BIRKHÄUSER

Panos J. Antsaklis
Anthony N. Michel

A Linear Systems Primer

Birkhäuser
Boston • Basel • Berlin

Panos J. Antsaklis
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
U.S.A.

Anthony N. Michel
Department of Electrical Engineering
University of Notre Dame
Notre Dame, IN 46556
U.S.A.

Cover design by Mary Burgess.

Mathematics Subject Classification (2000): 34A30, 34H05, 93-XX, 93-01, 93Axx, 93A30, 93Bxx, 93B03, 93B05, 93B07, 93B10, 93B11, 93B12, 93B15, 93B17, 93B18, 93B20, 93B25, 93B50, 93B55, 93B60, 93Cxx, 93C05, 93C15, 93C35, 93C55, 93C57, 93C62, 93Dxx, 93D05, 93D15, 93D20, 93D25, 93D30

Library of Congress Control Number: 2007905134

ISBN-13: 978-0-8176-4460-4 e-ISBN-13: 978-0-8176-4661-5

Printed on acid-free paper.

www.birkhauser.com (MP)

To our Families

To
Melinda and our daughter Lily
and to my parents
Dr. Ioannis and Marina Antsaklis
—Panos J. Antsaklis

To
Leone and our children
Mary, Kathy, John,
Tony, and Pat
—Anthony N. Michel

And to our Students

Preface

Brief Description

The purpose of this book is to provide an introduction to system theory with emphasis on control theory. It is intended to be the textbook of a typical one-semester course introduction to systems primarily for first-year graduate students in engineering, but also in mathematics, physics, and the rest of the sciences. Prerequisites for such a course include undergraduate-level differential equations and linear algebra, Laplace transforms, and modeling ideas of, say, electric circuits and simple mechanical systems. These topics are typically covered in the usual undergraduate curricula in engineering and sciences. The goal of this text is to provide a clear understanding of the fundamental concepts of systems and control theory, to highlight appropriately the principal results, and to present material sufficiently broad so that the reader will emerge with a clear picture of the dynamical behavior of linear systems and their advantages and limitations.

Organization and Coverage

This primer covers essential concepts and results in systems and control theory. Since a typical course that uses this book may serve students with different educational experiences, from different disciplines and from different educational systems, the first chapters are intended to build up the understanding of the dynamical behavior of systems as well as provide the necessary mathematical background. Internal and external system descriptions are described in detail, including state variable, impulse response and transfer function, polynomial matrix, and fractional representations. Stability, controllability, observability, and realizations are explained with the emphasis always being on fundamental results. State feedback, state estimation, and eigenvalue assignment are discussed in detail. All stabilizing feedback controllers are also parameterized using polynomial and fractional system representations. The emphasis in this primer is on time-invariant systems, both continuous and discrete time. Although time-varying systems are studied in the first chapter, for a full coverage the reader is encouraged to consult the companion book titled *Linear Systems*[1] that offers detailed descriptions and additional material, including all the proofs of the results presented in this book. In fact, this primer is based on the more complete treatment of *Linear Systems*, which can also serve as a reference for researchers in the field. This primer focuses more on course use of the material, with emphasis on a presentation that is more transparent, without sacrificing rigor, and emphasizes those results that are considered to be fundamental in systems and control and are accepted as important and essential topics of the subject.

Contents

In a typical course on Linear Systems, the depth of coverage will vary depending on the goals set for the course and the background of the students. We typically cover the material in the first three chapters in about six to seven weeks or about half of the semester; we spend about four to five weeks covering Chapters 4–8 on stability, controllability, and realizations; and we spend the remaining time in the course on state feedback, state estimation, and feedback control presented in Chapters 9–10. This book contains over 175 examples and almost 160 exercises. A Solutions Manual is available to course instructors from the publisher. Answers to selected exercises are given at the end of this book.

By the end of Chapter 3, the students should have gained a good understanding of the role of inputs and initial conditions in the response of systems that are linear and time-invariant and are described by state-variable internal descriptions for both continuous- and discrete-time systems; should have brushed up and acquired background in differential and difference equations, matrix algebra, Laplace and z transforms, vector spaces, and linear transformations; should have gained understanding of linearization and the generality and limitations of the linear models used; should have become familiar with eigenvalues, system modes, and stability of an equilibrium; should have an understanding of external descriptions, impulse responses, and transfer functions; and should have learned how sampled data system descriptions are derived.

Depending on the background of the students, in Chapter 1, one may want to define the initial value problem, discuss examples, briefly discuss existence and uniqueness of solutions of differential equations, identify methods to solve linear differential equations, and derive the state transition matrix. Next, in Chapter 2, one may wish to discuss the system response, introduce the impulse response, and relate it to the state-space descriptions for both continuous- and discrete-time cases. In Chapter 3, one may consider to study in detail the response of the systems to inputs and initial conditions. Note that it is

[1] P.J. Antsaklis and A.N. Michel, *Linear Systems*, Birkhäuser, Boston, MA, 2006.

possible to start the coverage of the material with Chapter 3 going back to Chapters 1 and 2 as the need arises.

A convenient way to decide the particular topics from each chapter that need to be covered is by reviewing the Summary and Highlights sections at the end of each chapter.

The Lyapunov stability of an equilibrium and the input/output stability of linear time-invariant systems, along with stability, controllability and observability, are fundamental system properties and are covered in Chapters 4 and 5. Chapter 6 describes useful forms of the state space representations such as the Kalman canonical form and the controller form. They are used in the subsequent chapters to provide insight into the relations between input and output descriptions in Chapter 7. In that chapter the polynomial matrix representation, an alternative internal description, is also introduced. Based on the results of Chapters 5–7, Chapter 8 discusses realizations of transfer functions. Chapter 9 describes state feedback, pole assignment, optimal control, as well as state observers and optimal state estimation. Chapter 10 characterizes all stabilizing controllers and discusses feedback problems using matrix fractional descriptions of the transfer functions.

Depending on the interest and the time constraints, several topics may be omitted completely without loss of continuity. These topics may include, for example, parts of Section 6.4 on controller and observer forms, Section 7.4 on poles and zeros, Section 7.5 on polynomial matrix descriptions, some of the realization algorithms in Section 8.4, sections in Chapter 9 on state feedback and state observers, and all of Chapter 10.

The appendix collects selected results on linear algebra, fields, vector spaces, eigenvectors, the Jordan canonical form, and normed linear spaces, and it addresses numerical analysis issues that arise when computing solutions of equations.

Simulating the behavior of dynamical systems, performing analysis using computational models, and designing systems using digital computers, although not central themes of this book, are certainly encouraged and often required in the examples and in the Exercise sections in each chapter. One could use one of several software packages specifically designed to perform such tasks that come under the label of control systems and signal processing, and work in different operating system environments; or one could also use more general computing languages such as C, which is certainly a more tedious undertaking. Such software packages are readily available commercially and found in many university campuses. In this book we are not endorsing any particular one, but we are encouraging students to make their own informed choices.

Acknowledgments

We are indebted to our students for their feedback and constructive suggestions during the evolution of this book. We are also grateful to colleagues

who provided useful feedback regarding what works best in the classroom in their particular institutions. Special thanks go to Eric Kuehner for his expert preparation of the manuscript. This project would not have been possible without the enthusiastic support of Tom Grasso, Birkhäuser's Computational Sciences and Engineering Editor, who thought that such a companion primer to *Linear Systems* was an excellent idea. We would also like to acknowledge the help of Regina Gorenshteyn, Associate Editor at Birkhäuser.

It was a pleasure writing this book. Our hope is that students enjoy reading it and learn from it. It was written for them.

Notre Dame, IN
Spring 2007

Panos J. Antsaklis
Anthony N. Michel

Contents

Preface . . . vii

1 System Models, Differential Equations, and Initial-Value Problems . . . 1

1.1 Introduction . . . 1

1.2 Preliminaries . . . 6

1.2.1 Notation . . . 7

1.2.2 Continuous Functions . . . 7

1.3 Initial-Value Problems . . . 8

1.3.1 Systems of First-Order Ordinary Differential Equations 9

1.3.2 Classification of Systems of First-Order Ordinary Differential Equations . . . 10

1.3.3 nth-Order Ordinary Differential Equations . . . 11

1.4 Examples of Initial-Value Problems . . . 13

1.5 Solutions of Initial-Value Problems: Existence, Continuation, Uniqueness, and Continuous Dependence on Parameters . . . 17

1.6 Systems of Linear First-Order Ordinary Differential Equations 20

1.6.1 Linearization . . . 21

1.6.2 Examples . . . 24

1.7 Linear Systems: Existence, Uniqueness, Continuation, and Continuity with Respect to Parameters of Solutions . . . 27

1.8 Solutions of Linear State Equations . . . 28

1.9 Summary and Highlights . . . 32

1.10 Notes . . . 33

References . . . 33

Exercises . . . 34

2 An Introduction to State-Space and Input–Output Descriptions of Systems . . . 47

2.1 Introduction . . . 47

2.2 State-Space Description of Continuous-Time Systems . . . 47

2.3 State-Space Description of Discrete-Time Systems 50
2.4 Input–Output Description of Systems 56
2.4.1 External Description of Systems: General Considerations 56
2.4.2 Linear Discrete-Time Systems 60
2.4.3 The Dirac Delta Distribution 65
2.4.4 Linear Continuous-Time Systems 68
2.5 Summary and Highlights 71
2.6 Notes 73
References 73
Exercises 74

3 Response of Continuous- and Discrete-Time Systems 77
3.1 Introduction 77
3.2 Solving $\dot{x} = Ax$ and $\dot{x} = Ax + g(t)$: The State Transition Matrix $\Phi(t, t_0)$ 78
3.2.1 The Fundamental Matrix 78
3.2.2 The State Transition Matrix 82
3.2.3 Nonhomogeneous Equations 84
3.3 The Matrix Exponential e^{At}, Modes, and Asymptotic Behavior of $\dot{x} = Ax$ 85
3.3.1 Properties of e^{At} 85
3.3.2 How to Determine e^{At} 86
3.3.3 Modes, Asymptotic Behavior, and Stability 94
3.4 State Equation and Input–Output Description of Continuous-Time Systems 100
3.4.1 Response of Linear Continuous-Time Systems 100
3.4.2 Transfer Functions 102
3.4.3 Equivalence of State-Space Representations 105
3.5 State Equation and Input–Output Description of Discrete-Time Systems 108
3.5.1 Response of Linear Discrete-Time Systems 108
3.5.2 The Transfer Function and the z-Transform 112
3.5.3 Equivalence of State-Space Representations 115
3.5.4 Sampled-Data Systems 116
3.5.5 Modes, Asymptotic Behavior, and Stability 121
3.6 An Important Comment on Notation 126
3.7 Summary and Highlights 127
3.8 Notes 129
References 130
Exercises 131

4 Stability 141
4.1 Introduction 141
4.2 The Concept of an Equilibrium 142
4.3 Qualitative Characterizations of an Equilibrium 144

4.4 Lyapunov Stability of Linear Systems 148
4.5 The Lyapunov Matrix Equation 153
4.6 Linearization 164
4.7 Input–Output Stability 170
4.8 Discrete-Time Systems 173
4.8.1 Preliminaries 173
4.8.2 Linear Systems 176
4.8.3 The Lyapunov Matrix Equation 179
4.8.4 Linearization 185
4.8.5 Input–Output Stability 186
4.9 Summary and Highlights 188
4.10 Notes 189
References 190
Exercises 191

5 Controllability and Observability: Fundamental Results 195
5.1 Introduction 195
5.2 A Brief Introduction to Reachability and Observability 195
5.2.1 Reachability and Controllability 196
5.2.2 Observability and Constructibility 200
5.2.3 Dual Systems 203
5.3 Reachability and Controllability 204
5.3.1 Continuous-Time Time-Invariant Systems 205
5.3.2 Discrete-Time Systems 213
5.4 Observability and Constructibility 218
5.4.1 Continuous-Time Time-Invariant Systems 219
5.4.2 Discrete-Time Time-Invariant Systems 225
5.5 Summary and Highlights 230
5.6 Notes 232
References 232
Exercises 233

6 Controllability and Observability: Special Forms 237
6.1 Introduction 237
6.2 Standard Forms for Uncontrollable and Unobservable Systems 237
6.2.1 Standard Form for Uncontrollable Systems 238
6.2.2 Standard Form for Unobservable Systems 241
6.2.3 Kalman's Decomposition Theorem 244
6.3 Eigenvalue/Eigenvector Tests for Controllability and Observability 248
6.4 Controller and Observer Forms 250
6.4.1 Controller Forms 251
6.4.2 Observer Forms 263

6.5 Summary and Highlights 269
6.6 Notes 271
References 272
Exercises 272

7 Internal and External Descriptions: Relations and Properties 277
7.1 Introduction 277
7.2 Relations Between State-Space and Input–Output Descriptions 277
7.3 Relations Between Lyapunov and Input–Output Stability 281
7.4 Poles and Zeros 282
7.4.1 Smith and Smith–McMillan Forms 283
7.4.2 Poles 284
7.4.3 Zeros 286
7.4.4 Relations Between Poles, Zeros, and Eigenvalues of A . . 290
7.5 Polynomial Matrix and Matrix Fractional Descriptions of Systems 292
7.5.1 A Brief Introduction to Polynomial and Fractional Descriptions 294
7.5.2 Coprimeness and Common Divisors 298
7.5.3 Controllability, Observability, and Stability 303
7.5.4 Poles and Zeros 304
7.6 Summary and Highlights 306
7.7 Notes 308
References 308
Exercises 309

8 Realization Theory and Algorithms 313
8.1 Introduction 313
8.2 State-Space Realizations of External Descriptions 313
8.2.1 Continuous-Time Systems 314
8.2.2 Discrete-Time Systems 315
8.3 Existence and Minimality of Realizations 316
8.3.1 Existence of Realizations 316
8.3.2 Minimality of Realizations 317
8.3.3 The Order of Minimal Realizations 321
8.3.4 Minimality of Realizations: Discrete-Time Systems 323
8.4 Realization Algorithms 324
8.4.1 Realizations Using Duality 324
8.4.2 Realizations in Controller/Observer Form 326
8.4.3 Realizations with Matrix A Diagonal 339
8.4.4 Realizations Using Singular-Value Decomposition 341
8.5 Polynomial Matrix Realizations 343
8.6 Summary and Highlights 345

8.7 Notes 346
References 346
Exercises 346

9 State Feedback and State Observers 351
9.1 Introduction 351
9.2 Linear State Feedback 352
9.2.1 Continuous-Time Systems 352
9.2.2 Eigenvalue Assignment 355
9.2.3 The Linear Quadratic Regulator (LQR): Continuous-Time Case 369
9.2.4 Input–Output Relations 372
9.2.5 Discrete-Time Systems 376
9.2.6 The Linear Quadratic Regulator (LQR): Discrete-Time Case 377
9.3 Linear State Observers 378
9.3.1 Full-Order Observers: Continuous-Time Systems 378
9.3.2 Reduced-Order Observers: Continuous-Time Systems . . 383
9.3.3 Optimal State Estimation: Continuous-Time Systems . . 385
9.3.4 Full-Order Observers: Discrete-Time Systems 387
9.3.5 Reduced-Order Observers: Discrete-Time Systems 391
9.3.6 Optimal State Estimation: Discrete-Time Systems 391
9.4 Observer-Based Dynamic Controllers 392
9.4.1 State-Space Analysis 393
9.4.2 Transfer Function Analysis 397
9.5 Summary and Highlights 400
9.6 Notes 403
References 404
Exercises 405

10 Feedback Control Systems 411
10.1 Introduction 411
10.2 Interconnected Systems 411
10.2.1 Systems Connected in Parallel and in Series 411
10.2.2 Systems Connected in Feedback Configuration 413
10.3 Parameterization of All Stabilizing Feedback Controllers 422
10.3.1 Stabilizing Feedback Controllers Using Polynomial MFDs 423
10.3.2 Stabilizing Feedback Controllers Using Proper and Stable MFDs 426
10.4 Two Degrees of Freedom Controllers 431
10.4.1 Internal Stability 432
10.4.2 Response Maps 435
10.4.3 Controller Implementations 439
10.4.4 Some Control Problems 445

10.5 Summary and Highlights 447
10.6 Notes 449
References 451
Exercises 452

A Appendix 455
A.1 Vector Spaces 455
A.1.1 Fields 455
A.1.2 Vector Spaces 456
A.2 Linear Independence and Bases 460
A.2.1 Linear Subspaces 460
A.2.2 Linear Independence 461
A.2.3 Linear Independence of Functions of Time 462
A.2.4 Bases 463
A.3 Linear Transformations 464
A.3.1 Linear Equations 465
A.3.2 Representation of Linear Transformations by Matrices . 466
A.3.3 Solving Linear Algebraic Equations 469
A.4 Equivalence and Similarity 471
A.4.1 Change of Bases: Vector Case 471
A.4.2 Change of Bases: Matrix Case 472
A.4.3 Equivalence and Similarity of Matrices 473
A.5 Eigenvalues and Eigenvectors 474
A.5.1 Characteristic Polynomial 474
A.5.2 The Cayley–Hamilton Theorem and Applications 475
A.5.3 Minimal Polynomials 477
A.6 Diagonal and Jordan Canonical Form of Matrices 478
A.7 Normed Linear Spaces 483
A.8 Some Facts from Matrix Algebra 486
A.9 Numerical Considerations 487
A.9.1 Solving Linear Algebraic Equations 488
A.9.2 Singular Values and Singular Value Decomposition 491
A.9.3 Least-Squares Problem 496
A.10 Notes 497
References 497

Solutions to Selected Exercises 499

Index 505

1

System Models, Differential Equations, and Initial-Value Problems

1.1 Introduction

The dynamical behavior of systems can be understood by studying their mathematical descriptions. The flight path of an airplane subject to certain engine thrust, rudder and elevator angles, and particular wind conditions, or the behavior of an automobile on cruise control when climbing a certain hill, can be predicted using mathematical descriptions of the pertinent behavior. Mathematical equations, typically differential or difference equations, are used to describe the behavior of processes and to predict their responses to certain inputs. Although computer simulation is an excellent tool for verifying predicted behavior, and thus for enhancing our understanding of processes, it is certainly not an adequate substitute for generating the information captured in a mathematical model, when such a model is available.

This chapter develops mathematical descriptions for linear continuous-time and linear discrete-time finite-dimensional systems. Since such systems are frequently the result of a linearization process of nonlinear systems, or the result of the modeling process of physical systems in which the nonlinear effects have been suppressed or neglected, the origins of these linear systems are frequently nonlinear systems. For this reason, here and in Chapter 4, when we deal with certain qualitative aspects (such as existence, uniqueness, continuation, and continuity with respect to parameters of solutions of system equations, stability of an equilibrium, and so forth), we consider linear as well as nonlinear system models, although the remainder of the book deals exclusively with linear systems.

In this chapter, mathematical models and classification of models are discussed in the remainder of this Introduction, Section 1.1. In Section 1.2, we provide some of the notation used and recall certain facts concerning continuous functions. In Section 1.3 we present the initial-value problem and we give several specific examples in Section 1.4. In Section 1.5 we present results that ensure the existence, continuation, and uniqueness of solutions of initial-value problems and results that ensure that the solutions of inital-value problems depend continuously on initial conditions and system parameters. In this section we also present the Method of Successive Approximations to determine solutions of intial-value problems. The results in Section 1.5 pertain to differential equations that in general are nonlinear. In Section 1.6 we address linearization of such equations and we provide several specific examples.

We utilize the results of Section 1.5 to establish in Section 1.7 conditions for the existence, uniqueness, continuation, and continuity with respect to initial conditions and parameters of solutions of initial-value problems determined by *linear ordinary* differential equations.

In Section 1.8 we determine the solutions of linear ordinary differential equations and introduce for the first time the notions of state and state transition matrix. We also present the variations of constants formula for solving linear nonhomogeneous ordinary differential equations, and we introduce the notions of homogeneous and particular solutions.

Summarizing, the purpose of Sections 1.3 to 1.8 is to provide material dealing with ordinary differential equations and initial-value problems that is essential in the study of continuous-time finite-dimensional systems. This material will enable us to introduce the state-space equations representation of continuous-time finite-dimensional systems. This introduction will be accomplished in the next chapter.

Physical Processes, Models, and Mathematical Descriptions

A systematic study of (physical) *phenomena* usually begins with a *modeling process*. Examples of models include diagrams of electric circuits consisting of interconnections of resistors, inductors, capacitors, transistors, diodes, voltage or current sources, and so on; mechanical circuits consisting of interconnections of point masses, springs, viscous dampers (dashpots), applied forces, and so on; verbal characterizations of economic and societal systems; among others. Next, appropriate *laws* or *principles* are invoked to generate *equations* that describe the models (e.g., Kirchhoff's current and voltage laws, Newton's laws, conservation laws, and so forth). When using an expression such as "we consider a *system* described by ordinary differential equations," we will have in mind a phenomenon described by an appropriate set of ordinary differential equations (not the description of the physical phenomenon itself).

A physical process (physical system) will typically give rise to several different models, depending on what questions are being asked. For instance, in the study of the voltage-current characteristics of a transistor (the physical process), one may utilize a circuit (the model) that is valid at low frequencies or a circuit (a second model) that is valid at high frequencies; alternatively, if semiconductor impurities are of interest, a third model, quite different from the preceding two, is appropriate.

Over the centuries, a great deal of progress has been made in developing mathematical descriptions of physical phenomena (using models of such phenomena). In doing so, we have invoked laws (or principles) of physics, chemistry, biology, economics, and so on, to derive mathematical expressions (usually equations) that characterize the evolution (in time) of the variables of interest. The availability of such mathematical descriptions enables us to make use of the vast resources offered by the many areas of applied and pure mathematics to conduct qualitative and quantitative studies of the behavior of processes. *A given model of a physical process may give rise to several different mathematical descriptions.* For example, when applying Kirchhoff's voltage and current laws to the low-frequency transistor model mentioned earlier, one can derive a set of differential and algebraic equations, a set consisting of only differential equations, or a set of integro-differential equations, and so forth. *This process of mathematical modeling, "from a physical phenomenon to a model to a mathematical description," is essential in science and engineering.* To capture phenomena of interest accurately and in tractable mathematical form is a demanding task, as can be imagined, and requires a thorough understanding of the physical process involved. For this reason, the mathematical description of complex electrical systems, such as power systems, is typically accomplished by electrical engineers, the equations of flight dynamics of an aircraft are derived by aeronautical engineers, the equations of chemical processes are arrived at by chemists and chemical engineers, and the equations that characterize the behavior of economic systems are provided by economists. In most nontrivial cases, this type of modeling process is close to an art form since *a good mathematical description must be detailed enough to accurately describe the phenomena of interest and at the same time simple enough to be amenable to analysis.* Depending on the applications on hand, a given mathematical description of a process may be further simplified before it is used in analysis and especially in design procedures. For example, using the finite element method, one can derive a set of first-order differential equations that describe the motion of a space antenna. Typically, such mathematical descriptions contain hundreds of differential equations. Whereas all these equations are quite useful in simulating the motion of the antenna, a lower order model is more suitable for the control design that, for example, may aim to counteract the effects of certain disturbances. Simpler mathematical models are required mainly because of our inability to deal effectively with hundreds of variables and their interactions. In such simplified mathematical descriptions, only those variables (and their interactions) that have significant effects on the phenomena of interest are included.

A point that cannot be overemphasized is that *the mathematical descriptions we will encounter characterize processes only approximately.* Most often, this is the case because the complexity of physical systems defies exact mathematical formulation. In many other cases, however, it is our own choice that a mathematical description of a given process approximate the actual phenomena by only a certain desired degree of accuracy. As discussed earlier, this is done in the interest of mathematical simplicity. For example, in the description of RLC circuits, one could use nonlinear differential equations that take into consideration parasitic effects in the capacitors; however, most often it suffices to use linear ordinary differential equations with constant coefficients to describe the voltage-current relations of such circuits, since typically such a description provides an adequate approximation and since it is much easier to work with linear rather than nonlinear differential equations.

In this book it will generally be assumed that the mathematical description of a system in question is given. In other words, we assume that the modeling of the process in question has taken place and that equations describing the process are given. Our main objective will be to present a theory of an important class of systems—finite-dimensional linear systems—by studying the equations representing such systems.

Classification of Systems

For our purposes, a comprehensive classification of systems is not particularly illuminating. However, an enumeration of the more common classes of systems encountered in engineering and science may be quite useful, if for no other reason than to show that the classes of systems considered in this book, although very important, are quite specialized.

As pointed out earlier, the particular set of equations describing a given system will in general depend on the effects one wishes to capture. Thus, one can speak of *lumped parameter* or *finite-dimensional systems* and *distributed parameter* or *infinite-dimensional systems*; *continuous-time* and *discrete-time systems*; *linear* and *nonlinear systems*; *time-varying* and *time-invariant systems*; *deterministic* and *stochastic systems*; appropriate combinations of the above, called *hybrid systems*; and perhaps others.

The appropriate mathematical settings for finite-dimensional systems are finite-dimensional vector spaces, and for infinite-dimensional systems they are most often infinite-dimensional linear spaces. Continuous-time finite-dimensional systems are usually described by ordinary differential equations or certain kinds of integral equations, whereas discrete-time finite-dimensional systems are usually characterized by ordinary difference equations or discrete-time counterparts to those integral equations. Equations used to describe infinite-dimensional systems include partial differential equations, Volterra integro-differential equations, functional differential equations, and so forth. Hybrid system descriptions involve two or more different types of equations. Nondeterministic systems are described by stochastic counterparts to those equations (e.g., Ito differential equations).

In a broader context, not addressed in this book, most of the systems described by the equations enumerated generate *dynamical systems*. It has become customary in the engineering literature to use the term “dynamical system” rather loosely, and it has even been applied to cases where the original definition does not exactly fit. (For a discussion of general dynamical systems, refer, e.g., to Michel et al [5].) We will address in this book dynamical systems determined by ordinary differential equations or ordinary difference equations, considered next.

Finite-Dimensional Systems

The dynamical systems we will be concerned with are *continuous-time* and *discrete-time finite-dimensional systems*—primarily *linear systems*. However, since such systems are frequently a consequence of a linearization process, it is important when dealing with fundamental qualitative issues that we have an understanding of the origins of such linear systems. In particular, when dealing with questions of existence and uniqueness of solutions of the equations describing a class of systems, and with stability properties of such systems, we may consider nonlinear models as well.

Continuous-time finite-dimensional dynamical systems that we will consider are described by equations of the form

$$\dot{x}_i = f_i(t, x_1, \dots, x_n, u_1, \dots, u_m), \qquad i = 1, \dots, n, \tag{1.1a}$$

$$y_i = g_i(t, x_1, \dots, x_n, u_1, \dots, u_m), \qquad i = 1, \dots, p, \tag{1.1b}$$

where u_i, $i = 1, \dots, m$, denote *inputs* or *stimuli*; y_i, $i = 1, \dots, p$, denote *outputs* or *responses*; x_i, $i = 1, \dots, n$, denote *state variables*; t denotes *time*; $\dot{x}_i$ denotes the time derivative of x_i; f_i, $i = 1, \dots, n$, are real-valued functions of $1 + n + m$ real variables; and g_i, $i = 1, \dots, p$, are real-valued functions of $1 + n + m$ real variables. A complete description of such systems will usually also require a set of *initial conditions* $x_i(t_0) = x_{i0}$, $i = 1, \dots, n$, where t_0 denotes *initial time*. We will elaborate later on restrictions that need to be imposed on the f_i, g_i, and u_i and on the origins of the term "state variables."

Equations (1.1a) and (1.1b) can be represented in vector form as

$$\dot{x} = f(t, x, u), \tag{1.2a}$$

$$y = g(t, x, u), \tag{1.2b}$$

where x is the *state vector* with components x_i, u is the *input vector* with components u_i, y is the *output vector* with components y_i, and f and g are vector-valued functions with components f_i and g_i, respectively. We call (1.2a) a *state equation* and (1.2b) an *output equation*.

Important special cases of (1.2a) and (1.2b) are the *linear time-varying state equation and output equation* given by

$$\dot{x} = A(t)x + B(t)u, \tag{1.3a}$$

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

where $A, B, C,$ and D are real $n \times n, n \times m, p \times n,$ and $p \times m$ matrices, respectively, whose elements are time-varying. Restrictions on these matrices will be provided later.

Linear time-invariant state and output equations given by

$$\dot{x} = Ax + Bu, \tag{1.4a}$$

$$y = Cx + Du \tag{1.4b}$$

constitute important special cases of (1.3a) and (1.3b), respectively.

Equations (1.3) and (1.4) may arise in the modeling process, or they may be a consequence of *linearization* of (1.1).

Discrete-time finite-dimensional dynamical systems are described by equations of the form

$$x_i(k+1) = f_i(k, x_1(k), \dots, x_n(k), u_1(k), \dots, u_m(k)) \quad i = 1, \dots, n, \tag{1.5a}$$
$$y_i(k) = g_i(k, x_1(k), \dots, x_n(k), u_1(k), \dots, u_m(k)) \quad i = 1, \dots, p, \tag{1.5b}$$

or in vector form,

$$x(k+1) = f(k, x(k), u(k)), \tag{1.6a}$$
$$y(k) = g(k, x(k), u(k)), \tag{1.6b}$$

where k is an integer that denotes *discrete time* and all other symbols are defined as before. A complete description of such systems involves a set of *initial conditions* $x(k_0) = x_{k_0}$, where k_0 denotes *initial time*. The corresponding linear time-varying and time-invariant state and output equations are given by

$$x(k+1) = A(k)x(k) + B(k)u(k), \tag{1.7a}$$
$$y(k) = C(k)x(k) + D(k)u(k) \tag{1.7b}$$

and

$$x(k+1) = Ax(k) + Bu(k), \tag{1.8a}$$
$$y(k) = Cx(k) + Du(k), \tag{1.8b}$$

respectively, where all symbols in (1.7) and (1.8) are defined as in (1.3) and (1.4), respectively.

This type of system characterization is called *state-space description* or *state-variable description* or *internal description* of finite-dimensional systems. Another way of describing continuous-time and discrete-time finite-dimensional dynamical systems involves operators that establish a relationship between the system inputs and outputs. Such characterization is called *input–output description* or *external description* of a system. In Chapter 2, we will address both the state-variable description and the input–output description of finite-dimensional systems. Before we can do this, however, we will require some background material concerning ordinary differential equations.

1.2 Preliminaries

We will employ a consistent notation and use certain facts from the calculus, analysis, and linear algebra. We will summarize this type of material, as needed, in various sections. This is the first such section.

1.2.1 Notation

Let V and W be *sets*. Then $V \cup W, V \cap W, V - W$, and $V \times W$ denote the *union, intersection, difference*, and *Cartesian product* of V and W, respectively. If V is a *subset* of W, we write $V \subset W$; if x is an *element* of V, we write $x \in V$; and if x is not an element of V, we write $x \notin V$. We let $V', \partial V, \bar{V}$, and int V denote the *complement, boundary, closure*, and *interior* of V, respectively.

Let ϕ denote the *empty set*, R the *real numbers*, $R^+ = \{x \in R : x \geq 0\}$ (i.e., R^+ denotes the set of nonnegative real numbers), Z the *integers*, and $Z^+ = \{x \in Z : x \geq 0\}$.

We will let $J \subset R$ denote open, closed, or half-open *intervals*. Thus, for $a, b \in R$, $a \leq b$, J may be of the form $J = (a, b) = \{x \in R : a < x < b\}$, $J = [a, b] = \{x \in R : a \leq x \leq b\}$, $J = [a, b) = \{x \in R : a \leq x < b\}$, or $J = (a, b] = \{x \in R : a < x \leq b\}$.

Let R^n denote the real n-space. If $x \in R^n$, then

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

and $x^T = (x_1, \ldots, x_n)$ denotes the *transpose* of the vector x. Also, let $R^{m \times n}$ denote the set of $m \times n$ real matrices. If $A \in R^{m \times n}$, then

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

and $A^T = [a_{ji}] \in R^{n \times m}$ denotes the *transpose* of the matrix A.

Similarly, we let C^n denote the set of n-vectors with complex components and $C^{m \times n}$ denote the set of $m \times n$ matrices with complex elements.

Let $f : V \to W$ denote a *mapping* or *function* from a set V into a set W, and denote by $D(f)$ and $R(f)$ the *domain* and the range of f, respectively. Also, let $f^{-1} : R(f) \to D(f)$, if it exists, denote the *inverse* of f.

1.2.2 Continuous Functions

First, let $J \subset R$ denote an open interval and consider a function $f : J \to R$. Recall that f is said to be *continuous at the point* $t_0 \in J$ if $\lim_{t \to t_0} f(t) = f(t_0)$ exists; i.e., if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $|f(t) - f(t_0)| < \epsilon$ whenever $|t - t_0| < \delta$ and $t \in J$. The function f is said to be *continuous on* J, or simply *continuous*, if it is continuous at each point in J.

In the above definition, δ depends on the choice of t_0 and ϵ; i.e., $\delta = \delta(\epsilon, t_0)$. If at *each* $t_0 \in J$ it is true that there is a $\delta > 0$, independent of t_0 [i.e., $\delta = \delta(\epsilon)$], such that $|f(t) - f(t_0)| < \epsilon$ whenever $|t - t_0| < \delta$ and $t \in J$, then f is said to be *uniformly continuous* (on J).

Let

$$C(J,R) \triangleq \{f : J \to R \mid f \text{ is continuous on } J\}.$$

Now suppose that J contains one or both endpoints. Then continuity is interpreted as being one-sided at these points. For example, if $J = [a,b]$, then $f \in C(J,R)$ will mean that $f \in C((a,b),R)$ and that $\lim_{t\to a^+} f(t) = f(a)$ and $\lim_{t\to b^-} f(t) = f(b)$ exist.

With k any positive integer, and with J an open interval, we will use the notation

$$\begin{aligned} C^k(J,R) \triangleq \{&f : J \to R \mid \text{the derivative } f^{(j)} \text{ exists on } J \text{ and} \\ &f^{(j)} \in C(J,R) \text{ for } j = 0,1,\ldots,k, \text{ where } f^{(0)} \triangleq f\} \end{aligned}$$

and we will call f in this case a C^k*-function*. Also, we will call f a *piecewise* C^k*-function* if $f \in C^{k-1}(J,R)$ and $f^{(k-1)}$ has continuous derivatives for all $t \in J$, with the possible exception of a finite set of points where $f^{(k)}$ may have jump discontinuities. As before, when J contains one or both endpoints, then the existence and continuity of derivatives is one-sided at these points.

For any subset D of the n-space R^n with nonempty interior, we can define $C(D,R)$ and $C^k(D,R)$ in a similar manner as before. Thus, $f \in C(D,R)$ indicates that at every point $x_0 = (x_{10},\ldots,x_{n0})^T \in D, \lim_{x\to x_0} f(x) = f(x_0)$ exists, or equivalently, at every $x_0 \in D$ it is true that for every $\epsilon > 0$ there exists a $\delta = \delta(\epsilon, x_0) > 0$ such that $|f(x) - f(x_0)| < \epsilon$ whenever $|x_1 - x_{10}| + \cdots + |x_n - x_{n0}| < \delta$ and $x \in D$. Also, we define $C^k(D,R)$ as

$$\begin{aligned} C^k(D,R) \triangleq \Big\{&f : D \to R \,\Big|\, \frac{\partial^j f}{\partial x_1^{i_1} \ldots \partial x_n^{i_n}} \in C(D,R), \quad i_1 + \cdots + i_n = j, \\ &j = 1,\ldots,k, \text{ and } f \in C(D,R)\Big\} \end{aligned}$$

(i.e., $i_1,\ldots,i_n$ take on all possible positive integer values such that their sum is j). When D contains its boundary (or part of its boundary), then the continuity of f and the existence and continuity of partial derivatives of f, $\frac{\partial^j f}{\partial x_1^{i_1} \ldots \partial x_n^{i_n}}$, $i_1 + \cdots + i_n = j$, $j = 1,\ldots,k$, will have to be interpreted in the appropriate way at the boundary points.

Recall that if $K \subset R^n$, $K \neq \phi$, and K is *compact* (i.e., K is closed and bounded), and if $f \in C(K,R)$, then f is uniformly continuous (on K) and f attains its maximum and minimum on K.

Finally, let D be a subset of R^n with nonempty interior and let $f : D \to R^m$. Then $f = (f_1,\ldots,f_m)^T$ where $f_i : D \to R$, $i = 1,\ldots,m$. We say that $f \in C(D,R^m)$ if $f_i \in C(D,R)$, $i = 1,\ldots,m$, and that for some positive integer k, $f \in C^k(D,R^m)$ if $f_i \in C^k(D,R)$, $i = 1,\ldots,m$.

1.3 Initial-Value Problems

In this section we make precise the meaning of several concepts that arise in the study of continuous-time finite-dimensional dynamical systems.

1.3.1 Systems of First-Order Ordinary Differential Equations

Let $D \subset R^{n+1}$ denote a *domain*, i.e., an open, nonempty, and connected subset of R^{n+1}. We call R^{n+1} the (t, x)-*space*, and we denote elements of R^{n+1} by (t, x) and elements of R^n by $x = (x_1, \ldots, x_n)^T$. Next, we consider the functions $f_i \in C(D, R)$, $i = 1, \ldots, n$, and if x_i is a function of t, we let $x_i^{(n)} = \frac{d^n x_i}{dt^n}$ denote the nth derivative of x_i with respect to t (provided that it exists). In particular, when $n = 1$, we usually write

$$x_i^{(1)} = \dot{x}_i = \frac{dx_i}{dt}.$$

We call the system of equations given by

$$\dot{x}_i = f_i(t, x_1, \ldots, x_n), \quad i = 1, \ldots, n, \tag{1.9}$$

a *system of n first-order ordinary differential equations*. By a *solution* of the system of equations (1.9), we shall mean n continuously differentiable functions $\phi_1, \ldots, \phi_n$ defined on an interval $J = (a, b)$ [i.e., $\phi \in C^1(J, R^n)$] such that $(t, \phi_1(t), \ldots, \phi_n(t)) \in D$ for all $t \in J$ and such that

$$\dot{\phi}_i(t) = f_i(t, \phi_1(t), \ldots, \phi_n(t)), \quad i = 1, \ldots, n,$$

for all $t \in J$.

Next, we let $(t_0, x_{10}, \ldots, x_{n0}) \in D$. Then the *initial-value problem* associated with (1.9) is given by

$$\begin{aligned} \dot{x}_i &= f_i(t, x_1, \ldots, x_n), \quad && i = 1, \ldots, n, \\ x_i(t_0) &= x_{i0}, && i = 1, \ldots, n. \end{aligned} \tag{1.10}$$

A set of functions $\{\phi_1, \ldots, \phi_n\}$ is a *solution* of the initial-value problem (1.10) if $\{\phi_1, \ldots, \phi_n\}$ is a solution of (1.9) on some interval J containing t_0 and if $(\phi_1(t_0), \ldots, \phi_n(t_0)) = (x_{10}, \ldots, x_{n0})$.

In Figure 1.1 the solution of a hypothetical initial-value problem is depicted graphically when $n = 1$. Note that $\dot{\phi}(\tau) = f(\tau, \tilde{x}) = \tan \alpha$, where α is the slope of the line L that is tangent to the plot of the curve $\phi(t)$ vs. t, at the point $(\tau, \tilde{x})$.

In dealing with systems of equations, we will utilize the vector notation $x = (x_1, \ldots, x_n)^T$, $x_0 = (x_{10}, \ldots, x_{n0})^T$, $\phi = (\phi_1, \ldots, \phi_n)^T$, $f(t, x) = (f_1(t, x_1, \ldots, x_n), \ldots, f_n(t, x_1, \ldots, x_n))^T = (f_1(t, x), \ldots, f_n(t, x))^T$, $\dot{x} = (\dot{x}_1, \ldots, \dot{x}_n)^T$, and $\int_{t_0}^t f(s, \phi(s))ds = [\int_{t_0}^t f_1(s, \phi(s))ds, \ldots, \int_{t_0}^t f_n(s, \phi(s))ds]^T$.

With the above notation we can express the system of first-order ordinary differential equations (1.9) by

$$\dot{x} = f(t, x) \tag{1.11}$$

and the initial-value problem (1.10) by

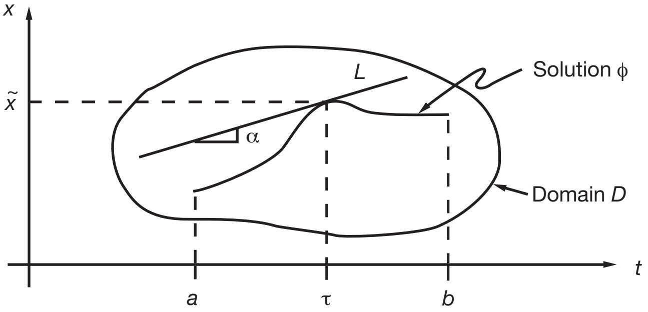


Figure 1.1. Solution of an initial-value problem when $n = 1$

$$\dot{x} = f(t, x), \quad x(t_0) = x_0. \tag{1.12}$$

We leave it to the reader to prove that the initial-value problem (1.12) can be equivalently expressed by the *integral equation*

$$\phi(t) = x_0 + \int_{t_0}^{t} f(s, \phi(s))ds, \tag{1.13}$$

where ϕ denotes a solution of (1.12).

1.3.2 Classification of Systems of First-Order Ordinary Differential Equations

Systems of first-order ordinary differential equations have been classified in many ways. We enumerate here some of the more important cases.

If in (1.11), $f(t, x) \equiv f(x)$ for all $(t, x) \in D$, then

$$\dot{x} = f(x). \tag{1.14}$$

We call (1.14) an *autonomous system* of first-order ordinary differential equations.

If $(t + T, x) \in D$ whenever $(t, x) \in D$ and if $f(t, x) = f(t + T, x)$ for all $(t, x) \in D$, then (1.11) assumes the form

$$\dot{x} = f(t, x) = f(t + T, x). \tag{1.15}$$

We call such an equation a *periodic system* of first-order differential equations with *period* T. The smallest $T > 0$ for which (1.15) is true is called the *least period* of this system of equations.

When in (1.11), $f(t, x) = A(t)x$, where $A(t) = [a_{ij}(t)]$ is a real $n \times n$ matrix with elements a_{ij} that are defined and at least piecewise continuous on a t-interval J, then we have

$$\dot{x} = A(t)x \tag{1.16}$$

and refer to (1.16) as a *linear homogeneous system* of first-order ordinary differential equations.

If for (1.16), $A(t)$ is defined for all real t, and if there is a $T > 0$ such that $A(t) = A(t+T)$ for all t, then we have

$$\dot{x} = A(t)x = A(t+T)x. \tag{1.17}$$

This system is called a *linear periodic system* of first-order ordinary differential equations.

