with λ_2. Again to simplify discussion, we assume that $\mathbf{J}_1$ has three Jordan blocks and $\mathbf{J}_2$ has two Jordan blocks or

$$\mathbf{J}_1 = \text{diag}\,(\mathbf{J}_{11}, \mathbf{J}_{12}, \mathbf{J}_{13}) \qquad \mathbf{J}_2 = \text{diag}\,(\mathbf{J}_{21}, \mathbf{J}_{22})$$

The row of $\mathbf{B}$ corresponding to the *last row* of $\mathbf{J}_{ij}$ is denoted by $\mathbf{b}_{lij}$. The column of $\mathbf{C}$ corresponding to the *first column* of $\mathbf{J}_{ij}$ is denoted by $\mathbf{c}_{fij}$.

Theorem 6.8

1. The state equation in (6.47) is controllable if and only if the three row vectors $\{\mathbf{b}_{l11}, \mathbf{b}_{l12}, \mathbf{b}_{l13}\}$ are linearly independent and the two row vectors $\{\mathbf{b}_{l21}, \mathbf{b}_{l22}\}$ are linearly independent.

2. The state equation in (6.47) is observable if and only if the three column vectors $\{\mathbf{c}_{f11}, \mathbf{c}_{f12}, \mathbf{c}_{f13}\}$ are linearly independent and the two column vectors $\{\mathbf{c}_{f21}, \mathbf{c}_{f22}\}$ are linearly independent.

We discuss first the implications of this theorem. If a state equation is in Jordan form, then the controllability of the state variables associated with one eigenvalue can be checked independently from those associated with different eigenvalues. The controllability of the state variables associated with the same eigenvalue depends only on the rows of $\mathbf{B}$ corresponding to the last row of all Jordan blocks associated with the eigenvalue. All other rows of $\mathbf{B}$ play no role in determining the controllability. Similar remarks apply to the observability part except that the columns of $\mathbf{C}$ corresponding to the first column of all Jordan blocks determine the observability. We use an example to illustrate the use of Theorem 6.8.

EXAMPLE 6.10 Consider the Jordan-form state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \mathbf{u} \tag{6.48}$$

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

The matrix $\mathbf{J}$ has two distinct eigenvalues λ_1 and λ_2. There are three Jordan blocks, with order 2, 1, and 1, associated with λ_1. The rows of $\mathbf{B}$ corresponding to the last row of the three Jordan blocks are [1 0 0], [0 1 0], and [1 1 1]. The three rows are linearly independent. There is only one Jordan block, with order 3, associated with λ_2. The row of $\mathbf{B}$ corresponding to the last row of the Jordan block is [1 1 1], which is nonzero and is therefore linearly independent. Thus we conclude that the state equation in (6.48) is controllable.

The conditions for (6.48) to be observable are that the three columns [1 1 1]′, [2 1 2]′, and [0 2 3]′ are linearly independent (they are) and the one column [0 0 0]′ is linearly independent (it is not). Therefore the state equation is not observable.

Before proving Theorm 6.8, we draw a block diagram to show how the conditions in the theorem arise. The inverse of $(s\mathbf{I} - \mathbf{J})$ is of the form shown in (3.49), whose entries consist of only $1/(s - \lambda_i)^k$. Using (3.49), we can draw a block diagram for (6.48) as shown in Fig. 6.9. Each chain of blocks corresponds to one Jordan block in the equation. Because (6.48) has four Jordan blocks, the figure has four chains. The output of each block can be assigned as a state variable as shown in Fig. 6.10. Let us consider the last chain in Fig. 6.9. If $\mathbf{b}_{f21} = \mathbf{0}$, the state variable x_{f21} is not connected to the input and is not controllable no matter what values $\mathbf{b}_{221}$ and $\mathbf{b}_{121}$ assume. On the other hand, if $\mathbf{b}_{f21}$ is nonzero, then all state variables in the chain are controllable. If there are two or more chains associated with the same eigenvalue, then we require the linear independence of the first gain vectors of those chains. The chains associated with different eigenvalues can be checked separately. All discussion applies to the observability part except that the column vector $\mathbf{c}_{fij}$ plays the role of the row vector $\mathbf{b}_{fij}$.

Proof of Theorem 6.8 We prove the theorem by using the condition that the matrix $[\mathbf{A} - s\mathbf{I}\ \ \mathbf{B}]$ or $[s\mathbf{I} - \mathbf{A}\ \ \mathbf{B}]$ has full row rank at every eigenvalue of $\mathbf{A}$. In order not to be overwhelmed by notation, we assume $[s\mathbf{I} - \mathbf{J}\ \ \mathbf{B}]$ to be of the form

$$\begin{bmatrix} s-\lambda_1 & -1 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{111} \\ 0 & s-\lambda_1 & -1 & 0 & 0 & 0 & 0 & \mathbf{b}_{211} \\ 0 & 0 & s-\lambda_1 & 0 & 0 & 0 & 0 & \mathbf{b}_{f11} \\ 0 & 0 & 0 & s-\lambda_1 & -1 & 0 & 0 & \mathbf{b}_{112} \\ 0 & 0 & 0 & 0 & s-\lambda_1 & 0 & 0 & \mathbf{b}_{f12} \\ 0 & 0 & 0 & 0 & 0 & s-\lambda_2 & -1 & \mathbf{b}_{121} \\ 0 & 0 & 0 & 0 & 0 & 0 & s-\lambda_2 & \mathbf{b}_{f21} \end{bmatrix} \tag{6.49}$$

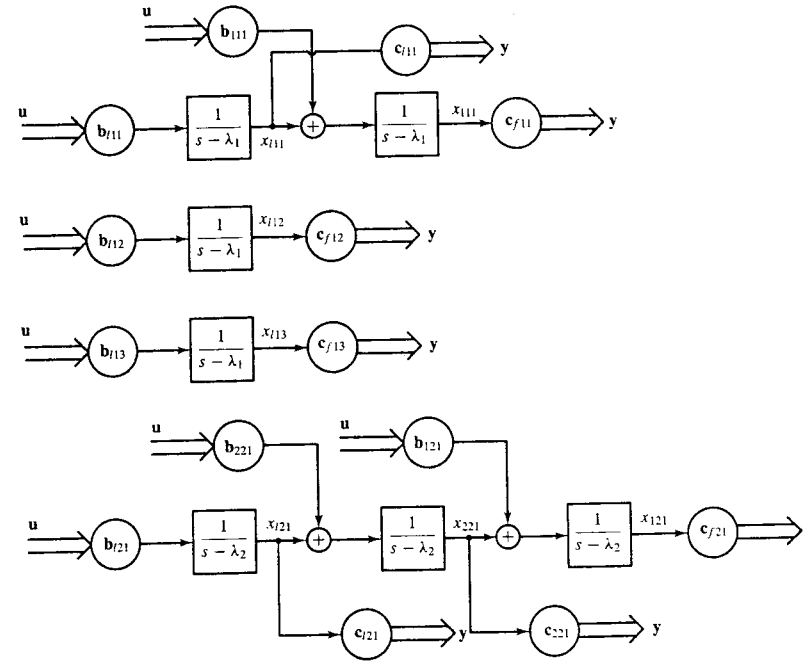


Figure 6.9 Block diagram of (6.48).