Next, if in (1.11), $f(t,x) = A(t)x + g(t)$, where $A(t)$ is as defined in (1.16), and $g(t) = [g_1(t), \ldots, g_n(t)]^T$ is a real n-vector with elements g_i that are defined and at least piecewise continuous on a t-interval J, then we have

$$\dot{x} = A(t)x + g(t). \tag{1.18}$$

In this case we speak of a *linear nonhomogeneous system* of first-order ordinary differential equations.

Finally, if in (1.11), $f(t,x) = Ax$, where $A = [a_{ij}] \in R^{n\times n}$, then we have

$$\dot{x} = Ax. \tag{1.19}$$

This type of system is called a *linear, autonomous, homogeneous* system of first-order ordinary differential equations.

1.3.3 nth-Order Ordinary Differential Equations

Thus far we have been concerned with systems of first-order ordinary differential equations. It is also possible to characterize initial-value problems by means of nth-order ordinary differential equations. To this end we let h be a real function that is defined and continuous on a domain D of the real $(t, y, \ldots, y_n)$-space [i.e., $D \subset R^{n+1}$, D is a domain, and $h \in C(D, R)$]. Then

$$y^{(n)} = h(t, y, y^{(1)}, \ldots, y^{(n-1)}) \tag{1.20}$$

is an nth-*order ordinary differential equation.*

A *solution* of (1.20) is a function $\phi \in C^n(J, R)$ that satisfies $(t, \phi(t), \phi^{(1)}(t), \ldots, \phi^{(n-1)}(t)) \in D$ for all $t \in J$ and

$$\phi^{(n)}(t) = h(t, \phi(t), \phi^{(1)}(t), \ldots, \phi^{(n-1)}(t))$$

for all $t \in J$, where $J = (a, b)$ is a t-interval.

Now for a given $(t_0, x_{10}, \ldots, x_{n0}) \in D$, the *initial -value problem* for (1.20) is

$$\begin{aligned} y^{(n)} &= h(t, y, y^{(1)}, \ldots, y^{(n-1)}), \\ y(t_0) &= x_{10}, \ldots, y^{(n-1)}(t_0) = x_{n0}. \end{aligned} \tag{1.21}$$

A function ϕ is a *solution* of (1.21) if ϕ is a solution of (1.20) on some interval containing t_0 and if $\phi(t_0) = x_{10}, \dots, \phi^{(n-1)}(t_0) = x_{n0}$.

As in the case of systems of first-order ordinary differential equations, we can point to several important special cases. Specifically, we consider equations of the form

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y^{(1)} + a_0(t)y = g(t), \tag{1.22}$$

where $a_i \in C(J, R)$, $i = 0, 1, \dots, n-1$, and $g \in C(J, R)$. We refer to (1.22) as a *linear nonhomogeneous ordinary differential equation of order* n.

If in (1.22) we let $g(t) \equiv 0$, then

$$y^{(n)} + a_{n-1}(t)y^{(n-1)} + \cdots + a_1(t)y^{(1)} + a_0(t)y = 0. \tag{1.23}$$

We call (1.23) a *linear homogeneous ordinary differential equation of order* n.

If in (1.23) we have $a_i(t) \equiv a_i$, $i = 0, 1, \dots, n-1$, then

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y^{(1)} + a_0y = 0, \tag{1.24}$$

and we call (1.24) a *linear, autonomous, homogeneous ordinary differential equation of order* n.

As in the case of systems of first-order ordinary differential equations, we can define *periodic* and *linear periodic ordinary differential equations of order* n in the obvious way.

It turns out that the theory of nth-order ordinary differential equations can be reduced to the theory of a system of n first-order ordinary differential equations. To demonstrate this, we let $y = x_1, y^{(1)} = x_2, \dots, y^{(n-1)} = x_n$ in (1.21). We now obtain the system of first-order ordinary differential equations

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 \\ &\vdots \\ \dot{x}_n &= h(t, x_1, \dots, x_n) \end{aligned} \tag{1.25}$$

that is defined for all $(t, x_1, \dots, x_n) \in D$. Assume that $\phi = (\phi_1, \dots, \phi_n)^T$ is a solution of (1.25) on an interval J. Since $\phi_2 = \dot{\phi}_1, \phi_3 = \dot{\phi}_2, \dots, \phi_n = \phi_1^{(n-1)}$, and since

$$\begin{aligned} h(t, \phi_1(t), \dots, \phi_n(t)) &= h(t, \phi_1(t), \phi_1^{(1)}(t), \dots, \phi_1^{(n-1)}(t)) \\ &= \phi_1^{(n)}(t), \end{aligned}$$

it follows that the first component ϕ_1 of the vector ϕ is a solution of (1.20) on the interval J. Conversely, if ϕ_1 is a solution of (1.20) on J, then the vector $(\phi, \phi^{(1)}, \dots, \phi^{(n-1)})^T$ is clearly a solution of (1.25). Moreover, if $\phi_1(t_0) = x_{10}, \dots, \phi_1^{(n-1)}(t_0) = x_{n0}$, then the vector ϕ satisfies $\phi(t_0) = x_0 = (x_{10}, \dots, x_{n0})^T$.

1.4 Examples of Initial-Value Problems

We now give several specific examples of initial-value problems.

Example 1.1. The mechanical system of Figure 1.2 consists of two point masses M_1 and M_2 that are acted upon by viscous damping forces (determined by viscous damping constants B, B_1, and B_2), spring forces (specified by the spring constants K, K_1, and K_2), and external forces f_1 and f_2. The initial displacements of M_1 and M_2 at $t_0 = 0$ are given by $y_1(0)$ and $y_2(0)$, respectively, and their initial velocities are given by $\dot{y}_1(0)$ and $\dot{y}_2(0)$. The arrows in Figure 1.2 indicate positive directions of displacement for M_1 and M_2.

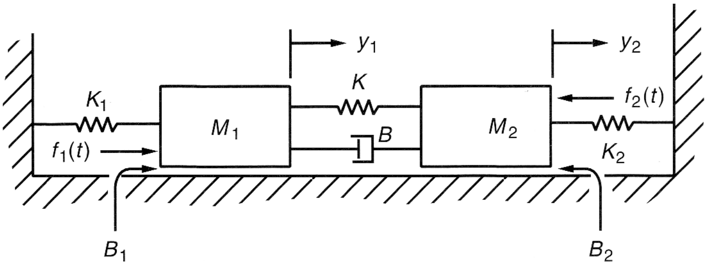


Figure 1.2. An example of a mechanical circuit

Newton's second law yields the following coupled second-order ordinary differential equations that describe the motions of the masses in Figure 1.2 (letting $y^{(2)} = d^2y/dt^2 = \ddot{y}$),

$$\begin{aligned} M_1\ddot{y}_1 + (B + B_1)\dot{y}_1 + (K + K_1)y_1 - B\dot{y}_2 - Ky_2 &= f_1(t) \\ M_2\ddot{y}_2 + (B + B_2)\dot{y}_2 + (K + K_2)y_2 - B_1\dot{y}_1 - Ky_1 &= -f_2(t) \end{aligned} \tag{1.26}$$

with initial data $y_1(0), y_2(0), \dot{y}_1(0)$, and $\dot{y}_2(0)$.

Letting $x_1 = y_1, x_2 = \dot{y}_1, x_3 = y_2$, and $x_4 = \dot{y}_2$, we can express (1.26) equivalently by the system of first-order ordinary differential equations

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{-(K_1+K)}{M_1} & \frac{-(B_1+B)}{M_1} & \frac{K}{M_1} & \frac{B}{M_1} \\ 0 & 0 & 0 & 1 \\ \frac{K}{M_2} & \frac{B}{M_2} & \frac{-(K+K_2)}{M_2} & \frac{-(B+B_2)}{M_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \\ &\quad + \begin{bmatrix} 0 \\ \frac{1}{M_1} f_1(t) \\ 0 \\ \frac{-1}{M_2} f_2(t) \end{bmatrix} \end{aligned} \tag{1.27}$$

with initial data given by $x(0) = (x_1(0),\, x_2(0),\, x_3(0),\, x_4(0))^T$.

Example 1.2. Using the node voltages v_1, v_2, and v_3 and applying Kirchhoff's current law, we can describe the behavior of the electric circuit given in Figure 1.3 by the system of first-order ordinary differential equations

$$\begin{bmatrix} \dot v_1 \\ \dot v_2 \\ \dot v_3 \end{bmatrix} = \begin{bmatrix} -\frac{1}{C_1}\left(\frac{1}{R_1}+\frac{1}{R_2}\right) & \frac{1}{R_2C_1} & 0 \\ -\frac{1}{C_1}\left(\frac{1}{R_1}+\frac{1}{R_2}\right) & -\left(\frac{R_2}{L}-\frac{1}{R_2C_1}\right) & \frac{R_2}{L} \\ \frac{1}{R_2C_2} & -\frac{1}{R_2C_2} & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} + \begin{bmatrix} \frac{v}{R_1C_1} \\ \frac{v}{R_1C_1} \\ 0 \end{bmatrix}. \quad (1.28)$$

To complete the description of this circuit, we specify the initial data at $t_0 = 0$, given by $v_1(0), v_2(0)$, and $v_3(0)$.

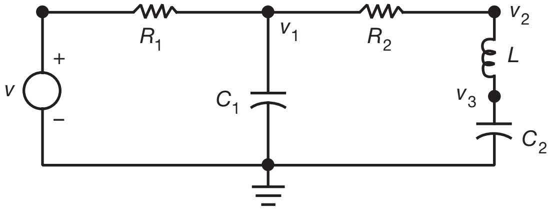


Figure 1.3. An example of an electric circuit

Example 1.3. Figure 1.4 represents a simplified model of an armature voltage-controlled dc servomotor consisting of a stationary field and a rotating armature and load. We assume that all effects of the field are negligible in the description of this system. The various parameters and variables in Figure 1.4 are e_a = externally applied armature voltage, i_a = armature current, R_a = resistance of the armature winding, L_a = armature winding inductance, e_m = back-emf voltage induced by the rotating armature winding, B = viscous damping due to bearing friction, J = moment of inertia of the armature and load, and θ = shaft position. The back-emf voltage (with the polarity as shown) is given by

$$e_m = K_\theta \dot\theta, \quad (1.29)$$

where $K_\theta > 0$ is a constant, and the torque T generated by the motor is given by

$$T = K_T i_a. \quad (1.30)$$

Application of Newton's second law and Kirchhoff's voltage law yields

$$J\ddot\theta + B\dot\theta = T(t) \quad (1.31)$$

and

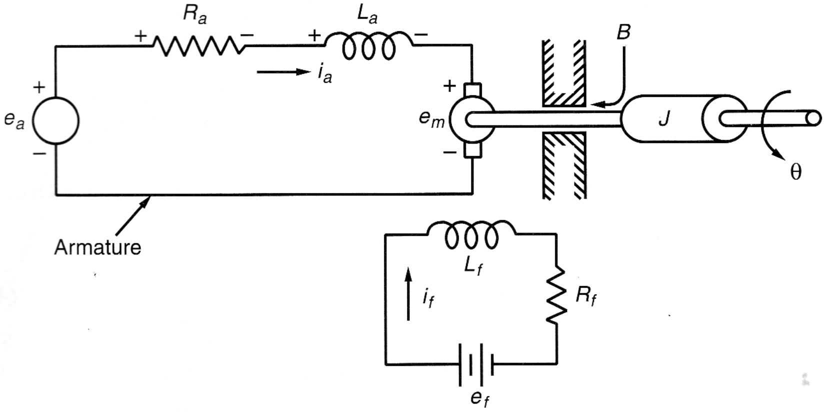


Figure 1.4. An example of an electro-mechanical system circuit

$$L_a \frac{di_a}{dt} + R_a i_a + e_m = e_a. \tag{1.32}$$

Combining (1.29) to (1.32) and letting $x_1 = \theta, x_2 = \dot{\theta}$, and $x_3 = i_a$ yields the system of first-order ordinary differential equations

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -B/J & K_T/J \\ 0 & -K_\theta/L_a & -R_a/L_a \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ e_a/L_a \end{bmatrix}. \tag{1.33}$$

A suitable set of initial data for (1.33) is given by $t_0 = 0$ and $(x_1(0), x_2(0), x_3(0))^T = (\theta(0), \dot{\theta}(0), i_a(0))^T$.

Example 1.4. A much studied ordinary differential equation is given by

$$\ddot{x} + f(x)\dot{x} + g(x) = 0, \tag{1.34}$$

where $f \in C^1(R, R)$ and $g \in C^1(R, R)$.

When $f(x) \geq 0$ for all $x \in R$ and $xg(x) > 0$ for all $x \neq 0$, then (1.34) is called the *Lienard Equation*. This equation can be used to represent, e.g., RLC circuits with nonlinear circuit elements.

Another important special case of (1.34) is the *van der Pol Equation* given by

$$\ddot{x} - \epsilon(1 - x^2)\dot{x} + x = 0, \tag{1.35}$$

where $\epsilon > 0$ is a parameter. This equation has been used to represent certain electronic oscillators.

If in (1.34), $f(x) \equiv 0$, we obtain

$$\ddot{x} + g(x) = 0. \tag{1.36}$$

When $xg(x) > 0$ for all $x \neq 0$, then (1.36) represents various models of so-called "mass on a nonlinear spring." In particular, if $g(x) = k(1 + a^2x^2)x$, where $k > 0$ and $a^2 > 0$ are parameters, then g represents the restoring force of a *hard spring*. If $g(x) = k(1 - a^2x^2)x$, where $k > 0$ and $a^2 > 0$ are parameters, then g represents the restoring force of a *soft spring*. Finally, if $g(x) = kx$, then g represents the restoring force of a *linear spring*. (See Figures 1.5 and 1.6.)

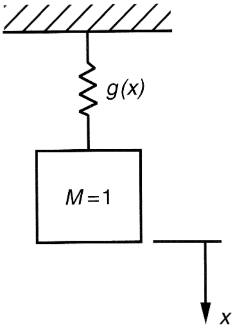
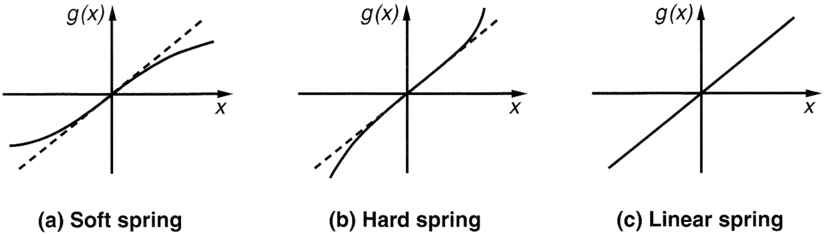


Figure 1.5. Mass on a nonlinear spring



(a) Soft spring

(b) Hard spring

(c) Linear spring

Figure 1.6. Mass on a nonlinear spring

For another special case of (1.34), let $f(x) \equiv 0$ and $g(x) = k \sin x$, where $k > 0$ is a parameter. Then (1.34) assumes the form

$$\ddot{x} + k \sin x = 0. \tag{1.37}$$

This equation describes the motion of a point mass moving in a circular path about the axis of rotation normal to a constant gravitational field, as shown in Figure 1.7. The parameter k depends on the radius l of the circular path, the gravitational acceleration g, and the mass. The symbol x denotes the angle of deflection measured from the vertical. The present model is called a *simple pendulum.*

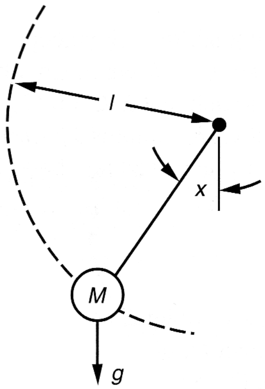


Figure 1.7. Model of a simple pendulum

Letting $x_1 = x$ and $x_2 = \dot{x}$, the second-order ordinary differential equation (1.34) can be represented by the system of first-order ordinary differential equations given by

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -f(x_1)x_2 - g(x_1). \end{aligned} \tag{1.38}$$

The required initial data for (1.38) are given by $x_1(0)$ and $x_2(0)$.

1.5 Solutions of Initial-Value Problems: Existence, Continuation, Uniqueness, and Continuous Dependence on Parameters

The following examples demonstrate that it is necessary to impose restrictions on the right-hand side of equation (1.11) to ensure the existence and uniqueness of solutions of the initial-value problem (1.12).

Example 1.5. For the initial-value problem,

$$\dot{x} = g(x), \quad x(0) = 0, \tag{1.39}$$

where $x \in R$, and

$$g(x) = \begin{cases} 1, & x = 0, \\ 0, & x \neq 0, \end{cases}$$

there exists no differentiable function ϕ that satisfies (1.39). Hence, *no solution* exists for this initial-value problem (in the sense defined in this chapter).

Example 1.6. The initial-value problem

$$\dot{x} = x^{1/3}, \quad x(t_0) = 0, \tag{1.40}$$

where $x \in R$, has at least two solutions given by $\phi_1(t) = [\frac{2}{3}(t - t_0)]^{3/2}$ and $\phi_2(t) = 0$ for $t \geq t_0$.

Example 1.7. The initial-value problem

$$\dot{x} = ax, \quad x(t_0) = x_0, \tag{1.41}$$

where $x \in R$, has a unique solution given by $\phi(t) = e^{a(t-t_0)}x(t_0)$ for $t \geq t_0$.

The following result provides a set of sufficient conditions for the *existence of solutions* of initial-value problem (1.12).

Theorem 1.8. *Let $f \in C(D, R^n)$. Then for any $(t_0, x_0) \in D$, the initial-value problem (1.12) has a solution defined on $[t_0, t_0 + c)$ for some $c > 0$.* ■

For a proof of Theorem 1.8, which is called the *Cauchy–Peano Existence Theorem*, refer to [1, Section 1.6].

The next result provides a set of sufficient conditions for the *uniqueness of solutions* for the initial-value problem (1.12).

Theorem 1.9. *Let $f \in C(D, R^n)$. Assume that for every compact set $K \subset D$, f satisfies the* Lipschitz condition

$$\| f(t, x) - f(t, y) \| \leq L_K \| x - y \| \tag{1.42}$$

for all $(t, x), (t, y) \in K$ where $L_K > 0$ is a constant depending only on K. Then (1.12) has at most one solution on any interval $[t_0, t_0 + c)$, $c > 0$. ■

For a proof of Theorem 1.9, refer to [1, Section 1.8]. In particular, if $f \in C^1(D, R^n)$, then the local Lipschitz condition (1.42) is automatically satisfied.

Now let ϕ be a solution of (1.11) on an interval J. By a *continuation* or *extension* of ϕ, we mean an extension ϕ_0 of ϕ to a larger interval J_0 in such a way that the extension solves (1.11) on J_0. Then ϕ is said to be *continued* or *extended* to the larger interval J_0. When no such continuation is possible, then ϕ is called *noncontinuable*.

Example 1.10. The scalar differential equation

$$\dot{x} = x^2 \tag{1.43}$$

has a solution $\phi(t) = \frac{1}{1-t}$ defined on $J = (-1, 1)$. This solution is continuable to the left to $-\infty$ and is not continuable to the right.

Example 1.11. The differential equation

$$\dot{x} = x^{1/3}, \tag{1.44}$$

where $x \in R$, has a solution $\psi(t) \equiv 0$ on $J = (-\infty, 0)$. This solution is continuable to the right in more than one way. For example, both $\psi_1(t) \equiv 0$ and $\psi_2(t) = (\frac{2t}{3})^{3/2}$ are solutions of (1.44) for $t \geq 0$.

In the next result, ∂D denotes the boundary of a domain D and ∂J denotes the boundary of an interval J.

Theorem 1.12. *If $f \in C(D, R^n)$ and if ϕ is a solution of (1.11) on an open interval J, then ϕ can be continued to a maximal open interval $J^* \supset J$ in such a way that $(t, \phi(t))$ tends to ∂D as $t \to \partial J^*$ when ∂D is not empty and $|t| + |\phi(t)| \to \infty$ if ∂D is empty. The extended solution ϕ^* on J^* is noncontinuable.* ■

For a proof of Theorem 1.12, refer to [1, Section 1.7].

When $D = J \times R^n$ for some open interval J and f satisfies a Lipschitz condition there (with respect to x), we have the following very useful *continuation* result.

Theorem 1.13. *Let $f \in C(J \times R^n, R^n)$ for some open interval $J \subset R$ and let f satisfy a Lipschitz condition on $J \times R^n$ (with respect to x). Then for any $(t_0, x_0) \in J \times R^n$, the initial-value problem (1.12) has a* unique *solution that exists on the* entire *interval J.* ■

For a proof of Theorem 1.13, refer to [1, Section 1.8].

In the next result we address initial-value problems that exhibit dependence on some parameter $\lambda \in G \subset R^m$ given by

$$\begin{aligned} \dot{x} &= f(t, x, \lambda), \\ x(\tau) &= \xi_\lambda, \end{aligned} \tag{1.45}$$

where $f \in C(J \times R^n \times G, R^n), J \subset R$ is an open interval, and ξ_λ depends continuously on λ.

Theorem 1.14. *Let $f \in C(J \times R^n \times G, R^n)$, where $J \subset R$ is an open interval and $G \subset R^m$. Assume that for each pair of compact subsets $J_0 \subset J$ and $G_0 \subset G$, there exists a constant $L = L_{J_0,G_0} > 0$ such that for all $(t,\lambda) \in J_0 \times G_0$, $x, y \in R^n$, the Lipschitz condition*

$$\| f(t,x,\lambda) - f(t,y,\lambda) \| \leq L \| x - y \| \tag{1.46}$$

is true. Then the initial-value problem (1.45) has a unique solution $\phi(t,\tau,\lambda)$, where $\phi \in C(J \times J \times G, R^n)$. Furthermore, if D is a set such that for all $\lambda_0 \in D$ there exists $\epsilon > 0$ such that $\overline{[\lambda_0 - \epsilon, \lambda_0 + \epsilon] \cap D} \subset D$, then $\phi(t,\tau,\lambda) \to \phi(t,\tau_0,\lambda_0)$ uniformly for $t_0 \in J_0$ as $(\tau,\lambda) \to (\tau_0,\lambda_0)$, where J_0 is any compact subset of J. (Recall that the upper bar denotes closure of a set.) ■

For a proof of Theorem 1.14, refer to [1, Section 1.9].

Note that Theorem 1.14 applies in the case of Example 1.7 and that the solution $\phi(t)$ of (1.41) depends continuously on the parameter a and the initial conditions $x(t_0) = x_0$.

When Theorem 1.9 is satisfied, it is possible to approximate the unique solutions of the initial-value problem (1.12) arbitrarily closely, using the *method of successive approximations* (also known as *Picard iterations*). Let $f \in C(D, R^n)$, let $K \subset D$ be a compact set, and let $(t_0, x_0) \in K$. *Successive approximations* for (1.12), or equivalently for (1.13), are defined as

$$\begin{aligned} \phi_0(t) &= x_0, \\ \phi_{m+1}(t) &= x_0 + \int_{t_0}^{t} f(s, \phi_m(s))ds, \quad m = 0, 1, 2, \ldots \end{aligned} \tag{1.47}$$

for $t_0 \leq t \leq t_0 + c$, for some $c > 0$.

Theorem 1.15. *If $f \in C(D, R^n)$ and if f is Lipschitz continuous on some compact set $K \subset D$ with constant L (with respect to x), then the successive approximations $\phi_m, m = 0, 1, 2, \ldots$ given in (1.47) exist on $[t_0, t_0 + c]$, are continuous there, and converge uniformly, as $m \to \infty$, to the unique solution ϕ of (1.12). (Thus, for every $\epsilon > 0$ there exists $N = N(\epsilon)$ such that for all $t \in [t_0, t_0 + c]$, $\| \phi(t) - \phi_m(t) \| < \epsilon$ whenever $m > N(\epsilon)$.)* ■

For the proof of Theorem 1.15, refer to [1, Section 1.8].

1.6 Systems of Linear First-Order Ordinary Differential Equations

In this section we will address linear ordinary differential equations of the form

$$\dot{x} = A(t)x + g(t) \tag{1.48}$$

and

$$\dot{x} = A(t)x \tag{1.49}$$

and

$$\dot{x} = Ax + g(t) \tag{1.50}$$

and

$$\dot{x} = Ax, \tag{1.51}$$

where $x \in R^n, A(t) = [a_{ij}(t)] \in C(R, R^{n \times n}), g \in C(R, R^n)$, and $A \in R^{n \times n}$.

Linear equations of the type enumerated above may arise in a natural manner in the modeling process of physical systems (see Section 1.4 for specific examples) or in the process of linearizing equations of the form (1.11) or (1.14) or some other kind of form.

1.6.1 Linearization

We consider the system of first-order ordinary differential equations given by

$$\dot{x} = f(t, x), \tag{1.52}$$

where $f : R \times D \to R^n$ and $D \subset R^n$ is some domain.

Linearization About a Solution ϕ

If $f \in C^1(R \times D, R^n)$ and if ϕ is a given solution of (1.52) defined for all $t \in R$, then we can *linearize* (1.52) about ϕ in the following manner. Define $\delta x = x - \phi(t)$ so that

$$\begin{aligned}\frac{d(\delta x)}{dt} \triangleq \delta\dot{x} &= f(t, x) - f(t, \phi(t)) \\ &= f(t, \delta x + \phi(t)) - f(t, \phi(t)) \\ &= \frac{\partial f}{\partial x}(t, \phi(t))\delta x + F(t, \delta x),\end{aligned} \tag{1.53}$$

where $\frac{\partial f}{\partial x}(t, x)$ denotes the *Jacobian matrix* of $f(t, x) = (f_1(t, x), \ldots, f_n(t, x))^T$ with respect to $x = (x_1, \ldots, x_n)^T$; i.e.,

$$\frac{\partial f}{\partial x}(t, x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(t, x) & \cdots & \frac{\partial f_1}{\partial x_n}(t, x) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(t, x) & \cdots & \frac{\partial f_n}{\partial x_n}(t, x) \end{bmatrix} \tag{1.54}$$

and

$$F(t, \delta x) \triangleq [f(t, \delta x + \phi(t)) - f(t, \phi(t)] - \frac{\partial f}{\partial x}(t, \phi(t))\delta x. \tag{1.55}$$

It turns out that $F(t, \delta x)$ is $o(\| \delta x \|)$ as $\| \delta x \| \to 0$ uniformly in t on compact subsets of R; i.e., for any compact subset $I \subset R$, we have

$$\lim_{\|\delta x\|\to 0}\left(\sup_{t\in I}\frac{\| F(t,\delta x) \|}{\| \delta x \|}\right) = 0.$$

For a proof of this assertion, we refer the reader to [1, Section 1.11].

Letting

$$\frac{\partial f}{\partial x}(t,\phi(t)) = A(t),$$

we obtain from (1.53) the equation

$$\frac{d(\delta x)}{dt} \triangleq \delta\dot{x} = A(t)\delta x + F(t,\delta x). \tag{1.56}$$

Associated with (1.56) we have the linear differential equation

$$\dot{z} = A(t)z, \tag{1.57}$$

called the *linearized equation* of (1.52) about the solution ϕ.

In applications, the linearization (1.57) of (1.52), about a given solution ϕ, is frequently used as a means of approximating a nonlinear process by a linear one (in the vicinity of ϕ). In Chapter 4, where we will study the stability properties of equilibria of (1.52) [which are specific kinds of solutions of (1.52)], we will show under what conditions it makes sense to deduce qualitative properties of a nonlinear process from its linearization.

Of special interest is the case when in (1.52), f is independent of t, i.e.,

$$\dot{x} = f(x) \tag{1.58}$$

and ϕ is a constant solution of (1.58), say, $\phi(t) = x_0$ for all $t \in R$. Under these conditions we have

$$\frac{d(\delta x)}{dt} \triangleq \delta\dot{x} = A\delta x + F(\delta x), \tag{1.59}$$

where

$$\lim_{\|\delta x\|\to 0}\frac{\| F(\delta x) \|}{\| \delta x \|} = 0 \tag{1.60}$$

and A denotes the Jacobian $\frac{\partial f}{\partial x}(x_0)$. Again, associated with (1.59) we have the linear differential equation

$$\dot{z} = Az,$$

called the *linearized equation* of (1.58) about the solution $\phi(t) \equiv x_0$.

Linearization About a Solution ϕ and an Input ψ

We can generalize the above to equations of the form

$$\dot{x} = f(t,x,u), \tag{1.61}$$

where $f : R \times D_1 \times D_2 \to R^n$ and $D_1 \subset R^n, D_2 \subset R^m$ are some domains. If $f \in C^1(R \times D_1 \times D_2, R^n)$ and if $\phi(t)$ is a given solution of (1.61) that we assume to exist for all $t \in R$ and that is determined by the initial condition x_0 and the *given specific function* $\psi \in C(R, R^m)$, i.e.,

$$\dot{\phi}(t) = f(t, \phi(t), \psi(t)), \quad t \in R,$$

then we can linearize (1.61) in the following manner. Define $\delta x = x - \phi(t)$ and $\delta u = u - \psi(t)$. Then

$$\begin{aligned} \frac{d(\delta x)}{dt} = \delta\dot{x} = \dot{x} - \dot{\phi}(t) &= f(t,x,u) - f(t,\phi(t),\psi(t)) \\ &= f(t, \delta x + \phi(t), \delta u + \psi(t)) - f(t,\phi(t),\psi(t)) \\ &= \frac{\partial f}{\partial x}(t,\phi(t),\psi(t))\delta x + \frac{\partial f}{\partial u}(t,\phi(t),\psi(t))\delta u \\ &\quad + F_1(t,\delta x, u) + F_2(t, \delta u), \end{aligned} \tag{1.62}$$

where

$$F_1(t,\delta x,u) = f(t, \delta x + \phi(t), u) - f(t,\phi(t),u) - \frac{\partial f}{\partial x}(t,\phi(t),\psi(t))\delta x$$

is $o(||\delta x||)$ as $\| \delta x \| \to 0$, uniformly in t on compact subsets of R for fixed u [i.e., for fixed u and for any compact subset $I \subset R$, $\lim\limits_{\|\delta x\| \to 0} \left(\sup_{t \in I} \frac{\|F_1(t,\delta x,u)\|}{\|\delta x\|} \right) = 0$], where

$$F_2(t,\delta u) = f(t,\phi(t),\delta u + \psi(t)) - f(t,\phi(t),\psi(t)) - \frac{\partial f}{\partial u}(t,\phi(t),\psi(t))\delta u$$

is $o(\| \delta u \|)$ as $\| \delta u \| \to 0$, uniformly in t on compact subsets of R [i.e., for any compact subset $I \subset R, \lim_{\|\delta u\| \to 0} \left(\sup_{t \in I} \frac{\|F_2(t,\delta u)\|}{\|\delta u\|} \right) = 0$], and where $\frac{\partial f}{\partial x}(\cdot)$ and $\frac{\partial f}{\partial u}(\cdot)$ denote the Jacobian matrix of f with respect to x and the Jacobian matrix of f with respect to u, respectively.

Letting

$$\frac{\partial f}{\partial x}(t,\phi(t),\psi(t)) = A(t) \text{ and } \frac{\partial f}{\partial u}(t,\phi(t),\psi(t)) = B(t),$$

we obtain from (1.62),

$$\frac{d(\delta x)}{dt} = \delta\dot{x} = A(t)\delta x + B(t)\delta u + F_1(t,\delta x,u) + F_2(t,\delta u). \tag{1.63}$$

Associated with (1.63), we have

$$\dot{z} = A(t)z + B(t)v. \tag{1.64}$$

We call (1.64) the *linearized equation* of (1.61) about the solution ϕ and the input function ψ.

As in the case of the linearization of (1.52) by (1.49), the linearization (1.64) of system (1.61) about a given solution ϕ and a given input ψ is often used in attempting to capture the qualitative properties of a nonlinear process by a linear process (in the vicinity of ϕ and ψ). In doing so, great care must be exercised to avoid erroneous conclusions.

The motivation of linearization is of course very obvious: much more is known about linear ordinary differential equations than about nonlinear ones. For example, the explicit forms of the solutions of (1.51) and (1.50) are known; the structures of the solutions of (1.49), (1.48), and (1.64) are known; the qualitative properties of the solutions of linear equations are known; and so forth.

1.6.2 Examples

We now consider some specific cases.

Example 1.16. We consider the *simple pendulum* discussed in Example 1.4 and described by the equation

$$\ddot{x} + k\sin x = 0, \tag{1.65}$$

where $k > 0$ is a constant. Letting $x_1 = x$ and $x_2 = \dot{x}$, (1.65) can be expressed as

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -k\sin x_1. \end{aligned} \tag{1.66}$$

It is easily verified that $\phi_1(t) \equiv 0$ and $\phi_2(t) \equiv 0$ is a solution of (1.66). Letting $f_1(x_1, x_2) = x_2$ and $f_2(x_1, x_2) = -k\sin x_1$, the Jacobian of $f(x_1, x_2) = (f_1(x_1, x_2), f_2(x_1, x_2))^T$ evaluated at $(x_1, x_2)^T = (0, 0)^T$ is given by

$$J(0) \triangleq A = \begin{bmatrix} 0 & 1 \\ -k\cos x_1 & 0 \end{bmatrix}_{\left[\begin{smallmatrix} x_1=0 \\ x_2=0 \end{smallmatrix}\right]} = \begin{bmatrix} 0 & 1 \\ -k & 0 \end{bmatrix}.$$

The linearized equation of (1.66) about the solution $\phi_1(t) \equiv 0, \phi_2(t) \equiv 0$ is given by

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -k & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Example 1.17. The system of equations

$$\begin{aligned} \dot{x}_1 &= ax_1 - bx_1x_2 - cx_1^2, \\ \dot{x}_2 &= dx_2 - ex_1x_2 - fx_2^2 \end{aligned} \tag{1.67}$$

describes the growth of two competing species (e.g., two species of small fish) that prey on each other (e.g., the adult members of one species prey on the young members of the other species, and vice versa). In (1.67) a, b, c, d, e, and f are positive parameters and it is assumed that $x_1 \geq 0$ and $x_2 \geq 0$. For (1.67), $\phi_1(t) = \phi_1(t, 0, 0) \equiv 0$ and $\phi_2(t) = \phi_2(t, 0, 0) \equiv 0, t \geq 0$, is a solution of (1.67). A simple computation yields

$$A = \frac{\partial f}{\partial x}(0) = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix},$$

and thus the system of equations

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \end{bmatrix} = \begin{bmatrix} a & 0 \\ 0 & d \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$$

constitutes the linearized equation of (1.67) about the solution $\phi_1(t) = 0, \phi_2(t) = 0, t \geq 0$.

Example 1.18. Consider a unit mass subjected to an inverse square law force field, as depicted in Figure 1.8. In this figure, r denotes radius and θ denotes angle, and it is assumed that the unit mass (representing, e.g., a satellite) can thrust in the radial and in the tangential directions with thrusts u_1 and u_2, respectively. The equations that govern this system are given by

$$\begin{aligned} \ddot{r} &= r\dot{\theta}^2 - \frac{k}{r^2} + u_1, \\ \ddot{\theta} &= \frac{-2\dot{\theta}\dot{r}}{r} + \frac{1}{r}u_2. \end{aligned} \tag{1.68}$$

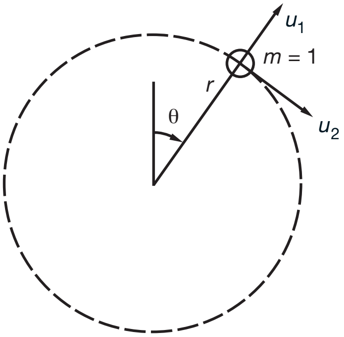


Figure 1.8. A unit mass subjected to an inverse square law force field

When $r(0) = r_0, \dot{r}(0) = 0, \theta(0) = \theta_0, \dot{\theta}(0) = \omega_0$, and $u_1(t) \equiv 0$, $u_2(t) \equiv 0$ for $t \geq 0$, it is easily verified that the system of equations (1.68) has as a solution the circular orbit given by

$$\begin{aligned} r(t) &\equiv r_0 = \text{ constant}, \\ \dot{\theta}(t) &= \omega_0 = \text{ constant} \end{aligned} \tag{1.69}$$

for all $t \geq 0$, which implies that

$$\theta(t) = \omega_0 t + \theta_0, \tag{1.70}$$

where $\omega_0 = (k/r_0^3)^{1/2}$.

If we let $x_1 = r$, $x_2 = \dot{r}$, $x_3 = \theta$, and $x_4 = \dot{\theta}$, the equations of motion (1.68) assume the form

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= x_1 x_4^2 - \frac{k}{x_1^2} + u_1, \\ \dot{x}_3 &= x_4, \\ \dot{x}_4 &= -\frac{2x_2 x_4}{x_1} + \frac{u_2}{x_1}. \end{aligned} \tag{1.71}$$

The linearized equation of (1.71) about the solution (1.70) [with $u_1(t) \equiv 0, u_2(t) \equiv 0$] is given by

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega_0^2 & 0 & 0 & 2r_0\omega_0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{-2\omega_0}{r_0} & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & \frac{1}{r_0} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}.$$

Example 1.19. In this example we consider systems described by equations of the form

$$\dot{x} + Af(x) + Bg(x) = u, \tag{1.72}$$

where $x \in R^n, A = [a_{ij}] \in R^{n \times n}, B = [b_{ij}] \in R^{n \times n}$ with $a_{ii} > 0$, $b_{ii} > 0$, $1 \leq i \leq n$, $f, g \in C^1(R^n, R^n)$, $u \in C(R^+, R^n)$, and $f(x) = 0$, $g(x) = 0$ if and only if $x = 0$.

Equation (1.72) can be used to model a great variety of physical systems. In particular, (1.72) has been used to model a large class of integrated circuits consisting of (nonlinear) transistors and diodes, (linear) capacitors and resistors, and current and voltage sources. (Figure 1.9 gives a symbolic representation of such circuits.) For such circuits, we assume that $f(x) = [f_1(x_1), \ldots, f_n(x_n)]^T$.

If $u(t) = 0$ for all $t \geq 0$, then $\phi_i(t) = 0$, $t \geq 0$, $1 \leq i \leq n$, is a solution of (1.72).

The system of equations (1.72) can be expressed equivalently as

$$\dot{x}_i = -\sum_{j=1}^{n} \left[a_{ij} \frac{f_j(x_j)}{x_j} + b_{ij} \frac{g_j(x_j)}{x_j} \right] x_j + u_i, \tag{1.73}$$

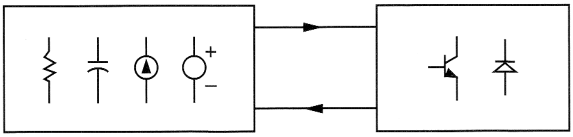


Figure 1.9. Integrated circuit

$i = 1, \ldots, n$. The linearized equation of (1.73) about the solution $\phi_i(t) = 0$, and the input $u_i(t) = 0$, $t \geq 0$, $i = 1, \ldots, n$, is given by

$$\dot{z}_i = -\sum_{j=1}^{n} \left[a_{ij} f_j'(0) + b_{ij} g_j'(0)\right] z_j + v_i, \tag{1.74}$$

where $f_j'(0) = \frac{df_j}{dx_j}(0)$ and $g_j'(0) = \frac{dg_j}{dx_j}(0), i = 1, \ldots, n$.

1.7 Linear Systems: Existence, Uniqueness, Continuation, and Continuity with Respect to Parameters of Solutions

In this section we address nonhomogeneous systems of first-order ordinary differential equations given by

$$\dot{x} = A(t)x + g(t), \tag{1.75}$$

where $x \in R^n, A(t) = [a_{ij}(t)]$ is a real $n \times n$ matrix and g is a real n-vector-valued function.

Theorem 1.20. *Suppose that $A \in C(J, R^{n \times n})$ and $g \in C(J, R^n)$, where J is some open interval. Then for any $t_0 \in J$ and any $x_0 \in R^n$, equation (1.75) has a unique solution satisfying $x(t_0) = x_0$. This solution exists on the* entire *interval J and is continuous in (t, t_0, x_0).*

Proof. The function $f(t,x) = A(t)x + g(t)$ is continuous in (t,x), and moreover, for any compact subinterval $J_0 \subset J$, there is an $L_0 \geq 0$ such that

$$\begin{aligned} \| f(t,x) - f(t,y) \|_1 &= \| A(t)(x-y) \|_1 \leq \| A(t) \|_1 \| x - y \|_1 \\ &\leq \left(\sum_{i=1}^{n} \max_{1 \leq j \leq n} |a_{ij}(t)| \right) \| x - y \|_1 \leq L_0 \| x - y \|_1 \end{aligned}$$

for all $(t,x), (t,y) \in J_0 \times R^n$, where L_0 is defined in the obvious way. Therefore, f satisfies a Lipschitz condition on $J_0 \times R^n$.

If $(t_0, x_0) \in J_0 \times R^n$, then the continuity of f implies the existence of solutions (Theorem 1.8), whereas the Lipschitz condition implies the uniqueness of solutions (Theorem 1.9). These solutions exist for the entire interval J_0 (Theorem 1.13). Since this argument holds for *any* compact subinterval $J_0 \subset J$, the solutions exist and are unique for all $t \in J$. Furthermore, the solutions are continuous with respect to t_0 and x_0 (Theorem 1.14 modified for the case where A and g do not depend on any parameters λ). ■

For the case when in (1.75) the matrix A and the vector g depend continuously on parameters λ and μ, respectively, it is possible to modify Theorem 1.20, and its proof, in the obvious way to show that the unique solutions of the system of equations

$$\dot{x} = A(t, \lambda)x + g(t, \mu) \tag{1.76}$$

are continuous in λ and μ as well. [Assume that $A \in C(J \times R^l, R^{n\times n})$ and $g \in C(J \times R^m, R^n)$ and follow a procedure that is similar to the proof of Theorem 1.20.]

1.8 Solutions of Linear State Equations

In this section we determine the specific form of the solutions of systems of linear first-order ordinary differential equations. We will revisit this topic in much greater detail in Chapter 3.

Homogeneous Equations

We begin by considering linear homogeneous systems

$$\dot{x} = A(t)x, \tag{1.77}$$

where $A \in C(R, R^{n\times n})$. By Theorem 1.20, for every $x_0 \in R^n$, (1.77) has a unique solution that exists for all $t \in R$. We will now use Theorem 1.15 to derive an expression for the solution $\phi(t, t_0, x_0)$ for (1.77) for $t \in R$ with $\phi(t_0, t_0, x_0) = x_0$. In this case the successive approximations given in (1.47) assume the form

$$\begin{aligned}
\phi_0(t, t_0, x_0) &= x_0, \\
\phi_1(t, t_0, x_0) &= x_0 + \int_{t_0}^{t} A(s)x_0 ds, \\
\phi_2(t, t_0, x_0) &= x_0 + \int_{t_0}^{t} A(s)\phi_1(s, t_0, x_0)ds, \\
&\cdots \\
\phi_m(t, t_0, x_0) &= x_0 + \int_{t_0}^{t} A(s)\phi_{m-1}(s, t_0, x_0)ds,
\end{aligned}$$

or

$$\begin{aligned}\phi_m(t,t_0,x_0) &= x_0 + \int_{t_0}^{t} A(s_1)x_0 ds_1 + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)x_0 ds_2 ds_1 + \cdots \\ &\quad + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)\cdots\int_{t_0}^{s_{m-1}} A(s_m)x_0 ds_m\cdots ds_1 \\ &= \Bigg[I + \int_{t_0}^{t} A(s_1)ds_1 + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)ds_2 ds_1 + \cdots \\ &\quad + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)\cdots\int_{t_0}^{s_{m-1}} A(s_m)ds_m\cdots ds_1\Bigg] x_0,\end{aligned} \tag{1.78}$$

where I denotes the $n \times n$ identity matrix. By Theorem 1.15, the sequence $\{\phi_m\}, m = 0,1,2,\dots$ determined by (1.78) converges uniformly, as $m \to \infty$, to the unique solution $\phi(t,t_0,x_0)$ of (1.77) on compact subsets of R. We thus have

$$\phi(t,t_0,x_0) = \Phi(t,t_0)x_0, \tag{1.79}$$

where

$$\begin{aligned}\Phi(t,t_0) &= I + \int_{t_0}^{t} A(s_1)ds_1 + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)ds_2 ds_1 \\ &\quad + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)\int_{t_0}^{s_2} A(s_3)ds_3 ds_2 ds_1 + \cdots \\ &\quad + \int_{t_0}^{t} A(s_1)\int_{t_0}^{s_1} A(s_2)\dots\int_{t_0}^{s_{m-1}} A(s_m)ds_m ds_{m-1}\cdots ds_1 + \cdots .\end{aligned} \tag{1.80}$$

Expression (1.80) is called the *Peano–Baker series*.

From expression (1.80) we immediately note that

$$\Phi(t,t) = I. \tag{1.81}$$

Furthermore, by differentiating expression (1.80) with respect to time andsubstituting into (1.77), we obtain that

$$\dot{\Phi}(t,t_0) = A(t)\Phi(t,t_0). \tag{1.82}$$

From (1.79) it is clear that once the initial data are specified and once the $n \times n$ matrix $\Phi(t,t_0)$ is known, the entire behavior of system (1.77) evolving in time t is known. This has motivated the *state* terminology: $x(t_0) = x_0$ is the *state of the system (1.77) at time* t_0, $\phi(t,t_0,x_0)$ is the *state of the system (1.77) at time* t, the solution ϕ is called the *state vector* of (1.77), the components of ϕ are called the *state variables* of (1.77), and the matrix $\Phi(t,t_0)$ that maps $x(t_0)$ into $\phi(t,t_0,x_0)$ is called the *state transition matrix* for (1.77). Also, the vector space containing the state vectors is called the *state space* for (1.77).

We can specialize the preceding discussion to linear systems of equations

$$\dot{x} = Ax. \tag{1.83}$$

In this case the mth term in (1.80) assumes the form

$$\begin{aligned}\int_{t_0}^{t} A(s_1) \int_{t_0}^{s_1} A(s_2) \int_{t_0}^{s_2} A(s_3) \dots \int_{t_0}^{s_{m-1}} A(s_m) ds_m \cdots ds_1 \\ = A^m \int_{t_0}^{t} \int_{t_0}^{s_1} \int_{t_0}^{s_2} \dots \int_{t_0}^{s_{m-1}} 1 ds_m \cdots ds_1 = \frac{A^m (t-t_0)^m}{m!},\end{aligned}$$

and expression (1.78) for ϕ_m assumes now the form

$$\phi_m(t, t_0, x_0) = \left[I + \sum_{k=1}^{m} \frac{A^k (t-t_0)^k}{k!} \right] x_0.$$

We conclude once more from Theorem 1.15 that $\{\phi_m\}$ converges uniformly as $m \to \infty$ to the unique solution $\phi(t, t_0, x_0)$ of (1.83) on compact subsets of R. We have

$$\begin{aligned}\phi(t, t_0, x_0) &= \left[I + \sum_{k=1}^{\infty} \frac{A^k (t-t_0)^k}{k!} \right] x_0 \\ &= \Phi(t, t_0) x_0 \triangleq \Phi(t - t_0) x_0,\end{aligned} \tag{1.84}$$

where $\Phi(t - t_0)$ denotes the state transition matrix for (1.83). [Note that by writing $\Phi(t, t_0) = \Phi(t - t_0)$, we have used a slight abuse of notation.] By making the analogy with the scalar $e^a = 1 + \sum_{k=1}^{\infty} \frac{a^k}{k!}$, usage of the notation

$$e^A = I + \sum_{k=1}^{\infty} \frac{A^k}{k!} \tag{1.85}$$

should be clear. We call e^A a *matrix exponential.* In Chapter 3 we will explore several ways of determining e^A for a given A.

Nonhomogeneous Equations

Next, we consider linear nonhomogeneous systems of ordinary differential equations

$$\dot{x} = A(t)x + g(t), \tag{1.86}$$

where $A \in C(R, R^{n \times n})$ and $g \in C(R, R^n)$. Again, by Theorem 1.20, for every $x_0 \in R^n$, (1.86) has a unique solution that exists for all $t \in R$. Instead of *deriving* the complete solution of (1.86) for a given set of initial data

$x(t_0) = x_0$, we will *guess* the solution and verify that it indeed satisfies (1.86). To this end, let us assume that the solution is of the form

$$\phi(t, t_0, x_0) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, s)g(s)ds, \tag{1.87}$$

where $\Phi(t, t_0)$ denotes the state transition matrix for (1.77).

To show that (1.87) is indeed the solution of (1.86), we first let $t = t_0$. In view of (1.81) and (1.87), we have $\phi(t_0, t_0, x_0) = x_0$. Next, by differentiating both sides of (1.87) and by using (1.81), (1.82), and (1.87), we have

$$\begin{aligned}\dot{\phi}(t, t_0, x_0) &= \dot{\Phi}(t, t_0)x_0 + \Phi(t, t)g(t) + \int_{t_0}^{t} \dot{\Phi}(t, s)g(s)ds \\ &= A(t)\Phi(t, t_0)x_0 + g(t) + \int_{t_0}^{t} A(t)\Phi(t, s)g(s)ds \\ &= A(t)[\Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, s)g(s)ds] + g(t) \\ &= A(t)\phi(t, t_0, x_0) + g(t);\end{aligned}$$

i.e., $\phi(t, t_0, x_0)$ given in (1.87) satisfies (1.86). Therefore, $\phi(t, t_0, x_0)$ is the unique solution of (1.86). Equation (1.87) is called the *variation of constants formula*, which is discussed further in Chapter 3. In the exercise section of Chapter 3 (refer to Exercise 3.13), we ask the reader (with hints) to *derive* the variation of constants formula (1.87), using a *change of variables.*

We note that when $x_0 = 0$, (1.87) reduces to

$$\phi(t, t_0, 0) \triangleq \phi_p(t) = \int_{t_0}^{t} \Phi(t, s)g(s)ds \tag{1.88}$$

and when $x_0 \neq 0$ but $g(t) = 0$ for all $t \in R$, (1.87) reduces to

$$\phi(t, t_0, x_0) \triangleq \phi_h(t) = \Phi(t, t_0)x_0. \tag{1.89}$$

Therefore, the *total solution* of (1.86) may be viewed as consisting of a component that is due to the initial conditions (t_0, x_0) and another component that is due to the *forcing term* $g(t)$. This type of separation is in general possible only in linear systems of differential equations. We call ϕ_p a *particular solution* of the nonhomogeneous system (1.86) and ϕ_h the *homogeneous solution.*

From (1.87) it is clear that for given initial conditions $x(t_0) = x_0$ and given forcing term $g(t)$, the behavior of system (1.86), summarized by ϕ, is known for all t. Thus, $\phi(t, t_0, x_0)$ specifies the *state vector*of (1.86) at time t. The components ϕ_i of ϕ, $i = 1, \ldots, n$, represent the *state variables* for (1.86), and the vector space that contains the state vectors is the *state space* for (1.86).

Before closing this section, it should be pointed out that in applications the matrix $A(t)$ and the vector $g(t)$ in (1.86) may be only *piecewise continuous* rather than continuous, as assumed above [i.e., $A(t)$ and $g(t)$ may have (at most) a finite number of discontinuities over any finite time interval]. In such cases, the derivative of x with respect to t [i.e., the right-hand side in (1.86)] will be discontinuous at a finite number of instants over any finite time interval; however, the state itself, x, will still be continuous at these instants [i.e., the solutions of (1.86) will still be continuous over R]. In such cases, all the results presented concerning existence, uniqueness, continuation of solutions, and so forth, as well as the explicit expressions of solutions of (1.86), are either still valid or can be modified in the obvious way. For example, should $g(t)$ experience a discontinuity at, say, $t_1 > t_0$, then expression (1.87) will be modified to read as follows:

$$\phi(t,t_0,x_0) = \Phi(t,t_0)x_0 + \int_{t_0}^{t} \Phi(t,s)g(s)ds, \quad t_0 \le t < t_1, \tag{1.90}$$

$$\phi(t,t_1,x_1) = \Phi(t,t_1)x_1 + \int_{t_1}^{t} \Phi(t,s)g(s)ds, \quad t \ge t_1, \tag{1.91}$$

where $x_1 = \lim_{t\to t_1^-} \phi(t,t_0,x_0)$.

1.9 Summary and Highlights

- *Initial-value problem*

$$\dot{x} = f(t,x), \quad x(t_0) = x_0 \tag{1.12}$$

or

$$\phi(t) = x_0 + \int_{t_0}^{t} f(s,\phi(s))ds, \tag{1.13}$$

where $\phi(t)$ is a solution of (1.12).

- *Successive approximations*

$$\begin{aligned} \phi_0(t) &= x_0, \\ \phi_{m+1}(t) &= x_0 + \int_{t_0}^{t} f(s,\phi_m(s))ds, \quad m = 0,1,2,\ldots. \end{aligned} \tag{1.47}$$

Under certain conditions (see Theorem 1.15) ϕ_m, $m = 1,2$, converges uniformly (on compact sets) as $m \to \infty$ to the unique solution of (1.12).

- *Linearization*

Given is $\dot{x} = f(t,x)$ and a solution ϕ. The Jacobian matrix is

$$\frac{\partial f}{\partial x}(t,x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(t,x) & \cdots & \frac{\partial f_1}{\partial x_n}(t,x) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(t,x) & \cdots & \frac{\partial f_n}{\partial x_n}(t,x) \end{bmatrix}. \tag{1.54}$$

For $A(t) = \frac{\partial f}{\partial x}(t,\phi(t))$,

$$\dot{z} = A(t)z \tag{1.57}$$

is the linearized equation about the solution ϕ.

- *Existence and uniqueness of solutions* of

$$\dot{x} = A(t)x + g(t). \tag{1.75}$$

See Theorem 1.20.

- The *solution* of

$$\dot{x} = A(t)x + g(t), \tag{1.86}$$

with $x(t_0) = x_0$, is given by the variation of constants formula

$$\phi(t, t_0, x_0) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, s)g(s)ds, \tag{1.87}$$

where the state transition matrix $\Phi(t, t_0)$ is given by

$$\Phi(t, t_0) = I + \int_{t_0}^{t} A(s_1)ds_1 + \int_{t_0}^{t} A(s_1) \int_{t_0}^{s_1} A(s_2)ds_2ds_1 + \cdots \tag{1.80}$$

the Peano–Baker series.

- In the *time-invariant case* $\dot{x} = Ax$,

$$\begin{aligned} \phi(t, t_0, x_0) &= \left[I + \sum_{k=1}^{\infty} \frac{A^k(t - t_0)^k}{k!} \right] x_0 \\ &= \Phi(t, t_0)x_0 \triangleq \Phi(t - t_0)x_0 \\ &= e^{A(t-t_0)}x_0, \end{aligned} \tag{1.84}$$

where

$$e^A = I + \sum_{k=1}^{\infty} \frac{A^k}{k!}. \tag{1.85}$$

1.10 Notes

For a classic reference on ordinary differential equations, see Coddington and Levinson [3]. Other excellent sources include Brauer and Nohel [2], Hartman [4], and Simmons [7]. Our treatment of ordinary differential equations in this chapter was greatly influenced by Miller and Michel [6].

References

1. P.J. Antsaklis and A.N. Michel, *Linear Systems*, Birkhäuser, Boston, MA, 2006.
2. F. Brauer and J.A. Nohel, *Qualitative Theory of Ordinary Differential Equations*, Benjamin, New York, NY, 1969.
3. E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, NY, 1955.

4. P. Hartman, *Ordinary Differential Equations*, Wiley, New York, NY, 1964.
5. A.N. Michel, K. Wang, and B. Hu, *Qualitative Theory of Dynamical Systems, 2nd Edition*, Marcel Dekker, New York, NY, 2001.
6. R.K. Miller and A.N. Michel, *Ordinary Differential Equations*, Academic Press, New York, NY, 1982.
7. G.F. Simmons, *Differential Equations*, McGraw-Hill, New York, NY, 1972.

Exercises

1.1. (*Hamiltonian dynamical systems*) *Conservative dynamical systems*, also called *Hamiltonian dynamical systems*, are those systems that contain no energy-dissipating elements. Such systems with n degrees of freedom can be characterized by means of a *Hamiltonian function* $H(p, q)$, where $q^T = (q_1, \dots, q_n)$ denotes n generalized position coordinates and $p^T = (p_1, \dots, p_n)$ denotes n generalized momentum coordinates. We assume that $H(p, q)$ is of the form

$$H(p, q) = T(q, \dot{q}) + W(q), \tag{1.92}$$

where T denotes the kinetic energy and W denotes the potential energy of the system. These energy terms are obtained from the path-independent line integrals

$$T(q, \dot{q}) = \int_0^{\dot{q}} p(q, \xi)^T d\xi = \int_0^{\dot{q}} \sum_{i=1}^{n} p_i(q, \xi) d\xi_i, \tag{1.93}$$

$$W(q) = \int_0^{q} f(\eta)^T d\eta = \int_0^{q} \sum_{i=1}^{n} f_i(\eta) d\eta_i, \tag{1.94}$$

where f_i, $i = 1, \dots, n$, denote generalized potential forces.

For the integral (1.93) to be path-independent, it is necessary and sufficient that

$$\frac{\partial p_i(q, \dot{q})}{\partial \dot{q}_j} = \frac{\partial p_j(q, \dot{q})}{\partial \dot{q}_i}, \quad i, j = 1, \dots, n. \tag{1.95}$$

A similar statement can be made about (1.94).

Conservative dynamical systems are described by the system of $2n$ ordinary differential equations

$$\begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i}(p, q), \qquad i = 1, \dots, n, \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i}(p, q), \quad i = 1, \dots, n. \end{aligned} \tag{1.96}$$

Note that if we compute the derivative of $H(p, q)$ with respect to t for (1.96) [along the solutions $q_i(t), p_i(t)$, $i = 1, \dots, n$], then we obtain, by the chain rule,

$$\begin{aligned}\frac{dH}{dt}(p(t),q(t)) &= \sum_{i=1}^{n}\frac{\partial H}{\partial p_i}(p,q)\dot{p}_i + \sum_{i=1}^{n}\frac{\partial H}{\partial q_i}(p,q)\dot{q}_i\\ &= \sum_{i=1}^{n} -\frac{\partial H}{\partial p_i}(p,q)\frac{\partial H}{\partial q_i}(p,q) + \sum_{i=1}^{n}\frac{\partial H}{\partial q_i}(p,q)\frac{\partial H}{\partial p_i}(p,q)\\ &= -\sum_{i=1}^{n}\frac{\partial H}{\partial p_i}(p,q)\frac{\partial H}{\partial q_i}(p,q) + \sum_{i=1}^{n}\frac{\partial H}{\partial p_i}(p,q)\frac{\partial H}{\partial q_i}(p,q) \equiv 0.\end{aligned}$$

In other words, in a conservative system (1.96), the Hamiltonian, i.e., the total energy, will be constant along the solutions (1.96). This constant is determined by the initial data $(p(0), q(0))$.

(a) In Figure 1.10, M_1 and M_2 denote point masses; K_1, K_2, K denote spring constants; and x_1, x_2 denote the displacements of the masses M_1 and M_2. Use the Hamiltonian formulation of dynamical systems described above to derive a system of first-order ordinary differential equations that characterize this system. Verify your answer by using Newton's second law of motion to derive the same system of equations. Assume that $x_1(0)$, $\dot{x}_1(0)$, $x_2(0)$, and $\dot{x}_2(0)$ are given.

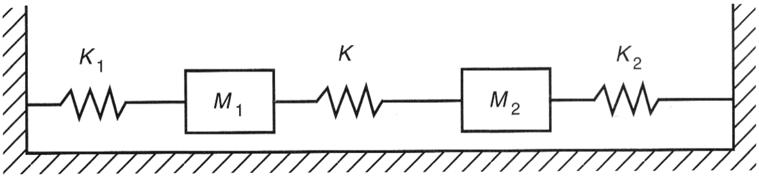


Figure 1.10. Example of a conservative dynamical system

(b) In Figure 1.11, a point mass M is moving in a circular path about the axis of rotation normal to a constant gravitational field (this is called the *simple pendulum problem*). Here l is the radius of the circular path, g is the gravitational acceleration, and θ denotes the angle of deflection measured from the vertical. Use the Hamiltonian formulation of dynamical systems described above to derive a system of first-order ordinary differential equations that characterize this system. Verify your answer by using Newton's second law of motion to derive the same system of equations. Assume that $\theta(0)$ and $\dot{\theta}(0)$ are given.

(c) Determine a system of first-order ordinary differential equations that characterizes the two-link pendulum depicted in Figure 1.12. Assume that $\theta_1(0)$, $\theta_2(0)$, $\dot{\theta}_1(0)$, and $\dot{\theta}_2(0)$ are given.

1.2. (*Lagrange's equation*) If a dynamical system contains elements that dissipate energy, such as viscous friction elements in mechanical systems and

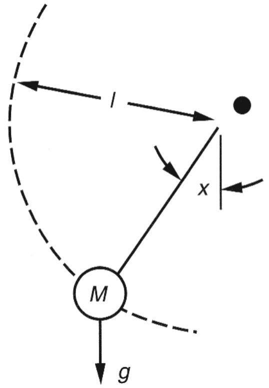


Figure 1.11. Simple pendulum

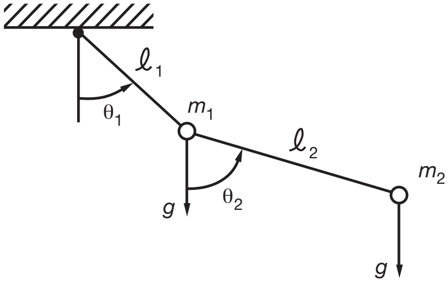


Figure 1.12. Two link pendulum

resistors in electric circuits, then we can use *Lagrange's equation* to describe such systems. (In the following, we use some of the same notation used in Exercise 1.1.) For a system with n degrees of freedom, this equation is given by

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}(q,\dot{q})\right) - \frac{\partial L}{\partial q}(q,\dot{q}) + \frac{\partial D}{\partial \dot{q}_i}(\dot{q}) = f_i, \quad i = 1,\dots,n, \tag{1.97}$$

where $q^T = (q_1,\dots,q_n)$ denotes the generalized position vector. The function $L(q,\dot{q})$ is called the *Lagrangian* and is defined as

$$L(q,\dot{q}) = T(q,\dot{q}) - W(q),$$

i.e., the difference between the kinetic energy T and the potential energy W.

The function $D(\dot{q})$ denotes *Rayleigh's dissipation function*, which we shall assume to be of the form

$$D(\dot{q}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \beta_{ij}\dot{q}_i\dot{q}_j,$$

where $[\beta_{ij}]$ is a positive semidefinite matrix (i.e., $[\beta_{ij}]$ is symmetric and all of its eigenvalues are nonnegative). The dissipation function D represents one-half the rate at which energy is dissipated as heat. It is produced by friction in mechanical systems and by resistance in electric circuits.

Finally, f_i in (1.97) denotes an applied force and includes all external forces associated with the q_i coordinate. The force f_i is defined as being positive when it acts to increase the value of the coordinate q_i.

(a) In Figure 1.13, M_1 and M_2 denote point masses; K_1, K_2, K denote spring constants; y_1, y_2 denote the displacements of masses M_1 and M_2, respectively; and B_1, B_2, B denote viscous damping coefficients. Use the Lagrange formulation of dynamical systems described above to derive two second-order differential equations that characterize this system. Transform these equations into a system of first-order ordinary differential equations. Verify your answer by using Newton's second law of motion to derive the same system equations. Assume that $y_1(0)$, $\dot{y}(0)$, $y_2(0)$, and $\dot{y}(0)$ are given.

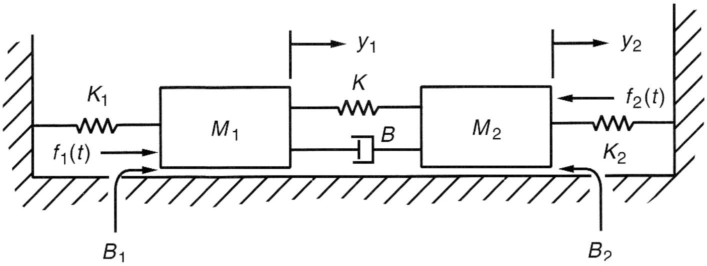


Figure 1.13. An example of a mechanical system with energy dissipation

(b) Consider the capacitor microphone depicted in Figure 1.14. Here we have a capacitor constructed from a fixed plate and a moving plate with mass M. The moving plate is suspended from the fixed frame by a spring with a spring constant K and has some damping expressed by the damping constant B. Sound waves exert an external force $f(t)$ on the moving plate. The output voltage v_s, which appears across the resistor R, will reproduce electrically the sound-wave patterns that strike the moving plate.

When $f(t) \equiv 0$ there is a charge q_0 on the capacitor. This produces a force of attraction between the plates that stretches the spring. When sound waves exert a force on the moving plate, there will be a resulting motion displacement x that is measured from the equilibrium position. The distance between the plates will then be $x_0 - x$, and the charge on the plates will be $q_0 + q$.

When displacements are small, the expression for the capacitance is given approximately by

$$C = \frac{\epsilon A}{x_0 - x}$$

with $C_0 = \epsilon A / x_0$, where $\epsilon > 0$ is the dielectric constant for air and A is the area of the plate.

Use the Lagrange formulation of dynamical systems to derive two second-order ordinary differential equations that characterize this system. Transform these equations into a system of first-order ordinary differential equations. Verify your answer by using Newton's laws of motion and Kirchhoff's voltage/current laws. Assume that $x(0), \dot{x}(0), q(0)$, and $\dot{q}(0)$ are given.

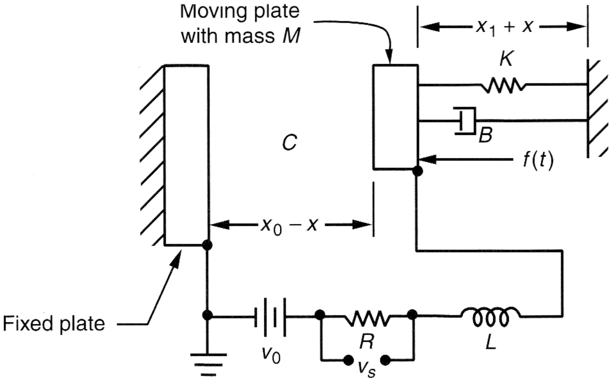


Figure 1.14. Capacitor microphone

(c) Use the Lagrange formulation to derive a system of first-order differential equations for the system given in Example 1.3.

1.3. Find examples of initial-value problems for which (a) no solutions exist; (b) more than one solution exists; (c) one or more solutions exist, but cannot be continued for all $t \in R$; and (d) unique solutions exist for all $t \in R$.

1.4. (*Numerical solution of ordinary differential equations—Euler's method*) An approximation to the solution of the *scalar* initial-value problem

$$\dot{y} = f(t, y), \quad y(t_0) = y_0 \tag{1.98}$$

is given by *Euler's method*,

$$y_{k+1} = y_k + hf(t_k, y_k), \quad k = 0, 1, 2, \ldots, \tag{1.99}$$

where $h = t_{k+1} - t_k$ is the (constant) integration step. The interpretation of this method is that the area below the solution curve is approximated by a sequence of sums of rectangular areas. This method is also called the *forward rectangular rule* (of integration).

(a) Use Euler's method to determine the solution of the initial-value problem

$$\dot{y} = 3y, \quad y(t_0) = 5, \quad t_0 = 0, \quad t_0 \le t \le 10.$$

(b) Use Euler's method to determine the solution of the initial-value problem

$$\ddot{y} = t(\dot{y})^2 - y^2, \quad y(t_0) = 1, \quad \dot{y}(t_0) = 0, \quad t_0 = 0, t_0 \le t \le 10.$$

Hint: In both cases, use $h = 0.2$. For part (b), let $y = x_1$, $\dot{x}_1 = x_2$, $\dot{x}_2 = tx_2^2 - x_1^2$, and apply (1.99), appropriately adjusted to the vector case. In both cases, plot y_k vs. $t_k, k = 0, 1, 2, \ldots$.
Remark:. Euler's method yields arbitrarily close approximations to the solutions of (1.98), by making h sufficiently small, *assuming infinite (computer) word length.* In practice, however, where truncation errors (quantization) and round-off errors (finite precision operations) are a reality, extremely small values of h may lead to numerical instabilities. Therefore, Euler's method is of limited value as a means of solving initial-value problems numerically.

1.5. (*Numerical solution of ordinary differential equations—Runge–Kutta methods*) The Runge–Kutta family of integration methods are among the most widely used techniques to solve initial-value problems (1.98). A simple version is given by

$$y_{i+1} = y_i + k,$$

where

$$k = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

with

$$\begin{aligned} k_1 &= hf(t_i, y_i), \\ k_2 &= hf(t_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1), \\ k_3 &= hf(t_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2), \\ k_4 &= hf(t_i + h, y_i + k_3), \end{aligned}$$

and $t_{i+1} = t_i + h, y(t_0) = y_0$.

The idea of this method is to probe ahead (in time) by one-half or by a whole step h to determine the values of the derivative at several points, and then to form a weighted average.

Runge–Kutta methods can also be applied to higher order ordinary differential equations. For example, after a change of variables, suppose that a second-order differential equation has been changed to a system of two first-order differential equations, say,

$$\begin{aligned} \dot{x}_1 &= f_1(t, x_1, x_2), \quad x_1(t_0) = x_{10}, \\ \dot{x}_2 &= f_2(t, x_1, x_2), \quad x_2(t_0) = x_{20}. \end{aligned} \tag{1.100}$$

In solving (1.100), a simple version of the Runge–Kutta method is given by

$$y_{i+1} = y_i + \underline{k},$$

where

$$y_i = (x_{1i}, x_{2i})^T \text{ and } \underline{k} = (k, l)^T$$

with

$$k = \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4), \quad l = \frac{1}{6}(l_1 + 2l_2 + 2l_3 + l_4)$$

and

$$\begin{aligned}
k_1 &= hf_1(t_i, x_{1i}, x_{2i}), & l_1 &= hf_2(t_i, x_{1i}, x_{2i}),\\
k_2 &= hf_1(t_i + \tfrac{1}{2}h, x_{1i} + \tfrac{1}{2}k_1, x_{2i} + \tfrac{1}{2}l_1), & l_2 &= hf_2(t_i + \tfrac{1}{2}h, x_{1i} + \tfrac{1}{2}k_1, x_{2i} + \tfrac{1}{2}l_1),\\
k_3 &= hf_1(t_i + \tfrac{1}{2}h, x_{1i} + \tfrac{1}{2}k_2, x_{2i} + \tfrac{1}{2}l_2), & l_3 &= hf_2(t_i + \tfrac{1}{2}h, x_{1i} + \tfrac{1}{2}k_2, x_{2i} + \tfrac{1}{2}l_2),\\
k_4 &= hf_1(t_i + h, x_{1i} + k_3, x_{2i} + l_3), & l_4 &= hf_2(t_i + h, x_{1i} + k_3, x_{2i} + l_3).
\end{aligned}$$

Use the Runge–Kutta method described above to obtain numerical solutions to the initial-value problems given in parts (a) and (b) of Exercise 1.4. Plot your data.
Remark: Since Runge–Kutta methods do not use past information, they constitute attractive starting methods for more efficient numerical integration schemes (e.g., predictor–corrector methods) . We note that since there are no built-in accuracy measures in the Runge–Kutta methods, significant computational efforts are frequently expended to achieve a desired accuracy.

1.6. (*Numerical solution of ordinary differential equations—Predictor–Corrector methods*) A common predictor–corrector technique for solving initial-value problems determined by ordinary differential equations, such as (1.98), is the *Milne* method, which we now summarize. In this method, $\dot{y}_{i-1}$ denotes the value of the first derivative at time t_{i-1}, where t_i is the time for the ith iteration step, $\dot{y}_{i-2}$ is similarly defined, and y_{i+1} represents the value of y to be determined. The details of the Milne method are:

1. $y_{i+1,p} = y_{i-3} + \frac{4h}{3}(2\dot{y}_{i-2} - \dot{y}_{i-1} + 2\dot{y}_i)$ (*predictor*)
2. $\dot{y}_{i+1,p} = f(t_{i+1}, y_{i+1,p})$
3. $y_{i+1,c} = y_{i-1} + \frac{h}{3}(\dot{y}_{i-1} + 4\dot{y}_i + \dot{y}_{i+1,p})$ (*corrector*)
4. $\dot{y}_{i+1,c} = f(t_{i+1}, y_{i+1,c})$
5. $y_{i+1,c} = y_{i-1} + \frac{h}{3}(\dot{y}_{i-1} + 4\dot{y}_i + \dot{y}_{i+1,c})$ (*iterating corrector*)

The first step is to obtain a predicted value of y_{i+1} and then to substitute $y_{i+1,p}$ into the given differential equation to obtain a predicted value of $\dot{y}_{i+1,p}$, as indicated in the second equation above. This predicted value, $\dot{y}_{i+1,p}$ is then used in the second equation, the corrector equation, to obtain a corrected value of y_{i+1}. The corrected value, $y_{i+1,c}$ is next substituted into the differential equation to obtain an improved value of $\dot{y}_{i+1}$, and so on. If necessary, an iteration process involving the fourth and fifth equations continues until successive values of y_{i+1} differ by less than the value of some desirable tolerance. With y_{i+1} determined to the desired accuracy, the method steps forward one h increment.

A more complicated predictor–corrector method that is more reliable than the Milne method is the *Adams–Bashforth–Moulton* method, the essential equations of which are

$$y_{i+1,p} = y_i + \frac{h}{24}(55\dot{y}_i - 59\dot{y}_{i-1} + 37\dot{y}_{i-2} - 9\dot{y}_{i-3}),$$
$$y_{i+1,c} = y_i + \frac{h}{24}(9\dot{y}_{i+1} + 19\dot{y}_i - 5\dot{y}_{i-1} + \dot{y}_{i-2}),$$

where in the corrector equation, $\dot{y}_{i+1}$ denotes the predicted value.

The application of predictor–corrector methods to systems of first-order ordinary differential equations is straightforward. For example, the application of the Milne method to the second-order system in (1.100) yields from the predictor step

$$x_{k,i+1,p} = x_{k,i-3} + \frac{4h}{3}(2\dot{x}_{k,i-2} - \dot{x}_{k,i-1} + 2\dot{x}_{k,i}), \quad k = 1, 2.$$

Then

$$\dot{x}_{k,i+1,p} = f_k(t_{i+1}, x_{1,i+1,p}, x_{2,i+1,p}), \quad k = 1, 2,$$

and the corrector step assumes the form

$$x_{k,i+1,c} = x_{k,i-1} + \frac{h}{3}(\dot{x}_{k,i-1} + 4\dot{x}_{k,i} + \dot{x}_{k,i+1}), \quad k = 1, 2,$$

and

$$\dot{x}_{k,i+1,c} = f_k(t_{i+1}, x_{1,i+1,c}, x_{2,i+1,c}), \quad k = 1, 2.$$

Use the Milne method and the Adams–Bashforth–Moulton method described above to obtain numerical solutions to the initial-value problems given in parts (a) and (b) of Exercise 1.4. To initiate the algorithm, refer to the Remark in Exercise 1.5.

Remark. Derivations and convergence properties of numerical integration schemes, such as those discussed here and in Exercises 1.4 and 1.5, can be found in many of the standard texts on numerical analysis.

1.7. Use Theorem 1.15 to solve the initial-value problem $\dot{x} = ax + t, x(0) = x_0$ for $t \geq 0$. Here $a \in R$.

1.8. Consider the initial-value problem

$$\dot{x} = Ax, \quad x(0) = x_0, \tag{1.101}$$

where $x \in R^2$ and $A \in R^{2\times 2}$. Let λ_1, λ_2 denote the eigenvalues of A; i.e., λ_1 and λ_2 are the roots of the equation $\det(A - \lambda I) = 0$, where det denotes determinant, λ is a scalar, and I denotes the 2×2 identity matrix. Make specific choices of A to obtain the following cases:

1. $\lambda_1 > 0, \lambda_2 > 0$, and $\lambda_1 \neq \lambda_2$
2. $\lambda_1 < 0, \lambda_2 < 0$, and $\lambda_1 \neq \lambda_2$
3. $\lambda_1 = \lambda_2 > 0$
4. $\lambda_1 = \lambda_2 < 0$
5. $\lambda_1 > 0, \lambda_2 < 0$
6. $\lambda_1 = \alpha + i\beta, \lambda_2 = \alpha - i\beta, i = \sqrt{-1}, \alpha > 0$
7. $\lambda_1 = \alpha + i\beta, \lambda_2 = \alpha - i\beta, \alpha < 0$
8. $\lambda_1 = i\beta, \lambda_2 = -i\beta$

Using t as a parameter, plot $\phi_2(t, 0, x_0)$ vs. $\phi_1(t, 0, x_0)$ for $0 \leq t \leq t_f$ for every case enumerated above. Here $[\phi_1(t, t_0, x_0), \phi_2(t, t_0, x_0)]^T = \phi(t, t_0, x_0)$ denotes the solution of (1.101). On your plots, indicate increasing time t by means of arrows. Plots of this type are called *trajectories* for (1.101), and sufficiently many plots (using different initial conditions and sufficiently large t_f) make up a *phase portrait* for (1.101). Generate a phase portrait for each case given above.

1.9. Write two first-order ordinary differential equations for the *van der Pol Equation* (1.35) by choosing $x_1 = x$ and $x_2 = \dot{x}_1$. Determine by simulation *phase portraits* (see Exercise 1.8) for this example for the cases $\epsilon = 0.05$ and $\epsilon = 10$ (refer also to Exercises 1.5 and 1.6 for numerical methods for solving differential equations). The periodic solution to which the trajectories of (1.35) tend to is an example of a *limit cycle*.

1.10. Consider a system whose state-space description is given by

$$\begin{aligned} \dot{x} &= -k_1 k_2 \sqrt{x} + k_2 u(t), \\ y &= k_1 \sqrt{x}. \end{aligned}$$

Linearize this system about the nominal solution

$$u_0 \equiv 0, \quad 2\sqrt{x_0(t)} = 2\sqrt{k} - k_1 k_2 t,$$

where $x_0(0) = k$.

1.11. For (1.36) consider the *hard, linear, and soft spring models* given by

$$\begin{aligned} g(x) &= k(1 + a^2 x^2)x, \\ g(x) &= kx, \\ g(x) &= k(1 - a^2 x^2)x, \end{aligned}$$

respectively, where $k > 0$ and $a^2 > 0$. Write two first-order ordinary differential equations for (1.36) by choosing $x_1 = x$ and $x_2 = \dot{x}$. Pick specific values for k and a^2. Determine by simulation *phase portraits* (see Exercise 1.8) for this example for the above three cases.

1.12. (a) Show that $x^T = (0, 0)$ is a solution of the system of equations

$$\begin{aligned} \dot{x}_1 &= x_1^2 + x_2^2 + x_2 \cos x_1, \\ \dot{x}_2 &= (1 + x_1)x_1 + (1 + x_2)x_2 + x_1 \sin x_2. \end{aligned}$$

Linearize this system about the point $x^T = (0, 0)$. By means of computer simulations, compare solutions corresponding to different initial conditions in the vicinity of the origin of the above system of equations and its linearization.

(b) Linearize the (bilinear control) system

$$\ddot{x} + (3 + \dot{x}^2)\dot{x} + (1 + x + x^2)u = 0$$

about the solution $x = 0, \dot{x} = 0$, and the input $u(t) \equiv 0$. As in part (a), compare (by means of computer simulations) solutions of the above equation with corresponding solutions of its linearization.

(c) In the circuit given in Figure 1.15, $v_i(t)$ is a voltage source and the nonlinear resistor obeys the relation $i_R = 1.5v_R^3$ [$v_i(t)$ is the circuit input and $v_R(t)$ is the circuit output]. Derive the differential equation for this circuit. Linearize this differential equation for the case when the circuit operates about the point $v_i = 14$.

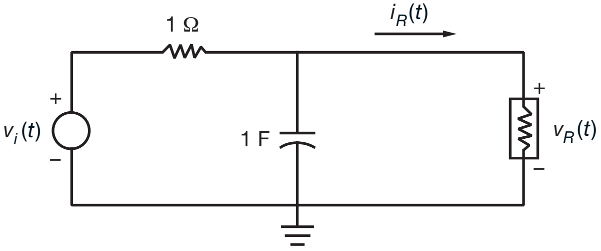


Figure 1.15. Nonlinear circuit

1.13. (*Inverted pendulum*) The inverted pendulum on a moving carriage subjected to an external force $\mu(t)$ is depicted in Figure 1.16.

The moment of inertia with respect to the center of gravity is J, and the coefficient of friction of the carriage (see Figure 1.16) is F. From Figure 1.17 we obtain the following equations for the dynamics of this system

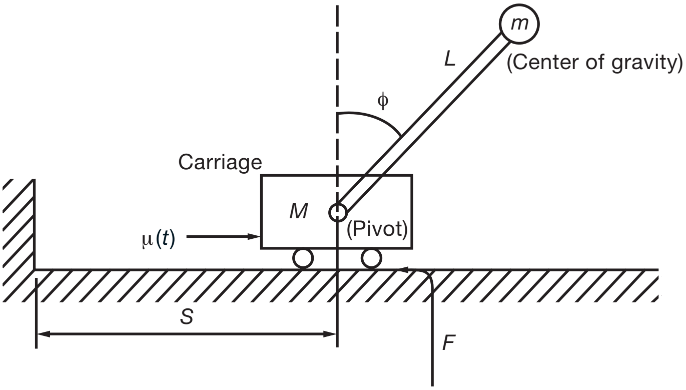


Figure 1.16. Inverted pendulum

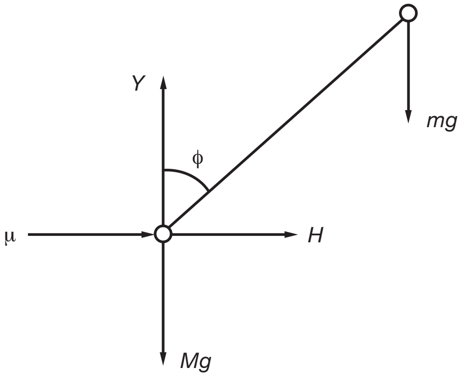


Figure 1.17. Force diagram of the inverted pendulum

$$m\frac{d^2}{dt^2}(S + L\sin\phi) \triangleq H, \tag{1.102a}$$

$$m\frac{d^2}{dt^2}(L\cos\phi) \triangleq Y - mg, \tag{1.102b}$$

$$J\frac{d^2\phi}{dt^2} = LY\sin\phi - LH\cos\phi, \tag{1.102c}$$

$$M\frac{d^2S}{dt^2} = \mu(t) - H - F\frac{dS}{dt}. \tag{1.102d}$$

Assuming that $m << M$, (1.102d) reduces to

$$M\frac{d^2S}{dt^2} = \mu(t) - F\frac{dS}{dt}. \tag{1.102e}$$

Eliminating H and Y from (1.102a) to (1.102c), we obtain

$$(J + mL^2)\ddot{\phi} = mgL\sin\phi - mL\ddot{S}\cos\phi. \tag{1.102f}$$

Thus, the system of Figure 1.16 is described by the equations

$$\begin{aligned} \ddot{\phi} - (g/L')\sin\phi + (1/L')\ddot{S}\cos\phi = 0, \\ M\ddot{S} + F\dot{S} = \mu(t), \end{aligned} \tag{1.102g}$$

where

$$L' = \frac{J + mL^2}{mL}$$

denotes the effective pendulum length.

Linearize system (1.102g) about $\phi = 0$.

1.14. (*Simple pendulum*) A system of first-order ordinary differential equations that characterize the simple pendulum considered in Exercise 1.1b is given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -\frac{g}{l}\sin x_1 \end{bmatrix},$$

where $x_1 \triangleq \theta$ and $x_2 \triangleq \dot{\theta}$ with $x_1(0) = \theta(0)$ and $x_2(0) = \dot{\theta}(0)$ specified. A linearized model of this system about the solution $x = [0, 0]^T$ is given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{g}{l} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Let $g = 10$ (m/sec^2) and $l = 1$ (m).

(a) For the case when $x(0) = [\theta_0, 0]^T$ with $\theta_0 = \pi/18$, $\pi/12$, $\pi/6$, and $\pi/3$, plot the states for $t \geq 0$ for the nonlinear model.

(b) Repeat (a) for the linear model.

(c) Compare the results in (a) and (b).

2

An Introduction to State-Space and Input–Output Descriptions of Systems

2.1 Introduction

State-space representations provide detailed descriptions of the internal behavior of a system, whereas input–output descriptions of systems emphasize external behavior and a system's interaction with this behavior.

In this chapter we address the state-space description of systems, which is an internal description of systems, and the input–output description of systems, also called the external description of systems. We will address continuous-time systems described by ordinary differential equations and discrete-time systems described by ordinary difference equations. We will emphasize linear systems. For such systems, the input–output descriptions involve the convolution integral for the continuous-time case and the convolution sum for the discrete-time case.