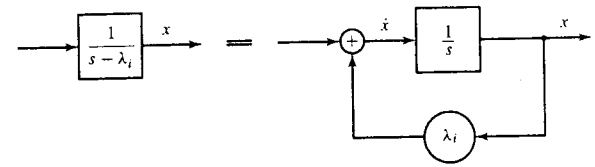


Figure 6.10 Internal structure of $1/(s - \lambda_i)$.

The Jordan-form matrix $\mathbf{J}$ has two distinct eigenvalues λ_1 and λ_2. There are two Jordan blocks associated with λ_1 and one associated with λ_2. If $s = \lambda_1$, (6.49) becomes

$$\begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{111} \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & \mathbf{b}_{211} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{f11} \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & \mathbf{b}_{112} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{f12} \\ 0 & 0 & 0 & 0 & 0 & \lambda_1-\lambda_2 & -1 & \mathbf{b}_{121} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1-\lambda_2 & \mathbf{b}_{f21} \end{bmatrix} \tag{6.50}$$

The rank of the matrix will not change by elementary column operations. We add the product of the second column of (6.50) by $\mathbf{b}_{111}$ to the last block column. Repeating the process for the third and fifth columns, we can obtain

$$\begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{111} \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{112} \\ 0 & 0 & 0 & 0 & 0 & \lambda_1 - \lambda_2 & -1 & \mathbf{b}_{121} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1 - \lambda_2 & \mathbf{b}_{121} \end{bmatrix}$$

Because λ_1 and λ_2 are distinct, $\lambda_1 - \lambda_2$ is nonzero. We add the product of the seventh column and $-\mathbf{b}_{121}/(\lambda_1 - \lambda_2)$ to the last column and then use the sixth column to eliminate its right-hand-side entries to yield

$$\begin{bmatrix} 0 & -1 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{111} \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{b}_{112} \\ 0 & 0 & 0 & 0 & 0 & \lambda_1 - \lambda_2 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1 - \lambda_2 & \mathbf{0} \end{bmatrix} \tag{6.51}$$

It is clear that the matrix in (6.51) has full row rank if and only if $\mathbf{b}_{111}$ and $\mathbf{b}_{112}$ are linearly independent. Proceeding similarly for each eigenvalue, we can establish Theorem 6.8. Q.E.D.

Consider an n-dimensional Jordan-form state equation with p inputs and q outputs. If there are m, with $m > p$, Jordan blocks associated with the same eigenvalue, then m number of $1 \times p$ row vectors can never be linearly independent and the state equation can never be controllable. Thus a necessary condition for the state equation to be controllable is $m \leq p$. Similarly, a necessary condition for the state equation to be observable is $m \leq q$. For the single-input or single-output case, we then have the following corollaries.

Corollary 6.8

A single-input Jordan-form state equation is controllable if and only if there is only one Jordan block associated with each distinct eigenvalue and every entry of $\mathbf{B}$ corresponding to the last row of each Jordan block is different from zero.

Corollary 6.O8

A single-output Jordan-form state equation is observable if and only if there is only one Jordan block associated with each distinct eigenvalue and every entry of $\mathbf{C}$ corresponding to the first column of each Jordan block is different from zero.

EXAMPLE 6.11 Consider the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 10 \\ 9 \\ 0 \\ 1 \end{bmatrix} u \tag{6.52}$$

$$y = [1 \ 0 \ 0 \ 2]\mathbf{x}$$

There are two Jordan blocks, one with order 3 and associated with eigenvalue 0, the other with order 1 and associated with eigenvalue -2. The entry of $\mathbf{B}$ corresponding to the last row of the first Jordan block is zero; thus the state equation is not controllable. The two entries of $\mathbf{C}$ corresponding to the first column of both Jordan blocks are different from zero; thus the state equation is observable.

6.6 Discrete-Time State Equations

Consider the n-dimensional p-input q-output state equation

$$\begin{aligned} \mathbf{x}[k+1] &= \mathbf{A}\mathbf{x}[k] + \mathbf{B}\mathbf{u}[k] \\ \mathbf{y}[k] &= \mathbf{C}\mathbf{x}[k] \end{aligned} \tag{6.53}$$

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are, respectively, $n \times n$, $n \times p$, and $q \times n$ real constant matrices.

Definition 6.D1 *The discrete-time state equation (6.53) or the pair* $(\mathbf{A}, \mathbf{B})$ *is said to be* controllable *if for any initial state* $\mathbf{x}(0) = \mathbf{x}_0$ *and any final state* $\mathbf{x}_1$, *there exists an input sequence of finite length that transfers* $\mathbf{x}_0$ *to* $\mathbf{x}_1$. *Otherwise the equation or* $(\mathbf{A}, \mathbf{B})$ *is said to be* uncontrollable.

Theorem 6.D1

The following statements are equivalent:

1. The n-dimensional pair $(\mathbf{A}, \mathbf{B})$ is controllable.
2. The $n \times n$ matrix

$$\mathbf{W}_{dc}[n-1] = \sum_{m=0}^{n-1} (\mathbf{A})^m \mathbf{B}\mathbf{B}'(\mathbf{A}')^m \tag{6.54}$$

   is nonsingular.
3. The $n \times np$ *controllability matrix*

$$C_d = [\mathbf{B} \ \mathbf{AB} \ \mathbf{A}^2\mathbf{B} \ \cdots \ \mathbf{A}^{n-1}\mathbf{B}] \tag{6.55}$$

   has rank n (full row rank). The matrix can be generated by calling `ctrb` in MATLAB.
4. The $n \times (n + p)$ matrix $[\mathbf{A} - \lambda\mathbf{I} \ \mathbf{B}]$ has full row rank at every eigenvalue, λ, of $\mathbf{A}$.
5. If, in addition, all eigenvalues of $\mathbf{A}$ have magnitudes less than 1, then the unique solution of

$$\mathbf{W}_{dc} - \mathbf{A}\mathbf{W}_{dc}\mathbf{A}' = \mathbf{B}\mathbf{B}' \tag{6.56}$$

is positive definite. The solution is called the discrete *controllability Gramian* and can be obtained by using the MATLAB function `dgram`. The discrete Gramian can be expressed as

$$\mathbf{W}_{dc} = \sum_{m=0}^{\infty} \mathbf{A}^m \mathbf{B}\mathbf{B}'(\mathbf{A}')^m \tag{6.57}$$

The solution of (6.53) at $k = n$ was derived in (4.20) as

$$x[n] = \mathbf{A}^n \mathbf{x}[0] + \sum_{m=0}^{n-1} \mathbf{A}^{n-1-m}\mathbf{B}\mathbf{u}[m]$$

which can be written as

$$\mathbf{x}[n] - \mathbf{A}^n\mathbf{x}[0] = [\mathbf{B}\ \ \mathbf{A}\mathbf{B}\ \ \cdots\ \ \mathbf{A}^{n-1}\mathbf{B}] \begin{bmatrix} \mathbf{u}[n-1] \\ \mathbf{u}[n-2] \\ \vdots \\ \mathbf{u}[0] \end{bmatrix} \tag{6.58}$$

It follows from Theorem 3.1 that for any $\mathbf{x}[0]$ and $\mathbf{x}[n]$, an input sequence exists if and only if the controllability matrix has full row rank. This shows the equivalence of (1) and (3). The matrix $\mathbf{W}_{dc}[n-1]$ can be written as

$$\mathbf{W}_{dc}[n-1] = [\mathbf{B}\ \ \mathbf{A}\mathbf{B}\ \ \cdots\ \ \mathbf{A}^{n-1}\mathbf{B}] \begin{bmatrix} \mathbf{B}' \\ \mathbf{B}'\mathbf{A}' \\ \vdots \\ \mathbf{B}'(\mathbf{A}')^{n-1} \end{bmatrix}$$

The equivalence of (2) and (3) then follows Theorem 3.8. Note that $\mathbf{W}_{dc}[m]$ is always positive semidefinite. If it is nonsingular or, equivalently, positive definite, then (6.53) is controllable. The proof of the equivalence of (3) and (4) is identical to the continuous-time case. Condition (5) follows Condition (2) and Theorem 5.D6. We see that establishing Theorem 6.D1 is considerably simpler than establishing Theorem 6.1.

There is one important difference between the continuous- and discrete-time cases. If a continuous-time state equation is controllable, the input can transfer any state to any other state in any nonzero time interval, no matter how small. If a discrete-time state equation is controllable, an input sequence of length n can transfer any state to any other state. If we compute the controllability index μ as defined in (6.15), then the transfer can be achieved using an input sequence of length μ. If an input sequence is shorter than μ, it is not possible to transfer any state to any other state.

Definition 6.D2 *The discrete-time state equation (6.53) or the pair* $(\mathbf{A}, \mathbf{C})$ *is said to be* observable *if for any unknown initial state* $\mathbf{x}[0]$*, there exists a finite integer* $k_1 > 0$ *such that the knowledge of the input sequence* $\mathbf{u}[k]$ *and output sequence* $\mathbf{y}[k]$ *from* $k = 0$ *to* k_1 *suffices to determine uniquely the initial state* $\mathbf{x}[0]$*. Otherwise, the equation is said to be* unobservable.

Theorem 6.DO1

The following statements are equivalent:

1. The n-dimensional pair $(\mathbf{A}, \mathbf{C})$ is observable.
2. The $n \times n$ matrix

$$\mathbf{W}_{do}[n-1] = \sum_{m=0}^{n-1} (\mathbf{A}')^m \mathbf{C}'\mathbf{C}\mathbf{A}^m \tag{6.59}$$

is nonsingular or, equivalently, positive definite.

3. The $nq \times n$ *observability matrix*

$$O_d = \begin{bmatrix} \mathbf{C} \\ \mathbf{C}\mathbf{A} \\ \vdots \\ \mathbf{C}\mathbf{A}^{n-1} \end{bmatrix} \tag{6.60}$$

has rank n (full column rank). The matrix can be generated by calling `obsv` in MATLAB.

4. The $(n+q) \times n$ matrix

$$\begin{bmatrix} \mathbf{A} - \lambda\mathbf{I} \\ \mathbf{B} \end{bmatrix}$$

has full column rank at every eigenvalue, λ, of $\mathbf{A}$.

5. If, in addition, all eigenvalues of $\mathbf{A}$ have magnitudes less than 1, then the unique solution of

$$\mathbf{W}_{do} - \mathbf{A}'\mathbf{W}_{do}\mathbf{A} = \mathbf{C}'\mathbf{C} \tag{6.61}$$

is positive definite. The solution is called the discrete *observability Gramian* and can be expressed as

$$\mathbf{W}_{do} = \sum_{m=0}^{\infty} (\mathbf{A}')^m \mathbf{C}'\mathbf{C}\mathbf{A}^m \tag{6.62}$$

This can be proved directly or indirectly using the duality theorem. We mention that all other properties—such as controllability and observability indices, Kalman decomposition, and Jordan-form controllability and observability conditions—discussed for the continuous-time case apply to the discrete-time case without any modification. The controllability index and observability index, however, have simple interpretations in the discrete-time case. The controllability index is the shortest input sequence that can transfer any state to any other state. The observability index is the shortest input and output sequences needed to determine the initial state uniquely.

6.6.1 Controllability to the Origin and Reachability

In the literature, there are three different controllability definitions:

1. Transfer any state to any other state as adopted in Definition 6.D1.
2. Transfer any state to the zero state, called controllability to the origin.

3. Transfer the zero state to any state, called controllability from the origin or, more often, *reachability*.

In the continuous-time case, because $e^{\mathbf{A}t}$ is nonsingular, the three definitions are equivalent. In the discrete-time case, if $\mathbf{A}$ is nonsingular, the three definitions are again equivalent. But if $\mathbf{A}$ is singular, then (1) and (3) are equivalent, but not (2) and (3). The equivalence of (1) and (3) can easily be seen from (6.58). We use examples to discuss the difference between (2) and (3). Consider

$$\mathbf{x}[k+1] = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{x}[k] + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} u[k] \tag{6.63}$$

Its controllability matrix has rank 0 and the equation is not controllable as defined in (1) or not reachable as defined in (3). The matrix $\mathbf{A}$ has the form shown in (3.40) and has the property $\mathbf{A}^k = \mathbf{0}$ for $k \geq 3$. Thus we have

$$\mathbf{x}[3] = \mathbf{A}^3\mathbf{x}[0] = \mathbf{0}$$

for any initial state $\mathbf{x}[0]$. Thus every state propagates to the zero state whether or not an input sequence is applied. Thus the equation is controllable to the origin. A different example follows. Consider

$$\mathbf{x}[k+1] = \begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{x}[k] + \begin{bmatrix} -1 \\ 0 \end{bmatrix} u[k] \tag{6.64}$$

Its controllability matrix

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

has rank 1 and the equation is not reachable. However, for any $x_1[0] = \alpha$ and $x_2[0] = \beta$, the input $u[0] = 2\alpha + \beta$ transfers $\mathbf{x}[0]$ to $\mathbf{x}[1] = \mathbf{0}$. Thus the equation is controllable to the origin. Note that the A-matrices in (6.63) and (6.64) are both singular. The definition adopted in Definition 6.D1 encompasses the other two definitions and makes the discussion simple. For a thorough discussion of the three definitions, see Reference [4].