This chapter is organized into three parts. In the first of these (Section 2.2), we develop the state-space description of continuous-time systems, whereas in the second part (Section 2.3), we present the state-space representation of discrete-time systems. In the third part (Section 2.4), we address the input–output description of both continuous-time and discrete-time systems. Required background material for this chapter includes certain essentials in ordinary differential equations and linear algebra. This material can be found in Chapter 1 and the appendix, respectively.

2.2 State-Space Description of Continuous-Time Systems

Let us consider once more systems described by equations of the form

$$\dot{x} = f(t, x, u), \tag{2.1a}$$

$$y = g(t, x, u), \tag{2.1b}$$

where $x \in R^n, y \in R^p, u \in R^m, f : R \times R^n \times R^m \to R^n$, and $g : R \times R^n \times R^m \to R^p$. Here t denotes time and u and y denote system *input* and system *output*, respectively. Equation (2.1a) is called the *state equation*, (2.1b) is called the *output equation*, and (2.1a) and (2.1b) constitute the *state-space description* of continuous-time finite-dimensional systems.

The system input may be a function of t only (i.e., $u : R \to R^m$), or as in the case of *feedback control systems*, it may be a function of t and x (i.e., $u : R \times R^n \to R^m$). In either case, for a *given* (i.e., specified) u, we let $f(t, x, u) = F(t, x)$ and rewrite (2.1a) as

$$\dot{x} = F(t, x). \tag{2.2}$$

Now according to Theorems 1.13 and 1.14, if $F \in C(R \times R^n, R^n)$ and if for any compact subinterval $J_0 \subset R$ there is a constant L_{J_0} such that $\| F(t, x) - F(t, \tilde{x}) \| \leq L_{J_0} \| x - \tilde{x} \|$ for all $t \in J_0$ and for all $x, \tilde{x} \in R^n$, then the following statements are true:

1. For any $(t_0, x_0) \in R \times R^n$, (2.2) has a unique solution $\phi(t, t_0, x_0)$ satisfying $\phi(t_0, t_0, x_0) = x_0$ that exists for all $t \in R$.
2. The solution ϕ is continuous in t, t_0, and x_0.
3. If F depends continuously on parameters (say, $\lambda \in R^l$) and if x_0 depends continuously on λ, the solution ϕ is continuous in λ as well.

Thus, if the above conditions are satisfied, then for a given t_0, x_0, and u, (2.1a) will have a unique solution that exists for $t \in R$. Therefore, as already discussed in Section 1.8, $\phi(t, t_0, x_0)$ characterizes the *state* of the system at time t. Moreover, under these conditions, the system will have a unique *response* for $t \in R$, determined by (2.1b). We usually assume that $g \in C(R \times R^n \times R^m, R^p)$ or that $g \in C^1(R \times R^n \times R^m, R^p)$.

An important special case of (2.1) is systems described by linear time-varying equations of the form

$$\dot{x} = A(t)x + B(t)u, \tag{2.3a}$$

$$y = C(t)x + D(t)u, \tag{2.3b}$$

where $A \in C(R, R^{n\times n}), B \in C(R, R^{n\times m}), C \in C(R, R^{p\times n})$, and $D \in C(R, R^{p\times m})$. Such equations may arise in the modeling process of a physical system, or they may be a consequence of a linearization process, as discussed in Section 1.6.

By applying the results of Section 1.7, we see that for every initial condition $x(t_0) = x_0$ and for every given input $u : R \to R^m$, system (2.3a) possesses a unique solution that exists for all $t \in R$ and that is continuous in (t, t_0, x_0). Moreover, if A and B depend continuously on parameters, say, $\lambda \in R^l$, then the solutions will be continuous in the parameters as well. Indeed, in accordance with (1.87), this solution is given by

$$\phi(t, t_0, x_0) = \Phi(t, t_0)x_0 + \int_{t_0}^{t} \Phi(t, s)B(s)u(s)ds, \tag{2.4}$$

where $\varPhi(t, t_0)$ denotes the state transition matrix of the system of equations

$$\dot{x} = A(t)x. \tag{2.5}$$

By using (2.3b) and (2.4) we obtain the *system response* as

$$y(t) = C(t)\varPhi(t, t_0)x_0 + C(t)\int_{t_0}^{t} \varPhi(t, s)B(s)u(s)ds + D(t)u(t). \tag{2.6}$$

When in (2.3), $A(t) \equiv A, B(t) \equiv B, C(t) \equiv C$, and $D(t) \equiv D$, we have the important linear time-invariant case given by

$$\dot{x} = Ax + Bu, \tag{2.7a}$$
$$y = Cx + Du. \tag{2.7b}$$

In accordance with (1.84), (1.85), (1.87), and (2.4), the solution of (2.7a) is given by

$$\phi(t, t_0, x_0) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds \tag{2.8}$$

and the response of the system is given by

$$y(t) = Ce^{A(t-t_0)}x_0 + C\int_{t_0}^{t} e^{A(t-s)}Bu(s)ds + Du(t). \tag{2.9}$$

Linearity

We have referred to systems described by the linear equations (2.3) [resp., (2.7)] as *linear systems.* In the following discussion, we establish precisely in what sense this linearity is to be understood. To this end, for (2.3) we first let y_1 and y_2 denote system outputs that correspond to system inputs given by u_1 and u_2, respectively, *under the condition that* $x_0 = 0$. By invoking (2.6), it is clear that the system output corresponding to the system input $u = \alpha_1 u_1 + \alpha_2 u_2$, where α_1 and α_2 are real scalars, is given by $y = \alpha_1 y_1 + \alpha_2 y_2$; i.e.,

$$\begin{aligned} y(t) &= C(t)\int_{t_0}^{t} \varPhi(t, s)B(s)[\alpha_1 u_1(s) + \alpha_2 u_2(s)]ds + D(t)[\alpha_1 u_1(t) + \alpha_2 u_2(t)] \\ &= \alpha_1 C(t)\int_{t_0}^{t} \varPhi(t, s)B(s)u_1(s)ds + \alpha_2 C(t)\int_{t_0}^{t} \varPhi(t, s)B(s)u_2(s)ds \\ &\quad + \alpha_1 D(t)u_1(t) + \alpha_2 D(t)u_2(t) \\ &= \alpha_1 y_1(t) + \alpha_2 y_2(t). \end{aligned} \tag{2.10}$$

Next, for (2.3) we let y_1 and y_2 denote system outputs that correspond to initial conditions $x_0^{(1)}$ and $x_0^{(2)}$, respectively, *under the condition that* $u(t) = 0$

for all $t \in R$. Again, by invoking (2.6), it is clear that the system output corresponding to the initial condition $x_0 = \alpha_1 x_0^{(1)} + \alpha_2 x_0^{(2)}$, where α_1 and α_2 are real scalars, is given by $y = \alpha_1 y_1 + \alpha_2 y_2$; i.e.,

$$\begin{aligned} y(t) &= C(t)\Phi(t,t_0)[\alpha_1 x_0^{(1)} + \alpha_2 x_0^{(2)}] \\ &= \alpha_1 C(t)\Phi(t,t_0)x_0^{(1)} + \alpha_2 C(t)\Phi(t,t_0)x_0^{(2)} \\ &= \alpha_1 y_1(t) + \alpha_2 y_2(t). \end{aligned} \tag{2.11}$$

Equations (2.10) and (2.11) show that for systems described by the linear equations (2.3) [and, hence, by (2.7)], a *superposition principle* holds in terms of the input u and the corresponding output y of the system under the assumption of zero initial conditions, and in terms of the initial conditions x_0 and the corresponding output y under the assumption of zero input. It is important to note, however, that such a superposition principle will in general not hold under conditions that combine nontrivial inputs and nontrivial initial conditions. For example, with $x_0 \neq 0$ given, and with inputs u_1 and u_2 resulting in corresponding outputs y_1 and y_2 in (2.3), it does not follow that the input $\alpha_1 u_1 + \alpha_2 u_2$ will result in an output $\alpha_1 y_1 + \alpha_2 y_2$.

2.3 State-Space Description of Discrete-Time Systems

State-Space Representation

The state-space description of discrete-time finite-dimensional dynamical systems is given by equations of the form

$$x_i(k+1) = f_i(k, x_1(k), \dots, x_n(k), u_1(k), \dots, u_m(k)) \quad i = 1, \dots, n, \tag{2.12a}$$

$$y_i(k) = g_i(k, x_1(k), \dots, x_n(k), u_1(k), \dots, u_m(k)) \quad i = 1, \dots, p, \tag{2.12b}$$

for $k = k_0, k_0+1, \dots$, where k_0 is an integer. (In the following discussion, we let Z denote the set of integers and we let Z^+ denote the set of nonnegative integers.) Letting $x(k)^T = (x_1(k), \dots, x_n(k)), f(\cdot)^T = (f_1(\cdot), \dots, f_n(\cdot)), u(k)^T = (u_1(k), \dots, u_m(k)), y(k)^T = (y_1(k), \dots, y_p(k))$, and $g(\cdot)^T = (g_1(\cdot), \dots, g_m(\cdot))$, we can rewrite (2.12) more compactly as

$$x(k+1) = f(k, x(k), u(k)), \tag{2.13a}$$

$$y(k) = g(k, x(k), u(k)). \tag{2.13b}$$

Throughout this section we will assume that $f : Z \times R^n \times R^m \to R^n$ and $g : Z \times R^n \times R^m \to R^p$.

Since f is a function, for given $k_0, x(k_0) = x_0$, and for given $u(k), k = k_0, k_0+1, \dots$, (2.13a) possesses a unique solution $x(k)$ that exists for all $k = k_0, k_0+1, \dots$. Furthermore, under these conditions, $y(k)$ is uniquely defined for $k = k_0, k_0+1, \dots$.

As in the case of continuous-time finite-dimensional systems [see (2.1)], k_0 denotes *initial time*, k denotes *time*, $u(k)$ denotes the system *input* (evaluated at time k), $y(k)$ denotes the system *output* or system *response* (evaluated at time k), $x(k)$ characterizes the *state* (evaluated at time k), $x_i(k)$, $i = 1, \dots, n$, denote the *state variables*, (2.13a) is called the *state equation*, and (2.13b) is called the *output equation*.

A moment's reflection should make it clear that in the case of discrete-time finite-dimensional dynamical systems described by (2.13), questions concerning existence, uniqueness, and continuation of solutions are not an issue, as was the case in continuous-time systems. Furthermore, continuity with respect to initial data $x(k_0) = x_0$, or with respect to system parameters, is not an issue either, provided that $f(\cdot)$ and $g(\cdot)$ have appropriate continuity properties.

In the case of continuous-time systems described by ordinary differential equations [see (2.1)], we allow time t to evolve "forward" and "backward." Note, however, that in the case of discrete-time systems described by (2.13), we restrict the evolution of time k in the forward direction to ensure uniqueness of solutions. (We will revisit this issue in more detail in Chapter 3.)

Special important cases of (2.13) are *linear time-varying systems* given by

$$x(k+1) = A(k)x(k) + B(k)u(k), \tag{2.14a}$$
$$y(k) = C(k)x(k) + D(k)u(k), \tag{2.14b}$$

where $A : Z \to R^{n\times n}$, $B : Z \to R^{n\times m}$, $C : Z \to R^{p\times n}$, and $D : Z \to R^{p\times m}$. When $A(k) \equiv A, B(k) \equiv B, C(k) \equiv C$, and $D(k) \equiv D$, we have *linear time-invariant systems* given by

$$x(k+1) = Ax(k) + Bu(k), \tag{2.15a}$$
$$y(k) = Cx(k) + Du(k). \tag{2.15b}$$

As in the case of continuous-time finite-dimensional dynamical systems, many qualitative properties of discrete-time finite-dimensional systems can be studied in terms of *initial-value problems* given by

$$x(k+1) = f(k, x(k)), \quad x(k_0) = x_0, \tag{2.16}$$

where $x \in R^n$, $f : Z \times R^n \to R^n, k_0 \in Z$, and $k = k_0, k_0 + 1, \cdots$. We call the equation

$$x(k+1) = f(k, x(k)), \tag{2.17}$$

a *system of first-order ordinary difference equations*. Special important cases of (2.17) include *autonomous systems* described by

$$x(k+1) = f(x(k)), \tag{2.18}$$

periodic systems given by

$$x(k+1) = f(k, x(k)) = f(k+K, x(k)) \tag{2.19}$$

for fixed $K \in Z^+$ and for all $k \in Z$, *linear homogeneous systems* given by

$$x(k+1) = A(k)x(k), \tag{2.20}$$

linear periodic systems characterized by

$$x(k+1) = A(k)x(k) = A(k+K)x(k) \tag{2.21}$$

for fixed $K \in Z^+$ and for all $k \in Z$, *linear nonhomogeneous systems*

$$x(k+1) = A(k)x(k) + g(k), \tag{2.22}$$

and *linear, autonomous, homogeneous systems* characterized by

$$x(k+1) = Ax(k). \tag{2.23}$$

In these equations all symbols used are defined in the obvious way by making reference to the corresponding systems of ordinary differential equations (see Subsection 1.3.2).

Difference Equations of Order n

Thus far we have addressed systems of first-order difference equations. As in the continuous-time case, it is also possible to characterize initial-value problems by nth-order ordinary difference equations, say,

$$y(k+n) = h(k, y(k), y(k+1), \ldots, y(k+n-1)), \tag{2.24}$$

where $h : Z \times R^n \to R, n \in Z^+, k = k_0, k_0+1, \ldots$. By specifying an *initial time* $k_0 \in Z$ and by specifying $y(k_0), y(k_0+1), \ldots, y(k_0+n-1)$, we again have an initial-value problem given by

$$\begin{aligned} y(k+n) &= h(k, y(k), y(k+1), \ldots, y(k+n-1)), \\ y(k_0) &= x_{10}, \ldots, y(k_0+n-1) = x_{n0}. \end{aligned} \tag{2.25}$$

We call (2.24) an *nth-order ordinary difference equation*, and we note once more that in the case of initial-value problems described by such equations, there are no difficult issues involving the existence, uniqueness, and continuation of solutions.

We can reduce the study of (2.25) to the study of initial-value problems determined by systems of first-order ordinary difference equations. To accomplish this, we let in (2.25) $y(k) = x_1(k), y(k+1) = x_2(k), \ldots, y(k+n-1) = x_n(k)$. We now obtain the system of first-order ordinary difference equations

$$\begin{aligned} x_1(k+1) &= x_2(k), \\ &\cdots \\ x_{n-1}(k+1) &= x_n(k), \\ x_n(k+1) &= h(k, x_1(k), \ldots, x_n(k)). \end{aligned} \tag{2.26}$$

Equations (2.26), together with the initial data $x_0^T = (x_{10}, \dots, x_{n0})$, are equivalent to the initial-value problem (2.25) in the sense that these two problems will generate identical solutions [and in the sense that the transformation of (2.25) into (2.26) can be reversed unambiguously and uniquely].

As in the case of systems of first-order ordinary difference equations, we can point to several important special cases of nth-order ordinary difference equations, including equations of the form

$$y(k+n) + a_{n-1}(k)y(k+n-1) + \cdots + a_1(k)y(k+1) + a_0(k)y(k) = g(k), \quad (2.27)$$

$$y(k+n) + a_{n-1}(k)y(k+n-1) + \cdots + a_1(k)y(k+1) + a_0(k)y(k) = 0, \quad (2.28)$$

and

$$y(k+n) + a_{n-1}y(k+n-1) + \cdots + a_1y(k+1) + a_0y(k) = 0. \quad (2.29)$$

We call (2.27) a *linear nonhomogeneous ordinary difference equation of order n*, we call (2.28) a *linear homogeneous ordinary difference equation of order n*, and we call (2.29) a *linear, autonomous, homogeneous ordinary difference equation of order n*. As in the case of systems of first-order ordinary difference equations, we can define *periodic* and *linear periodic ordinary difference equations of order n* in the obvious way.

Solutions of State Equations

Returning now to linear homogeneous systems

$$x(k+1) = A(k)x(k), \quad (2.30)$$

we observe that

$$\begin{aligned} x(k+2) &= A(k+1)x(k+1) = A(k+1)A(k)x(k) \\ &\cdots \\ x(n) &= A(n-1)A(n-2)\cdots A(k+1)A(k)x(k) \\ &= \prod_{j=k}^{n-1} A(j)x(k); \end{aligned}$$

i.e., the state of the system at time n is related to the state at time k by means of the $n \times n$ matrix $\prod_{j=k}^{n-1} A(j)$ (as can easily be proved by induction). This suggests that the *state transition matrix* for (2.30) is given by

$$\Phi(n,k) = \prod_{j=k}^{n-1} A(j), \quad n > k, \quad (2.31)$$

and that

$$\Phi(k,k) = I. \quad (2.32)$$

As in the continuous-time case, the solution to the initial-value problem

$$\begin{aligned} x(k+1) &= A(k)x(k) \\ x(k_0) &= x_{k_0}, \quad k_0 \in Z, \end{aligned} \tag{2.33}$$

is now given by

$$x(n) = \Phi(n, k_0)x_{k_0} = \prod_{j=k_0}^{n-1} A(j)x_{k_0}, n > k_0. \tag{2.34}$$

Continuing, let us next consider initial-value problems determined by linear nonhomogeneous systems (2.22),

$$\begin{aligned} x(k+1) &= A(k)x(k) + g(k), \\ x(k_0) &= x_{k_0}. \end{aligned} \tag{2.35}$$

Then

$$\begin{aligned} x(k_0+1) &= A(k_0)x(k_0) + g(k_0), \\ x(k_0+2) &= A(k_0+1)x(k_0+1) + g(k_0+1) \\ &= A(k_0+1)A(k_0)x(k_0) + A(k_0+1)g(k_0) + g(k_0+1), \\ x(k_0+3) &= A(k_0+2)x(k_0+2) + g(k_0+2) \\ &= A(k_0+2)A(k_0+1)A(k_0)x(k_0) + A(k_0+2)A(k_0+1)g(k_0) \\ &\quad + A(k_0+2)g(k_0+1) + g(k_0+2) \\ &= \Phi(k_0+3, k_0)x_{k_0} + \Phi(k_0+3, k_0+1)g(k_0) \\ &\quad + \Phi(k_0+3, k_0+2)g(k_0+1) + \Phi(k_0+3, k_0+3)g(k_0+2), \end{aligned}$$

and so forth. For $k \geq k_0 + 1$, we easily obtain the expression for the solution of (2.35) as

$$x(k) = \Phi(k, k_0)x_{k_0} + \sum_{j=k_0}^{k-1} \Phi(k, j+1)g(j). \tag{2.36}$$

In the time-invariant case

$$\begin{aligned} x(k+1) &= Ax(k) + g(k), \\ x(k_0) &= x_{k_0}, \end{aligned} \tag{2.37}$$

the solution is again given by (2.36) where now the state transition matrix

$$\Phi(k, k_0) = A^{k-k_0}, \quad k \geq k_0, \tag{2.38}$$

in view of (2.31) and (2.32). The solution of (2.37) is then

$$x(k) = A^{k-k_0}x_{k_0} + \sum_{j=k_0}^{k-1} A^{k-(j+1)}g(j), \quad k > k_0. \tag{2.39}$$

We note that when $x_{k_0} = 0$, (2.36) reduces to

$$x_p(k) = \sum_{j=k_0}^{k-1} \Phi(k, j+1) g(j), \tag{2.40}$$

and when $x_{k_0} \neq 0$ but $g(k) \equiv 0$, then (2.36) reduces to

$$x_h(k) = \Phi(k, k_0) x_{k_0}. \tag{2.41}$$

Therefore, the *total solution* of (2.35) consists of the sum of its *particular solution*, $x_p(k)$, and its *homogeneous solution*, $x_h(k)$.

System Response

Finally, we observe that in view of (2.14b) and (2.36), the *system response* of the system (2.14), is of the form

$$\begin{aligned} y(k) = C(k)\Phi(k, k_0)x_{k_0} + C(k) \sum_{j=k_0}^{k-1} \Phi(k, j+1) B(j) u(j) \\ + D(k)u(k), \quad k > k_0, \end{aligned} \tag{2.42}$$

and

$$y(k_0) = C(k_0)x_{k_0} + D(k_0)u(k_0). \tag{2.43}$$

In the time-invariant case, in view of (2.39), the system response of the system (2.15) is

$$y(k) = CA^{k-k_0}x_{k_0} + C \sum_{j=k_0}^{k-1} A^{k-(j+1)} B(j) u(j) + Du(k), \quad k > k_0, \tag{2.44}$$

and

$$y(k_0) = Cx_{k_0} + Du(k_0). \tag{2.45}$$

Discrete-time systems, as discussed above, arise in several ways, including the *numerical solution* of ordinary differential equations (see, e.g., our discussion in Exercise 1.4 of *Euler's method*); the representation of *sampled-data systems* at discrete points in time (which will be discussed in further detail in Chapter 3); in the modeling process of systems that are defined only at discrete points in time (e.g., digital computer systems); and so forth.

As a specific example of a discrete-time system we consider a *second-order section digital filter* in *direct form*,

$$\begin{aligned} x_1(k+1) &= x_2(k), \\ x_2(k+1) &= ax_1(k) + bx_2(k) + u(k), \end{aligned} \tag{2.46a}$$

$$y(k) = x_1(k), \tag{2.46b}$$

$k \in Z^+$, where $x_1(k)$ and $x_2(k)$ denote the state variables, $u(k)$ denotes the input, and $y(k)$ denotes the output of the digital filter. We depict system (2.46) in block diagram form in Figure 2.1.

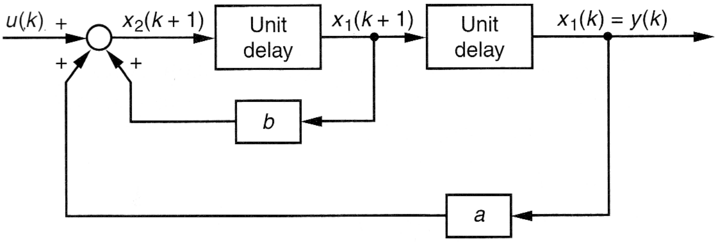


Figure 2.1. Second-order section digital filter in direct form

2.4 Input–Output Description of Systems

This section consists of four subsections. First we consider rather general aspects of the input–output description of systems. Because of their simplicity, we address the characterization of linear discrete-time systems next. In the third subsection we provide a foundation for the impulse response of linear continuous-time systems. Finally, we address the external description of linear continuous-time systems.

2.4.1 External Description of Systems: General Considerations

The state-space representation of systems presupposes knowledge of the *internal structure* of the system. When this structure is unknown, it may still be possible to arrive at a system description—an *external description*—that relates system inputs to system outputs. In linear system theory, a great deal of attention is given to relating the internal description of systems (the state representation) to the external description (the input–output description).

In the present context, we view *system inputs* and *system outputs* as elements of two real vector spaces U and Y, respectively, and we view a system as being represented by an operator T that relates elements of U to elements of Y. For $u \in U$ and $y \in Y$ we will assume that $u : R \to R^m$ and $y : R \to R^p$ in the case of *continuous-time systems*, and that $u : Z \to R^m$ and $y : Z \to R^p$ in the case of *discrete-time systems*. If $m = p = 1$, we speak of a *single-input/single-output (SISO) system*. Systems for which $m > 1$, $p > 1$, are called *multi-input/multi-output (MIMO) systems*. For continuous-time systems we define vector addition (on U) and multiplication of vectors by scalars (on U) as

$$(u_1 + u_2)(t) = u_1(t) + u_2(t) \tag{2.47}$$

and

$$(\alpha u)(t) = \alpha u(t) \tag{2.48}$$

for all $u_1, u_2 \in U, \alpha \in R$, and $t \in R$. We similarly define vector addition and multiplication of vectors by scalars on Y. Furthermore, for discrete-time systems we define these operations on U and Y analogously. In this case the elements of U and Y are real sequences that we denote, e.g., by $u = \{u_k\}$ or $u = \{u(k)\}$. (It is easily verified that under these rather general conditions, U and Y satisfy all the axioms of a vector space, both for the continuous-time case and the discrete-time case.) In the continuous-time case as well as in the discrete-time case the system is represented by $T : U \to Y$, and we write

$$y = T(u). \tag{2.49}$$

In the subsequent development, we will impose restrictions on the vector spaces U, Y, and on the operator T, as needed.

Linearity. If T is a linear operator, the system is called a *linear system.* In this case we have

$$\begin{aligned} y &= T(\alpha_1 u_1 + \alpha_2 u_2) \\ &= \alpha_1 T(u_1) + \alpha_2 T(u_2) \\ &= \alpha_1 y_1 + \alpha_2 y_2 \end{aligned} \tag{2.50}$$

for all $\alpha_1, \alpha_2 \in R$ and $u_1, u_2 \in U$ where $y_i = T(u_i) \in Y$, $i = 1, 2$, and $y \in Y$. Equation (2.50) represents the well-known *principle of superposition* of linear systems.

With or Without Memory. We say that a system is *memoryless*, or *without memory,* if its output for each value of the independent variable (t or k) is dependent only on the input evaluated at the same value of the independent variable [e.g., $y(t_1)$ depends only on $u(t_1)$ and $y(k_1)$ depends only on $u(k_1)$]. An example of such a system is the resistor circuit shown in Figure 2.2, where the current $i(t) = u(t)$ denotes the system input at time t and the voltage across the resistor, $v(t) = Ri(t) = y(t)$, denotes the system output at time t.

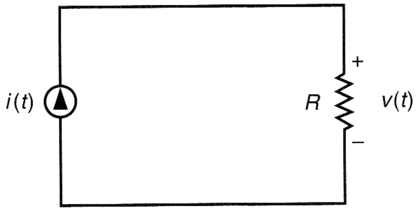


Figure 2.2. Resistor circuit

A system that is not memoryless is said to have memory. An example of a continuous-time *system with memory* is the capacitor circuit shown in Figure 2.3, where the current $i(t) = u(t)$ represents the system input at time t and the voltage across the capacitor,

$$y(t) = v(t) = \frac{1}{C}\int_{-\infty}^{t} i(\tau)d\tau,$$

denotes the system output at time t. Another example of a continuous-time system with memory is described by the scalar equation

$$y(t) = u(t-1), \quad t \in R,$$

and an example of a discrete-time system with memory is characterized by the scalar equation

$$y(n) = \sum_{k=-\infty}^{n} x(k), \quad n, k \in Z.$$

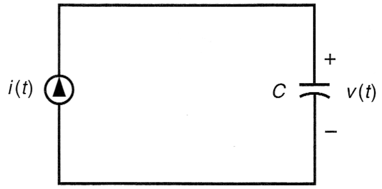


Figure 2.3. Capacitor circuit

Causality. A system is said to be *causal* if its output at any time, say t_1 (or k_1), depends only on values of the input evaluated for $t \le t_1$ (for $k \le k_1$). Thus, $y(t_1)$ depends only on $u(t), t \le t_1$ [or $y(k_1)$ depends only on $u(k), k \le k_1$]. Such a system is referred to as being *nonanticipative* since the system output does not anticipate future values of the input.

To make the above concept a bit more precise, we define the function $u_\tau : R \to R^m$ for $u \in U$ by

$$u_\tau(t) = \begin{cases} u(t), & t \le \tau, \\ 0, & t > \tau, \end{cases}$$

and we similarly define the function $y_\tau : R \to R^p$ for $y \in Y$. A system that is represented by the mapping $y = T(u)$ is said to be *causal* if and only if

$$(T(u))_\tau = (T(u_\tau))_\tau \quad \text{for all } \tau \in R, \text{ for all } u \in U.$$

Equivalently, this system is causal if and only if for $u, v \in U$ and $u_\tau = v_\tau$ it is true that

$$(T(u))_\tau = (T(v))_\tau \quad \text{for all } \tau \in R.$$

For example, the discrete-time system described by the scalar equation

$$y(n) = u(n) - u(n+1), \quad n \in Z,$$

is *not causal.* Neither is the continuous-time system characterized by the scalar equation

$$y(t) = x(t+1), \quad t \in R.$$

It should be pointed out that systems that are not causal are by no means useless. For example, causality is *not* of fundamental importance in image-processing applications where the independent variable is not time. Even when time is the independent variable, noncausal systems may play an important role. For example, in the processing of data that have been recorded (such as speech, meteorological data, demographic data, and stock market fluctuations), one is not constrained to processing the data causally. An example of this would be the smoothing of data over a time interval, say, by means of the system

$$y(n) = \frac{1}{2M+1} \sum_{k=-M}^{M} u(n-k).$$

Time-Invariance. A system is said to be *time-invariant* if a time shift in the input signal causes a corresponding time shift in the output signal. To make this concept more precise, for fixed $\alpha \in R$, we introduce the *shift operator* $Q_\alpha : U \to U$ as

$$Q_\alpha u(t) = u(t-\alpha), \quad u \in U, t \in R.$$

A system that is represented by the mapping $y = T(u)$ is said to be *time-invariant* if and only if

$$TQ_\alpha(u) = Q_\alpha(T(u)) = Q_\alpha(y)$$

for any $\alpha \in R$ and any $u \in U$. If a system is not time-invariant, it is said to be *time-varying*.

For example, a system described by the relation

$$y(t) = \cos u(t)$$

is time-invariant. To see this, consider the inputs $u_1(t)$ and $u_2(t) = u_1(t-t_0)$. Then

$$y_1(t) = \cos u_1(t), \quad y_2(t) = \cos u_2(t) = \cos u_1(t-t_0)$$

and

$$y_1(t-t_0) = \cos u_1(t-t_0) = y_2(t).$$

As a second example, consider a system described by the relation

$$y(n) = nu(n)$$

and consider two inputs $u_1(n)$ and $u_2(n) = u_1(n-n_0)$. Then

$$y_1(n) = nu_1(n) \quad \text{and} \quad y_2(n) = nu_2(n) = nu_1(n - n_0).$$

However, if we shift the output $y_1(n)$ by n_0, we obtain

$$y_1(n - n_0) = (n - n_0)u_1(n - n_0) \neq y_2(n).$$

Therefore, this system is not time-invariant.

2.4.2 Linear Discrete-Time Systems

In this subsection we investigate the representation of linear discrete-time systems. We begin our discussion by considering SISO systems.

In the following, we employ the *discrete-time impulse* (or *unit pulse* or *unit sample*), which is defined as

$$\delta(n) = \begin{cases} 0, & n \neq 0, n \in Z, \\ 1, & n = 0. \end{cases} \tag{2.51}$$

Note that if $\{p(n)\}$ denotes the *unit step sequence*, i.e.,

$$p(n) = \begin{cases} 1, & n \geq 0, n \in Z, \\ 0, & n < 0, n \in Z, \end{cases} \tag{2.52}$$

then

$$\delta(n) = p(n) - p(n - 1)$$

and

$$p(n) = \begin{cases} \sum_{k=0}^{\infty} \delta(n - k), & n \geq 0, \\ 0, & n < 0. \end{cases} \tag{2.53}$$

Furthermore, note that an arbitrary sequence $\{x(n)\}$ can be expressed as

$$x(n) = \sum_{k=-\infty}^{\infty} x(k)\delta(n - k). \tag{2.54}$$

We can easily show that a transformation $T : U \to Y$ determined by the equation

$$y(n) = \sum_{k=-\infty}^{\infty} h(n, k)u(k), \tag{2.55}$$

where $y \triangleq \{y(k)\} \in Y$, $u \triangleq \{u(k)\} \in U$, and $h : Z \times Z \to R$, is a linear transformation. Also, we note that for (2.55) to make any sense, we need to impose restrictions on $\{h(n, k)\}$ and $\{u(k)\}$. For example, if for every fixed n, $\{h(n, k)\} \in l_2$ and $\{u(k)\} \in l_2 = U$, then it follows from the Hölder Inequality (resp., Schwarz Inequality), see Section A.7, that (2.55) is well defined. There are of course other conditions that one might want to impose on (2.55).

For example, if for every fixed n, $\sum_{k=-\infty}^{\infty} |h(n,k)| < \infty$ (i.e., for every fixed $n, \{h(n,k)\} \in l_1$) and if $\sup_{k\in Z} |u(k)| < \infty$ (i.e., $\{u(k)\} \in l_\infty$), then (2.55) is also well defined.

We shall now elaborate on the suitability of (2.55) to represent linear discrete-time systems. To this end, we will agree once and for all that, in the ensuing discussion, all assumptions on $\{h(n,k)\}$ and $\{u(k)\}$ are satisfied that ensure that (2.55) is well defined.

We will view $y \in Y$ and $u \in U$ as system outputs and system inputs, respectively, and we will let $T : U \to Y$ denote a linear transformation that relates u to y. We first consider the case when $u(k) = 0$ for $k < k_0$, k, $k_0 \in Z$. Also, we assume that for $k > n \geq k_0$, the inputs $u(k)$ do not contribute to the system output at time n (i.e., the system is *causal*). Under these assumptions, and in view of the linearity of T, and by invoking the representation of signals by (2.54), we obtain for $y = \{y(n)\}$, $n \in Z$, the expression $y(n) = T(\sum_{k=-\infty}^{\infty} u(k)\delta(n-k)) = T(\sum_{k=k_0}^{n} u(k)\delta(n-k)) = \sum_{k=k_0}^{n} u(k)T(\delta(n-k)) = \sum_{k=k_0}^{n} h(n,k)u(k)$, $n \geq k_0$, and $y(n) = 0$, $n < k_0$, where $T(\delta(n-k)) \triangleq (T\delta)(n-k) \triangleq h(n,k)$ represents the response of T to a unit pulse (resp., discrete-time impulse or unit sample) occurring at $n = k$.

When the assumptions in the preceding discussion are no longer valid, then a different argument than the one given above needs to be used to arrive at the system representation. Indeed, for *infinite sums*, the interchanging of the order of the summation operation $\sum$ with the linear transformation T is no longer valid. We refer the reader to a paper by I. W. Sandberg ("A Representation Theorem for Linear Systems," *IEEE Transactions on Circuits and Systems—I*, Vol. 45, No. 5, pp. 578–580, May 1998) for a derivation of the representation of general linear discrete-time systems. In that paper it is shown that an extra term needs to be added to the right-hand side of equation (2.55), even in the representation of *general*, linear, time-invariant, causal, discrete-time systems. [In the proof, the Hahn–Banach Theorem (which is concerned with the extension of bounded linear functionals) is employed and the extra required term is given by $\lim_{l\to\infty} T(\sum_{k=-\infty}^{-c_l-1} u(k)\delta(n-k) + \sum_{k=c_l+1}^{\infty} u(k)\delta(n-k)]$ with $c_l \to \infty$ as $l \to \infty$. For a statement and proof of the Hahn–Banach Theorem, refer, e.g., to A. N. Michel and C. J. Herget, *Applied Algebra and Functional Analysis*, Dover, New York, 1993, pp. 367–370.) In that paper it is also pointed out, however, that cases with such extra *nonzero* terms are not necessarily of importance in applications. In particular, if inputs and outputs are defined (to be nonzero) on just the non-negative integers, then for causal systems no additional term is needed (or more specifically, the extra term is zero), as seen in our earlier argument. In any event, *throughout this book we will concern ourselves with linear discrete-time systems that can be represented by equation (2.55)* for the single-input/single-output case (and appropriate generalizations for multi-input/multi-output cases).

Next, suppose that T represents a time-invariant system. This means that if $\{h(n,0)\}$ is the response to $\{\delta(n)\}$, then by time invariance, the response to $\{\delta(n-k)\}$ is simply $\{h(n-k,0)\}$. By a slight abuse of notation, we let $h(n-k,0) \triangleq h(n-k)$. Then (2.55) assumes the form

$$y(n) = \sum_{k=-\infty}^{\infty} u(k)h(n-k). \tag{2.56}$$

Expression (2.56) is called a *convolution sum* and is written more compactly as

$$y(n) = u(n) * h(n).$$

Now by a substitution of variables, we obtain for (2.56) the alternative expression

$$y(n) = \sum_{k=-\infty}^{\infty} h(k)u(n-k),$$

and therefore, we have

$$y(n) = u(n) * h(n) = h(n) * u(n);$$

i.e., the convolution operation $*$ commutes.

As a specific example, consider a linear, time-invariant, discrete-time system with unit impulse response given by

$$h(n) = \left\{ \begin{matrix} a^n, \; n \geq 0 \\ 0, \quad n < 0 \end{matrix} \right\} = a^n p(n), \quad 0 < a < 1,$$

where $p(n)$ is the unit step sequence given in (2.52). It is an easy matter to show that the response of this system to an input given by

$$u(n) = p(n) - p(n-N)$$

is

$$\begin{aligned} y(n) &= 0, n < 0, \\ y(n) &= \sum_{k=0}^{n} a^{n-k} = a^n \frac{1 - 1^{-(n+1)}}{1 - a^{-1}} = \frac{1 - a^{n+1}}{1 - a}, \quad 0 \leq n < N, \end{aligned}$$

and

$$y(n) = \sum_{k=0}^{N-1} a^{n-k} = a^n \frac{1 - a^{-N}}{1 - a^{-1}} = \frac{a^{n-N+1} - a^{n+1}}{1 - a}, \quad N \leq n.$$

Proceeding, with reference to (2.55) we note that $h(n,k)$ represents the system output at time n due to a δ-function input applied at time k. Now if system (2.55) is *causal*, then its output will be identically zero before an input is applied. Hence, a *linear system (2.55) is causal if and only if*

$$h(n,k) = 0 \quad \text{for all} \quad n < k.$$

Therefore, when the system (2.55) is causal, we have in fact

$$y(n) = \sum_{k=-\infty}^{n} h(n,k)u(k). \tag{2.57a}$$

We can rewrite (2.57a) as

$$\begin{aligned} y(n) &= \sum_{k=-\infty}^{k_0-1} h(n,k)u(k) + \sum_{k=k_0}^{n} h(n,k)u(k) \\ &\triangleq y(k_0-1) + \sum_{k=k_0}^{n} h(n,k)u(k). \end{aligned} \tag{2.57b}$$

We say that the discrete-time system described by (2.55) is *at rest* at $k = k_0 \in Z$ if $u(k) = 0$ for $k \geq k_0$ implies that $y(k) = 0$ for $k \geq k_0$. Accordingly, if system (2.55) is known to be at rest at $k = k_0$, we have

$$y(n) = \sum_{k=k_0}^{\infty} h(n,k)u(k).$$

Furthermore, if system (2.55) is known to be causal and at rest at $k = k_0$, its input–output description assumes the form [in view of (2.57b)]

$$y(n) = \sum_{k=k_0}^{n} h(n,k)u(k). \tag{2.58}$$

If now, in addition, system (2.55) is also time-invariant, (2.58) becomes

$$y(n) = \sum_{k=k_0}^{n} h(n-k)u(k) = \sum_{k=k_0}^{n} h(k)u(n-k), \tag{2.59}$$

which is a convolution sum. [Note that in (2.59) we have slightly abused the notation for $h(\cdot)$, namely that $h(n-k) = h(n-k,0)(= h(n,k)$.]

Next, turning to linear, discrete-time, *MIMO systems*, we can generalize (2.55) to

$$y(n) = \sum_{k=-\infty}^{\infty} H(n,k)u(k), \tag{2.60}$$

where $y : Z \to R^p$, $u : Z \to R^m$, and

$$H(n,k) = \begin{bmatrix} h_{11}(n,k) & h_{12}(n,k) & \cdots & h_{1m}(n,k) \\ h_{21}(n,k) & h_{22}(n,k) & \cdots & h_{2m}(n,k) \\ \cdots & \cdots & \cdots & \cdots \\ h_{p1}(n,k) & h_{p2}(n,k) & \cdots & h_{pm}(n,k) \end{bmatrix}, \tag{2.61}$$

where $h_{ij}(n,k)$ represents the system response at time n of the ith component of y due to a discrete-time impulse δ applied at time k at the jth component of u, whereas the inputs at all other components of u are being held zero. The matrix H is called the *discrete-time unit impulse response matrix* of the system.

Similarly, it follows that the system (2.60) is *causal* if and only if

$$H(n,k) = 0 \quad \text{for all} \quad n < k,$$

and that the input–output description of linear, discrete-time, causal systems is given by

$$y(n) = \sum_{k=-\infty}^{n} H(n,k)u(k). \tag{2.62}$$

A discrete-time system described by (2.60) is said to be *at rest at* $k = k_0 \in Z$ if $u(k) = 0$ for $k \geq k_0$ implies that $y(k) = 0$ for $k \geq k_0$. Accordingly, if system (2.60) is known to be at rest at $k = k_0$, we have

$$y(n) = \sum_{k=k_0}^{\infty} H(n,k)u(k). \tag{2.63}$$

Moreover, if a linear discrete-time system that is at rest at k_0 is known to be causal, then its input–output description reduces to

$$y(n) = \sum_{k=k_0}^{n} H(n,k)u(k). \tag{2.64}$$

Finally, as in (2.56), it is easily shown that the unit impulse response $H(n,k)$ of a linear, *time-invariant*, discrete-time MIMO system depends only on the difference of n and k; i.e., by a slight abuse of notation we can write

$$H(n,k) = H(n-k,0) \triangleq H(n-k) \tag{2.65}$$

for all n and k. Accordingly, linear, time-invariant, causal, discrete-time MIMO systems that are at rest at $k = k_0$ are described by equations of the form

$$y(n) = \sum_{k=k_0}^{n} H(n-k)u(k). \tag{2.66}$$

We conclude by supposing that the system on hand is described by (2.14) under the assumption that $x(k_0) = 0$; i.e., the system is at rest at $k = k_0$. Then, according to (2.42) and (2.43), we obtain

$$H(n,k) = \begin{cases} C(n)\Phi(n,k+1)B(k), & n > k, \\ D(n), & n = k, \\ 0, & n < k. \end{cases} \tag{2.67}$$

Furthermore, for the time-invariant case, we obtain

$$H(n-k) = \begin{cases} CA^{n-(k+1)}B, & n > k, \\ D, & n = k, \\ 0, & n < k. \end{cases} \tag{2.68}$$

2.4.3 The Dirac Delta Distribution

For any linear time-invariant operator P from $C(R, R)$ to itself, we say that P admits an *integral representation* if there exists an integrable function (in the Riemann or Lebesgue sense), $g_p : R \to R$, such that for any $f \in C(R, R)$,

$$(Pf)(x) = (f * g_p)(x) \triangleq \int_{-\infty}^{\infty} f(\tau) g_p(x-\tau) d\tau.$$

We call g_p a *kernel of the integral representation of* P.

For the identity operator I [defined by $If = f$ for any $f \in C(R, R)$] an integral representation for which g_p is a function in the usual sense does not exist (see, e.g., Z. Szmydt, *Fourier Transformation and Linear Differential Equations*, D. Reidel Publishing Company, Boston, 1977). However, there exists a sequence of functions $\{\phi_n\}$ such that for any $f \in C(R, R)$,

$$(If)(x) = f(x) = \lim_{n\to\infty} (f * \phi_n)(x). \tag{2.69}$$

To establish (2.69) we make use of functions $\{\phi_n\}$ given by

$$\phi_n(x) = \begin{cases} n(1 - n|x|), & \text{if } |x| \le \frac{1}{n}, \\ 0, & \text{if } |x| > \frac{1}{n}, \end{cases}$$

$n = 1, 2, 3, \ldots$. A plot of ϕ_n is depicted in Figure 2.4. In Antsaklis and Michel [1], the following useful property of ϕ_n is proved.

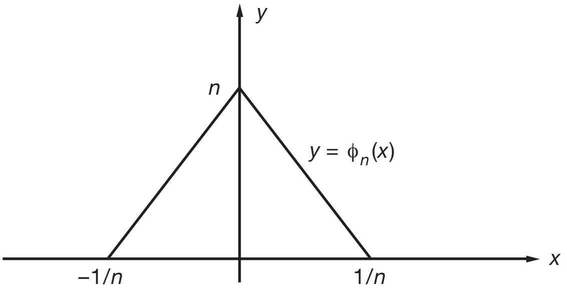


Figure 2.4. Generation of n delta distribution

Lemma 2.1. *Let f be a continuous real-valued function defined on R, and let ϕ_n be defined as above (Figure 2.4). Then for any $a \in R$,*

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} f(\tau)\phi_n(a-\tau)d\tau = f(a). \tag{2.70}$$

■

The above result, when applied to (2.69), now allows us to *define* a *generalized function* δ (also called a *distribution*) as the kernel of a *formal* or *symbolic* integral representation of the identity operator I; i.e.,

$$\begin{aligned} f(x) &= \lim_{n\to\infty} \int_{-\infty}^{\infty} f(\tau)\phi_n(x-\tau)d\tau && (2.71)\\ &\triangleq \int_{-\infty}^{\infty} f(\tau)\delta(x-\tau)d\tau && (2.72)\\ &= f * \delta(x). && (2.73)\end{aligned}$$

It is emphasized that the expression (2.72) *is not an integral at all* (in the Riemann or Lebesgue sense) *but only a symbolic representation.* The *generalized function* δ is called the *unit impulse* or the *Dirac delta distribution.*

In applications we frequently encounter functions $f \in C(R^+, R)$. If we extend f to be defined on all of R by letting $f(x) = 0$ for $x < 0$, then (2.70) becomes

$$\lim_{n\to\infty} \int_{0}^{\infty} f(\tau)\phi_n(a-\tau)d\tau = f(a) \tag{2.74}$$

for any $a > 0$, where we have used the fact that in the proof of Lemma 2.1, we need f to be continuous only in a neighborhood of a (refer to [1]). Therefore, for $f \in C(R^+, R)$, (2.71) to (2.74) yield

$$\lim_{n\to\infty} \int_{0}^{\infty} f(\tau)\phi_n(t-\tau)d\tau \triangleq \int_{0}^{\infty} f(\tau)\delta(t-\tau)d\tau = f(t) \tag{2.75}$$

for any $t > 0$. Since the ϕ_n are even functions, we have $\phi_n(t-\tau) = \phi_n(\tau - t)$, which allows for the representation $\delta(t-\tau) = \delta(\tau - t)$. We obtain from (2.75) that

$$\lim_{n\to\infty} \int_{0}^{\infty} f(\tau)\phi_n(\tau-t)d\tau \triangleq \int_{0}^{\infty} f(\tau)\delta(\tau-t)d\tau = f(t)$$

for any $t > 0$. Changing the variable $\tau' = \tau - t$, we obtain

$$\lim_{n\to\infty} \int_{-t}^{\infty} f(\tau'+t)\phi_n(\tau')d\tau' \triangleq \int_{-t}^{\infty} f(\tau'+t)\delta(\tau')d\tau' = f(t)$$

for any $t > 0$. Taking the limit $t \to 0^+$, we obtain

$$\lim_{n\to\infty} \int_{0^-}^{\infty} f(\tau'+t)\phi_n(\tau')d\tau' \triangleq \int_{0^-}^{\infty} f(\tau')\delta(\tau')d\tau' = f(0), \tag{2.76}$$

where $\int_{0-}^{\infty} f(\tau')\delta(\tau')d\tau'$ is not an integral but a symbolic representation of $\lim_{n\to\infty}\int_{0-}^{\infty} f(\tau'+t)\phi_n(\tau')d\tau'$.

Now let s denote a complex variable. If in (2.75) and (2.76) we let $f(\tau) = e^{-s\tau}, \tau > 0$, then we obtain the *Laplace transform*

$$\lim_{n\to\infty}\int_{0-}^{\infty} e^{-s\tau}\phi_n(\tau)d\tau \triangleq \int_{0-}^{\infty} e^{-s\tau}\delta(\tau)d\tau = 1. \tag{2.77}$$

Symbolically we denote (2.77) by

$$\mathcal{L}(\delta) = 1, \tag{2.78}$$

and we say that the Laplace transform of the unit impulse function or the Dirac delta distribution is equal to one.

Next, we point out another important property of δ. Consider a (time-invariant) operator P and assume that P admits an integral representation with kernel g_P. If in (2.75) we let $f = g_P$, we have

$$\lim_{n\to\infty}(P\phi_n)(t) = g_P(t), \tag{2.79}$$

and we write this (symbolically) as

$$P\delta = g_P. \tag{2.80}$$

This shows that the impulse response of a linear, time-invariant, continuous-time system with integral representation is equal to the kernel of the integral representation of the system.

Next, for any linear time-varying operator P from $C(R,R)$ to itself, we say that P admits an *integral representation* if there exists an integrable function (in the Riemann or Lebesgue sense), $g_P : R \times R \to R$, such that for any $f \in C(R,R)$,

$$(Pf)(\eta) = \int_{-\infty}^{\infty} f(\tau)g_P(\eta,\tau)d\tau. \tag{2.81}$$

Again, we call g_P a *kernel of the integral representation of* P. It turns out that *the impulse response of a linear, time-varying, continuous-time system with integral representation is again equal to the kernel of the integral representation of the system.* To see this, we first observe that if $h \in C(R\times R, R)$, and if in Lemma 2.1 we replace $f \in C(R,R)$ by h, then all the ensuing relationships still hold, with obvious modifications. In particular, as in (2.71), we have for all $t \in R$,

$$\lim_{n\to\infty}\int_{-\infty}^{\infty} h(t,\tau)\phi_n(\eta-\tau)d\tau \triangleq \int_{-\infty}^{\infty} h(t,\tau)\delta(\eta-\tau)d\tau = h(t,\eta). \tag{2.82}$$

Also, as in (2.75), we have

$$\lim_{n\to\infty}\int_{0}^{\infty} h(t,\tau)\phi_n(\eta-\tau)d\tau \triangleq \int_{0}^{\infty} h(t,\tau)\delta(\eta-\tau)d\tau = h(t,\eta) \tag{2.83}$$

for $\eta > 0$.

Now let $h(t,\tau) = g_P(t,\tau)$. Then (2.82) yields

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} g_P(t,\tau)\phi_n(\eta-\tau)d\tau \triangleq \int_{-\infty}^{\infty} g_P(t,\tau)\delta(\eta-\tau)d\tau = g_P(t,\eta), \quad (2.84)$$

which establishes our assertion. The common interpretation of (2.84) is that $g_P(t,\eta)$ represents the response of the system at time t due to an impulse applied at time η.

2.4.4 Linear Continuous-Time Systems

We let P denote a linear time-varying operator from $C(R,R^m) \triangleq U$ to $C(R,R^p) = Y$, and we assume that P admits an *integral representation* given by

$$y(t) = (Pu)(t) = \int_{-\infty}^{\infty} H_P(t,\tau)u(\tau)d\tau, \quad (2.85)$$

where $H_P : R \times R \to R^{p\times m}, u \in U$, and $y \in Y$ and where H_P is assumed to be integrable. This means that each element of H_P, $h_{P_{ij}} : R \times R \to R$ is integrable (in the Riemann or Lebesgue sense).

Now let y_1 and y_2 denote the response of system (2.85) corresponding to the input u_1 and u_2, respectively, let α_1 and α_2 be real scalars, and let y denote the response of system (2.85) corresponding to the input $\alpha_1 u_1 + \alpha_2 u_2 = u$. Then

$$\begin{aligned} y = P(u) = P(\alpha_1 u_1 + \alpha_2 u_2) &= \int_{-\infty}^{\infty} H_P(t,\tau)[\alpha_1 u_1(\tau) + \alpha_2 u_2(\tau)]d\tau \\ &= \alpha_1 \int_{-\infty}^{\infty} H_P(t,\tau)u_1(\tau)d\tau + \alpha_2 \int_{-\infty}^{\infty} H_P(t,\tau)u_2(\tau)d\tau \\ &= \alpha_1 P(u_1) + \alpha_2 P(u_2) = \alpha_1 y_1 + \alpha_2 y_2, \end{aligned} \quad (2.86)$$

which shows that system (2.85) is indeed a *linear* system in the sense defined in (2.50).

Next, we let all components of $u(\tau)$ in (2.85) be zero, except for the jth component. Then the ith component of $y(t)$ in (2.85) assumes the form

$$y_i(t) = \int_{-\infty}^{\infty} h_{P_{ij}}(t,\tau)u_j(\tau)d\tau. \quad (2.87)$$

According to the results of the previous subsection [see (2.84)], $h_{P_{ij}}(t,\tau)$ denotes the response of the ith component of the output of system (2.85), measured at time t, due to an impulse applied to the jth component of the input of system (2.85), applied at time τ, whereas all of the remaining components of the input are zero. Therefore, we call $H_P(t,\tau) = [h_{P_{ij}}(t,\tau)]$ the *impulse response matrix* of system (2.85).

Now suppose that it is known that system (2.85) is *causal.* Then its output will be identically zero before an input is applied. It follows that system (2.85) is causal if and only if

$$H_P(t,\tau) = 0 \quad \text{for all } t < \tau.$$

Therefore, when system (2.85) is causal, we have in fact that

$$y(t) = \int_{-\infty}^{t} H_P(t,\tau)u(\tau)d\tau. \tag{2.88}$$

We can rewrite (2.88) as

$$\begin{aligned} y(t) &= \int_{-\infty}^{t_0} H_P(t,\tau)u(\tau)d\tau + \int_{t_0}^{t} H_P(t,\tau)u(\tau)d\tau \\ &\triangleq y(t_0) + \int_{t_0}^{t} H_P(t,\tau)u(\tau)d\tau. \end{aligned} \tag{2.89}$$

We say that the continuous-time system (2.85) is *at rest at* $t = t_0$ if $u(t) = 0$ for $t \geq t_0$ implies that $y(t) = 0$ for $t \geq t_0$. Note that our problem formulation mandates that the system be at rest at $t_0 = -\infty$. Also, note that if a system (2.85) is known to be causal and to be at rest at $t = t_0$, then according to (2.89) we have

$$y(t) = \int_{t_0}^{t} H_P(t,\tau)u(\tau)d\tau. \tag{2.90}$$

Next, suppose that it is known that the system (2.85) is *time-invariant.* This means that if in (2.87) $h_{P_{ij}}(t,\tau)$ is the response y_i at time t due to an impulse applied at time τ at the jth component of the input [i.e., $u_j(\tau) = \delta(t)$], with all other input components set to zero, then a $-\tau$ time shift in the input [i.e., $u_j(t-\tau) = \delta(t-\tau)$] will result in a corresponding $-\tau$ time shift in the response, which results in $h_{P_{ij}}(t-\tau, 0)$. Since this argument holds for all $t, \tau \in R$ and for all $i = 1, \ldots, p$, and $j = 1, \ldots, m$, we have $H_P(t,\tau) = H_P(t-\tau, 0)$. If we define (using a slight abuse of notation) $H_P(t-\tau, 0) = H_P(t-\tau)$, then (2.85) assumes the form

$$y(t) = \int_{-\infty}^{\infty} H_P(t-\tau)u(\tau)d\tau. \tag{2.91}$$

Note that (2.91) is consistent with the definition of the integral representation of a linear time-invariant operator introduced in the previous subsection.

The right-hand side of (2.91) is the familiar *convolution integral* of H_P and u and is written more compactly as

$$y(t) = (H_P * u)(t). \tag{2.92}$$

We note that since $H_P(t-\tau)$ represents responses at time t due to impulse inputs applied at time τ, then $H_P(t)$ represents responses at time t due to impulse function inputs applied at $\tau = 0$. Therefore, a linear time-invariant system (2.91) is causal if and only if $H_P(t) = 0$ for all $t < 0$.

If it is known that the linear time-invariant system (2.91) is causal and is at rest at t_0, then we have

$$y(t) = \int_{t_0}^{t} H_P(t-\tau)u(\tau)d\tau = \int_{t_0}^{t} H_P(\tau)u(t-\tau)d\tau. \tag{2.93}$$

In this case it is customary to choose, without loss of generality, $t_0 = 0$. We thus have

$$y(t) = \int_{0}^{t} H_P(t-\tau)u(\tau)d\tau, \quad t \geq 0. \tag{2.94}$$

If we take the Laplace transform of both sides of (2.94), provided it exists, we obtain

$$\hat{y}(s) = \widehat{H}_P(s)\hat{u}(s), \tag{2.95}$$

where $\hat{y}(s) = [\hat{y}_1(s), \ldots, \hat{y}_p(s)]^T$, $\widehat{H}_P(s) = [\hat{h}_{P_{ij}}(s)]$, $\hat{u}(s) = [\hat{u}_1(s), \ldots, \hat{u}_m(s)]^T$ where the $\hat{y}_i(s)$, $\hat{u}_j(s)$, and $\hat{h}_{P_{ij}}(s)$ denote the Laplace transforms of $y_i(t)$, $u_j(t)$, and $h_{Pij}(t)$, respectively [see Chapter 3 for more details concerning Laplace transforms]. Consistent with (2.78), we note that $\widehat{H}_P(s)$ represents the Laplace transform of the impulse response matrix $H_P(t)$. We call $\widehat{H}_P(s)$ a *transfer function matrix*.

Now suppose that the input–output relation of a system is specified by the state and output equations (2.3), repeated here as

$$\dot{x} = A(t)x + B(t)u, \tag{2.96a}$$

$$y = C(t)x + D(t)u. \tag{2.96b}$$

If we assume that $x(t_0) = 0$ so that the system is at rest at $t_0 = 0$, we obtain for the response of this system,

$$y(t) = \int_{t_0}^{t} C(t)\Phi(t,\tau)B(\tau)u(\tau)d\tau + D(t)u(t) \tag{2.97}$$

$$= \int_{t_0}^{t} [C(t)\Phi(t,\tau)B(\tau) + D(t)\delta(t-\tau)]u(\tau)d\tau, \tag{2.98}$$

where in (2.98) we have made use of the interpretation of δ given in Subsection 2.4.3. Comparing (2.98) with (2.90), we conclude that the impulse response matrix for system (2.96) is given by

$$H_P(t,\tau) = \begin{cases} C(t)\Phi(t,\tau)B(\tau) + D(t)\delta(t-\tau), & t \geq \tau, \\ 0, & t < \tau. \end{cases} \tag{2.99}$$

Finally, for time-invariant systems described by the state and output equations (2.7), repeated here as

$$\dot{x} = Ax + Bu, \tag{2.100a}$$

$$y = Cx + Du, \tag{2.100b}$$

we obtain for the impulse response matrix the expression

$$H_P(t-\tau) = \begin{cases} Ce^{A(t-\tau)}B + D\delta(t-\tau), & t \geq \tau, \\ 0, & t < \tau, \end{cases} \tag{2.101}$$

or, as is more commonly written,

$$H_P(t) = \begin{cases} Ce^{At}B + D\delta(t), & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{2.102}$$

We will pursue the topics of this section further in Chapter 3.

2.5 Summary and Highlights

Internal Descriptions

- The *response of the time-varying continuous-time system*

$$\dot{x} = A(t)x + B(t)u, \quad y = C(t)x + D(t)u, \tag{2.3}$$

with $x(t_0) = x_0$ is given by

$$y(t) = C(t)\Phi(t, t_0)x_0 + C(t)\int_{t_0}^{t} \Phi(t,s)B(s)u(s)ds + D(t)u(t). \tag{2.6}$$

- The *response of the time-invariant continuous-time system*

$$\dot{x} = Ax + Bu, \quad y = Cx + Du, \tag{2.7}$$

is given by

$$y(t) = Ce^{A(t-t_0)}x_0 + C\int_{t_0}^{t} e^{A(t-s)}Bu(s)ds + Du(t). \tag{2.9}$$

- The *response of the discrete-time system*

$$x(k+1) = A(k)x(k) + B(k)u(k), \quad y(k) = C(k)x(k) + D(k)u(k), \tag{2.14}$$

with $x(k_0) = x_{k_0}$ is given by

$$y(k) = C(k)\Phi(k,k_0)x_{k_0} + C(k)\sum_{j=k_0}^{k-1}\Phi(k,j+1)B(j)u(j) \\ + D(k)u(k), \quad k > k_0 \tag{2.42}$$

and

$$y(k_0) = C(k_0)x_{k_0} + D(k_0)u(k_0), \tag{2.43}$$

where the state transition matrix

$$\Phi(k,k_0) = \prod_{j=k_0}^{k-1} A(j), \quad k > k_0, \tag{2.31}$$

$$\Phi(k_0,k_0) = I. \tag{2.32}$$

In the time-invariant case

$$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k), \tag{2.15}$$

with $x(k) = x_{k_0}$, the system response is given by

$$y(k) = CA^{k-k_0}x_{k_0} + C\sum_{j=k_0}^{k-1} A^{k-(j+1)}B(j)u(j) + Du(k), \quad k > k_0, \tag{2.44}$$

and

$$y(k_0) = Cx_{k_0} + Du(k_0). \tag{2.45}$$

External Descriptions

- Properties: *Linearity* (2.50); with *memory; causality; time-invariance*
- *The input–output description* of a *linear, discrete-time, causal, time-invariant* system that is at rest at $k = k_0$ is given by

$$y(n) = \sum_{k=k_0}^{n} h(n-k)u(k) = \sum_{k=k_0}^{n} h(k)u(n-k). \tag{2.59}$$

  $h(n-k)(= h(n-k,0))$ is the discrete-time unit impulse response of the system.
- For the *discrete-time, time-invariant system*

$$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k),$$

  the discrete-time unit impulse response (for the MIMO case) is

$$H(n-k) = \begin{cases} CA^{n-(k+1)}B, & n > k, \\ D, & n = k, \\ 0, & n < k. \end{cases} \tag{2.68}$$

- *The unit impulse* (*Dirac delta distribution*) $\delta(t)$ satisfies

$$\int_a^b f(\tau)\delta(t-\tau)d\tau = f(t),$$

where $a < t < b$ [see (2.75)].
- *The input–output description of a linear, continuous-time, causal, time-invariant system* that is at rest at $t = t_0$ is given by

$$y(t) = \int_{t_0}^t H_P(t-\tau)u(\tau)d\tau = \int_{t_0}^t H_P(\tau)u(t-\tau)d\tau. \tag{2.93}$$

$H_P(t-\tau)(= H_P(t-\tau, 0))$ is the continuous-time unit impulse response of the system.
- For the *time-invariant system*

$$\dot{x} = Ax + Bu \quad y = Cx + Du, \tag{2.100}$$

the continuous-time unit impulse response is

$$H_P(t-\tau) = \begin{cases} Ce^{A(t-\tau)}B + D\delta(t-\tau), & t \geq \tau, \\ 0, & t < \tau. \end{cases} \tag{2.101}$$

2.6 Notes

An original standard reference on linear systems is by Zadeh and Desoer [7]. Of the many excellent texts on this subject, the reader may want to refer to Brockett [2], Kailath [5], and Chen [3]. For more recent texts on linear systems, consult, e.g., Rugh [6] and DeCarlo [4]. The presentation in this book relies mostly on the recent text by Antsaklis and Michel [1].

References

1. P.J. Antsaklis and A.N. Michel, *Linear Systems*, Birkhäuser, Boston, MA, 2006.
2. R.W. Brockett, *Finite Dimensional Linear Systems*, Wiley, New York, NY, 1970.
3. C.T. Chen, *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, NY, 1984.
4. R.A. DeCarlo, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
5. T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
6. W.J. Rugh, *Linear System Theory, Second Edition*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
7. L.A. Zadeh and C.A. Desoer, *Linear System Theory - The State Space Approach*, McGraw-Hill, New York, NY, 1963.

Exercises

2.1. (a) For the mechanical system given in Exercise 1.2a, we view f_1 and f_2 as making up the system input vector, and y_1 and y_2 the system output vector. Determine a state-space description for this system.

(b) For the same mechanical system, we view $f_1 + 5f_2$ as the (scalar-valued) system input and we view $8y_1 + 10y_2$ as the (scalar-valued) system output. Determine a state-space description for this system.

(c) For part (a), determine the input–output description of the system.

(d) For part (b), determine the input–output description of the system.

2.2. In Example 1.3, we view e_a and θ as the system input and output, respectively.

(a) Detemine a state-space representation for this system.

(b) Determine the input–output description of this system.

2.3. For the second-order section digital filter in direct form, given in Figure 2.1, determine the input–output description, where $x_1(k)$ and $u(k)$ denote the output and input, respectively.

2.4. In the circuit of Figure 2.5, $v_i(t)$ and $v_0(t)$ are voltages (at time t) and R_1 and R_2 are resistors. There is also an ideal diode that acts as a short circuit when v_i is positive and as an open circuit when v_i is negative. We view v_i and v_0 as the system input and output, respectively.

(a) Determine an input–output description of this system.

(b) Is this system linear? Is it time-varying or time-invariant? Is it causal? Explain your answers.

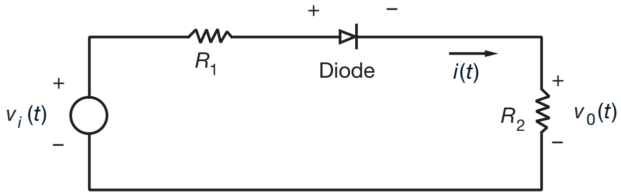


Figure 2.5. Diode circuit

2.5. We consider the *truncation operator* given by

$$y(t) = T_\tau(u(t))$$

as a system, where $\tau \in R$ is fixed, u and y denote system input and output, respectively, t denotes time, and $T_\tau(\cdot)$ is specified by

$$T_\tau(u(t)) = \begin{cases} u(t) & t \le \tau, \\ 0 & t > \tau. \end{cases}$$

Is this system causal? Is it linear? Is it time-invariant? What is its impulse response?

2.6. We consider the *shift operator* given by

$$y(t) = Q_\tau(u(t)) = u(t - \tau)$$

as a system, where $\tau \in R$ is fixed, u and y denote system input and system output, respectively, and t denotes time. Is this system causal? Is it linear? Is it time-invariant? What is its impulse response?

2.7. Consider the system whose input–output description is given by

$$y(t) = \min\{u_1(t), u_2(t)\},$$

where $u(t) = [u_1(t), u_2(t)]^T$ denotes the system input and $y(t)$ is the system output. Is this system linear?

2.8. Suppose it is known that a linear system has impulse response given by $h(t, \tau) = \exp(-|t - \tau|)$. Is this system causal? Is it time-invariant?

2.9. Consider a system with input–output description given by

$$y(k) = 3u(k + 1) + 1, \quad k \in Z,$$

where y and u denote the output and input, respectively (recall that Z denotes the integers). Is this system causal? Is it linear?

2.10. Use expression (2.54),

$$x(n) = \sum_{k=-\infty}^{\infty} x(k)\delta(n - k),$$

and $\delta(n) = p(n) - p(n - 1)$ to express the system response $y(n)$ due to any input $u(k)$, as a function of the unit step response of the system [i.e., due to $u(k) = p(k)$].

3

Response of Continuous- and Discrete-Time Systems

3.1 Introduction

In system theory it is important to clearly understand how inputs and initial conditions affect the response of a system. Many reasons exist for this. For example, in control theory, it is important to be able to select an input that will cause the system output to satisfy certain properties [e.g., to remain bounded (stability) or to follow a given trajectory (tracking)]. This is in stark contrast to the study of ordinary differential equations, where it is usually assumed that the forcing function (input) is given.

The goal of this chapter is to study the response of linear systems in greater detail than was done in Chapter 2. To this end, solutions of linear ordinary differential equations are reexamined, this time with an emphasis on characterizing all solutions using bases (of the solution vector space) and on determining such solutions. For convenience, certain results from Chapter 2 are repeated. We will find it convenient to treat continuous-time and discrete-time cases separately. Whereas in Chapter 2, certain fundamental issues that include input–output system descriptions, causality, linearity, and time-invariance are emphasized, here we will address in greater detail impulse (and pulse) response and transfer functions for continuous-time systems and discrete-time systems.

In Chapters 1 and 2 we addressed linear as well as nonlinear systems that may be time-varying or time-invariant. We considered this level of generality since this may be mandated during the modeling process of the systems. However, in the analysis and synthesis of such systems, simplified models involving linear time-invariant systems usually suffice. Accordingly, in the remainder of this book, we will emphasize linear, time-invariant continuous-time and discrete-time systems.

In this chapter, in Section 3.2, we further study linear systems of ordinary differential equations with constant coefficients. Specifically, in this section, we develop a general characterization of the solutions of such equations and we study the properties of the solutions by investigating the properties of fundamental matrices and state transition matrices. In Section 3.3 we address several methods of determining the state transition matrix and we study the asymptotic behavior of the solutions of such systems. In Sections 3.4 and 3.5, we further investigate the properties of the state representations and the input–output representations of continuous-time and discrete-time finite-dimensional systems. Specifically, in these sections we study equivalent representations of such systems, we investigate the properties of transfer function matrices, and for the discrete-time case we also address sampled data systems and the asymptotic behavior of the system response of time-invariant systems.

3.2 Solving $\dot{x} = Ax$ and $\dot{x} = Ax + g(t)$: The State Transition Matrix $\Phi(t, t_0)$

In this section we consider systems of linear homogeneous ordinary differential equations with constant coefficients.

$$\dot{x} = Ax \tag{3.1}$$

and linear nonhomogeneous ordinary differential equations

$$\dot{x} = Ax + g(t). \tag{3.2}$$

In Theorem 1.20 of Chapter 1 it was shown that these systems of equations, subject to initial conditions $x(t_0) = x_0$, possess unique solutions for every $(t_0, x_0) \in D$, where $D = \{(t, x) : t \in J = (a, b),\ x \in R^n\}$ and where it is assumed that $A \in R^{n\times n}$ and $g \in C(J, R^n)$. These solutions exist over the entire interval $J = (a, b)$, and they depend continuously on the initial conditions. Typically, we will assume that $J = (-\infty, \infty)$. We note that $\phi(t) \equiv 0$, for all $t \in J$, is a solution of (3.1), with $\phi(t_0) = 0$. We call this the *trivial solution.* As in Chapter 1 (refer to Section 1.8), we recall that the preceding statements are also true when $g(t)$ is piecewise continuous on J.

In the sequel, we sometimes will encounter the case where A is in Jordan canonical form that may have entries in the complex plane C. For this reason, we will allow $D = \{(t, x) : t \in J = (a, b),\ x \in R^n \text{ (or } x \in C^n)\}$ and $A \in R^{n\times n}$ [or $A \in C^{n\times n}$], as needed. For the case of real vectors, the field of scalars for the x-space will be the field of real numbers ($F = R$), whereas for the case of complex vectors, the field of scalars for the x-space will be the field of complex numbers ($F = C$). For the latter case, the theory concerning the existence and uniqueness of solutions for (3.1), as presented in Chapter 1, carries over and can be modified in the obvious way.

3.2.1 The Fundamental Matrix

Solution Space

In our first result we will make use of several facts concerning vector spaces, bases, and linear spaces, which are addressed in the appendix.

Theorem 3.1. *The set of solutions of (3.1) on the interval J forms an n-dimensional vector space.*

Proof. Let V denote the set of all solutions of (3.1) on J. Let $\alpha_1, \alpha_2 \in F$ ($F = R$ or $F = C$), and let $\phi_1, \phi_2 \in V$. Then $\alpha_1\phi_1 + \alpha_2\phi_2 \in V$ since $\frac{d}{dt}[\alpha_1\phi_1 + \alpha_2\phi_2] = \alpha_1\frac{d}{dt}\phi_1(t) + \alpha_2\frac{d}{dt}\phi_2(t) = \alpha_1 A\phi_1(t) + \alpha_2 A\phi_2(t) = A[\alpha_1\phi_1(t) + \alpha_2\phi_2(t)]$ for all $t \in J$. [Note that in this time-invariant case, it can be assumed without loss of generality, that $J = (-\infty, \infty)$.] This shows that V is a vector space.

To complete the proof of the theorem, we must show that V is of dimension n. To accomplish this, we must find n linearly independent solutions $\phi_1, \dots, \phi_n$ that span V. To this end, we choose a set of n linearly independent vectors $x_0^1, \dots, x_0^n$ in the n-dimensional x-space (i.e., in R^n or C^n). By the existence results in Chapter 1, if $t_0 \in J$, then there exist n solutions $\phi_1, \dots, \phi_n$ of (3.1) such that $\phi_1(t_0) = x_0^1, \dots, \phi_n(t_0) = x_0^n$. We first show that these solutions are linearly independent. If on the contrary, these solutions are linearly dependent, there exist scalars $\alpha_1, \dots, \alpha_n \in F$, not all zero, such that $\sum_{i=1}^n \alpha_i\phi_i(t) = 0$ for all $t \in J$. This implies in particular that $\sum_{i=1}^n \alpha_i\phi_i(t_0) = \sum_{i=1}^n \alpha_i x_0^i = 0$. But this contradicts the assumption that $\{x_0^1, \dots, x_0^n\}$ is a linearly independent set. Therefore, the solutions $\phi_1, \dots, \phi_n$ are linearly independent.

To conclude the proof, we must show that the solutions $\phi_1, \dots, \phi_n$ span V. Let ϕ be any solution of (3.1) on the interval J such that $\phi(t_0) = x_0$. Then there exist unique scalars $\alpha_1, \dots, \alpha_n \in F$ such that

$$x_0 = \sum_{i=1}^n \alpha_i x_0^i,$$

since, by assumption, the vectors $x_0^1, \dots, x_0^n$ form a basis for the x-space. Now

$$\psi = \sum_{i=1}^n \alpha_i\phi_i$$

is a solution of (3.1) on J such that $\psi(t_0) = x_0$. But by the uniqueness results of Chapter 1, we have that

$$\phi = \psi = \sum_{i=1}^n \alpha_i\phi_i.$$

Since ϕ was chosen arbitrarily, it follows that $\phi_1, \dots, \phi_n$ span V. ■

Fundamental Matrix and Properties

Theorem 3.1 enables us to make the following definition.

Definition 3.2. *A set of n linearly independent solutions of (3.1) on J, $\{\phi_1, \dots, \phi_n\}$, is called a* fundamental set of solutions *of (3.1), and the $n \times n$ matrix*

$$\Phi = [\phi_1, \phi_2, \ldots, \phi_n] = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2n} \\ \vdots & \vdots & & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{nn} \end{bmatrix}$$

is called a fundamental matrix *of (3.1).* ■

We note that there are infinitely many different fundamental sets of solutions of (3.1) and, hence, infinitely many different fundamental matrices for (3.1). Clearly $[\phi_1, \phi_2, \ldots, \phi_n]$ is a basis of the solution space. We now study some of the basic properties of a fundamental matrix.

In the next result, $X = [x_{ij}]$ denotes an $n \times n$ matrix, and the derivative of X with respect to t is defined as $\dot{X} = [\dot{x}_{ij}]$. Let A be the $n \times n$ matrix given in (3.1). We call the system of n^2 equations

$$\dot{X} = AX \tag{3.3}$$

a *matrix differential equation.*

Theorem 3.3. *A fundamental matrix Φ of (3.1) satisfies the matrix equation (3.3) on the interval J.*

Proof. We have

$$\dot{\Phi} = [\dot{\phi}_1, \dot{\phi}_2, \ldots, \dot{\phi}_n] = [A\phi_1, A\phi_2, \ldots, A\phi_n] = A[\phi_1, \phi_2, \ldots, \phi_n] = A\Phi.$$

■

The next result is called *Abel's formula.*

Theorem 3.4. *If Φ is a solution of the matrix equation (3.3) on an interval J and τ is any point of J, then*

$$\det \Phi(t) = \det \Phi(\tau) \exp \left[\int_{\tau}^{t} trA ds \right]$$

for every $t \in J$. [$trA = tr[a_{ij}]$ denotes the trace of A; i.e., $trA = \sum_{j=1}^{n} a_{jj}$.] ■

The proof of Theorem 3.4 is omitted. We refer the reader to [1] for a proof.

Since in Theorem 3.4 τ is arbitrary, it follows that either $\det \Phi(t) \neq 0$ for all $t \in J$ or $\det \Phi(t) = 0$ for each $t \in J$. The next result provides a test on whether an $n \times n$ matrix $\Phi(t)$ is a fundamental matrix of (3.1).

Theorem 3.5. *A solution Φ of the matrix equation (3.3) is a fundamental matrix of (3.1) if and only if its determinant is nonzero for all $t \in J$.*

Proof. If $\Phi = [\phi_1, \phi_2, \ldots, \phi_n]$ is a fundamental matrix for (3.1), then the columns of Φ, $\phi_1, \ldots, \phi_n$, form a linearly independent set. Now let ϕ be a nontrivial solution of (3.1). Then by Theorem 3.1 there exist unique scalars $\alpha_1, \ldots, \alpha_n \in F$, not all zero, such that $\phi = \sum_{j=1}^{n} \alpha_j \phi_j = \Phi a$, where $a^T = (\alpha_1, \ldots, \alpha_n)$. Let $t = \tau \in J$. Then $\phi(\tau) = \Phi(\tau)a$, which is a system of n linear algebraic equations. By construction, this system of equations has a unique solution for any choice of $\phi(\tau)$. Therefore, $\det \Phi(\tau) \neq 0$. It now follows from Theorem 3.4 that $\det \Phi(t) \neq 0$ for any $t \in J$.

Conversely, let Φ be a solution of (3.3) and assume that $\det \Phi(t) \neq 0$ for all $t \in J$. Then the columns of Φ are linearly independent for all $t \in J$. Hence, Φ is a fundamental matrix of (3.1). ■

It is emphasized that a matrix may have identically zero determinant over some interval, even though its columns are linearly independent. For example, the columns of the matrix

$$\Phi(t) = \begin{bmatrix} 1 & t & t^2 \\ 0 & 1 & t \\ 0 & 0 & 0 \end{bmatrix}$$

are linearly independent, and yet $\det \Phi(t) = 0$ for all $t \in (-\infty, \infty)$. In accordance with Theorem 3.5, the above matrix cannot be a fundamental solution of the matrix equation (3.3) for any matrix A.

Theorem 3.6. *If Φ is a fundamental matrix of (3.1) and if C is any nonsingular constant $n \times n$ matrix, then ΦC is also a fundamental matrix of (3.1). Moreover, if Ψ is any other fundamental matrix of (3.1), then there exists a constant $n \times n$ nonsingular matrix P such that $\Psi = \Phi P$.*

Proof. For the matrix ΦC we have $\frac{d}{dt}(\Phi C) = \dot{\Phi} C = [A\Phi]C = A(\Phi C)$, and therefore, ΦC is a solution of the matrix equation (3.3). Furthermore, since $\det \Phi(t) \neq 0$ for $t \in J$ and $\det C \neq 0$, it follows that $\det[\Phi(t)C] = [\det \Phi(t)](\det C) \neq 0, t \in J$. By Theorem 3.5, ΦC is a fundamental matrix.

Next, let Ψ be any other fundamental matrix of (3.1) and consider the product $\Phi^{-1}(t)\Psi$. [Notice that since $\det \Phi(t) \neq 0$ for all $t \in J$, then $\Phi^{-1}(t)$ exists for all $t \in J$.] Also, consider $\Phi\Phi^{-1} = I$ where I denotes the $n \times n$ identity matrix. Differentiating both sides, we obtain $\left(\frac{d}{dt}\Phi\right)\Phi^{-1} + \Phi\left(\frac{d}{dt}\Phi^{-1}\right) = 0$ or $\frac{d}{dt}\Phi^{-1} = -\Phi^{-1}\left(\frac{d}{dt}\Phi\right)\Phi^{-1}$. Therefore, we can compute $\frac{d}{dt}(\Phi^{-1}\Psi) = \Phi^{-1}\left(\frac{d}{dt}\Psi\right) + \left(\frac{d}{dt}\Phi^{-1}\right)\Psi = \Phi^{-1}A\Psi - [\Phi^{-1}\left(\frac{d}{dt}\Phi\right)\Phi^{-1}]\Psi = \Phi^{-1}A\Psi - (\Phi^{-1}A\Phi\Phi^{-1})\Psi = \Phi^{-1}A\Psi - \Phi^{-1}A\Psi = 0$. Hence, $\Phi^{-1}\Psi = P$ or $\Psi = \Phi P$. ■

Example 3.7. It is easily verified that the system of equations

$$\begin{aligned} \dot{x}_1 &= 5x_1 - 2x_2 \\ \dot{x}_2 &= 4x_1 - x_2 \end{aligned} \tag{3.4}$$

has two linearly independent solutions given by $\phi_1(t) = (e^{3t}, e^{3t})^T$, $\phi_2(t) = (e^t, 2e^t)^T$, and therefore, the matrix

$$\Phi(t) = \begin{bmatrix} e^{3t} & e^t \\ e^{3t} & 2e^t \end{bmatrix} \tag{3.5}$$

is a fundamental matrix of (3.4).

Using Theorem 3.6 we can find the particular fundamental matrix Ψ of (3.4) that satisfies the initial condition $\Psi(0) = I$ by using $\Phi(t)$ given in (3.5). We have $\Psi(0) = I = \Phi(0)C$ or $C = \Phi^{-1}(0)$, and therefore,

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

and

$$\Psi(t) = \Phi C = \begin{bmatrix} (2e^{3t} - e^t) & (-e^{3t} + e^t) \\ (2e^{3t} - 2e^t) & (-e^{3t} + 2e^t) \end{bmatrix}.$$

3.2.2 The State Transition Matrix

In Chapter 1 we used the *method of successive approximations* (Theorem 1.15) to prove that for every $(t_0, x_0) \in J \times R^n$,

$$\dot{x} = A(t)x \tag{3.6}$$

possesses a unique solution of the form

$$\phi(t, t_0, x_0) = \Phi(t, t_0)x_0,$$

such that $\phi(t_0, t_0, x_0) = x_0$, which exists for all $t \in J$, where $\Phi(t, t_0)$ is the *state transition matrix* (see Section 1.8). We derived an expression for $\Phi(t, t_0)$ in series form, called the *Peano–Baker series* [see (1.80) of Chapter 1], and we showed that $\Phi(t, t_0)$ is the unique solution of the matrix differential equation

$$\frac{\partial}{\partial t}\Phi(t, t_0) = A(t)\Phi(t, t_0), \tag{3.7}$$

where

$$\Phi(t_0, t_0) = I \text{ for all } t \in J. \tag{3.8}$$

Of course, these results hold for (3.1) as well.

We now provide an alternative formulation of the state transition matrix, and we study some of the properties of such matrices. Even though much of the subsequent discussion applies to system (3.6), we will confine ourselves to system (3.1). In the following definition, we use the natural basis $\{e_1, e_2, \ldots, e_n\}$ (refer to Section A.2).

Definition 3.8. *A fundamental matrix Φ of (3.1) whose columns are determined by the linearly independent solutions $\phi_1, \dots, \phi_n$ with*

$$\phi_1(t_0) = e_1, \dots, \phi_n(t_0) = e_n, \quad t_0 \in J,$$

is called the state transition matrix Φ *for (3.1). Equivalently, if Ψ is* any *fundamental matrix of (3.1), then the matrix Φ determined by*

$$\Phi(t, t_0) \triangleq \Psi(t)\Psi^{-1}(t_0) \quad \textit{for all} \quad t, t_0 \in J, \tag{3.9}$$

is said to be the state *transition matrix of (3.1).* ■

We note that the state transition matrix of (3.1) is *uniquely* determined by the matrix A and is *independent* of the particular choice of the fundamental matrix. To show this, let Ψ_1 and Ψ_2 be two different fundamental matrices of (3.1). Then by Theorem 3.6 there exists a constant $n \times n$ nonsingular matrix P such that $\Psi_2 = \Psi_1 P$. Now by the definition of state transition matrix, we have $\Phi(t, t_0) = \Psi_2(t)[\Psi_2(t_0)]^{-1} = \Psi_1(t)PP^{-1}[\Psi_1(t_0)]^{-1} = \Psi_1(t)[\Psi_1(t_0)]^{-1}$. This shows that $\Phi(t, t_0)$ is independent of the fundamental matrix chosen.

Properties of the State Transition Matrix

In the following discussion, we summarize some of the properties of state transition matrix.

Theorem 3.9. *Let $t_0 \in J$, let $\phi(t_0) = x_0$, and let $\Phi(t, t_0)$ denote the state transition matrix for (3.1) for all $t \in J$. Then the following statements are true:*

(i) *$\Phi(t, t_0)$ is the unique solution of the matrix equation $\frac{\partial}{\partial t}\Phi(t, t_0) = A\Phi(t, t_0)$ with $\Phi(t_0, t_0) = I$, the $n \times n$ identity matrix.*

(ii) *$\Phi(t, t_0)$ is nonsingular for all $t \in J$.*

(iii) *For any $t, \sigma, \tau \in J$, we have $\Phi(t, \tau) = \Phi(t, \sigma)\Phi(\sigma, \tau)$ (semigroup property).*

(iv) *$[\Phi(t, t_0)]^{-1} \triangleq \Phi^{-1}(t, t_0) = \Phi(t_0, t)$ for all $t, t_0 \in J$.*

(v) *The unique solution $\phi(t, t_0, x_0)$ of (3.1), with $\phi(t_0, t_0, x_0) = x_0$ specified, is given by*

$$\phi(t, t_0, x_0) = \Phi(t, t_0)x_0 \textit{ for all } t \in J. \tag{3.10}$$

Proof. (i) For any fundamental matrix of (3.1), say Ψ, we have, by definition, $\Phi(t, t_0) = \Psi(t)\Psi^{-1}(t_0)$, independent of the choice of Ψ. Therefore, $\frac{\partial}{\partial t}\Phi(t, t_0) = \dot{\Psi}(t)\Psi^{-1}(t_0) = A\Psi(t)\Psi^{-1}(t_0) = A\Phi(t, t_0)$. Furthermore, $\Phi(t_0, t_0) = \Psi(t_0)\Psi^{-1}(t_0) = I$.