6.7 Controllability After Sampling

Consider a continuous-time state equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{6.65}$$

If the input is piecewise constant or

$$u[k] := u(kT) = u(t) \qquad \text{for } kT \leq t < (k+1)T$$

then the equation can be described, as developed in (4.17), by

$$\bar{\mathbf{x}}[k+1] = \bar{\mathbf{A}}\bar{\mathbf{x}}[k] + \bar{\mathbf{B}}\mathbf{u}[k] \tag{6.66}$$

with

$$\bar{\mathbf{A}} = e^{\mathbf{A}T} \qquad \bar{\mathbf{B}} = \left(\int_0^T e^{\mathbf{A}t}dt\right)\mathbf{B} =: \mathbf{MB} \tag{6.67}$$

The question is: If (6.65) is controllable, will its sampled equation in (6.66) be controllable? This problem is important in designing so-called dead-beat sampled-data systems and in computer control of continuous-time systems. The answer to the question depends on the sampling period T and the location of the eigenvalues of $\mathbf{A}$. Let λ_i and $\bar{\lambda}_i$ be, respectively, the eigenvalues of $\mathbf{A}$ and $\bar{\mathbf{A}}$. We use Re and Im to denote the real part and imaginary part. Then we have the following theorem.

Theorem 6.9

Suppose (6.65) is controllable. A sufficient condition for its discretized equation in (6.66), with sampling period T, to be controllable is that $|\text{Im}[\lambda_i - \lambda_j]| \neq 2\pi m/T$ for $m = 1, 2, \ldots$, whenever $\text{Re}[\lambda_i - \lambda_j] = 0$. For the single-input case, the condition is necessary as well.

First we remark on the conditions. If $\mathbf{A}$ has only real eigenvalues, then the discretized equation with any sampling period $T > 0$ is always controllable. Suppose $\mathbf{A}$ has complex conjugate eigenvalues $\alpha \pm j\beta$. If the sampling period T does not equal any integer multiple of π/β, then the discretized state equation is controllable. If $T = m\pi/\beta$ for some integer m, then the discretized equation *may not* be controllable. The reason is as follows. Because $\bar{\mathbf{A}} = e^{\mathbf{A}T}$, if λ_i is an eigenvalue of $\mathbf{A}$, then $\bar{\lambda}_i := e^{\lambda_i T}$ is an eigenvalue of $\bar{\mathbf{A}}$ (Problem 3.19). If $T = m\pi/\beta$, the two distinct eigenvalues $\lambda_1 = \alpha + j\beta$ and $\lambda_2 = \alpha - j\beta$ of $\mathbf{A}$ become a repeated eigenvalue $-e^{\alpha T}$ or $e^{\alpha T}$ of $\bar{\mathbf{A}}$. This will cause the discretized equation to be uncontrollable, as we will see in the proof. We show Theorem 6.9 by assuming $\mathbf{A}$ to be in Jordan form. This is permitted because controllability is invariant under any equivalence transformation.

⇒ ***Proof of Theorem 6.9*** To simplify the discussion, we assume $\mathbf{A}$ to be of the form

$$\mathbf{A} = \text{diag}(\mathbf{A}_{11}, \mathbf{A}_{12}, \mathbf{A}_{21}) = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix} \tag{6.68}$$

In other words, $\mathbf{A}$ has two distinct eigenvalues λ_1 and λ_2. There are two Jordan blocks, one with order 3 and one with order 1, associated with λ_1 and only one Jordan block of order 2 associated with λ_2. Using (3.48), we have

$$\bar{\mathbf{A}} = \text{diag}(\bar{\mathbf{A}}_{11}, \bar{\mathbf{A}}_{12}, \bar{\mathbf{A}}_{21}) = \begin{bmatrix} e^{\lambda_1 T} & Te^{\lambda_1 T} & T^2e^{\lambda_1 T}/2 & 0 & 0 & 0 \\ 0 & e^{\lambda_1 T} & Te^{\lambda_1 T} & 0 & 0 & 0 \\ 0 & 0 & e^{\lambda_1 T} & 0 & 0 & 0 \\ 0 & 0 & 0 & e^{\lambda_1 T} & 0 & 0 \\ 0 & 0 & 0 & 0 & e^{\lambda_2 T} & Te^{\lambda_2 T} \\ 0 & 0 & 0 & 0 & 0 & e^{\lambda_2 T} \end{bmatrix} \tag{6.69}$$

This is not in Jordan form. Because we will use Theorem 6.8, which is also applicable to the discrete-time case without any modification, to prove Theorem 6.9, we must transform $\bar{\mathbf{A}}$ in (6.69) into Jordan form. It turns out that the Jordan form of $\bar{\mathbf{A}}$ equals the one in (6.68) if λ_i is replaced by $\bar{\lambda}_i := e^{\lambda_i T}$ (Problem 3.17). In other words, there exists a nonsingular triangular matrix $\mathbf{P}$ such that the transformation $\bar{\mathbf{x}} = \mathbf{P}\bar{\mathbf{x}}$ will transform (6.66) into

$$\bar{\mathbf{x}}[k+1] = \mathbf{P}\bar{\mathbf{A}}\mathbf{P}^{-1}\bar{\mathbf{x}}[k] + \mathbf{PMBu}[k] \tag{6.70}$$

with $\mathbf{P}\bar{\mathbf{A}}\mathbf{P}^{-1}$ in the Jordan form in (6.68) with λ_i replaced by $\bar{\lambda}_i$. Now we are ready to establish Theorem 6.9.

First we show that $\mathbf{M}$ in (6.67) is nonsingular. If $\mathbf{A}$ is of the form shown in (6.68), then $\mathbf{M}$ is block diagonal and triangular. Its diagonal entry is of form

$$m_{ii} := \int_0^T e^{\lambda_i \tau} d\tau = \begin{cases} (e^{\lambda_i T} - 1)/\lambda_i & \text{if } \lambda_i \neq 0 \\ T & \text{if } \lambda_i = 0 \end{cases} \tag{6.71}$$

Let $\lambda_i = \alpha_i + j\beta_i$. The only way for $m_{ii} = 0$ is $\alpha_i = 0$ and $\beta_i T = 2\pi m$. In this case, $-j\beta_i$ is also an eigenvalue and the theorem requires that $2\beta_i T \neq 2\pi m$. Thus we conclude $m_{ii} \neq 0$ and $\mathbf{M}$ is nonsingular and triangular.