(ii) For any fundamental matrix of (3.1) we have that $\det \Psi(t) \neq 0$ for all $t \in J$. Therefore, $\det \Phi(t, t_0) = \det[\Psi(t)\Psi^{-1}(t_0)] = \det \Psi(t) \det \Psi^{-1}(t_0) \neq 0$ for all $t, t_0 \in J$.

(iii) For any fundamental matrix Ψ of (3.1) and for the state transition matrix Φ of (3.1), we have $\Phi(t,\tau) = \Psi(t)\Psi^{-1}(\tau) = \Psi(t)\Psi^{-1}(\sigma)\Psi(\sigma)\Psi^{-1}(\tau) = \Phi(t,\sigma)\Phi(\sigma,\tau)$ for any $t,\sigma,\tau \in J$.

(iv) Let Ψ be any fundamental matrix of (3.1), and let Φ be the state transition matrix of (3.1). Then $[\Phi(t,t_0)]^{-1} = [\Psi(t)\Psi(t_0)^{-1}]^{-1} = \Psi(t_0)\Psi^{-1}(t) = \Phi(t_0,t)$ for any $t, t_0 \in J$.

(v) By the results established in Chapter 1, we know that for every $(t_0,x_0) \in D$, (3.1) has a unique solution $\phi(t)$ for all $t \in J$ with $\phi(t_0) = x_0$. To verify (3.10), we note that $\dot{\phi}(t) = \frac{\partial \Phi}{\partial t}(t,t_0)x_0 = A\Phi(t,t_0)x_0 = A\phi(t)$. ■

In Chapter 1 we pointed out that the state transition matrix $\Phi(t,t_0)$ maps the solution (state) of (3.1) at time t_0 to the solution (state) of (3.1) at time t. Since there is no restriction on t relative to t_0 (i.e., we may have $t < t_0$, $t = t_0$, or $t > t_0$), we can "move forward or backward" in time. Indeed, given the solution (state) of (3.1) at time t, we can solve the solution (state) of (3.1) at time t_0. Thus, $x(t_0) = x_0 = [\Phi(t,t_0)]^{-1}\phi(t,t_0,x_0) = \Phi(t_0,t)\phi(t,t_0,x_0)$. This *reversibility in time* is possible because $\Phi^{-1}(t,t_0)$ always exists. [In the case of discrete-time systems described by difference equations, this reversibility in time does in general not exist (refer to Section 3.5).]

3.2.3 Nonhomogeneous Equations

In Section 1.8, we proved the following result [refer to (1.87) to (1.89)].

Theorem 3.10. *Let $t_0 \in J$, let $(t_0,x_0) \in D$, and let $\Phi(t,t_0)$ denote the state transition matrix for (3.1) for all $t \in J$. Then the unique solution $\phi(t,t_0,x_0)$ of (3.2) satisfying $\phi(t_0,t_0,x_0) = x_0$ is given by*

$$\phi(t,t_0,x_0) = \Phi(t,t_0)x_0 + \int_{t_0}^{t} \Phi(t,\eta)g(\eta)d\eta. \tag{3.11}$$

■

As pointed out in Section 1.8, when $x_0 = 0$, (3.11) reduces to the *zero state response*

$$\phi(t,t_0,0) \triangleq \phi_p(t) = \int_{t_0}^{t} \Phi(t,s)g(s)ds, \tag{3.12}$$

and when $x_0 \neq 0$, but $g(t) \equiv 0$, (3.11) reduces to the *zero input response*

$$\phi(t,t_0,x_0) \triangleq \phi_h(t) = \Phi(t,t_0)x_0 \tag{3.13}$$

and the solution of (3.2) may be viewed as consisting of a component that is due to the initial data x_0 and another component that is due to the forcing term $g(t)$. We recall that ϕ_p is called a *particular solution* of the nonhomogeneous system (3.2), whereas ϕ_h is called the *homogeneous solution.*

3.3 The Matrix Exponential e^{At}, Modes, and Asymptotic Behavior of $\dot{x} = Ax$

In the time-invariant case $\dot{x} = Ax$, the state transition matrix $\Phi(t, t_0)$ equals the matrix exponential $e^{A(t-t_0)}$, which is studied in the following discussion.

Let $D = \{(t, x) : t \in R,\ x \in R^n\}$. In view of the results of Section 1.8, it follows that for every $(t_0, x_0) \in D$, the unique solution of (3.1) with $x(0) = x_0$ specified is given by

$$\begin{aligned}\phi(t, t_0, x_0) &= \left(I + \sum_{k=1}^{\infty} \frac{A^k (t - t_0)^k}{k!} \right) x_0 \\ &= \Phi(t, t_0) x_0 \triangleq \Phi(t - t_0) x_0 \triangleq e^{A(t-t_0)} x_0, \end{aligned} \tag{3.14}$$

where $\Phi(t - t_0) = e^{A(t-t_0)}$ denotes the state transition matrix for (3.1). [By writing $\Phi(t, t_0) = \Phi(t - t_0)$, we are using a slight abuse of notation.]

In arriving at (3.14) we invoked Theorem 1.15 of Chapter 1 in Section 1.5, to show that the sequence $\{\phi_m\}$, where

$$\phi_m(t, t_0, x_0) = \left(I + \sum_{k=1}^{m} \frac{A^k (t - t_0)^k}{k!} \right) x_0 \triangleq S_m(t - t_0) x_0, \tag{3.15}$$

converges uniformly and absolutely as $m \to \infty$ to the unique solution $\phi(t, t_0, x_0)$ of (3.1) given by (3.14) on compact subsets of R. In the process of arriving at this result, we also proved the following results.

Theorem 3.11. *Let A be a constant $n \times n$ matrix (which may be real or complex), and let $S_m(t)$ denote the partial sum of matrices defined by*

$$S_m(t) = I + \sum_{k=1}^{m} \frac{t^k}{k!} A^k. \tag{3.16}$$

Then each element of the matrix $S_m(t)$ converges absolutely and uniformly on any finite t interval $(-a, a), a > 0$, as $m \to \infty$. Furthermore, $\dot{S}_m(t) = AS_{m-1}(t) = S_{m-1}(t)A$, and thus, the limit of $S_m(t)$ as $t \to \infty$ is a C^1 function on R. Moreover, this limit commutes with A. ■

3.3.1 Properties of e^{At}

In view of Theorem 3.11, the following definition makes sense (see also Section 1.8).

Definition 3.12. *Let A be a constant $n \times n$ matrix (which may be real or complex). We define e^{At} to be the matrix*

$$e^{At} = I + \sum_{k=1}^{\infty} \frac{t^k}{k!} A^k \tag{3.17}$$

for any $-\infty < t < \infty$, and we call e^{At} a matrix exponential. ■

We are now in a position to provide the following characterizations of e^{At}.

Theorem 3.13. *Let $J = R, t_0 \in J$, and let A be a given constant matrix for (3.1). Then*

(i) $\Phi(t) \triangleq e^{At}$ *is a fundamental matrix for all* $t \in J$.
(ii) The state transition matrix for (3.1) is given by $\Phi(t, t_0) = e^{A(t-t_0)} \triangleq \Phi(t - t_0)$, $t \in J$.
(iii) $e^{At_1}e^{At_2} = e^{A(t_1+t_2)}$ *for all* $t_1, t_2 \in J$.
(iv) $Ae^{At} = e^{At}A$ *for all* $t \in J$.
(v) $(e^{At})^{-1} = e^{-At}$ *for all* $t \in J$.

Proof. By (3.17) and Theorem 3.11 we have that $\frac{d}{dt}[e^{At}] = \lim_{m\to\infty} AS_m(t) = \lim_{m\to\infty} S_m(t)A = Ae^{At} = e^{At}A$. Therefore, $\Phi(t) = e^{At}$ is a solution of the matrix equation $\dot{\Phi} = A\Phi$. Next, observe that $\Phi(0) = I$. It follows from Theorem 3.4 that $\det[e^{At}] = e^{trace(At)} \neq 0$ for all $t \in R$. Therefore, by Theorem 3.5 $\Phi(t) = e^{At}$ is a fundamental matrix for $\dot{x} = Ax$. We have proved parts (i) and (iv).

To prove (iii), we note that in view of Theorem 3.9(iii), we have for any $t_1, t_2 \in R$ that $\Phi(t_1, t_2) = \Phi(t_1, 0)\Phi(0, t_2)$. By Theorem 3.9(i) we see that $\Phi(t, t_0)$ solves (3.1) with $\Phi(t_0, t_0) = I$. It was just proved that $\Psi(t) \triangleq e^{A(t-t_0)}$ is also a solution. By uniqueness, it follows that $\Phi(t, t_0) = e^{A(t-t_0)}$. For $t = t_1, t_0 = -t_2$, we therefore obtain $e^{A(t_1+t_2)} = \Phi(t_1, -t_2) = \Phi(t_1)\Phi(-t_2)^{-1}$, and for $t = t_1, t_0 = 0$, we have $\Phi(t_1, 0) = e^{At_1} = \Phi(t_1)$. Also, for $t = 0, t_0 = -t_2$, we obtain $\Phi(0, -t_2) = e^{t_2 A} = \Phi(-t_2)^{-1}$. Therefore, $e^{A(t_1+t_2)} = e^{At_1}e^{At_2}$ for all $t_1, t_2 \in R$.

Finally, to prove (ii), we note that by (iii) we have $\Phi(t, t_0) \triangleq e^{A(t-t_0)} = I + \sum_{k=1}^{\infty} \frac{(t-t_0)^k}{k!}A^k = \Phi(t - t_0)$ is a fundamental matrix for $\dot{x} = Ax$ with $\Phi(t_0, t_0) = I$. Therefore, it is its state transition matrix. ■

We conclude this section by stating the solution of $\dot{x} = Ax + g(t)$,

$$\begin{aligned}
\phi(t, t_0, x_0) &= \Phi(t - t_0)x_0 + \int_{t_0}^{t} \Phi(t - s)g(s)ds \\
&= e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}g(s)ds \\
&= e^{A(t-t_0)}x_0 + e^{At}\int_{t_0}^{t} e^{-As}g(s)ds, \qquad (3.18)
\end{aligned}$$

for all $t \in R$. In arriving at (3.18), we have used expression (1.87) of Chapter 1 and the fact that in this case, $\Phi(t, t_0) = e^{A(t-t_0)}$.

3.3.2 How to Determine e^{At}

We begin by considering the specific case

$$A = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}. \tag{3.19}$$

From (3.17) it follows immediately that

$$e^{At} = I + tA = \begin{bmatrix} 1 & \alpha t \\ 0 & 1 \end{bmatrix}. \tag{3.20}$$

As another example, we consider

$$A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{3.21}$$

where $\lambda_1, \lambda_2 \in R$. Again, from (3.17) it follows that

$$\begin{aligned} e^{At} &= \begin{bmatrix} 1 + \sum_{k=1}^{\infty} \frac{t^k}{k!}\, \lambda_1^k & 0 \\ 0 & 1 + \sum_{k=1}^{\infty} \frac{t^k}{k!}\, \lambda_2^k \end{bmatrix} \\ &= \begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix}. \end{aligned} \tag{3.22}$$

Unfortunately, in general it is much more difficult to evaluate the matrix exponential than the preceding examples suggest. In the following discussion, we consider several methods of evaluating e^{At}.

The Infinite Series Method

In this case we evaluate the partial sum $S_m(t)$ (see Theorem 3.11)

$$S_m(t) = I + \sum_{k=1}^{m} \frac{t^k}{k!}\, A^k$$

for some fixed t, say, t_1, and for $m = 1, 2, \ldots$ until no significant changes occur in succeeding sums. This yields the matrix e^{At_1}. This method works reasonably well if the smallest and largest real parts of the eigenvalues of A are not widely separated.

In the same spirit as above, we could use any of the vector differential solvers to solve $\dot{x} = Ax$, using the natural basis for R^n as n linearly independent initial conditions [i.e., using as initial conditions the vectors $e_1 = (1, 0, \ldots, 0)^T$, $e_2 = (0, 1, 0, \ldots, 0)^T, \ldots, e_n = (0, \ldots, 0, 1)^T$] and observing that in view of (3.14), the resulting solutions are the columns of e^{At} (with $t_0 = 0$).

Example 3.14. There are cases when the definition of e^{At} (in series form) directly produces a closed-form expression. This occurs for example when $A^k = 0$ for some k. In particular, if all the eigenvalues of A are at the origin, then $A^k = 0$ for some $k \leq n$. In this case, only a finite number of terms in (3.17) will be nonzero and e^{At} can be evaluated in closed form. This was precisely the case in (3.19).

The Similarity Transformation Method

Let us consider the initial-value problem

$$\dot{x} = Ax, \quad x(t_0) = x_0; \tag{3.23}$$

let P be a real $n \times n$ nonsingular matrix, and consider the transformation $x = Py$, or equivalently, $y = P^{-1}x$. Differentiating both sides with respect to t, we obtain $\dot{y} = P^{-1}\dot{x} = P^{-1}APy = Jy, y(t_0) = y_0 = P^{-1}x_0$. The solution of the above equation is given by

$$\psi(t, t_0, y_0) = e^{J(t-t_0)}P^{-1}x_0. \tag{3.24}$$

Using (3.24) and $x = Py$, we obtain for the solution of (3.23),

$$\phi(t, t_0, x_0) = Pe^{J(t-t_0)}P^{-1}x_0. \tag{3.25}$$

Now suppose that the similarity transformation P given above has been chosen in such a manner that

$$J = P^{-1}AP \tag{3.26}$$

is in Jordan canonical form (refer to Section A.6). We first consider the case when A has n linearly independent eigenvectors, say, v_i, that correspond to the eigenvalues λ_i (not necessarily distinct), $i = 1, \dots, n$. (Necessary and sufficient conditions for this to be the case are given in Sections A.5 and A.6. A sufficient condition for the eigenvectors v_i, $i = 1, \dots, n$, to be linearly independent is that the eigenvalues of $A, \lambda_1, \dots, \lambda_n$, be distinct.) Then P can be chosen so that $P = [v_1, \dots, v_n]$ and the matrix $J = P^{-1}AP$ assumes the form

$$J = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}. \tag{3.27}$$

Using the power series representation

$$e^{Jt} = I + \sum_{k=1}^{\infty} \frac{t^k J^k}{k!}, \tag{3.28}$$

we immediately obtain the expression

$$e^{Jt} = \begin{bmatrix} e^{\lambda_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n t} \end{bmatrix}. \tag{3.29}$$

Accordingly, the solution of the initial-value problem (3.23) is now given by

$$\phi(t, t_0, x_0) = P \begin{bmatrix} e^{\lambda_1 (t-t_0)} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_n (t-t_0)} \end{bmatrix} P^{-1} x_0. \tag{3.30}$$

In the general case when A has repeated eigenvalues, it is no longer possible to diagonalize A (see Section A.6). However, we can generate n linearly independent vectors $v_1, \ldots, v_n$ and an $n \times n$ similarity transformation $P = [v_1, \ldots, v_n]$ that takes A into the Jordan canonical form $J = P^{-1}AP$. Here J is in the block diagonal form given by

$$J = \begin{bmatrix} J_0 & & & 0 \\ & J_1 & & \\ & & \ddots & \\ 0 & & & J_s \end{bmatrix}, \tag{3.31}$$

where J_0 is a diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_k$ (not necessarily distinct), and each $J_i, i \geq 1$, is an $n_i \times n_i$ matrix of the form

$$J_i = \begin{bmatrix} \lambda_{k+i} & 1 & 0 & \cdots & 0 \\ 0 & \lambda_{k+i} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda_{k+i} \end{bmatrix}, \tag{3.32}$$

where λ_{k+i} need not be different from λ_{k+j} if $i \neq j$, and where $k + n_1 + \cdots + n_s = n$.

Now since for any square block diagonal matrix

$$C = \begin{bmatrix} C_1 & & 0 \\ & \ddots & \\ 0 & & C_l \end{bmatrix}$$

with $C_i, i = 1, \ldots, l$, square, we have that

$$C^k = \begin{bmatrix} C_1^k & & 0 \\ & \ddots & \\ 0 & & C_l^k \end{bmatrix},$$

it follows from the power series representation of e^{Jt} that

$$e^{Jt} = \begin{bmatrix} e^{J_0 t} & & & 0 \\ & e^{J_1 t} & & \\ & & \ddots & \\ 0 & & & e^{J_s t} \end{bmatrix}, \tag{3.33}$$

$t \in R$. As shown earlier, we have

$$e^{J_0 t} = \begin{bmatrix} e^{\lambda_1 t} & & 0 \\ & \ddots & \\ 0 & & e^{\lambda_k t} \end{bmatrix}. \tag{3.34}$$

For J_i, $i = 1, \dots, s$, we have

$$J_i = \lambda_{k+i} I_i + N_i, \tag{3.35}$$

where I_i denotes the $n_i \times n_i$ identity matrix and N_i is the $n_i \times n_i$ nilpotent matrix given by

$$N_i = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 \\ 0 & \cdots & \cdots & 0 \end{bmatrix}. \tag{3.36}$$

Since $\lambda_{k+i} I_i$ and N_i commute, we have that

$$e^{J_i t} = e^{\lambda_{k+i} t} e^{N_i t}. \tag{3.37}$$

Repeated multiplication of N_i by itself results in $N_i^k = 0$ for all $k \geq n_i$. Therefore, the series defining e^{tN_i} terminates, resulting in

$$e^{tJ_i} = e^{\lambda_{k+i} t} \begin{bmatrix} 1 & t & \cdots & t^{n_i - 1}/(n_i - 1)! \\ 0 & 1 & \cdots & t^{n_i - 2}/(n_i - 2)! \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}, \quad i = 1, \dots, s. \tag{3.38}$$

It now follows that the solution of (3.23) is given by

$$\phi(t, t_0, x_0) = P \begin{bmatrix} e^{J_0(t-t_0)} & 0 & \cdots & 0 \\ 0 & e^{J_1(t-t_0)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & e^{J_s(t-t_0)} \end{bmatrix} P^{-1} x_0. \tag{3.39}$$

Example 3.15. In system (3.23), let $A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}$. The eigenvalues of A are $\lambda_1 = -1$ and $\lambda_2 = 1$, and corresponding eigenvectors for A are given by $v_1 = (1, 0)^T$ and $v_2 = (1, 1)^T$, respectively. Then $P = [v_1, v_2] = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, $P^{-1} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$, and $J = P^{-1} A P = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, as expected. We obtain $e^{At} = P e^{Jt} P^{-1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} e^{-t} & 0 \\ 0 & e^t \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} e^t & e^t - e^{-t} \\ 0 & e^t \end{bmatrix}$.

Suppose next that in (3.23) the matrix A is either in *companion form* or that it has been transformed into this form via some suitable similarity transformation P, so that $A = A_c$, where

$$A_c = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}. \tag{3.40}$$

Since in this case we have $x_{i+1} = \dot{x}_i$, $i = 1, \dots, n-1$, it should be clear that in the calculation of e^{At} we need to determine, via some method, only the first row of e^{At}. We demonstrate this by means of a specific example.

Example 3.16. In system (3.23), assume that $A = A_c = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$, which is in companion form. To demonstrate the above observation, let us compute e^{At} by some other method, say diagonalization. The eigenvalues of A are $\lambda_1 = -1$ and $\lambda_2 = -2$, and a set of corresponding eigenvectors is given by $v_1 = (1, -1)^T$ and $v_2 = (1, -2)^T$. We obtain $P = [v_1, v_2] = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix}$, $P^{-1} = \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix}$ and $J = P^{-1} A_c P = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$, $e^{At} = P e^{Jt} P^{-1} = \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} e^{-t} & 0 \\ 0 & e^{-2t} \end{bmatrix} \begin{bmatrix} 2 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} (2e^{-t} - e^{-2t}) & (e^{-t} - e^{-2t}) \\ (-2e^{-t} + 2e^{-2t}) & (-e^{-t} + 2e^{-2t}) \end{bmatrix}$. Note that the second row of the above matrix is the derivative of the first row, as expected.

The Cayley–Hamilton Theorem Method

If $\alpha(\lambda) = \det(\lambda I - A)$ is the characteristic polynomial of an $n \times n$ matrix A, we have that $\alpha(A) = 0$, in view of the Cayley–Hamilton Theorem; i.e., every $n \times n$ matrix satisfies its characteristic equation (refer to Sections A.5 and A.6). Using this result, along with the series definition of the matrix exponential e^{At}, it is easily shown that

$$e^{At} = \sum_{i=0}^{n-1} \alpha_i(t) A^i \tag{3.41}$$

[Refer to Sections A.5 and A.6 for the details on how to determine the terms $\alpha_i(t)$.]

The Laplace Transform Method

We assume that the reader is familiar with the basics of the (one-sided) Laplace transform. If $f(t) = [f_1(t), \dots, f_n(t)]^T$, where $f_i : [0, \infty) \to R$, $i = 1, \dots, n$, and if each f_i is Laplace transformable, then we define the Laplace transform of the vector f component-wise; i.e., $\hat{f}(s) = [\hat{f}_1(s), \dots, \hat{f}_n(s)]^T$, where $\hat{f}_i(s) = \mathcal{L}[f_i(t)] \triangleq \int_0^\infty f_i(t)e^{-st}dt$.

We define the Laplace transform of a matrix $C(t) = [c_{ij}(t)]$ similarly. Thus, if each $c_{ij} : [0, \infty) \to R$ and if each c_{ij} is Laplace transformable, then the Laplace transform of $C(t)$ is defined as $\widehat{C}(s) = \mathcal{L}[c_{ij}(t)] = [\mathcal{L}c_{ij}(t)] = [\hat{c}_{ij}(s)]$.

Laplace transforms of some of the common time signals are enumerated in Table 3.1. Also, in Table 3.2 we summarize some of the more important properties of the Laplace transform. In Table 3.1, $\delta(t)$ denotes the *Dirac delta distribution* (see Subsection 2.4.3) and $p(t)$ represents the *unit step function.*

Table 3.1. Laplace transforms

$f(t)(t \geq 0)$	$\hat{f}(s) = \mathcal{L}[f(t)]$
$\delta(t)$	1
$p(t)$	$1/s$
$t^k/k!$	$1/s^{k+1}$
e^{-at}	$1/(s+a)$
$t^k e^{-at}$	$k!/(s+a)^{k+1}$
$e^{-at}\sin bt$	$b/[(s+a)^2 + b^2]$
$e^{-at}\cos bt$	$(s+a)/[(s+a)^2 + b^2]$

Table 3.2. Laplace transform properties

Time differentiation	$df(t)/dt$	$s\hat{f}(s) - f(0)$
	$d^k f(t)/dt^k$	$s^k\hat{f}(s) - [s^{k-1}f(0) + \cdots + f^{(k-1)}(0)]$
Frequency shift	$e^{-at}f(t)$	$\hat{f}(s+a)$
Time shift	$f(t-a)p(t-a), a > 0$	$e^{-as}\hat{f}(s)$
Scaling	$f(t/\alpha), \alpha > 0$	$\alpha\hat{f}(\alpha s)$
Convolution	$\int_0^t f(\tau)g(t-\tau)d\tau = f(t) * g(t)$	$\hat{f}(s)\hat{g}(s)$
Initial value	$\lim_{t \to 0^+} f(t) = f(0^+)$	$\lim_{s \to \infty} s\hat{f}(s)^\dagger$
Final value	$\lim_{t \to \infty} f(t)$	$\lim_{s \to 0} s\hat{f}(s)^\ddagger$

† If the limit exists.

‡ If $s\hat{f}(s)$ has no singularities on the imaginary axis or in the right half s plane.

Now consider once more the initial-value problem (3.23), letting $t_0 = 0$; i.e.,

$$\dot{x} = Ax, \quad x(0) = x_0. \tag{3.42}$$

Taking the Laplace transform of both sides of $\dot{x} = Ax$, and taking into account the initial condition $x(0) = x_0$, we obtain $s\hat{x}(s) - x_0 = A\hat{x}(s)$, $(sI - A)\hat{x}(s) = x_0$, or

$$\hat{x}(s) = (sI - A)^{-1}x_0. \tag{3.43}$$

It can be shown by analytic continuation that $(sI-A)^{-1}$ exists for all s, except at the eigenvalues of A. Taking the inverse Laplace transform of (3.43), we obtain the solution

$$\phi(t) = \mathcal{L}^{-1}[(sI - A)^{-1}]x_0 = \Phi(t, 0)x_0 = e^{At}x_0. \tag{3.44}$$

It follows from (3.42) and (3.44) that $\hat{\Phi}(s) = (sI - A)^{-1}$ and that

$$\Phi(t, 0) \triangleq \Phi(t - 0) = \Phi(t) = \mathcal{L}^{-1}[(sI - A)^{-1}] = e^{At}. \tag{3.45}$$

Finally, note that when $t_0 \neq 0$, we can immediately compute $\Phi(t, t_0) = \Phi(t - t_0) = e^{A(t-t_0)}$.

Example 3.17. In (3.42), let $A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}$. Then

$$(sI - A)^{-1} = \begin{bmatrix} s+1 & -2 \\ 0 & s-1 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{s+1} & \frac{2}{(s+1)(s-1)} \\ 0 & \frac{1}{s-1} \end{bmatrix} = \begin{bmatrix} \frac{1}{s+1} & \left(\frac{1}{s-1} - \frac{1}{s+1}\right) \\ 0 & \frac{1}{s-1} \end{bmatrix}.$$

Using Table 3.1, we obtain $\mathcal{L}^{-1}[(sI - A)^{-1}] = e^{At} = \begin{bmatrix} e^{-t} & (e^t - e^{-t}) \\ 0 & e^t \end{bmatrix}$.

Before concluding this subsection, we briefly consider initial-value problems described by

$$\dot{x} = Ax + g(t), \quad x(t_0) = x_0. \tag{3.46}$$

We wish to apply the Laplace transform method discussed above in solving (3.46). To this end we assume $t_0 = 0$ and we take the Laplace transform of both sides of (3.46) to obtain $s\hat{x}(s) - x_0 = A\hat{x}(s) + \hat{g}(s)$, $(sI - A)\hat{x}(s) = x_0 + \hat{g}(s)$, or

$$\begin{aligned} \hat{x}(s) &= (sI - A)^{-1}x_0 + (sI - A)^{-1}\hat{g}(s) \\ &= \hat{\Phi}(s)x_0 + \hat{\Phi}(s)\hat{g}(s) \\ &\triangleq \hat{\phi}_h(s) + \hat{\phi}_p(s). \end{aligned} \tag{3.47}$$

Taking the inverse Laplace transform of both sides of (3.47) and using (3.18) with $t_0 = 0$, we obtain $\phi(t) = \phi_h(t) + \phi_p(t) = \mathcal{L}^{-1}[(sI - A)^{-1}]x_0 + \mathcal{L}^{-1}[(sI - A)^{-1}\hat{g}(s)] = \Phi(t)x_0 + \int_0^t \Phi(t - \eta)g(\eta)d\eta$, where ϕ_h denotes the homogeneous solution and ϕ_p is the particular solution, as expected.

Example 3.18. Consider the initial-value problem given by

$$\begin{aligned} \dot{x}_1 &= -x_1 + x_2, \\ \dot{x}_2 &= -2x_2 + u(t), \end{aligned}$$

with $x_1(0) = -1$, $x_2(0) = 0$, and

$$u(t) = \begin{cases} 1 & \text{for } t > 0, \\ 0 & \text{for } t \leq 0. \end{cases}$$

It is easily verified that in this case

$$\begin{aligned}
\hat{\Phi}(s) &= \begin{bmatrix} \frac{1}{s+1} & \left(\frac{1}{s+1} - \frac{1}{s+2}\right) \\ 0 & \frac{1}{s+2} \end{bmatrix}, \\
\Phi(t) &= \begin{bmatrix} e^{-t} & (e^{-t} - e^{-2t}) \\ 0 & e^{-2t} \end{bmatrix}, \\
\phi_h(t) &= \begin{bmatrix} e^{-t} & (e^{-t} - e^{-2t}) \\ 0 & e^{-t} \end{bmatrix} \begin{bmatrix} -1 \\ 0 \end{bmatrix} = \begin{bmatrix} -e^{-t} \\ 0 \end{bmatrix}, \\
\hat{\phi}_p(s) &= \begin{bmatrix} \frac{1}{s+1} & \left(\frac{1}{s+1} - \frac{1}{s+2}\right) \\ 0 & \frac{1}{s+2} \end{bmatrix} \begin{bmatrix} 0 \\ \frac{1}{s} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\left(\frac{1}{s}\right) + \frac{1}{2}\left(\frac{1}{s+2}\right) - \frac{1}{s+1} \\ \frac{1}{2}\left(\frac{1}{s}\right) - \frac{1}{2}\left(\frac{1}{s+2}\right) \end{bmatrix}, \\
\phi_p(t) &= \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2t} - e^{-t} \\ \frac{1}{2} - \frac{1}{2}e^{-2t} \end{bmatrix},
\end{aligned}$$

and

$$\phi(t) = \phi_h(t) + \phi_p(t) = \begin{bmatrix} \frac{1}{2} - 2e^{-t} + \frac{1}{2}e^{-2t} \\ \frac{1}{2} - \frac{1}{2}e^{-2t} \end{bmatrix}.$$

3.3.3 Modes, Asymptotic Behavior, and Stability

In this subsection we study the qualitative behavior of the solutions of $\dot{x} = Ax$ by means of the modes of such systems, to be introduced shortly. Although we will not address the stability of systems in detail until Chapter 4, the results here will enable us to give some general stability characterizations for such systems.

Modes: General Case

We begin by recalling that the unique solution of

$$\dot{x} = Ax, \tag{3.48}$$

satisfying $x(0) = x_0$, is given by

$$\phi(t, 0, x_0) = \Phi(t, 0)x(0) = \Phi(t, 0)x_0 = e^{At}x_0. \tag{3.49}$$

We also recall that $\det(sI - A) = \prod_{i=1}^{\sigma}(s - \lambda_i)^{n_i}$, where $\lambda_1, \dots, \lambda_\sigma$ denote the σ distinct eigenvalues of A, where λ_i with $i = 1, \dots, \sigma$, is assumed to be repeated n_i times (i.e., n_i is the algebraic multiplicity of λ_i), and $\Sigma_{i=1}^{\sigma} n_i = n$.

To introduce the modes for (3.48), we must show that

$$\begin{aligned} e^{At} &= \sum_{i=1}^{\sigma} \sum_{k=0}^{n_i - 1} A_{ik} t^k e^{\lambda_i t} \\ &= \sum_{i=1}^{\sigma} [A_{i0} e^{\lambda_i t} + A_{i1} t e^{\lambda_i t} + \cdots + A_{i(n_i-1)} t^{n_i - 1} e^{\lambda_i t}], \end{aligned} \tag{3.50}$$

where

$$A_{ik} = \frac{1}{k!} \frac{1}{(n_i - 1 - k)!} \lim_{s \to \lambda_i} [[(s - \lambda_i)^{n_i} (sI - A)^{-1}]^{(n_i - 1 - k)}]. \tag{3.51}$$

In (3.51), $[\cdot]^{(l)}$ denotes the lth derivative with respect to s.

Equation (3.50) shows that e^{At} can be expressed as the sum of terms of the form $A_{ik} t^k e^{\lambda_i t}$, where $A_{ik} \in R^{n \times n}$. We call $A_{ik} t^k e^{\lambda_i t}$ *a mode of system (3.48)*. If an eigenvalue λ_i is repeated n_i times, there are n_i modes, $A_{ik} t^k e^{\lambda_i t}$, $k = 0, 1, \dots, n_i - 1$, in e^{At} associated with λ_i. Accordingly, the solution (3.49) of (3.48) is determined by the n modes of (3.48) corresponding to the n eigenvalues of A and by the initial condition $x(0)$. We note that by selecting $x(0)$ appropriately, modes can be combined or eliminated $[A_{ik} x(0) = 0]$, thus affecting the behavior of $\phi(t, 0, x_0)$.

To verify (3.50) we recall that $e^{At} = \mathcal{L}^{-1}[(sI - A)^{-1}]$ and we make use of the partial fraction expansion method to determine the inverse Laplace transform. As in the scalar case, it can be shown that

$$(sI - A)^{-1} = \sum_{i=1}^{\sigma} \sum_{k=0}^{n_i - 1} (k! A_{ik})(s - \lambda_i)^{-(k+1)}, \tag{3.52}$$

where the $(k! A_{ik})$ are the coefficients of the partial fractions ($k!$ is for scaling). It is known that these coefficients can be evaluated for each i by multiplying both sides of (3.52) by $(s - \lambda_i)^{n_i}$, differentiating $(n_i - 1 - k)$ times with respect to s, and then evaluating the resulting expression at $s = \lambda_i$. This yields (3.51). Taking the inverse Laplace transform of (3.52) and using the fact that $\mathcal{L}[t^k e^{\lambda_i t}] = k!(s - \lambda_i)^{-(k+1)}$ (refer to Table 3.1) results in (3.50).

When all n eigenvalues λ_i of A are distinct, then $\sigma = n, n_i = 1, i = 1, \dots, n$, and (3.50) reduces to the expression

$$e^{At} = \sum_{i=1}^{n} A_i e^{\lambda_i t}, \tag{3.53}$$

where

$$A_i = \lim_{s \to \lambda_i} [(s - \lambda_i)(sI - A)^{-1}]. \tag{3.54}$$

Expression (3.54) can also be derived directly, using a partial fraction expansion of $(sI - A)^{-1}$ given in (3.52).

Example 3.19. For (3.48) we let $A = \begin{bmatrix} 0 & 1 \\ -4 & -4 \end{bmatrix}$, for which the eigenvalue $\lambda_1 = -2$ is repeated twice; i.e., $n_1 = 2$. Applying (3.50) and (3.51), we obtain

$$e^{At} = A_{10} e^{\lambda_1 t} + A_{11} t e^{\lambda_1 t} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} e^{-2t} + \begin{bmatrix} 2 & 1 \\ -4 & -2 \end{bmatrix} t e^{-2t}.$$

Example 3.20. For (3.48) we let $A = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix}$, for which the eigenvalues are given by (the complex conjugate pair) $\lambda_1 = -\frac{1}{2} + j\frac{\sqrt{3}}{2}, \lambda_2 = -\frac{1}{2} - j\frac{\sqrt{3}}{2}$. Applying (3.53) and (3.54), we obtain

$$A_1 = \frac{1}{\lambda_1 - \lambda_2} \begin{bmatrix} \lambda_1 + 1 & 1 \\ -1 & \lambda_1 \end{bmatrix} = \frac{1}{j\sqrt{3}} \begin{bmatrix} \frac{1}{2} + j\frac{\sqrt{3}}{2} & 1 \\ -1 & -\frac{1}{2} + j\frac{\sqrt{3}}{2} \end{bmatrix}$$

$$A_2 = \frac{1}{\lambda_2 - \lambda_1} \begin{bmatrix} \lambda_2 + 1 & 1 \\ -1 & \lambda_2 \end{bmatrix} = \frac{1}{-j\sqrt{3}} \begin{bmatrix} \frac{1}{2} - j\frac{\sqrt{3}}{2} & 1 \\ -1 & -\frac{1}{2} - j\frac{\sqrt{3}}{2} \end{bmatrix}$$

[i.e., $A_1 = A_2^*$, where $(\cdot)^*$ denotes the complex conjugate of $(\cdot)$], and

$$\begin{aligned} e^{At} &= A_1 e^{\lambda_1 t} + A_2 e^{\lambda_2 t} = A_1 e^{\lambda_1 t} + A_1^* e^{\lambda_1^* t} \\ &= 2(Re\, A_1)(Re\, e^{\lambda_1 t}) - 2(Im\, A_1)(Im\, e^{\lambda_1 t}) \\ &= 2e^{-\frac{1}{2}t} \left[\begin{bmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{bmatrix} \cos \frac{\sqrt{3}}{2} t - \begin{bmatrix} -\frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{2\sqrt{3}} \end{bmatrix} \sin \left(\frac{\sqrt{3}}{2} t \right) \right]. \end{aligned}$$

The last expression involves only real numbers, as expected, since A and e^{At} are real matrices.

Example 3.21. For (3.48) we let $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, for which the eigenvalue $\lambda_1 = 1$ is repeated twice; i.e., $n_1 = 2$. Applying (3.50) and (3.51), we obtain

$$e^{At} = A_{10} e^{\lambda_1 t} + A_{11} t e^{\lambda_1 t} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} e^{t} + \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} t e^{t} = I e^{t}.$$

This example shows that not all modes of the system are necessarily present in e^{At}. What is present depends in fact on the number and dimensions of the individual blocks of the Jordan canonical form of A corresponding to identical eigenvalues. To illustrate this further, we let for (3.48), $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, where the two repeated eigenvalues $\lambda_1 = 1$ belong to the same Jordan block. Then $e^{At} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} e^t + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} te^t$.

Stability of an Equilibrium

In Chapter 4 we will study the *qualitative properties* of linear dynamical systems, including systems described by (3.48). This will be accomplished by studying the *stability properties* of such systems or, more specifically, the *stability properties* of an *equilibrium* of such systems.

If $\phi(t, 0, x_e)$ denotes the solution of system (3.48) with $x(0) = x_e$, then x_e is said to be an *equilibrium* of (3.48) if $\phi(t, 0, x_e) = x_e$ for all $t \geq 0$. Clearly, $x_e = 0$ is an equilibrium of (3.48). In discussing the qualitative properties, it is often customary to speak, somewhat loosely, of the *stability properties of system (3.48)*, rather than the stability properties of the equilibrium $x_e = 0$ of system (3.48).

We will show in Chapter 4 that the following qualitative characterizations of system (3.48) are actually *equivalent* to more fundamental qualitative characterizations of the equilibrium $x_e = 0$ of system (3.48):

1. The system (3.48) is said to be *stable* if all solutions of (3.48) are bounded for all $t \geq 0$ [i.e., for any $\phi(t, 0, x_0) = (\phi_1(t, 0, x_0), \ldots, \phi_n(t, 0, x_0))^T$ of (3.48), there exist constants M_i, $i = 1, \ldots, n$ (which in general will depend on the solution on hand) such that $|\phi_i(t, 0, x_0)| < M_i$ for all $t \geq 0$].
2. The system (3.48) is said to be *asymptotically stable* if it is stable and if all solutions of (3.48) tend to the origin as t tends to infinity [i.e., for any solution $\phi(t, 0, x_0) = (\phi_1(t, 0, x_0), \ldots, \phi_n(t, 0, x_0))^T$ of (3.48), we have $\lim_{t\to\infty} \phi_i(t, 0, x_0) = 0$, $i = 1, \ldots, n$].
3. The system (3.48) is said to be *unstable* if it is not stable.

By inspecting the modes of (3.48) given by (3.50), (3.51) and (3.53), (3.54), the following stability criteria for system (3.48) are now evident:

1. The system (3.48) is *asymptotically stable* if and only if all eigenvalues of A have negative real parts (i.e., $Re\lambda_j < 0$, $j = 1, \ldots, n$).
2. The system (3.48) is *stable* if and only if $Re\lambda_j \leq 0$, $j = 1, \ldots, n$, and for all eigenvalues with $Re\lambda_j = 0$ having multiplicity $n_j > 1$, it is true that
$$\lim_{s\to\lambda_j} [(s - \lambda_j)^{n_j} (sI - A)^{-1}]^{(n_j - 1 - k)} = 0, \quad k = 1, \ldots, n_j - 1. \qquad (3.55)$$
3. System (3.48) is *unstable* if and only if (2) is not true.

We note in particular that if $Re\lambda_j = 0$ and $n_j > 1$, then there will be modes $A_{jk}t^k$, $k = 0, \dots, n_j - 1$ that will yield terms in (3.50) whose norm will tend to infinity as $t \to \infty$, unless their coefficients are zero. This shows why the necessary and sufficient conditions for stability of (3.48) include condition (3.55).

Example 3.22. The systems in Examples 3.19 and 3.20 are asymptotically stable. A system (3.48) with $A = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}$ is stable, since the eigenvalues of A above are $\lambda_1 = 0$, $\lambda_2 = -1$. A system (3.48) with $A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ is unstable since the eigenvalues of A are $\lambda_1 = 1$, $\lambda_2 = -1$. The system of Example 3.21 is also unstable.

Modes: Distinct Eigenvalue Case

When the eigenvalues λ_i of A are distinct, there is an alternative way to (3.54) of computing the matrix coefficients A_i, expressed in terms of the corresponding right and left eigenvectors of A. This method offers great insight into questions concerning the presence or absence of modes in the response of a system. Specifically, if A has n distinct eigenvalues λ_i, then

$$e^{At} = \sum_{i=1}^{n} A_i e^{\lambda_i t}, \tag{3.56}$$

where

$$A_i = v_i \tilde{v}_i, \tag{3.57}$$

where $v_i \in R^n$ and $(\tilde{v}_i)^T \in R^n$ are right and left eigenvectors of A corresponding to the eigenvalue λ_i, respectively.

To prove the above assertions, we recall that $(\lambda_i I - A)v_i = 0$ and $\tilde{v}_i(\lambda_i I - A) = 0$. If $Q \triangleq [v_1, \dots, v_n]$, then the $\tilde{v}_i$ are the rows of

$$P = Q^{-1} = \begin{bmatrix} \tilde{v}_1 \\ \vdots \\ \tilde{v}_n \end{bmatrix}.$$

The matrix Q is of course nonsingular, since the eigenvalues λ_i, $i = 1, \dots, n$, are by assumption distinct and since the corresponding eigenvectors are linearly independent. Notice that $Q \operatorname{diag}[\lambda_1, \dots, \lambda_n] = AQ$ and that $\operatorname{diag}[\lambda_1, \dots, \lambda_n]P = PA$. Also, notice that $\tilde{v}_i v_j = \delta_{ij}$, where

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

In view of this, we now have $(sI - A)^{-1} = [sI - Q\,\mathrm{diag}[\lambda_1, \ldots, \lambda_n]Q^{-1}]^{-1} = Q[sI - \mathrm{diag}[\lambda_1, \cdots, \lambda_n]]^{-1}Q^{-1} = Q\,\mathrm{diag}[(s-\lambda_1)^{-1}, \ldots, (s-\lambda_n)^{-1}]Q^{-1} = \sum_{i=1}^{n} v_i\tilde{v}_i(s-\lambda_i)^{-1}$. If we take the inverse Laplace transform of the above expression, we obtain (3.56).

If we choose the initial value $x(0)$ for (3.48) to be colinear with an eigenvector v_j of A [i.e., $x(0) = \alpha v_j$ for some real $\alpha \neq 0$], then $e^{\lambda_j t}$ is the only mode that will appear in the solution ϕ of (3.48). This can easily be seen from our preceding discussion. In particular if $x(0) = \alpha v_j$, then (3.56) and (3.57) yield

$$\phi(t, 0, x(0)) = e^{At}x(0) = v_1\tilde{v}_1 x(0)e^{\lambda_1 t} + \cdots + v_n\tilde{v}_n x(0)e^{\lambda_n t} = \alpha v_j e^{\lambda_j t} \quad (3.58)$$

since $\tilde{v}_i v_j = 1$ when $i = j$, and $\tilde{v}_i v_j = 0$ otherwise.