If $\mathbf{A}$ is of the form shown in (6.68), then it is controllable if and only if the third and fourth rows of $\mathbf{B}$ are linearly independent and the last row of $\mathbf{B}$ is nonzero (Theorem 6.8). Under the condition in Theorem 6.9, the two eigenvalues $\bar{\lambda}_1 = e^{\lambda_1 T}$ and $\bar{\lambda}_2 = e^{\lambda_2 T}$ of $\bar{\mathbf{A}}$ are distinct. Thus (6.70) is controllable if and only if the third and fourth rows of $\mathbf{PMB}$ are linearly independent and the last row of $\mathbf{PMB}$ is nonzero. Because $\mathbf{P}$ and $\mathbf{M}$ are both triangular and nonsingular, $\mathbf{PMB}$ and $\mathbf{B}$ have the same properties on the linear independence of their rows. This shows the sufficiency of the theorem. If the condition in Theorem 6.9 is not met, then $\bar{\lambda}_1 = \bar{\lambda}_2$. In this case, (6.70) is controllable if the third, fourth, and last rows of $\mathbf{PMB}$ are linearly independent. This is still possible if $\mathbf{B}$ has three or more columns. Thus the condition is not necessary. In the single-input case, if $\bar{\lambda}_1 = \bar{\lambda}_2$, then (6.70) has two or more Jordan blocks associated with the same eigenvalue and (6.70) is, following Corollary 6.8, not controllable. This establishes the theorem. Q.E.D.

In the proof of Theorem 6.9, we have essentially established the theorem that follows.

Theorem 6.10

If a continuous-time linear time-invariant state equation is not controllable, then its discretized state equation, with any sampling period, is not controllable.

This theorem is intuitively obvious. If a state equation is not controllable using any input, it is certainly not controllable using only piecewise constant input.

Example 6.12 Consider the system shown in Fig. 6.11. Its input is sampled every T seconds and then kept constant using a hold circuit. The transfer function of the system is given as

$$\hat{g}(s) = \frac{s+2}{s^3 + 3s^2 + 7s + 5} = \frac{s+2}{(s+1)(s+1+j2)(s+1-j2)} \tag{6.72}$$

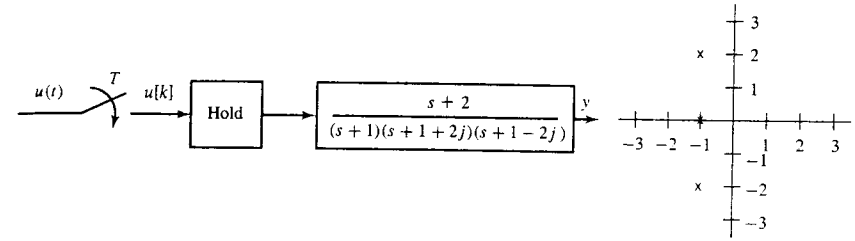


Figure 6.11 System with piecewise constant input.

Using (4.41), we can readily obtain the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} -3 & -7 & -5 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u \tag{6.73}$$

$$y = [0 \;\; 1 \;\; 2]\mathbf{x}$$

to describe the system. It is a controllable-form realization and is clearly controllable. The eigenvalues of $\mathbf{A}$ are -1, $-1 \pm j2$ and are plotted in Fig. 6.11. The three eigenvalues have the same real part; their differences in imaginary parts are 2 and 4. Thus the discretized state equation is controllable if and only if

$$T \neq \frac{2\pi m}{2} = \pi m \quad \text{and} \quad T \neq \frac{2\pi m}{4} = 0.5\pi m$$

for $m = 1, 2, \ldots$. The second condition includes the first condition. Thus we conclude that the discretized equation of (6.73) is controllable if and only if $T \neq 0.5m\pi$ for any positive integer m.

We use MATLAB to check the result for $m = 1$ or $T = 0.5\pi$. Typing

```
a=[-3 -7 -5;1 0 0;0 1 0];b=[1;0;0];
[ad,bd]=c2d(a,b,pi/2)
```

yields the discretized state equation as

$$\bar{\mathbf{x}}[k+1] = \begin{bmatrix} -0.1039 & 0.2079 & 0.5197 \\ -0.1390 & -0.4158 & -0.5197 \\ 0.1039 & 0.2079 & 0.3118 \end{bmatrix} \bar{\mathbf{x}}[k] + \begin{bmatrix} -0.1039 \\ 0.1039 \\ 0.1376 \end{bmatrix} u[k] \tag{6.74}$$

Its controllability matrix can be obtained by typing `ctrb(ad,bd)`, which yields

$$C_d = \begin{bmatrix} -0.1039 & 0.1039 & -0.0045 \\ 0.1039 & -0.1039 & 0.0045 \\ 0.1376 & 0.0539 & 0.0059 \end{bmatrix}$$

Its first two rows are clearly linearly dependent. Thus C_d does not have full row rank and (6.74) is not controllable as predicted by Theorem 6.9. We mention that if we type

`rank(ctrb(ad,bd))`, the result is 3 and (6.74) is controllable. This is incorrect and is due to roundoff errors. We see once again that the rank is very sensitive to roundoff errors.

What has been discussed is also applicable to the observability part. In other words, under the conditions in Theorem 6.9, if a continuous-time state equation is observable, its discretized equation is also observable.

6.8 LTV State Equations

Consider the n-dimensional p-input q-output state equation

$$\begin{aligned}\dot{\mathbf{x}} &= \mathbf{A}(t)\mathbf{x} + \mathbf{B}(t)\mathbf{u}\\ \mathbf{y} &= \mathbf{C}(t)\mathbf{x}\end{aligned} \tag{6.75}$$

The state equation is said to be controllable at t_0, if there exists a finite $t_1 > t_0$ such that for any $\mathbf{x}(t_0) = \mathbf{x}_0$ and any $\mathbf{x}_1$, there exists an input that transfers $\mathbf{x}_0$ to $\mathbf{x}_1$ at time t_1. Otherwise the state equation is uncontrollable at t_0. In the time-invariant case, if a state equation is controllable, then it is controllable at every t_0 and for every $t_1 > t_0$; thus there is no need to specify t_0 and t_1. In the time-varying case, the specification of t_0 and t_1 is crucial.

Theorem 6.11

The n-dimensional pair $(\mathbf{A}(t), \mathbf{B}(t))$ is controllable at time t_0 if and only if there exists a finite $t_1 > t_0$ such that the $n \times n$ matrix

$$\mathbf{W}_c(t_0, t_1) = \int_{t_0}^{t_1} \mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{B}'(\tau)\mathbf{\Phi}'(t_1, \tau)\, d\tau \tag{6.76}$$

where $\mathbf{\Phi}(t, \tau)$ is the state transition matrix of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$, is nonsingular.