Example 3.23. In (3.48) we let $A = \begin{bmatrix} -1 & 1 \\ 0 & 1 \end{bmatrix}$. The eigenvalues of A are given by $\lambda_1 = -1$ and $\lambda_2 = 1$ and $Q = [v_1, v_2] = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$, $Q^{-1} = \begin{bmatrix} \tilde{v}_1 \\ \tilde{v}_2 \end{bmatrix} = \begin{bmatrix} 1 & -1/2 \\ 0 & 1/2 \end{bmatrix}$. Then $e^{At} = v_1\tilde{v}_1 e^{\lambda_1 t} + v_2\tilde{v}_2 e^{\lambda_2 t} = \begin{bmatrix} 1 & -1/2 \\ 0 & 0 \end{bmatrix} e^{-t} + \begin{bmatrix} 0 & 1/2 \\ 0 & 1 \end{bmatrix} e^{t}$. If in particular we choose $x(0) = \alpha v_1 = (\alpha, 0)^T$, then $\phi(t, 0, x(0)) = e^{At}x(0) = \alpha(1, 0)^T e^{-t}$, which contains only the mode corresponding to the eigenvalue $\lambda_1 = -1$. Thus, for this particular choice of initial vector, the unstable behavior of the system is suppressed.

Remark

We conclude our discussion of modes and asymptotic behavior by briefly considering systems of linear, nonhomogeneous, ordinary differential equations $\dot{x} = Ax + g(t)$ in (3.2) for the special case where $g(t) = Bu(t)$,

$$\dot{x} = Ax + Bu, \quad (3.59)$$

where $B \in R^{n\times m}, u : R \to R^m$, and where it is assumed that the Laplace transform of u exists. Taking the Laplace transform of both sides of (3.59) and rearranging yields

$$\hat{x}(s) = (sI - A)^{-1}x(0) + (sI - A)^{-1}B\hat{u}(s). \quad (3.60)$$

By taking the inverse Laplace transform of (3.60), we see that the solution ϕ is the sum of modes that correspond to the singularities or poles of $(sI - A)^{-1}x(0)$ and of $(sI - A)^{-1}B\hat{u}(s)$. If in particular (3.48) is asymptotically stable (i.e., for $\dot{x} = Ax, Re\lambda_i < 0$, $i = 1, \ldots, n$) and if u in (3.59) is bounded (i.e., there is an M such that $|u_i(t)| < M$ for all $t \geq 0$, $i = 1, \ldots, m$), then it is easily seen that the solutions of (3.59) are bounded as well. Thus, the fact that the system (3.48) is asymptotically stable has repercussions on the asymptotic behavior of the solution of (3.59). Issues of this type will be addressed in greater detail in Chapter 4.

3.4 State Equation and Input–Output Description of Continuous-Time Systems

This section consists of three subsections. We first study the response of linear continuous-time systems. Next, we examine transfer functions of linear time-invariant systems, given the state equations of such systems. Finally, we explore the equivalence of internal representations of systems.

3.4.1 Response of Linear Continuous-Time Systems

We consider once more systems described by linear equations of the form

$$\dot{x} = Ax + Bu, \tag{3.61a}$$

$$y = Cx + Du, \tag{3.61b}$$

where $A \in R^{n\times n}$, $B \in R^{n\times m}$, $C \in R^{p\times n}, D \in R^{p\times m}$, and $u : R \to R^m$ is assumed to be continuous or piecewise continuous. We recall that in (3.61), x denotes the state vector, u denotes the system input, and y denotes the system output. From Section 2.2 we recall that for given initial conditions $t_0 \in R, x(t_0) = x_0 \in R^n$ and for a given input u, the unique solution of (3.61a) is given by

$$\phi(t, t_0, x_0) = \varPhi(t, t_0)x_0 + \int_{t_0}^{t} \varPhi(t, s)Bu(s)ds \tag{3.62}$$

for $t \in R$, where $\varPhi$ denotes the state transition matrix of A. Furthermore, by substituting (3.62) into (3.61b), we obtain, for all $t \in R$, the *total system response* given by

$$y(t) = C\varPhi(t, t_0)x_0 + C\int_{t_0}^{t} \varPhi(t, s)Bu(s)ds + Du(t). \tag{3.63}$$

Recall that the total response (3.63) may be viewed as consisting of the sum of two components, the *zero-input response* given by the term

$$\psi(t, t_0, x_0, 0) = C\varPhi(t, t_0)x_0 \tag{3.64}$$

and the *zero-state response* given by the term

$$\rho(t, t_0, 0, u) = C\int_{t_0}^{t} \varPhi(t, s)Bu(s)ds + Du(t). \tag{3.65}$$

The cause of the former is the initial condition x_0 [and can be obtained from (3.63) by letting $u(t) \equiv 0$], whereas for the latter the cause is the input u [and can be obtained by setting $x_0 = 0$ in (3.63)].

The zero-state response can be used to introduce the *impulse response* of the system (3.61). We recall from Subsection 2.4.3 that by using the Dirac delta distribution δ, we can rewrite (3.63) with $x_0 = 0$ as

$$\begin{aligned} y(t) &= \int_{t_0}^{t} [C\Phi(t,\tau)B + D\delta(t-\tau)]u(\tau)d\tau \\ &= \int_{t_0}^{t} H(t,\tau)u(\tau)d\tau, \end{aligned} \tag{3.66}$$

where $H(t,\tau)$ denotes the impulse response matrix of system (3.61) given by

$$H(t,\tau) = \begin{cases} C\Phi(t,\tau)B + D\delta(t-\tau), & t \geq \tau, \\ 0, & t < \tau. \end{cases} \tag{3.67}$$

Now recall that

$$\Phi(t,t_0) = e^{A(t-t_0)}. \tag{3.68}$$

The solution of (3.61a) is thus given by

$$\phi(t,t_0,x_0) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)}Bu(s)ds, \tag{3.69}$$

the *total response* of system (3.61) is given by

$$y(t) = Ce^{A(t-t_0)}x_0 + C\int_{t_0}^{t} e^{A(t-s)}Bu(s)ds + Du(t) \tag{3.70}$$

and the *zero-state response* of (3.61), is given by $y(t) = \int_{t_0}^{t} [Ce^{A(t-\tau)}B+ D\delta(t-\tau)]u(\tau)d\tau = \int_{t_0}^{t} H(t,\tau)u(\tau)d\tau = \int_{t_0}^{t} H(t-\tau)u(\tau)d\tau$, where the *impulse response matrix* H of system (3.61) is given by

$$H(t-\tau) = \begin{cases} Ce^{A(t-\tau)}B + D\delta(t-\tau), & t \geq \tau, \\ 0, & t < \tau, \end{cases} \tag{3.71}$$

or, as is more commonly written,

$$H(t) = \begin{cases} Ce^{At}B + D\delta(t), & t \geq 0, \\ 0, & t < 0. \end{cases} \tag{3.72}$$

At this point it may be worthwhile to consider some specific cases.

Example 3.24. In (3.61), let $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $C = [0,1], D = 0$ and consider the case when $t_0 = 0, x(0) = (1,-1)^T$, u is the unit step, and $t \geq 0$. We can easily compute the solution of (3.61a) as

$$\phi(t, t_0, x_0) = \phi_h(t, t_0, x_0) + \phi_p(t, t_0, x_0) = \begin{bmatrix} 1-t \\ -1 \end{bmatrix} + \begin{bmatrix} \frac{1}{2}t^2 \\ t \end{bmatrix}$$

with $t_0 = 0$ and for $t \geq 0$. The total system response $y(t) = Cx(t)$ is given by the sum of the zero-input response and the zero-state response, $y(t, t_0, x_0, u) = \psi(t, t_0, x_0, 0) + \rho(t, t_0, 0, u) = -1 + t$, $t \geq 0$.

Example 3.25. Consider the time-invariant system given above in Example 3.24. It is easily verified that in the present case

$$\Phi(t) = e^{At} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}.$$

Then $H(t, \tau) = Ce^{A(t-\tau)}B = 1$ for $t \geq \tau$ and $H(t, \tau) = 0$ for $t < \tau$. Thus, the response of this system to an impulse input for zero initial conditions is the unit step.

As one might expect, external descriptions of finite-dimensional linear systems are not as complete as internal descriptions of such systems. Indeed, the utility of impulse responses is found in the fact that they represent the input–output relations of a system quite well, assuming that the system is at rest. To describe other dynamic behavior, one needs in general additional information [e.g., the initial state vector (or perhaps the history of the system input since the last time instant when the system was at rest) as well as the internal structure of the system].

Internal descriptions, such as state-space representations, constitute more complete descriptions than external descriptions. However, the latter are simpler to apply than the former. Both types of representations are useful. It is quite straightforward to obtain external descriptions of systems from internal descriptions, as was demonstrated in this section. The reverse process, however, is not quite as straightforward. The process of determining an internal system description from an external description is called *realization* and will be addressed in Chapter 8. The principal issue in system realization is to obtain minimal order internal descriptions that model a given system, avoiding the generation of unnecessary dynamics.

3.4.2 Transfer Functions

Next, if as in [(2.95) in Chapter 2], we take the Laplace transform of (3.71), we obtain the input–output relation

$$\hat{y}(s) = \widehat{H}(s)\hat{u}(s). \tag{3.73}$$

We recall from Section 2.4 that $\widehat{H}(s)$ is called the *transfer function matrix* of system (3.61). We can evaluate this matrix in a straightforward manner by first taking the Laplace transform of both sides of (3.61a) and (3.61b) to obtain

$$s\hat{x}(s) - x(0) = A\hat{x}(s) + B\hat{u}(s), \tag{3.74}$$

$$\hat{y}(s) = C\hat{x}(s) + D\hat{u}(s). \tag{3.75}$$

Using (3.74) to solve for $\hat{x}(s)$, we obtain

$$\hat{x}(s) = (sI - A)^{-1}x(0) + (sI - A)^{-1}B\hat{u}(s). \tag{3.76}$$

Substituting (3.76) into (3.75) yields

$$\hat{y}(s) = C(sI - A)^{-1}x(0) + C(sI - A)^{-1}B\hat{u}(s) + D\hat{u}(s) \tag{3.77}$$

and

$$y(t) = \mathcal{L}^{-1}\hat{y}(s) = Ce^{At}x(0) + C\int_0^t e^{A(t-s)}Bu(s)ds + Du(t), \tag{3.78}$$

as expected.

If in (3.77) we let $x(0) = 0$, we obtain the Laplace transform of the zero-state response given by

$$\begin{aligned}\hat{y}(s) &= [C(sI - A)^{-1}B + D]\hat{u}(s)\\ &= \widehat{H}(s)\hat{u}(s),\end{aligned} \tag{3.79}$$

where $\widehat{H}(s)$ denotes the transfer function of system (3.61), given by

$$\widehat{H}(s) = C(sI - A)^{-1}B + D. \tag{3.80}$$

Recalling that $\mathcal{L}[e^{At}] = \Phi(s) = (sI - A)^{-1}$ [refer to (3.45)], we could of course have obtained (3.80) directly by taking the Laplace transform of $H(t)$ given in (3.73).

Example 3.26. In Example 3.24, let $t_0 = 0$ and $x(0) = 0$. Then

$$\begin{aligned}\widehat{H}(s) = C(sI - A)^{-1}B + D &= [0, 1]\begin{bmatrix} s & -1 \\ 0 & s \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ 1 \end{bmatrix}\\ &= [0, 1]\begin{bmatrix} 1/s & 1/s^2 \\ 0 & 1/s \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = 1/s\end{aligned}$$

and $H(t) = \mathcal{L}^{-1}\widehat{H}(s) = 1$ for $t \geq 0$, as expected (see Example 3.24).

Next, as in Example 3.24, let $x(0) = (1, -1)^T$ and let u be the unit step. Then $\hat{y}(s) = C(sI - A)^{-1}x(0) + \widehat{H}(s)\hat{u}(s) = [0, 1/s](1, -1)^T + (1/s)(1/s) = -1/s + 1/s^2$ and $y(t) = \mathcal{L}^{-1}[\hat{y}(s)] = -1 + t$ for $t \geq 0$, as expected (see Example 3.24).

We note that the eigenvalues of the matrix A in Example 3.26 are the roots of the equation $\det(sI - A) = s^2 = 0$, and are given by $s_1 = 0, s_2 = 0$, whereas the transfer function $\widehat{H}(s)$ in this example has only one pole (the zero of its denominator polynomial), located at the origin. It will be shown in Chapter 8 (on realization) that the *poles of the transfer function* $\widehat{H}(s)$ *(of a SISO system)* are in general a subset of the eigenvalues of A. In Chapter 5 we will introduce and study two important system theoretic concepts, called *controllability* and *observability*. We will show in Chapter 8 that the eigenvalues of A are precisely the poles of the transfer function $\widehat{H}(s) = C(sI - A)^{-1}B + D$ if and only if the system (3.61) is observable and controllable. This is demonstrated in the next example.

Example 3.27. In (3.61), let $A = \begin{bmatrix} 0 & 1 \\ -1 & -2 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $C = [-3, 3], D = 0$. The eigenvalues of A are the roots of the equation $\det(sI - A) = s^2 + 2s + 1 = (s+1)^2 = 0$ given by $s_1 = -1, s_2 = -1$, and the transfer function of this SISO system is given by

$$\begin{aligned}\widehat{H}(s) &= C(sI - A)^{-1}B + D = [-3, 3] \begin{bmatrix} s & -1 \\ 1 & s+2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= 3[-1, 1] \frac{1}{(s+1)^2} \begin{bmatrix} s+2 & 1 \\ -1 & s \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{3(s-1)}{(s+1)^2},\end{aligned}$$

with poles (the zeros of the denominator polynomial) also given by $s_1 = -1, s_2 = -1$.

If in Example 3.27 we replace $B = [0, 1]^T$ and $D = 0$ by $B = \begin{bmatrix} 0 & -1/2 \\ 1 & 1/2 \end{bmatrix}$ and $D = [0, 0]$, then we have a multi-input system whose transfer function is given by

$$\widehat{H}(s) = \left[\frac{3(s-1)}{(s+1)^2}, \frac{3}{(s+1)} \right].$$

The concepts of poles and zeros for MIMO systems (also called multivariable systems) will be introduced in Chapter 7. The determination of the poles of such systems is not as straightforward as in the case of SISO systems. It turns out that in the present case the poles of $\widehat{H}(s)$ are $s_1 = -1, s_2 = -1$, the same as the eigenvalues of A.

Before proceeding to our next topic, the equivalence of internal representations, an observation concerning the transfer function $\widehat{H}(s)$ of system (3.61), given by (3.80), $\widehat{H}(s) = C(sI - A)^{-1}B + D$ is in order. Since the numerator matrix polynomial of $(sI - A)^{-1}$ is of degree $(n - 1)$, while its denominator polynomial, the characteristic polynomial $\alpha(s)$ of A, is of degree n, it is clear that

$$\lim_{s\to\infty} \widehat{H}(s) = D,$$

a real-valued $m \times n$ matrix, and in particular, when the *direct link matrix* D in the output equation (3.61b) is zero, then

$$\lim_{s\to\infty} \widehat{H}(s) = 0,$$

the $m \times n$ matrix with zeros as its entries. In the former case (when $D \neq 0$), $\widehat{H}(s)$ is said to be a *proper transfer function*, whereas in the latter case (when $D = 0$), $\widehat{H}(s)$ is said to be a *strictly proper transfer function*.

When discussing the realization of transfer functions by state-space descriptions (in Chapter 8), we will study the properties of transfer functions in greater detail. In this connection, there are also systems that can be described by models corresponding to transfer functions $\widehat{H}(s)$ that are *not proper.* The differential equation representation of a differentiator (or an inductor) given by $y(t) = (d/dt)u(t)$ is one such example. Indeed, in this case the system cannot be represented by (3.61) and the transfer function, given by $\widehat{H}(s) = s$ is not proper. Such systems will be discussed in Chapter 10.

3.4.3 Equivalence of State-Space Representations

In Subsection 3.3.2 it was shown that when a linear, autonomous, homogeneous system of first-order ordinary differential equations $\dot{x} = Ax$ is subjected to an appropriately chosen similarity transformation, the resulting set of equations may be considerably easier to use and may exhibit latent properties of the system of equations. It is therefore natural that we consider a similar course of action in the case of the linear systems (3.61).

We begin by letting

$$\tilde{x} = Px, \tag{3.81}$$

where P is a real, nonsingular matrix (i.e., P is a similarity transformation). Consistent with what has been said thus far, we see that such transformations bring about a *change of basis* for the state space of system (3.61). Application of (3.81) to this system will result, as will be seen, in a system description of the same form as (3.61), but involving different state variables. We will say that the system (3.61), and the system obtained by subjecting (3.61) to the transformation (3.81), constitute *equivalent internal representations* of an underlying system. We will show that equivalent internal representations (of the same system) possess identical external descriptions, as one would expect, by showing that they have identical impulse responses and transfer function matrices. In connection with this discussion, two important notions called *zero-input equivalence* and *zero-state equivalence* of a system will arise in a natural manner.

If we differentiate both sides of (3.81), and if we apply $x = P^{-1}\tilde{x}$ to (3.61), we obtain the equivalent internal representation of (3.61) given by

$$\dot{\tilde{x}} = \widetilde{A}\tilde{x} + \widetilde{B}u, \tag{3.82a}$$

$$y = \widetilde{C}\tilde{x} + \widetilde{D}u, \tag{3.82b}$$

where

$$\widetilde{A} = PAP^{-1}, \quad \widetilde{B} = PB, \quad \widetilde{C} = CP^{-1}, \quad \widetilde{D} = D \tag{3.83}$$

and where $\tilde{x}$ is given by (3.81). It is now easily verified that the system (3.61) and the system (3.82) have the same external representation. Recall that for (3.61) and for (3.82), we have for the impulse response

$$H(t,\tau) \triangleq H(t-\tau,0) = \begin{cases} Ce^{A(t-\tau)}B + D\delta(t-\tau), & t \geq \tau, \\ 0, & t < \tau, \end{cases} \tag{3.84}$$

and

$$\widetilde{H}(t,\tau) \triangleq \widetilde{H}(t-\tau,0) = \begin{cases} \widetilde{C}e^{\widetilde{A}(t-\tau)}\widetilde{B} + \widetilde{D}\delta(t-\tau), & t \geq \tau, \\ 0, & t < \tau. \end{cases} \tag{3.85}$$

Recalling from Subsection 3.3.2 [see (3.25)] that

$$e^{\widetilde{A}(t-\tau)} = Pe^{A(t-\tau)}P^{-1}, \tag{3.86}$$

we obtain from (3.83)–(3.85) that $\widetilde{C}e^{\tilde{A}(t-\tau)}\widetilde{B}+\widetilde{D}\delta(t-\tau)=CP^{-1}Pe^{A(t-\tau)}P^{-1}PB+D\delta(t-\tau)=Ce^{A(t-\tau)}B+D\delta(t-\tau)$, which proves, in view of (3.84) and (3.85), that

$$\widetilde{H}(t,\tau) = H(t,\tau), \tag{3.87}$$

and this in turn shows that

$$\widehat{\widetilde{H}}(s) = \widehat{H}(s). \tag{3.88}$$

This last relationship can also be verified by observing that $\widehat{\widetilde{H}}(s) = \widetilde{C}(sI - \widetilde{A})^{-1}\widetilde{B} + \widetilde{D} = CP^{-1}(sI - PAP^{-1})^{-1}PB + D = CP^{-1}P(sI-A)^{-1}P^{-1}PB + D = C(sI-A)^{-1}B + D = \widehat{H}(s)$.

Next, recall that in view of (3.70) we have for (3.61) that

$$\begin{aligned} y(t) &= Ce^{A(t-t_0)}x_0 + \int_{t_0}^{t} H(t-\tau,0)u(\tau)d\tau \\ &= \psi(t,t_0,x_0,0) + \rho(t,t_0,0,u) \end{aligned} \tag{3.89}$$

and for (3.82) that

$$\begin{aligned} y(t) &= \widetilde{C}e^{\widetilde{A}(t-t_0)}\tilde{x}_0 + \int_{t_0}^{t} \widetilde{H}(t-\tau,0)u(\tau)d\tau \\ &= \widetilde{\psi}(t,t_0,\tilde{x}_0,0) + \tilde{\rho}(t,t_0,0,u) \end{aligned} \tag{3.90}$$

where ψ and $\widetilde{\psi}$ denote the zero-input response of (3.61) and (3.82), respectively, whereas ρ and $\tilde{\rho}$ denote the zero-state response of (3.61) and (3.82), respectively. The relations (3.89) and (3.90) give rise to the following concepts: Two state-space representations are *zero-state equivalent* if they give rise to the same impulse response (the same external description). Also, two state-space representations are *zero-input equivalent* if for any initial state vector for one representation there exists an initial state vector for the second representation such that the zero-input responses for the two representations are identical.

The following result is now clear: *If two state-space representations are equivalent, then they are both zero-state and zero-input equivalent.* They are clearly zero-state equivalent since $H(t,\tau) = \widetilde{H}(t,\tau)$. Also, in view of (3.89) and (3.90), we have $\widetilde{C}e^{\widetilde{A}(t-t_0)}\tilde{x}_0 = (CP^{-1})[Pe^{A(t-t_0)}P^{-1}]\tilde{x}_0 = Ce^{A(t-t_0)}x_0$, where (3.86) was used. Therefore, the two state representations are also zero-input equivalent.

The converse to the above result is in general not true, since there are representations that are both zero-state and zero-input equivalent, yet not equivalent. In Chapter 8, which deals with state-space realizations of transfer functions, we will consider this topic further.

Example 3.28. System (3.61) with

$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [-1, -5], \quad D = 1$$

has the transfer function

$$H(s) = C(sI - A)^{-1}B + D = \frac{-5s - 1}{s^2 + 3s + 2} + 1 = \frac{(s-1)^2}{(s+1)(s+2)}.$$

Using the similarity transformation

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

yields the equivalent representation of the system given by

$$\widetilde{A} = PAP^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}, \quad \widetilde{B} = PB = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad \widetilde{C} = CP^{-1} = [4, 9]$$

and $\widetilde{D} = D = 1$. Note that the columns of P^{-1}, given by $[1, -1]^T$ and $[1, -2]^T$, are eigenvectors of A corresponding to the eigenvalues $\lambda_1 = -1, \lambda_2 = -2$ of A; that is, P was chosen to diagonalize A. Notice that A (which is in companion form) has characteristic polynomial $s^2+3s+2 = (s+1)(s+2)$. Notice also that the eigenvectors given above are of the form $[1, \lambda_i]^T$, $i = 1, 2$. The transfer function of the equivalent representation of the system is now given by

$$\begin{aligned}\hat{\tilde{H}}(s) &= \widetilde{C}(sI - \widetilde{A})^{-1}\widetilde{B} + \widetilde{D} = [4, 0]\begin{bmatrix} \frac{1}{s+1} & 0 \\ 0 & \frac{1}{s+2} \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} + 1 \\ &= \frac{-5s - 1}{(s+1)(s+2)} + 1 = H(s).\end{aligned}$$

Finally, it is easily verified that $e^{\widetilde{A}t} = Pe^{At}P^{-1}$.

From the above discussion it should be clear that systems [of the form (3.61)] described by equivalent representations have identical behavior to the *outside world*, since both their zero-input and zero-state responses are the same. Their states, however, are in general not identical, but are related by the transformation $\tilde{x}(t) = Px(t)$.

3.5 State Equation and Input–Output Description of Discrete-Time Systems

In this section, which consists of five subsections, we address the state equation and input–output description of linear discrete-time systems. In the first subsection we study the response of linear time-invariant systems described by the difference equations (2.15) [or (1.8)]. In the second subsection we consider transfer functions for linear time-invariant systems, whereas in the third subsection we address the equivalence of the internal representations of time-invariant linear discrete-time systems [described by (2.15)]. Some of the most important classes of discrete-time systems include linear sampled-data systems that we develop in the fourth subsection. In the final part of this section, we address the modes and asymptotic behavior of linear time-invariant discrete-time systems.

3.5.1 Response of Linear Discrete-Time Systems

We consider once again systems described by linear time-invariant equations of the form

$$x(k+1) = Ax(k) + Bu(k), \tag{3.91a}$$

$$y(k) = Cx(k) + Du(k), \tag{3.91b}$$

where $A \in R^{n\times n}$, $B \in R^{n\times m}$, $C \in R^{p\times n}$, and $D \in R^{p\times m}$. We recall that in (3.91), x denotes the state vector, u denotes the system input, and y denotes the system output. For given initial conditions $k_0 \in Z, x(k_0) = x_{k_0} \in R^n$ and for a given input u, equation (3.91a) possesses a unique solution $x(k)$, which is defined for all $k \geq k_0$, and thus, the response $y(k)$ for (3.91b) is also defined for all $k \geq k_0$.

Associated with (3.91a) is the linear autonomous, homogeneous system of equations given by

$$x(k+1) = Ax(k). \tag{3.92}$$

We recall from Section 2.3 that the solution of the initial-value problem

$$x(k+1) = Ax(k), \quad x(k_0) = x_{k_0} \tag{3.93}$$

is given by

$$x(k) = \Phi(k, k_0)x_{k_0} = A^{k-k_0}x_{k_0}, \quad k > k_0, \tag{3.94}$$

where $\Phi(k, k_0)$ denotes the state transition matrix of (3.92) with

$$\Phi(k, k) = I \tag{3.95}$$

[refer to (2.31) to (2.34) in Chapter 2].

Common properties of the state transition matrix $\Phi(k, l)$, such as for example the *semigroup property* (forward in time) given by

$$\Phi(k, l) = \Phi(k, m)\Phi(m, l), \quad k \geq m \geq l,$$

can quite easily be derived from (3.94), (3.95). We caution the reader, however, that not all of the properties of the state transition matrix $\Phi(t, \tau)$ for continuous-time systems $\dot{x} = Ax$ carry over to the discrete-time case (3.92). In particular we recall that if for the continuous-time case we have $t > \tau$, then future values of the state ϕ at time t can be obtained from past values of the state ϕ at time τ, and vice versa, from the relationships $\phi(t) = \Phi(t, \tau)\phi(\tau)$ and $\phi(\tau) = \Phi^{-1}(t, \tau)\phi(t) = \Phi(\tau, t)\phi(t)$, i.e., for continuous-time systems a principle of *time reversibility exists*. This principle is in general not true for system (3.92), unless $A^{-1}(k)$ exists. The reason for this lies in the fact that $\Phi(k, l)$ will not be nonsingular if A is not nonsingular.

Example 3.29. In (3.94), let $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $x(0) = \begin{bmatrix} 1 \\ \alpha \end{bmatrix}$, $\alpha \in R$. The initial state $x(0)$ at $k_0 = 0$ for *any* $\alpha \in R$ will map into the state $x(1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Accordingly, in this case, time reversibility will not apply.

Example 3.30. In (3.93), let $A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}$. In view of (3.94) we have that $A^{(k-k_0)} = \begin{bmatrix} (-1)^{(k-k_0)} & 1-(-1)^{(k-k_0)} \\ 0 & 1 \end{bmatrix}$, $k \geq k_0$; i.e., $A^{(k-k_0)} = A$ when $(k-k_0)$ is odd, and $A^{(k-k_0)} = I$ when $(k-k_0)$ is even. Therefore, given $k_0 = 0$ and $x(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, then $x(k) = Ax(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $k = 1, 3, 5, \ldots$, and

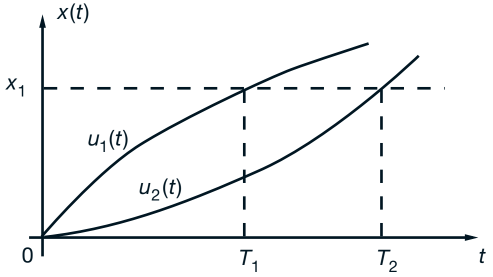


Figure 3.1. Plots of states for Example 3.30

$x(k) = Ix(0) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $k = 2, 4, 6, \ldots$. A plot of the states $x(k) = [x_1(k), x_2(k)]^T$ is given in Figure 3.1.

Continuing, we recall that the solutions of initial-value problems determined by linear nonhomogeneous systems (2.35) are given by expression (2.36). Utilizing (2.36), the solution of (3.91a) for given $x(k_0)$ and $u(k)$ is given as

$$x(k) = \Phi(k, k_0)x(k_0) + \sum_{j=k_0}^{k-1} \Phi(k, j+1)Bu(j), \quad k > k_0. \tag{3.96}$$

This expression in turn can be used to determine the system response for system (3.91) as

$$\begin{aligned} y(k) &= C\Phi(k, k_0)x(k_0) \\ &\quad + \sum_{j=k_0}^{k-1} C\Phi(k, j+1)Bu(j) + Du(k), \quad k > k_0, \\ y(k_0) &= Cx(k_0) + Du(k_0), \end{aligned} \tag{3.97}$$

or

$$\begin{aligned} y(k) &= CA^{(k-k_0)}x(k_0) + \sum_{j=k_0}^{k-1} CA^{k-(j+1)}Bu(j) + Du(k), \quad k > k_0, \\ y(k_0) &= Cx(k_0) + Du(k_0). \end{aligned} \tag{3.98}$$

Since the system (3.91) is time-invariant, we can let $k_0 = 0$ without loss of generality to obtain from (3.98) the expression

$$y(k) = CA^k x(0) + \sum_{j=0}^{k-1} CA^{k-(j+1)}Bu(j) + Du(k), \quad k > 0. \tag{3.99}$$

As in the continuous-time case, the *total system response* (3.97) may be viewed as consisting of two components, the *zero-input response*, given by

$$\psi(k) = C\Phi(k, k_0)x(k_0), \quad k > k_0,$$

and the *zero-state response*, given by

$$\left.\begin{aligned} \rho(k) &= \sum_{j=k_0}^{k-1} C\Phi(k, j+1)Bu(j) + Du(k), && k > k_0, \\ \rho(k_0) &= Du(k_0), && k = k_0. \end{aligned}\right\} \tag{3.100}$$

Finally, in view of (2.67), we recall that the (discrete-time) unit impulse response matrix of system (3.91) is given by

$$H(k,l) = \begin{cases} CA^{k-(l+1)}B, & k > l, \\ D, & k = l, \\ 0, & k < l, \end{cases} \tag{3.101}$$

and in particular, when $l = 0$ (i.e., when the pulse is applied at time $l = 0$),

$$H(k,0) = \begin{cases} CA^{k-1}B, & k > 0, \\ D, & k = 0, \\ 0, & k < 0. \end{cases} \tag{3.102}$$

Example 3.31. In (3.91), let

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C^T = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad D = 0.$$

We first determine A^k by using the Cayley–Hamilton Theorem (refer to Section A.5). To this end we compute the eigenvalues of A as $\lambda_1 = 0, \lambda_2 = -1$, we let $A^k = f(A)$, where $f(s) = s^k$, and we let $g(s) = \alpha_1 s + \alpha_0$. Then $f(\lambda_1) = g(\lambda_1)$, $\alpha_0 = 0$ and $f(\lambda_2) = g(\lambda_2)$, or $(-1)^k = -\alpha_1 + \alpha_0$. Therefore, $A^k = \alpha_1 A + \alpha_0 I = -(-1)^k \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 0 & (-1)^{k-1} \\ 0 & (-1)^k \end{bmatrix}$, $k = 1, 2, \ldots$, or $A^k = \begin{bmatrix} \delta(k) & (-1)^{k-1}p(k-1) \\ 0 & (-1)^k p(k) \end{bmatrix}$, $k = 0, 1, 2, \ldots$, where $A^0 = I$, and where $p(k)$ denotes the unit step given by

$$p(k) = \begin{cases} 1, & k \geq 0, \\ 0, & k < 0. \end{cases}$$

The above expression for A^k is now substituted into (3.98) to determine the response $y(k)$ for $k > 0$ for a given initial condition $x(0)$ and a given input $u(k), k \geq 0$. To determine the unit impulse response, we note that $H(k,0) = 0$ for $k < 0$ and $k = 0$. When $k > 0, H(k,0) = CA^{k-1}B = (-1)^{k-2}p(k-2)$ for $k > 0$ or $H(k,0) = 0$ for $k = 1$ and $H(k,0) = (-1)^{k-2}$ for $k = 2, 3, \ldots$.

3.5.2 The Transfer Function and the z-Transform

We assume that the reader is familiar with the concept and properties of the *one-sided z-transform* of a real-valued sequence $\{f(k)\}$, given by

$$\mathcal{Z}\{f(k)\} = \hat{f}(z) = \sum_{j=0}^{\infty} z^{-j} f(j). \tag{3.103}$$

An important property of this transform, which is useful in solving difference equations, is given by the relation

$$\begin{aligned}\mathcal{Z}\{f(k+1)\} &= \sum_{j=0}^{\infty} z^{-j} f(j+1) = \sum_{j=1}^{\infty} z^{-(j-1)} f(j) \\ &= z\left[\sum_{j=0}^{\infty} z^{-j} f(j) - f(0)\right] \\ &= z\left[\mathcal{Z}\{f(k)\} - f(0)\right] = z\hat{f}(z) - zf(0).\end{aligned} \tag{3.104}$$

If we take the z-transform of both sides of (3.91a), we obtain, in view of (3.104), $z\hat{x}(z) - zx(0) = A\hat{x}(z) + B\hat{u}(z)$ or

$$\hat{x}(z) = (zI - A)^{-1} zx(0) + (zI - A)^{-1} B\hat{u}(z). \tag{3.105}$$

Next, by taking the z-transform of both sides of (3.91b), and by substituting (3.105) into the resulting expression, we obtain

$$\hat{y}(z) = C(zI - A)^{-1} zx(0) + [C(zI - A)^{-1} B + D]\hat{u}(z). \tag{3.106}$$

The time sequence $\{y(k)\}$ can be recovered from its one-sided z-transform $\hat{y}(z)$ by applying the *inverse z-transform*, denoted by $\mathcal{Z}^{-1}[\hat{y}(z)]$.

In Table 3.3 we provide the one-sided z-transforms of some of the commonly used sequences, and in Table 3.4 we enumerate some of the more frequently encountered properties of the one-sided z-transform.

The *transfer function matrix* $\widehat{H}(z)$ of system (3.91) relates the z-transform of the output y to the z-transform of the input u under the assumption that $x(0) = 0$. We have

$$\hat{y}(z) = \widehat{H}(z)\hat{u}(z), \tag{3.107}$$

where

$$\widehat{H}(z) = C(zI - A)^{-1} B + D. \tag{3.108}$$

To relate $\widehat{H}(z)$ to the impulse response matrix $H(k, l)$, we notice that $\mathcal{Z}\{\delta(k - l)\} = z^{-l}$, where δ denotes the *discrete-time impulse* (or *unit pulse* or *unit sample*) defined in (2.51); i.e.,

$$\delta(k - l) = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases} \tag{3.109}$$

Table 3.3. Some commonly used z-transforms

$\{f(k)\}, \quad k \geq 0$	$\hat{f}(z) = \mathcal{Z}\{f(k)\}$
$\delta(k)$	1
$p(k)$	$1/(1 - z^{-1})$
k	$z^{-1}/(1 - z^{-1})^2$
k^2	$[z^{-1}(1 + z^{-1})]/(1 - z^{-1})^3$
a^k	$1/(1 - az^{-1})$
$(k + 1)a^k$	$1/(1 - az^{-1})^2$
$[(1/l!)(k + 1) \cdots (k + l)]a^k \quad l \geq 1$	$1/(1 - az^{-1})^{l+1}$
$a \cos \alpha k + b \sin \alpha k$	$\dfrac{a + z^{-1}(b \sin \alpha - a \cos \alpha)}{1 - 2z^{-1} \cos \alpha + z^{-2}}$

Table 3.4. Some properties of z-transforms

	$\{f(k)\}$, $k \geq 0$	$f(z)$
Time shift	$f(k + 1)$	$z\hat{f}(z) - zf(0)$
-Advance	$f(k + l) \quad l \geq 1$	$z^l \hat{f}(z) - z \sum_{i=1}^{l} z^{l-i} f(i - 1)$
Time shift	$f(k - 1)$	$z^{-1}\hat{f}(z) + f(-1)$
-Delay	$f(k - l) \quad l \geq 1$	$z^{-l}\hat{f}(z) + \sum_{i=1}^{l} z^{-l+i} f(-i)$
Scaling	$a^k f(k)$	$\hat{f}(z/a)$
	$kf(k)$	$-z(d/dz)\hat{f}(z)$
Convolution	$\sum_{l=0}^{\infty} f(l)g(k - l) = f(k) * g(k)$	$\hat{f}(z)\hat{g}(z)$
Initial value	$f(l)$ with $f(k) = 0, \quad k < l$	$\lim_{z\to\infty} z^l \hat{f}(z)^{\dagger}$
Final value	$\lim_{k\to\infty} f(k)$	$\lim_{z\to 1}(1 - z^{-1})\hat{f}(z)^{\ddagger}$

† If the limit exists.

‡ If $(1 - z^{-1})\hat{f}(z)$ has no singularities on or outside the unit circle.

This implies that the z-transform of a unit pulse applied at time zero is $\mathcal{Z}\{\delta(k)\} = 1$. It is not difficult to see now that $\{H(k, 0)\} = \mathcal{Z}^{-1}[\hat{y}(z)]$, where $\hat{y}(z) = \widehat{H}(z)\hat{u}(z)$ with $\hat{u}(z) = 1$. This shows that

$$\mathcal{Z}^{-1}[\widehat{H}(z)] = \mathcal{Z}^{-1}[C(zI - A)^{-1}B + D] = \{H(k, 0)\}, \tag{3.110}$$

where the unit impulse response matrix $H(k, 0)$ is given by (3.102).

The above result can also be derived directly by taking the z-transform of $\{H(k, 0)\}$ given in (3.102) (prove this). In particular, notice that the z-transform of $\{A^{k-1}\}$, $k = 1, 2, \ldots$ is $(zI - A)^{-1}$ since

$$\begin{aligned} \mathcal{Z}\{0, A^{k-1}\} &= \sum_{j=1}^{\infty} z^{-j} A^{j-1} = z^{-1} \sum_{j=0}^{\infty} z^{-j} A^{j} \\ &= z^{-1}(I + z^{-1}A + z^{-2}A^2 + \ldots) \\ &= z^{-1}(I - z^{-1}A)^{-1} = (zI - A)^{-1}. \end{aligned} \tag{3.111}$$

Above, the matrix determined by the expression $(1-\lambda)^{-1} = 1+\lambda+\lambda^2+\cdots$ was used. It is easily shown that the corresponding series involving A converges. Notice also that $\mathcal{Z}\{A^k\}, k = 0, 1, 2, \ldots$ is $z(zI - A)^{-1}$. This fact can be used to show that the inverse z-transform of (3.106) yields the time response (3.99), as expected.

We conclude this subsection with a specific example.

Example 3.32. In system (3.91), we let

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [1\ 0], \quad D = 0.$$

To verify that $\mathcal{Z}^{-1}[z(zI-A)^{-1}] = A^k$, we compute $z(zI-A)^{-1} = z\left[\begin{smallmatrix} z & -1 \\ 0 & z+1 \end{smallmatrix}\right]^{-1}$ $= z\begin{bmatrix} \frac{1}{z} & \frac{1}{z(z+1)} \\ 0 & \frac{1}{z+1} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{z+1} \\ 0 & \frac{z}{z+1} \end{bmatrix}$ and

$$\mathcal{Z}^{-1}[z(zI-A)^{-1}] = \begin{bmatrix} \delta(k) & (-1)^{k-1}p(k-1) \\ 0 & (-1)^k p(k) \end{bmatrix}$$

or

$$A^k = \begin{cases} \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right], & \text{when } k = 0, \\ \left[\begin{smallmatrix} 0 & (-1)^{k-1} \\ 0 & (-1)^k \end{smallmatrix}\right], & \text{when } k = 1, 2, \ldots, \end{cases}$$

as expected from Example 3.31.

Notice that

$$\begin{aligned} \mathcal{Z}^{-1}[(zI-A)^{-1}] &= \mathcal{Z}^{-1}\left[\begin{bmatrix} 1/z & 1/[z(z+1)] \\ 0 & 1/(z+1) \end{bmatrix}\right] \\ &= \begin{bmatrix} \delta(k-1)p(k-1) & \delta(k-1)p(k-1) - (-1)^{k-1}p(k-1) \\ 0 & (-1)^{k-1}p(k-1) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \text{ for } k = 0, \text{ and } \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ for } k = 1, \end{aligned}$$

and

$$\mathcal{Z}^{-1}[(zI-A)^{-1}] = \begin{bmatrix} 0 & -(-1)^{k-1} \\ 0 & (-1)^{k-1} \end{bmatrix} \text{ for } k = 2, 3, \ldots,$$

which is equal to $A^k, k \geq 0$, delayed by one unit; i.e., it is equal to $A^{k-1}, k = 1, 2, \ldots$, as expected.

Next, we consider the system response with $x(0) = 0$ and $u(k) = p(k)$. We have

$$\begin{aligned} y(k) &= \mathcal{Z}^{-1}[\hat{y}(z)] = \mathcal{Z}^{-1}[C(zI-A)^{-1}B \cdot \hat{u}(z)] \\ &= \mathcal{Z}^{-1}\left[\frac{1}{(z+1)(z-1)}\right] = \mathcal{Z}^{-1}\left[\tfrac{1/2}{z-1} - \tfrac{1/2}{z+1}\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}[(1)^{k-1} - (-1)^{k-1}]p(k-1) \\
&= \begin{cases} 0, & k = 0, \\ \frac{1}{2}(1 - (-1)^{k-1}), & k = 1, 2, \dots, \end{cases} \\
&= \begin{cases} 0, & k = 0, \\ 0, & k = 1, 3, 5, \dots, \\ 1, & k = 2, 4, 6, \dots . \end{cases}
\end{aligned}
$$

Note that if $x(0) = 0$ and $u(k) = \delta(k)$, then

$$
\begin{aligned}
y(k) &= \mathcal{Z}^{-1}[C(zI - A)^{-1}B] = \mathcal{Z}^{-1}\left[\frac{1}{z(z+1)}\right] \\
&= \delta(k-1)p(k-1) - (-1)^{k-1}p(k-1) \\
&= \begin{cases} 0, & k = 0, 1, \\ (-1)^{k-2}, & k = 2, 3, \dots, \end{cases}
\end{aligned}
$$

which is the unit impulse response of the system (refer to Example 3.31).