⇒ ***Proof:*** We first show that if $\mathbf{W}_c(t_0, t_1)$ is nonsingular, then (6.75) is controllable. The response of (6.75) at t_1 was computed in (4.57) as

$$\mathbf{x}(t_1) = \mathbf{\Phi}(t_1, t_0)\mathbf{x}_0 + \int_{t_0}^{t_1} \mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \tag{6.77}$$

We claim that the input

$$\mathbf{u}(t) = -\mathbf{B}'(t)\mathbf{\Phi}'(t_1, t)\mathbf{W}_c^{-1}(t_0, t_1)[\mathbf{\Phi}(t_1, t_0)\mathbf{x}_0 - \mathbf{x}_1] \tag{6.78}$$

will transfer $\mathbf{x}_0$ at time t_0 to $\mathbf{x}_1$ at time t_1. Indeed, substituting (6.78) into (6.77) yields

$$\begin{aligned}\mathbf{x}(t_1) &= \mathbf{\Phi}(t_1, t_0)\mathbf{x}_0 - \int_{t_0}^{t_1} \mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{B}'(\tau)\mathbf{\Phi}'(t_1, \tau)\, d\tau\\ &\quad \cdot \mathbf{W}_c^{-1}(t_0, t_1)[\mathbf{\Phi}(t_1, t_0)\mathbf{x}_0 - \mathbf{x}_1]\\ &= \mathbf{\Phi}(t_1, t_0)\mathbf{x}_0 - \mathbf{W}_c(t_0, t_1)\mathbf{W}_c^{-1}(t_0, t_1)[\mathbf{\Phi}(t_1, t_0)\mathbf{x}_0 - \mathbf{x}_1] = \mathbf{x}_1\end{aligned}$$

Thus the equation is controllable at t_0. We show the converse by contradiction. Suppose (6.75) is controllable at t_0 but $\mathbf{W}_c(t_0, t)$ is singular or, positive semidefinite, for all $t_1 > t_0$. Then there exists an $n \times 1$ nonzero constant vector $\mathbf{v}$ such that

$$\begin{aligned}\mathbf{v}'\mathbf{W}_c(t_0, t_1)\mathbf{v} &= \int_{t_0}^{t_1} \mathbf{v}'\mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{B}'(\tau)\mathbf{\Phi}'(t_1, \tau)\mathbf{v}\, d\tau\\ &= \int_{t_0}^{t_1} \|\mathbf{B}'(\tau)\mathbf{\Phi}'(t_1, \tau)\mathbf{v}\|^2 d\tau = 0\end{aligned}$$

which implies

$$\mathbf{B}'(\tau)\mathbf{\Phi}'(t_1, \tau)\mathbf{v} \equiv \mathbf{0} \quad \text{or} \quad \mathbf{v}'\mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau) \equiv \mathbf{0} \tag{6.79}$$

for all τ in $[t_0, t_1]$. If (6.75) is controllable, there exists an input that transfers the initial state $\mathbf{x}_0 = \mathbf{\Phi}(t_0, t_1)\mathbf{v}$ at t_0 to $\mathbf{x}(t_1) = \mathbf{0}$. Then (6.77) becomes

$$\mathbf{0} = \mathbf{\Phi}(t_1, t_0)\mathbf{\Phi}(t_0, t_1)\mathbf{v} + \int_{t_0}^{t_1} \mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau \tag{6.80}$$

Its premultiplication by $\mathbf{v}'$ yields

$$0 = \mathbf{v}'\mathbf{v} + \mathbf{v}' \int_{t_0}^{t_1} \mathbf{\Phi}(t_1, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)\, d\tau = \|\mathbf{v}\|^2 + 0$$

This contradicts the hypothesis $\mathbf{v} \neq \mathbf{0}$. Thus if $(\mathbf{A}(t), \mathbf{B}(t))$ is controllable at t_0, $\mathbf{W}_c(t_0, t_1)$ must be nonsingular for some finite $t_1 > t_0$. This establishes Theorem 6.11. Q.E.D.

In order to apply Theorem 6.11, we need knowledge of the state transition matrix, which, however, may not be available. Therefore it is desirable to develop a controllability condition without involving $\mathbf{\Phi}(t, \tau)$. This is possible if we have additional conditions on $\mathbf{A}(t)$ and $\mathbf{B}(t)$. Recall that we have assumed $\mathbf{A}(t)$ and $\mathbf{B}(t)$ to be continuous. Now we require them to be $(n-1)$ times continuously differentiable. Define $\mathbf{M}_0(t) = \mathbf{B}(t)$. We then define recursively a sequence of $n \times p$ matrices $\mathbf{M}_m(t)$ as

$$\mathbf{M}_{m+1}(t) := -\mathbf{A}(t)\mathbf{M}_m(t) + \frac{d}{dt}\mathbf{M}_m(t) \tag{6.81}$$

for $m = 0, 1, \ldots, n-1$. Clearly, we have

$$\mathbf{\Phi}(t_2, t)\mathbf{B}(t) = \mathbf{\Phi}(t_2, t)\mathbf{M}_0(t)$$

for any fixed t_2. Using

$$\frac{\partial}{\partial t}\mathbf{\Phi}(t_2, t) = -\mathbf{\Phi}(t_2, t)\mathbf{A}(t)$$

(Problem 4.17), we compute

$$\begin{aligned}\frac{\partial}{\partial t}[\mathbf{\Phi}(t_2, t)\mathbf{B}(t)] &= \frac{\partial}{\partial t}[\mathbf{\Phi}(t_2, t)]\mathbf{B}(t) + \mathbf{\Phi}(t_2, t)\frac{d}{dt}\mathbf{B}(t)\\ &= \mathbf{\Phi}(t_2, t)[-\mathbf{A}(t)\mathbf{M}_0(t) + \frac{d}{dt}\mathbf{M}_0(t)] = \mathbf{\Phi}(t_2, t)\mathbf{M}_1(t)\end{aligned}$$

Proceeding forward, we have

$$\frac{\partial^m}{\partial t^m}\Phi(t_2, t)\mathbf{B}(t) = \Phi(t_2, t)\mathbf{M}_m(t) \tag{6.82}$$

for $m = 0, 1, 2, \ldots$. The following theorem is sufficient but not necessary for (6.75) to be controllable.

Theorem 6.12

Let $\mathbf{A}(t)$ and $\mathbf{B}(t)$ be $n-1$ times continuously differentiable. Then the n-dimensional pair $(\mathbf{A}(t), \mathbf{B}(t))$ is controllable at t_0 if there exists a finite $t_1 > t_0$ such that

$$\text{rank } [\mathbf{M}_0(t_1) \;\; \mathbf{M}_1(t_1) \;\; \cdots \;\; \mathbf{M}_{n-1}(t_1)] = n \tag{6.83}$$

Proof: We show that if (6.83) holds, then $\mathbf{W}_c(t_0, t)$ is nonsingular for all $t \geq t_1$. Suppose not, that is, $\mathbf{W}_c(t_0, t)$ is singular or positive semidefinite for some $t_2 \geq t_1$. Then there exists an $n \times 1$ nonzero constant vector $\mathbf{v}$ such that

$$\begin{aligned}\mathbf{v}'\mathbf{W}_c(t_0, t_2)\mathbf{v} &= \int_{t_0}^{t_2} \mathbf{v}'\Phi(t_2, \tau)\mathbf{B}(\tau)\mathbf{B}'(\tau)\Phi'(t_2, \tau)\mathbf{v}\, d\tau \\ &= \int_{t_0}^{t_2} \|\mathbf{B}'(\tau)\Phi'(t_2, \tau)\mathbf{v}\|^2 d\tau = 0\end{aligned}$$

which implies

$$\mathbf{B}'(\tau)\Phi'(t_2, \tau)\mathbf{v} \equiv \mathbf{0} \quad \text{or} \quad \mathbf{v}'\Phi(t_2, \tau)\mathbf{B}(\tau) \equiv \mathbf{0} \tag{6.84}$$

for all τ in $[t_0, t_2]$. Its differentiations with respect to τ yield, as derived in (6.82),

$$\mathbf{v}'\Phi(t_2, \tau)\mathbf{M}_m(\tau) \equiv \mathbf{0}$$

for $m = 0, 1, 2, \ldots, n-1$, and all τ in $[t_0, t_2]$, in particular, at t_1. They can be arranged as

$$\mathbf{v}'\Phi(t_2, t_1)[\mathbf{M}_0(t_1) \;\; \mathbf{M}_1(t_1) \;\; \cdots \;\; \mathbf{M}_{n-1}(t_1)] = \mathbf{0} \tag{6.85}$$

Because $\Phi(t_2, t_1)$ is nonsingular, $\mathbf{v}'\Phi(t_2, t_1)$ is nonzero. Thus (6.85) contradicts (6.83). Therefore, under the condition in (6.83), $\mathbf{W}_c(t_0, t_2)$, for any $t_2 \geq t_1$, is nonsingular and $(\mathbf{A}(t), \mathbf{B}(t))$ is, following Theorem 6.11, controllable at t_0. Q.E.D.

EXAMPLE 6.13 Consider

$$\dot{\mathbf{x}} = \begin{bmatrix} t & -1 & 0 \\ 0 & -t & t \\ 0 & 0 & t \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u \tag{6.86}$$

We have $\mathbf{M}_0 = [0 \;\; 1 \;\; 1]'$ and compute

$$\mathbf{M}_1 = -\mathbf{A}(t)\mathbf{M}_0 + \frac{d}{dt}\mathbf{M}_0 = \begin{bmatrix} 1 \\ 0 \\ -t \end{bmatrix}$$

$$\mathbf{M}_2 = -\mathbf{A}(t)\mathbf{M}_1 + \frac{d}{dt}\mathbf{M}_1 = \begin{bmatrix} -t \\ t^2 \\ t^2 - 1 \end{bmatrix}$$

The determinant of the matrix

$$[\mathbf{M}_0 \; \mathbf{M}_1 \; \mathbf{M}_2] = \begin{bmatrix} 0 & 1 & -t \\ 1 & 0 & t^2 \\ 1 & -t & t^2 - 1 \end{bmatrix}$$

is $t^2 + 1$, which is nonzero for all t. Thus the state equation in (6.86) is controllable at every t.

EXAMPLE 6.14 Consider

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u \tag{6.87}$$

and

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} e^t \\ e^{2t} \end{bmatrix} u \tag{6.88}$$

Equation (6.87) is a time-invariant equation and is controllable according to Corollary 6.8. Equation (6.88) is a time-varying equation; the two entries of its B-matrix are nonzero for all t and one might be tempted to conclude that (6.88) is controllable. Let us check this by using Theorem 6.11. Its state transition matrix is

$$\Phi(t, \tau) = \begin{bmatrix} e^{t-\tau} & 0 \\ 0 & e^{2(t-\tau)} \end{bmatrix}$$

and

$$\Phi(t, \tau)\mathbf{B}(\tau) = \begin{bmatrix} e^{t-\tau} & 0 \\ 0 & e^{2(t-\tau)} \end{bmatrix}\begin{bmatrix} e^{\tau} \\ e^{2\tau} \end{bmatrix} = \begin{bmatrix} e^t \\ e^{2t} \end{bmatrix}$$

We compute

$$\begin{aligned}\mathbf{W}_c(t_0, t) &= \int_{t_0}^{t} \begin{bmatrix} e^t \\ e^{2t} \end{bmatrix}[e^t \;\; e^{2t}]d\tau = \begin{bmatrix} \int_{t_0}^{t} e^{2t}d\tau & \int_{t_0}^{t} e^{3t}d\tau \\ \int_{t_0}^{t} e^{3t}d\tau & \int_{t_0}^{t} e^{4t}d\tau \end{bmatrix} \\ &= \begin{bmatrix} e^{2t}(t - t_0) & e^{3t}(t - t_0) \\ e^{3t}(t - t_0) & e^{4t}(t - t_0) \end{bmatrix}\end{aligned}$$

Its determinant is identically zero for all t_0 and t. Thus (6.88) is not controllable at any t_0. From this example, we see that, in applying a theorem, every condition should be checked carefully; otherwise, we might obtain an erroneous conclusion.

We now discuss the observability part. The linear time-varying state equation in (6.75) is observable at t_0 if there exists a finite t_1 such that for any state $\mathbf{x}(t_0) = \mathbf{x}_0$, the knowledge of the input and output over the time interval $[t_0, t_1]$ suffices to determine uniquely the initial state $\mathbf{x}_0$. Otherwise, the state equation is said to be unobservable at t_0.

Theorem 6.O11

The pair $(\mathbf{A}(t), \mathbf{C}(t))$ is observable at time t_0 if and only if there exists a finite $t_1 > t_0$ such that the $n \times n$ matrix

$$\mathbf{W}_o(t_0, t_1) = \int_{t_0}^{t_1} \mathbf{\Phi}'(\tau, t_0)\mathbf{C}'(\tau)\mathbf{C}(\tau)\mathbf{\Phi}(\tau, t_0)\, d\tau \tag{6.89}$$

where $\mathbf{\Phi}(t, \tau)$ is the state transition matrix of $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$, is nonsingular.

Theorem 6.O12

Let $\mathbf{A}(t)$ and $\mathbf{C}(t)$ be $n-1$ times continuously differentiable. Then the n-dimensional pair $(\mathbf{A}(t), \mathbf{C}(t))$ is observable at t_0 if there exists a finite $t_1 > t_0$ such that

$$\operatorname{rank}\begin{bmatrix} \mathbf{N}_0(t_1) \\ \mathbf{N}_1(t_1) \\ \vdots \\ \mathbf{N}_{n-1}(t_1) \end{bmatrix} = n \tag{6.90}$$

where

$$\mathbf{N}_{m+1}(t) = \mathbf{N}_m(t)\mathbf{A}(t) + \frac{d}{dt}\mathbf{N}_m(t) \qquad m = 0, 1, \ldots, n-1$$

with

$$\mathbf{N}_0 = \mathbf{C}(t)$$

We mention that the duality theorem in Theorem 6.5 for time-invariant systems is not applicable to time-varying systems. It must be modified. See Problems 6.22 and 6.23.

PROBLEMS

6.1 Is the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -3 & -3 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u$$
$$y = [1 \;\; 2 \;\; 1]\mathbf{x}$$

controllable? Observable?

6.2 Is the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & -1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}\mathbf{u}$$
$$y = [1 \;\; 0 \;\; 1]\mathbf{x}$$

controllable? Observable?

6.3 Is it true that the rank of $[\mathbf{B} \;\; \mathbf{AB} \;\; \cdots \;\; \mathbf{A}^{n-1}\mathbf{B}]$ equals the rank of $[\mathbf{AB} \;\; \mathbf{A}^2\mathbf{B} \;\; \cdots \;\; \mathbf{A}^n\mathbf{B}]$? If not, under what conditon will it be true?

6.4 Show that the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}\mathbf{x} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{0} \end{bmatrix}\mathbf{u}$$

is controllable if and only if the pair $(\mathbf{A}_{22}, \mathbf{A}_{21})$ is controllable.

6.5 Find a state equation to describe the network shown in Fig. 6.1, and then check its controllability and observability.

6.6 Find the controllability index and observability index of the state equations in Problems 6.1 and 6.2.

6.7 What is the controllability index of the state equation

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Iu}$$

where $\mathbf{I}$ is the unit matrix?

6.8 Reduce the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} -1 & 4 \\ 4 & -1 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u \qquad y = [1 \;\; 1]\mathbf{x}$$

to a controllable one. Is the reduced equation observable?

6.9 Reduce the state equation in Problem 6.5 to a controllable and observable equation.

6.10 Reduce the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} \lambda_1 & 1 & 0 & 0 & 0 \\ 0 & \lambda_1 & 1 & 0 & 0 \\ 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 & 1 \\ 0 & 0 & 0 & 0 & \lambda_2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix} u$$
$$y = [0 \;\; 1 \;\; 1 \;\; 1 \;\; 0 \;\; 1]\mathbf{x}$$

to a controllable and observable equation.

6.11 Consider the n-dimensional state equation

$$\dot{\mathbf{x}} = \mathbf{Ax} + \mathbf{Bu}$$
$$\mathbf{y} = \mathbf{Cx} + \mathbf{Du}$$

The rank of its controllability matrix is assumed to be $n_1 < n$. Let $\mathbf{Q}_1$ be an $n \times n_1$ matrix whose columns are any n_1 linearly independent columns of the controllability matrix. Let $\mathbf{P}_1$ be an $n_1 \times n$ matrix such that $\mathbf{P}_1\mathbf{Q}_1 = \mathbf{I}_{n_1}$, where $\mathbf{I}_{n_1}$ is the unit matrix of order n_1. Show that the following n_1-dimensional state equation

$$\dot{\bar{\mathbf{x}}}_1 = \mathbf{P}_1\mathbf{A}\mathbf{Q}_1\bar{\mathbf{x}}_1 + \mathbf{P}_1\mathbf{Bu}$$
$$\bar{\mathbf{y}} = \mathbf{C}\mathbf{Q}_1\bar{\mathbf{x}}_1 + \mathbf{Du}$$

is controllable and has the same transfer matrix as the original state equation.

6.12 In Problem 6.11, the reduction procedure reduces to solving for $\mathbf{P}_1$ in $\mathbf{P}_1\mathbf{Q}_1 = \mathbf{I}$. How do you solve $\mathbf{P}_1$?

6.13 Develop a similar statement as in Problem 6.11 for an unobservable state equation.

6.14 Is the Jordan-form state equation controllable and observable?

$$\dot{\mathbf{x}} = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 2 & 1 & 0 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \\ 3 & 2 & 1 \\ -1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \mathbf{u}$$

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

6.15 Is it possible to find a set of b_{ij} and a set of c_{ij} such that the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \\ b_{51} & b_{52} \end{bmatrix} \mathbf{u}$$

$$\mathbf{y} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} \\ c_{21} & c_{22} & c_{23} & c_{24} & c_{25} \\ c_{31} & c_{32} & c_{33} & c_{34} & c_{35} \end{bmatrix}$$

is controllable? Observable?

6.16 Consider the state equation

$$\dot{\mathbf{x}} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \alpha_1 & \beta_1 & 0 & 0 \\ 0 & -\beta_1 & \alpha_1 & 0 & 0 \\ 0 & 0 & 0 & \alpha_2 & \beta_2 \\ 0 & 0 & 0 & -\beta_2 & \alpha_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} b_1 \\ b_{11} \\ b_{12} \\ b_{21} \\ b_{22} \end{bmatrix} \mathbf{u}$$

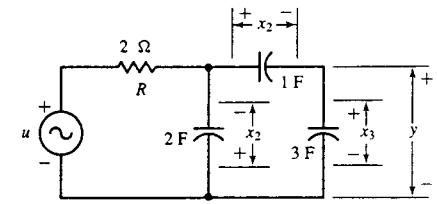
$$\mathbf{y} = [c_1 \quad c_{11} \quad c_{12} \quad c_{21} \quad c_{22}]$$

It is the modal form discussed in (4.28). It has one real eigenvalue and two pairs of complex conjugate eigenvalues. It is assumed that they are distinct. Show that the state equation is controllable if and only if $b_1 \neq 0$; $b_{i1} \neq 0$ or $b_{i2} \neq 0$ for $i = 1, 2$. It is observable if and only if $c_1 \neq 0$; $c_{i1} \neq 0$ or $c_{i2} \neq 0$ for $i = 1, 2$.

6.17 Find two- and three-dimensional state equations to describe the network shown in Fig. 6.12. Discuss their controllability and observability.

6.18 Check controllability and observability of the state equation obtained in Problem 2.19. Can you give a physical interpretation directly from the network?

Figure 6.12



6.19 Consider the continuous-time state equation in Problem 4.2 and its discretized equations in Problem 4.3 with sampling period $T = 1$ and π. Discuss controllability and observability of the discretized equations.

6.20 Check controllability and observability of

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 1 \\ 0 & t \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \qquad y = [0 \quad 1]\mathbf{x}$$

6.21 Check controllability and observability of

$$\dot{\mathbf{x}} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ e^{-t} \end{bmatrix} u \qquad y = [0 \quad e^{-t}]\mathbf{x}$$

6.22 Show that $(\mathbf{A}(t), \mathbf{B}(t))$ is controllable at t_0 if and only if $(-\mathbf{A}'(t), \mathbf{B}'(t))$ is observable at t_0.

6.23 For time-invariant systems, show that $(\mathbf{A}, \mathbf{B})$ is controllable if and only if $(-\mathbf{A}, \mathbf{B})$ is controllable. Is this true for time-varying systems?