3.5.3 Equivalence of State-Space Representations

Equivalent representations of linear discrete-time systems are defined in a manner analogous to the continuous-time case. For systems (3.91), we let P denote a real nonsingular $n \times n$ matrix and we define

$$\tilde{x}(k) = Px(k). \tag{3.112}$$

Substituting (3.112) into (3.91) yields the equivalent system representation

$$\tilde{x}(k+1) = \widetilde{A}\tilde{x}(k) + \widetilde{B}u(k), \tag{3.113a}$$

$$y(k) = \widetilde{C}\tilde{x}(k) + \widetilde{D}u(k), \tag{3.113b}$$

where

$$\widetilde{A} = P^{-1}AP, \quad \widetilde{B} = PB, \quad \widetilde{C} = CP^{-1}, \quad \widetilde{D} = D. \tag{3.114}$$

We note that the terms in (3.114) are identical to corresponding terms obtained for the case of linear continuous-time systems.

We conclude by noting that if $\widehat{H}(z)$ and $\widehat{\widetilde{H}}(z)$ denote the transfer functions of the unit impulse response matrices of system (3.91) and system (3.113), respectively, then it is easily verified that $\widehat{H}(z) = \widehat{\widetilde{H}}(z)$.

3.5.4 Sampled-Data Systems

Discrete-time dynamical systems arise in a variety of ways in the modeling process. There are systems that are inherently defined only at discrete points in time, and there are representations of continuous-time systems at discrete points in time. Examples of the former include digital computers and devices (e.g., digital filters) where the behavior of interest of a system is adequately described by values of variables at discrete-time instants (and what happens between the discrete instants of time is quite irrelevant to the problem on hand); inventory systems where only the inventory status at the end of each day (or month) is of interest; economic systems, such as banking, where, e.g., interests are calculated and added to savings accounts at discrete-time intervals only; and so forth. Examples of the latter include simulations of continuous-time processes by means of digital computers, making use of difference equations that approximate the differential equations describing the process in question; feedback control systems that employ digital controllers and give rise to sampled-data systems (as discussed further in the following); and so forth.

In providing a short discussion of sampled-data systems, we make use of the specific class of linear feedback control systems depicted in Figure 3.2. This system may be viewed as an interconnection of a subsystem S_1, called the *plant* (the object to be controlled), and a subsystem S_2, called the *digital controller*. The plant is described by the equations

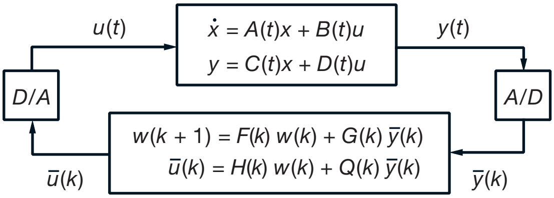

u(t)
ẋ = A(t)x + B(t)u
y = C(t)x + D(t)u
y(t)
D/A
A/D
w(k + 1) = F(k) w(k) + G(k) ȳ(k)
ū(k) = H(k) w(k) + Q(k) ȳ(k)
ū(k)
ȳ(k)

Figure 3.2. Digital control system

$$\dot{x} = A(t)x + B(t)u, \tag{3.115a}$$

$$y = C(t)x + D(t)u, \tag{3.115b}$$

where all symbols in (3.115) are defined as in (2.3) and where we assume that $t \geq t_0 \geq 0$.

Since our presentation pertains equally to the time-varying and time-invariant cases, we will first address the more general time-varying case. Next, we specialize our results to the time-invariant case.

The subsystem S_2 accepts the continuous-time signal $y(t)$ as its input, and it produces the piecewise continuous-time signal $u(t)$ as its output, where $t \geq t_0$. The continuous-time signal y is converted into a discrete-time signal $\{\bar{y}(k)\}$,

$k \geq k_0 \geq 0$, $k, k_0 \in Z$, by means of an analog-to-digital (A/D) converter and is processed according to a control algorithm given by the difference equations

$$w(k+1) = F(k)w(k) + G(k)\bar{y}(k), \tag{3.116a}$$
$$\bar{u}(k) = H(k)w(k) + Q(k)\bar{y}(k), \tag{3.116b}$$

where the $w(k), \bar{y}(k), \bar{u}(k)$ are real vectors and the $F(k), G(k), H(k)$, and $Q(k)$ are real, time-varying matrices with a consistent set of dimensions. Finally, the discrete-time signal $\{\bar{u}(k)\}$, $k \geq k_0 \geq 0$, is converted into the continuous-time signal u by means of a digital-to-analog (D/A) converter. To simplify our discussion, we assume in the following that $t_0 = k_0$.

An (ideal) A/D converter is a device that has as input a continuous-time signal, in our case y, and as output a sequence of real numbers, in our case $\{\bar{y}(k)\}, k = k_0, k_0 + 1, \ldots$, determined by the relation

$$\bar{y}(k) = y(t_k). \tag{3.117}$$

In other words, the (ideal) A/D converter is a device that *samples* an input signal, in our case $y(t)$, at times $t_0, t_1, \ldots$ producing the corresponding sequence $\{y(t_0), y(t_1), \ldots\}$.

A *D/A converter* is a device that has as input a discrete-time signal, in our case the sequence $\{\bar{u}(k)\}, k = k_0, k_0 + 1, \ldots$, and as output a continuous-time signal, in our case u, determined by the relation

$$u(t) = \bar{u}(k), t_k \leq t < t_{k+1}, \quad k = k_0, k_0 + 1, \ldots. \tag{3.118}$$

In other words, the D/A converter is a device that keeps its output constant at the last value of the sequence entered. We also call such a device a *zero-order hold*.

The system of Figure 3.2, as described above, is an example of a *sampled-data system*, since it involves truly *sampled data* (i.e., sampled signals), making use of an *ideal A/D converter*. In practice the digital controller S_2 uses *digital signals* as variables. In the scalar case, such signals are represented by real-valued sequences whose numbers belong to a subset of R consisting of a discrete set of points. (In the vector case, the previous statement applies to the components of the vector.) Specifically, in the present case, after the signal $y(t)$ has been sampled, it must be *quantized* (or *digitized*) to yield a *digital signal*, since only such signals are representable in a digital computer. If a computer uses, e.g., 8-bit words, then we can represent $2^8 = 256$ distinct levels for a variable, which determine the signal quantization. By way of a specific example, if we expect in the representation of a function a signal that varies from 9 to 25 volts, we may choose a 0.1-volt quantization step. Then 2.3 and 2.4 volts are represented by two different numbers (quantization levels); however, 2.315, 2.308, and 2.3 are all represented by the bit combination corresponding to 2.3. Quantization is an approximation and for short wordlengths may lead to significant errors. Problems associated with *quantization effects* will not be addressed in this book.

In addition to being a sampled-data system, the system represented by (3.115) to (3.118) constitutes a *hybrid system* as well, since it involves descriptions given by ordinary differential equations and ordinary difference equations. The analysis and synthesis of such systems can be simplified appreciably by replacing the description of subsystem S_1 (the plant) by a set of ordinary difference equations, valid only at discrete points in time $t_k, k = 0, 1, 2, \ldots$. [In terms of the blocks of Figure 3.2, this corresponds to considering the plant S_1, together with the D/A and A/D devices, to obtain a system with input $\bar{u}(k)$ and output $\bar{y}(k)$, as shown in Figure 3.3.] To accomplish this, we apply the variation of constants formula to (3.115a) to obtain

$$x(t) = \Phi(t, t_k)x(t_k) + \int_{t_k}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau, \tag{3.119}$$

where the notation $\phi(t, t_k, x(t_k)) = x(t)$ has been used. Since the input $u(t)$ is the output of the zero-order hold device (the D/A converter), given by (3.118), we obtain from (3.119) the expression

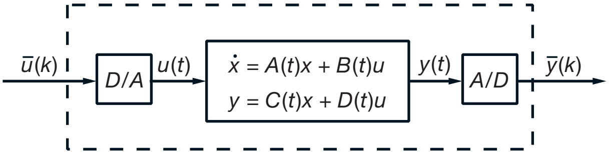


Figure 3.3. System described by (3.121) and (3.124)

$$x(t_{k+1}) = \Phi(t_{t+1}, t_k)x(t_k) + \left[\int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau)B(\tau)d\tau\right] u(t_k). \tag{3.120}$$

Since $\bar{x}(k) \triangleq x(t_k)$ and $\bar{u}(k) \triangleq u(t_k)$, we obtain a discrete-time version of the state equation for the plant, given by

$$\bar{x}(k+1) = \bar{A}(k)\bar{x}(k) + \bar{B}(k)\bar{u}(k), \tag{3.121}$$

where

$$\left.\begin{aligned} \bar{A}(k) &\triangleq \Phi(t_{k+1}, t_k), \\ \bar{B}(k) &\triangleq \int_{t_k}^{t_{k+1}} \Phi(t_{k+1}, \tau)B(\tau)d\tau. \end{aligned}\right\} \tag{3.122}$$

Next, we assume that the output of the plant is sampled at instants t'_k that do not necessarily coincide with the instants t_k at which the input to the plant is adjusted, and we assume that $t_k \le t'_k < t_{k+1}$. Then (3.115) and (3.119) yield

$$y(t'_k) = C(t'_k)\Phi(t'_k, t_k)x(t_k) + \left[C(t'_k)\int_{t_k}^{t'_k} \Phi(t'_k, \tau)B(\tau)d\tau\right] u(t_k) + D(t'_k)u(t_k). \tag{3.123}$$

Defining $\bar{y}(k) \triangleq y(t'_k)$, we obtain from (3.123),

$$\bar{y}(k) = \bar{C}(k)\bar{x}(k) + \bar{D}(k)\bar{u}(k), \tag{3.124}$$

where

$$\left.\begin{aligned} \bar{C}(k) &\triangleq C(t'_k)\Phi(t'_k, t_k), \\ \bar{D}(k) &\triangleq C(t'_k)\int_{t_k}^{t'_k} \Phi(t'_k, \tau)B(\tau)d\tau + D(t'_k). \end{aligned}\right\} \tag{3.125}$$

Summarizing, (3.121) and (3.124) constitute a state-space representation, valid at discrete points in time, of the plant [given by (3.115a)] and including the A/D and D/A devices [given by (3.117) and (3.118), see Figure 3.3]. Furthermore, the entire hybrid system of Figure 3.2, valid at discrete points in time, can now be represented by (3.121), (3.124), and (3.116).

Time-Invariant System With Constant Sampling Rate

We now turn to the case of the time-invariant plant, where $A(t) \equiv A, B(t) \equiv B, C(t) \equiv C$, and $D(t) \equiv D$, and we assume that $t_{k+1} - t_k = T$ and $t'_k - t_k = \alpha$ for all $k = 0, 1, 2, \ldots$. Then the expressions given in (3.121), (3.122), (3.124), and (3.125) assume the form

$$\bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}\bar{u}(k), \tag{3.126a}$$
$$\bar{y}(k) = \bar{C}\bar{x}(k) + \bar{D}\bar{u}(k), \tag{3.126b}$$

where

$$\left.\begin{aligned} \bar{A} &= e^{AT}, & \bar{B} &= \left(\int_0^T e^{A\tau}d\tau\right)B, \\ \bar{C} &= Ce^{A\alpha}, & \bar{D} &= C\left(\int_0^\alpha e^{A\tau}d\tau\right)B + D. \end{aligned}\right\} \tag{3.127}$$

If $t'_k = t_k$, or $\alpha = 0$, then $\bar{C} = C$ and $\bar{D} = D$.

In the preceding, T is called the *sampling period* and $1/T$ is called the *sampling rate*. Sampled-data systems are treated in great detail in texts dealing with digital control systems and with digital signal processing.

Example 3.33. In the control system of Figure 3.2, let

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [1, 0], \quad D = 0,$$

let T denote the sampling period, and assume that $\alpha = 0$. The discrete-time state-space representation of the plant, preceded by a zero-order hold

(D/A converter) and followed by a sampler [an (ideal) A/D converter], both sampling synchronously at a rate of 1/T, is given by $\bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}\bar{u}(k)$, $\bar{y}(k) = \bar{C}x(k)$, where

$$\begin{aligned}
\bar{A} &= e^{AT} = \sum_{j=1}^{\infty}(T^j/j!)A^j = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} T = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, \\
\bar{B} &= \left(\int_0^T e^{A\tau} d\tau\right) B = \left(\int_0^T \begin{bmatrix} 1 & \tau \\ 0 & 1 \end{bmatrix} d\tau\right) \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} T & T^2/2 \\ 0 & T \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} T^2/2 \\ T \end{bmatrix}, \\
\bar{C} &= C = [1\ 0].
\end{aligned}$$

The transfer function (relating $\bar{y}$ to $\bar{u}$) is given by

$$\begin{aligned}
\widehat{H}(z) &= \bar{C}(zI - \bar{A})^{-1}\bar{B} \\
&= [1\ 0] \begin{bmatrix} z-1 & -T \\ 0 & z-1 \end{bmatrix}^{-1} \begin{bmatrix} T^2/2 \\ T \end{bmatrix} \\
&= [1\ 0] \begin{bmatrix} 1/(z-1) & T/(z-1)^2 \\ 0 & 1/(z-1) \end{bmatrix} \begin{bmatrix} T^2/2 \\ T \end{bmatrix} \\
&= \frac{T^2}{2} \frac{(z+1)}{(z-1)^2}.
\end{aligned}$$

The transfer function of the continuous-time system (continuous-time description of the plant) is determined to be $\widehat{H}(s) = C(sI-A)^{-1}B = 1/s^2$, the double integrator.

The behavior of the system between the discrete instants, $t, t_k \leq t < t_{k+1}$, can be determined by using (3.119), letting $x(t_k) = x(k)$ and $u(t_k) = u(k)$.

An interesting observation, useful when calculating $\bar{A}$ and $\bar{B}$, is that both can be expressed in terms of a single series. In particular, $\bar{A} = e^{AT} = I + TA + (T^2/2!)A^2 + \cdots = I + TA\Psi(T)$, where $\Psi(T) = I + (T/2!)A + (T^2/3!)A^2 + \cdots = \sum_{j=0}^{\infty}(T^j/(j+1)!)A^j$. Then $\bar{B} = (\int_0^T e^{A\tau} d\tau)\ B = (\sum_{j=0}^{\infty}(T^{j+1}/(j+1)!)A^j)$ $B = T\Psi(T)B$. If $\Psi(T)$ is determined first, and then both $\bar{A}$ and $\bar{B}$ can easily be calculated.

Example 3.34. In Example 3.33, $\Psi(T) = I + TA = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$. Therefore, $\bar{A} = I + TA\Psi(T) = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$ and $\bar{B} = T\Psi(T)B = \begin{bmatrix} T^2/2 \\ T \end{bmatrix}$, as expected.

3.5.5 Modes, Asymptotic Behavior, and Stability

As in the case of continuous-time systems, we study in this subsection the qualitative behavior of the solutions of linear, autonomous, homogeneous ordinary difference equations

$$x(k+1) = Ax(k) \tag{3.128}$$

in terms of the modes of such systems, where $A \in R^{n\times n}$ and $x(k) \in R^n$ for every $k \in Z^+$. From before, the unique solution of (3.128) satisfying $x(0) = x_0$ is given by

$$\phi(k, 0, x_0) = A^k x_0. \tag{3.129}$$

Let $\lambda_1, \dots, \lambda_\sigma$, denote the σ distinct eigenvalues of A, where λ_i with $i = 1, \dots, \sigma$, is assumed to be repeated n_i times so that $\sum_{i=1}^{\sigma} n_i = n$. Then

$$\det(zI - A) = \prod_{i=1}^{\sigma}(z-\lambda_i)^{n_i}. \tag{3.130}$$

To introduce the modes for (3.128), we first derive the expressions

$$\begin{aligned} A^k &= \sum_{i=1}^{\sigma}[A_{i0}\lambda_i^k p(k) + \sum_{l=1}^{n_i-1} A_{il}k(k-1)\cdots(k-l+1)\lambda_i^{k-l}p(k-l)] \\ &= \sum_{i=1}^{\sigma}[A_{i0}\lambda_i^k p(k) + A_{i1}k\lambda_i^{k-1}p(k-1) + \cdots \\ &\quad + A_{i(n_i-1)}k(k-1)\cdots(k-n_i+2)\lambda_i^{k-(n_i-1)}p(k-n_i+1)], \end{aligned} \tag{3.131}$$

where

$$A_{il} = \frac{1}{l!}\,\frac{1}{(n_i-1-l)!}\lim_{z\to\lambda_i}\{[(z-\lambda_i)^{n_i}(zI-A)^{-1}]^{(n_i-1-l)}\}. \tag{3.132}$$

In (3.132), $[\cdot]^{(q)}$ denotes the qth derivative with respect to z, and in (3.131), $p(k)$ denotes the unit step [i.e., $p(k) = 0$ for $k < 0$ and $p(k) = 1$ for $k \geq 0$]. Note that if an eigenvalue λ_i of A is zero, then (3.131) must be modified. In this case,

$$\sum_{i=0}^{n_i-1} A_{il} l!\delta(k-l) \tag{3.133}$$

are the terms in (3.131) corresponding to the zero eigenvalue.

To prove (3.131), (3.132), we proceed as in the proof of (3.50), (3.51). We recall that $\{A^k\} = \mathcal{Z}^{-1}[z(zI-A)^{-1}]$ and we use the partial fraction expansion method to determine the z-transform. In particular, as in the proof of (3.50), (3.51), we can readily verify that

$$z(zI-A)^{-1} = z\sum_{i=1}^{\sigma}\sum_{l=0}^{n_i-1}(l!A_{il})(z-\lambda_i)^{-(l+1)}, \tag{3.134}$$

where the A_{il} are given in (3.132). We now take the inverse z-transform of both sides of (3.134). We first notice that

$$\begin{aligned}\mathcal{Z}^{-1}[z(z-\lambda_i)^{-(l+1)}] &= \mathcal{Z}^{-1}[z^{-l}z^{l+1}(z-\lambda_i)^{-(l+1)}] \\ &= \mathcal{Z}^{-1}[z^{-l}(1-\lambda_i z^{-1})^{-(l+1)}] = f(k-l)p(k-l) \\ &= \begin{cases} f(k-l), & \text{for } k \geq l, \\ 0, & \text{otherwise.}\end{cases}\end{aligned}$$

Referring to Tables 3.3 and 3.4 we note that $f(k)p(k) = \mathcal{Z}^{-1}[(1-\lambda_i z^{-1})^{-(l+1)}] = [\frac{1}{l!}(k+1)\cdots(k+l)]\lambda_i^k$ for $\lambda_i \neq 0$ and $l \geq 1$. Therefore, $\mathcal{Z}^{-1}[l!z(z-\lambda_i)^{-(l+1)}] = l!f(k-l)p(k-l) = k(k-1)\cdots(k-l+1)\lambda_i^{k-l}, l \geq 1$. For $l = 0$, we have $\mathcal{Z}^{-1}[(1-\lambda_i z^{-1})^{-1}] = \lambda_i^k$. This shows that (3.131) is true when $\lambda_i \neq 0$. Finally, if $\lambda_i = 0$, we note that $\mathcal{Z}^{-1}[l!z^{-l}] = l!\delta(k-l)$, which implies (3.133).

Note that one can derive several alternative but equivalent expressions for (3.131) that correspond to different ways of determining the inverse z-transform of $z(zI-A)^{-1}$ or of determining A^k via some other methods.

In complete analogy with the continuous-time case, we call the terms $A_{il}k(k-1)\cdots(k-l+1)\lambda_i^{k-l}$ the *modes of the system* (3.128). There are n_i modes corresponding to the eigenvalues $\lambda_i, l = 0,\ldots,n_i-1$, and the system (3.128) has a total of n modes.

It is particularly interesting to study the matrix $A^k, k = 0,1,2,\ldots$ using the Jordan canonical form of A, i.e., $J = P^{-1}AP$, where the similarity transformation P is constructed by using the generalized eigenvectors of A. We recall once more that $J = \text{diag}[J_1,\ldots,J_\sigma] \triangleq \text{diag}[J_i]$ where each $n_i \times n_i$ block J_i corresponds to the eigenvalue λ_i and where, in turn, $J_i = \text{diag}[J_{i1},\ldots,J_{il_i}]$ with J_{ij} being smaller square blocks, the dimensions of which depend on the length of the chains of generalized eigenvectors corresponding to J_i (refer to Subsection 3.3.2). Let J_{ij} denote a typical Jordan canonical form block. We shall investigate the matrix J_{ij}^k, since $A^k = P^{-1}J^kP = P^{-1}\,\text{diag}[J_{ij}^k]P$.

Let

$$J_{ij} = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 \\ 0 & 0 & \cdots & \cdots & \lambda_i \end{bmatrix} = \lambda_i I + N_i, \tag{3.135}$$

where

$$N_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & & & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 1 \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix}$$

and where we assume that J_{ij} is a $t \times t$ matrix. Then

$$\begin{aligned}(J_{ij})^k &= (\lambda_i I + N_i)^k \\ &= \lambda_i^k I + k\lambda_i^{k-1} N_i + \frac{k(k-1)}{2!}\lambda_i^{k-2} N_i^2 + \cdots + k\lambda_i N_i^{k-1} + N_i^k. \quad (3.136)\end{aligned}$$

Now since $N_i^k = 0$ for $k \geq t$, a typical $t\times t$ Jordan block J_{ij} will generate terms that involve only the scalars $\lambda_i^k, \lambda_i^{k-1}, \ldots, \lambda_i^{k-(t-1)}$. Since the largest possible block associated with the eigenvalue λ_i is of dimension $n_i \times n_i$, the expression of A^k in (3.131) should involve at most the terms $\lambda_i^k, \lambda_i^{k-1}, \ldots, \lambda_i^{k-(n_i-1)}$, which it does.

The above enables us to prove the following useful fact: Given $A \in R^{n\times n}$, there exists an integer $k \geq 0$ such that

$$A^k = 0 \qquad (3.137)$$

if and only if all the eigenvalues λ_i of A are at the origin. Furthermore, the smallest k for which (3.137) holds is equal to the dimension of the largest block J_{ij} of the Jordan canonical form of A.

The second part of the above assertion follows readily from (3.136). We ask the reader to prove the first part of the assertion.

We conclude by observing that when all n eigenvalues λ_i of A are distinct, then

$$A^k = \sum_{i=1}^{n} A_i \lambda_i^k, \quad k \geq 0, \qquad (3.138)$$

where

$$A_i = \lim_{z\to\lambda_i} [(z-\lambda_i)(zI-A)^{-1}]. \qquad (3.139)$$

If $\lambda_i = 0$, we use $\delta(k)$, the unit pulse, in place of λ_i^k in (3.138). This result is straightforward, in view of (3.131), (3.132).

Example 3.35. In (3.128) we let $A = \begin{bmatrix} 0 & 1 \\ -\frac{1}{4} & 1 \end{bmatrix}$. The eigenvalues of A are $\lambda_1 = \lambda_2 = \frac{1}{2}$, and therefore, $n_1 = 2$ and $\sigma = 1$. Applying (3.131), (3.132), we obtain

$$\begin{aligned}A^k &= A_{10}\lambda_1^k p(k) + A_{11} k \lambda_1^{k-1} p(k-1) \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \left(\frac{1}{2}\right)^k p(k) + \begin{bmatrix} -\frac{1}{2} & 1 \\ -\frac{1}{4} & \frac{1}{2} \end{bmatrix} (k) \left(\frac{1}{2}\right)^{k-1} p(k-1).\end{aligned}$$

Example 3.36. In (3.128) we let $A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}$. The eigenvalues of A are $\lambda_1 = -1, \lambda_2 = 1$ (so that $\sigma = 2$). Applying (3.138), (3.139), we obtain

$$A^k = A_{10}\lambda_1^k + A_{20}\lambda_2^k = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}(-1)^k + \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad k \geq 0.$$

Note that this same result was obtained by an entirely different method in Example 3.30.

Example 3.37. In (3.128) we let $A = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}$. The eigenvalues of A are $\lambda_1 = 0$, $\lambda_2 = -1$, and $\sigma = 2$. Applying (3.138), (3.139), we obtain

$$A_0 = \lim_{z \to 0}[z(zI - A)^{-1}] = \frac{1}{z+1}\begin{bmatrix} z+1 & 1 \\ 0 & z \end{bmatrix}|_{z=0} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$$

$$A_1 = \lim_{z \to -1}\left[(z+1)\frac{1}{z(z+1)}\begin{bmatrix} z+1 & 1 \\ 0 & z \end{bmatrix}\right] = \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix}$$

and

$$A^k = A_0\delta(k) + A_1(-1)^k = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}\delta(k) + \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix}(-1)^k, \quad k \geq 0.$$

As in the case of continuous-time systems described by (3.48), various notions of stability of an equilibrium for discrete-time systems described by linear, autonomous, homogeneous ordinary difference equations (3.128) will be studied in detail in Chapter 4. If $\phi(k, 0, x_e)$ denotes the solution of system (3.128) with $x(0) = x_e$, then x_e is said to be an *equilibrium* of (3.128) if $\phi(k, 0, x_e) = x_e$ for all $k \geq 0$. Clearly, $x_e = 0$ is an equilibrium of (3.128). In discussing the qualitative properties, it is customary to speak, somewhat informally, of the stability properties of (3.128), rather than the stability properties of the equilibrium $x_e = 0$ of system (3.128).

The concepts of *stability*, *asymptotic stability*, and *instability* of system (3.128) are now defined in an identical manner as in Subsection 3.3.3 for system (3.48), except that in this case continuous-time t ($t \in R^+$) is replaced by discrete-time k ($k \in Z^+$).

By inspecting the modes of system (3.128) [given by (3.131) and (3.132)], we can readily establish the following stability criteria:

1. The system (3.128) is *asymptotically stable* if and only if all eigenvalues of A are within the unit circle of the complex plane (i.e., $|\lambda_j| < 1$, $j = 1, \ldots, n$).

2. The system (3.128) is *stable* if and only if $|\lambda_j| \leq 1$, $j = 1, \dots, n$, and for all eigenvalues with $|\lambda_j| = 1$ having multiplicity $n_j > 1$, it is true that

$$\lim_{z \to \lambda_j} [[(z - \lambda_j)^{n_j} (zI - A)^{-1}]^{(n_j - 1 - l)}] = 0 \text{ for } l = 1, \dots, n_j - 1. \quad (3.140)$$

3. The system (3.128) is *unstable* if and only if (2) is not true.

Example 3.38. The system given in Example 3.35 is asymptotically stable. The system given in Example 3.36 is stable. In particular, note that the solution $\phi(k, 0, x(0)) = A^k x(0)$ for Example 3.36 is bounded.

When the eigenvalues λ_i of A are distinct, then as in the continuous-time case [refer to (3.56), (3.57)], we can readily show that

$$A^k = \sum_{j=1}^{n} A_j \lambda_j^k, \quad A_j = v_j \tilde{v}_j, \quad k \geq 0, \quad (3.141)$$

where the v_j and $\tilde{v}_j$ are right and left eigenvectors of A corresponding to λ_j, respectively. If $\lambda_j = 0$, we use $\delta(k)$, the unit pulse, in place of λ_j^k in (3.141).

In proving (3.141), we use the same approach as in the proof of (3.56), (3.57). We have $A^k = Q \operatorname{diag}[\lambda_1^k, \dots, \lambda_n^k] Q^{-1}$, where the columns of Q are the n right eigenvectors and the rows of Q^{-1} are the n left eigenvectors of A.

As in the continuous-time case [system (3.48)], the initial condition $x(0)$ for system (3.128) can be selected to be colinear with the eigenvector v_i to eliminate from the solution of (3.128) all modes except the ones involving λ_i^k.

Example 3.39. As in Example 3.36, we let $A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}$. Corresponding to the eigenvalues $\lambda_1 = -1$, $\lambda_2 = 1$, we have the right and left eigenvectors $v_1 = (1, 0)^T, v_2 = (1, 1)^T, \tilde{v}_1 = (1, -1)$, and $\tilde{v}_2 = (0, 1)$. Then

$$\begin{aligned} A^k &= [v_1 \ \tilde{v}_1] \lambda_1^k + [v_2 \ \tilde{v}_2] \lambda_2^k \\ &= \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} (-1)^k + \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} (1)^k, \quad k \geq 0. \end{aligned}$$

Choose $x(0) = \alpha(1, 0)^T = \alpha v_1$ with $\alpha \neq 0$. Then

$$\phi(k, 0, x(0)) = \begin{bmatrix} \alpha \\ 0 \end{bmatrix} (-1)^k,$$

which contains only the mode associated with $\lambda_1 = -1$.

We conclude our discussion of modes and asymptotic behavior by briefly considering the state equation

$$x(k+1) = Ax(k) + Bu(k), \tag{3.142}$$

where x, u, A, and B are as defined in (3.91a). Taking the $\mathcal{Z}$-transform of both sides of (3.142) and rearranging yields

$$\tilde{x}(z) = z(zI - A)^{-1}x(0) + (zI - A)^{-1}B\tilde{u}(z). \tag{3.143}$$

By taking the inverse $\mathcal{Z}$-transform of (3.143), we see that the solution ϕ of (3.142) is the sum of modes that correspond to the singularities or poles of $z(zI - A)^{-1}x(0)$ and of $(zI - A)^{-1}B\tilde{u}(z)$. If in particular, system (3.128) is asymptotically stable [i.e., for $x(k+1) = Ax(k)$, all eigenvalues λ_j of A are such that $|\lambda_j| < 1$, $j = 1, \dots, n$] and if $u(k)$ in (3.142) is bounded [i.e., there is an M such that $|u_i(k)| < M$ for all $k \geq 0, i = 1, \dots, m$], then it is easily seen that the solutions of (3.142) are bounded as well.

3.6 An Important Comment on Notation

Chapters 1–3 are primarily concerned with the basic (qualitative) properties of systems of first-order ordinary differential equations, such as, e.g., the system of equations given by

$$\dot{x} = Ax, \tag{3.144}$$

where $x \in R^n$ and $A \in R^{n\times n}$. In the arguments and proofs to establish various properties for such systems, we highlighted the solutions by using the ϕ-notation. Thus, the unique solution of (3.144) for a given set of initial data (t_0, x_0) was written as $\phi(t, t_0, x_0)$ with $\phi(t_0, t_0, x_0) = x_0$. A similar notation was used in the case of the equation given by

$$\dot{x} = f(t, x) \tag{3.145}$$

and the equations given by

$$\dot{x} = Ax + Bu, \tag{3.146a}$$

$$y = Cx + Du, \tag{3.146b}$$

where in (3.145) and in (3.146) all symbols are defined as in (1.11) (see Chapter 1) and as in (3.61) of this chapter, respectively.

In the study of control systems such as system (3.61), the center of attention is usually the control input u and the resulting evolution of the system state in the state space and the system output. In the development of control systems theory, the x-notation has been adopted to express the solutions of systems. Thus, the solution of (3.61a) is denoted by $x(t)$ [or $x(t, t_0, x_0)$ when t_0 and x_0 are to be emphasized] and the evolution of the system output y

in (3.61b) is denoted by $y(t)$. In all subsequent chapters, except Chapter 4, we will also follow this practice, employing the usual notation utilized in the control systems literature. In Chapter 4, which is concerned with the stability properties of systems, we will use the ϕ-notation when studying the Lyapunov stability of an equilibrium [such as system (3.144)] and the x-notation when investigating the input–output properties of control systems [such as system (3.61)].

3.7 Summary and Highlights

Continuous-Time Systems

- *The state transition matrix* $\Phi(t, t_0)$ of $\dot{x} = Ax$
$$\Phi(t, t_0) \triangleq \Psi(t)\Psi^{-1}(t_0), \tag{3.9}$$
where $\Psi(t)$ is any fundamental matrix of $\dot{x} = Ax$. See Definitions 3.8 and 3.2 and Theorem 3.9 for properties of $\Phi(t, t_0)$. In the present time-invariant case
$$\Phi(t, t_0) = e^{A(t-t_0)},$$
where
$$e^{At} = I + \sum_{k=1}^{\infty} \frac{t^k A^k}{k!} \tag{3.17}$$
is the matrix exponential. See Theorem 3.13 for properties.
- *Methods to evaluate* e^{At}. Via infinite series (3.17) and via similarity transformation
$$e^{At} = P^{-1} e^{Jt} P$$
with $J = P^{-1}AP$ [see (3.25)] where J is diagonal or in Jordan canonical form; via the Cayley–Hamilton Theorem [see (3.41)] and via the Laplace transform, where
$$e^{At} = \mathcal{L}^{-1}[(sI - A)^{-1}], \tag{3.45}$$
or via the system modes [see (3.50)], which simplify to
$$e^{At} = \sum_{i=1}^{n} A_i e^{\lambda_i t} \tag{3.53}$$
when the n eigenvalues of A, λ_i, are distinct. See also (3.56), (3.57).
- *Modes of the system.* e^{At} is expressed in terms of the modes $A_{ik} t^k e^{\lambda_i t}$ in (3.50). The distinct eigenvalue case is found in (3.53), (3.54) and in (3.56), (3.57).
- *The stability of an equilibrium* of $\dot{x} = Ax$ is defined and related to the eigenvalues of A using the expression for e^{At} in terms of the modes.

- Given $\dot{x} = Ax + Bu$,

  $$x(t) = e^{At}x(0) + \int_0^t e^{A(t-s)}Bu(s)s$$

  is its solution, the *variation of constants formula.*
  If in addition $y = Cx + Du$, then the *total response of the system* is

  $$y(t) = Ce^{A(t-t_0)}x_0 + C\int_{t_0}^t e^{A(t-s)}Bu(s)ds + Du(t). \tag{3.70}$$

  The *impulse response* is

  $$H(t) = \begin{cases} Ce^{At}B + D\delta(t), & t \geq \tau, \\ 0, & t < 0, \end{cases} \tag{3.72}$$

  and the *transfer function* is

  $$\widehat{H}(s) = C(sI - A)^{-1}B + D. \tag{3.80}$$

  Note that $H(s) = \mathcal{L}(H(t, 0))$.
- *Equivalent representations*

  $$\begin{aligned} \dot{\tilde{x}} &= \widetilde{A}\tilde{x} + \widetilde{B}u, \\ y &= \widetilde{C}\tilde{x} + \widetilde{D}u, \end{aligned} \tag{3.82}$$

  where

  $$\widetilde{A} = PAP^{-1}, \quad \widetilde{B} = PB, \quad \widetilde{C} = CP^{-1}, \quad \widetilde{D} = D \tag{3.83}$$

  is equivalent to $\dot{x} = Ax + Bu$, $y = Cx + Du$.

Discrete-Time Systems

- Consider the *discrete-time system*

  $$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k). \tag{3.91}$$

  Then

  $$y(k) = CA^k x(0) + \sum_{j=0}^{k-1} CA^{k-(j+1)}Bu(j) + Du(k), \quad k > 0. \tag{3.99}$$

  The *discrete-time unit impulse response* is

  $$H(k, 0) = \begin{cases} CA^{k-1}B & k \geq 0, \\ D & k = 0, \\ 0 & k < 0, \end{cases} \tag{3.102}$$

  and the *transfer function* is

  $$\widehat{H}(z) = C(zI - A)^{-1}B + D. \tag{3.108}$$

  Note that $\widehat{H}(z) = \mathcal{Z}\{H(k, 0)\}$.

- $A^k = \mathcal{Z}^{-1}(z(zI - A)^{-1})$. A^k may also be calculated using the Cayley–Hamilton theorem. Note that when all n eigenvalues of A, λ_i, are distinct then
$$A^k = \sum_{j=0}^{n} A_i \lambda_i^k, \quad k \geq 0, \tag{3.138}$$
$A_i \lambda_i^k$ are the modes of the system.
- The *stability of an equilibrium* of $x(k+1) = Ax(k)$ is defined and related to the eigenvalues of A using the expressions of A^k in terms of the modes.

Sampled Data Systems

- When $\dot{x} = Ax + Bu$, $y = Cx + Du$ is the system in Figure 3.3, the discrete-time description is
$$\begin{aligned} \bar{x}(k+1) &= \bar{A}\bar{x}(k) + \bar{B}\bar{u}(k), \\ \bar{y}(k) &= \bar{C}\bar{x}(k) + \bar{D}\bar{u}(k), \end{aligned} \tag{3.126}$$
with
$$\begin{aligned} &\bar{A} = e^{AT}, \quad \bar{B} = \left[\int_0^T e^{A\tau} d\tau\right] B, \\ &\bar{C} = C, \qquad \bar{D} = D, \end{aligned} \tag{3.127}$$
where T is the sampling period.

3.8 Notes

Our treatment of basic aspects of linear ordinary differential equations in Sections 3.2 and 3.3 follows along lines similar to the development of this subject given in Miller and Michel [8].

State-space and input–output representations of continuous-time systems and discrete-time systems, addressed in Sections 3.4 and 3.5, respectively, are addressed in a variety of textbooks, including Kailath [7], Chen [4], Brockett [3], DeCarlo [5], Rugh [11], and others. For further material on sampled-data systems, refer to Aström and Wittenmark [2] and to the early works on this subject that include Jury [6] and Ragazzini and Franklin [9].

Detailed treatments of the Laplace transform and the z-transform, discussed briefly in Sections 3.3 and 3.5, respectively, can be found in numerous texts on signals and linear systems, control systems, and signal processing.

In the presentation of the material in all the sections of this chapter, we have relied principally on Antsaklis and Michel [1].

The state representation of systems received wide acceptance in systems theory beginning in the late 1950s. This was primarily due to the work of R.

E. Kalman and others in filtering theory and quadratic control theory and due to the work of applied mathematicians concerned with the stability theory of dynamical systems. For comments and extensive references on some of the early contributions in these areas, refer to Kailath [7] and Sontag [12]. Of course, differential equations have been used to describe the dynamical behavior of artificial systems for many years. For example, in 1868 J. C. Maxwell presented a complete treatment of the behavior of devices that regulate the steam pressure in steam engines called flyball governors (Watt governors) to explain certain phenomena.

The use of state-space representations in the systems and control area opened the way for the systematic study of systems with multi-inputs and multi-outputs. Since the 1960s an alternative description is also being used to characterize time-invariant MIMO control systems that involves usage of polynomial matrices or differential operators. Some of the original references on this approach include Rosenbrock [10] and Wolovich [13]. This method, which corresponds to system descriptions by means of higher order ordinary differential equations (rather than systems of first-order ordinary differential equations, as is the case in the state-space description), is addressed in Sections 7.5 and 8.5 and in Chapter 10.

References

1. P.J. Antsaklis and A.N. Michel, *Linear Systems*, Birkhäuser, Boston, MA, 2006.
2. K.J. Aström and B. Wittenmark, *Computer-Controlled Systems. Theory and Design*, Prentice-Hall, Englewood Cliffs, NJ, 1990.
3. R.W. Brockett, *Finite Dimensional Linear Systems*, Wiley, New York, NY, 1970.
4. C.T. Chen, *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, NY, 1984.
5. R.A. DeCarlo, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
6. E.I. Jury, *Sampled-Data Control Systems*, Wiley, New York, NY, 1958.
7. T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
8. R.K. Miller and A.N. Michel, *Ordinary Differential Equations*, Academic Press, New York, NY, 1982.
9. J.R. Ragazzini and G.F. Franklin, *Sampled-Data Control Systems*, McGraw-Hill, New York, NY, 1958.
10. H.H. Rosenbrock, *State Space and Multivariable Theory*, Wiley, New York, NY, 1970.
11. W.J. Rugh, *Linear System Theory, Second Edition*, Prentice-Hall, Englewood Cliffs, NJ, 1996.
12. E.D. Sontag, *Mathematical Control Theory. Deterministic Finite Dimensional Systems*, TAM 6, Springer-Verlag, New York, NY, 1990.
13. W.A. Wolovich, *Linear Multivariable Systems*, Springer-Verlag, New York, NY, 1974.

Exercises

For the first 12 exercises, the reader may want to refer to the appendix, which contains appropriate material on matrices and linear algebra.

3.1. (a) Let $(V, F) = (R^3, R)$. Determine the representation of $v = (1, 4, 0)^T$ with respect to the basis $v^1 = (1, -1, 0)^T, v^2 = (1, 0, -1)^T$, and $v^3 = (0, 1, 0)^T$.

(b) Let $V = F^3$, and let F be the field of rational functions. Determine the representation of $\tilde{v} = (s+2, 1/s, -2)^T$ with respect to the basis $\{v^1, v^2, v^3\}$ given in (a).

3.2. Find the relationship between the two bases $\{v^1, v^2, v^3\}$ and $\{\bar{v}^1, \bar{v}^2, \bar{v}^3\}$ (i.e., find the matrix of $\{\bar{v}^1, \bar{v}^2, \bar{v}^3\}$ with respect to $\{v^1, v^2, v^3\}$) where $v^1 = (2, 1, 0)^T, v^2 = (1, 0, -1)^T, v^3 = (1, 0, 0)^T, \bar{v}^1 = (1, 0, 0)^T, \bar{v}^2 = (0, 1, -1)$, and $\bar{v}^3 = (0, 1, 1)$. Determine the representation of the vector $e_2 = (0, 1, 0)^T$ with respect to both of the above bases.

3.3. Let $\alpha \in R$ be fixed. Show that the set of all vectors $(x, \alpha x)^T, x \in R$, determines a vector space of dimension one over $F = R$, where vector addition and multiplication of vectors by scalars is defined in the usual manner. Determine a basis for this space.

3.4. Show that the set of all real $n \times n$ matrices with the usual operation of matrix addition and the usual operation of multiplication of matrices by scalars constitutes a vector space over the reals [denoted by $(R^{n\times n}, R)$]. Determine the dimension and a basis for this space. Is the above statement still true if $R^{n\times n}$ is replaced by $R^{m\times n}$, the set of real $m \times n$ matrices? Is the above statement still true if $R^{n\times n}$ is replaced by the set of nonsingular matrices? Justify your answers.

3.5. Let $v^1 = (s^2, s)^T$ and $v^2 = (1, 1/s)^T$. Is the set $\{v^1, v^2\}$ linearly independent over the field of rational functions? Is it linearly independent over the field of real numbers?

3.6. Determine the rank of the following matrices, carefully specifying the field:

$$\text{(a)} \begin{bmatrix} j \\ 3j \\ -1 \end{bmatrix}, \quad \text{(b)} \begin{bmatrix} 1 & 4 & -5 \\ 7 & 0 & 2 \end{bmatrix}, \quad \text{(c)} \begin{bmatrix} (s+4) & -2 \\ (s^2-1) & 6 \\ 0 & 2s+3 \\ s & -s+4 \end{bmatrix}, \quad \text{(d)} \left(\frac{s+1}{s^2}\right),$$

where $j = \sqrt{-1}$.

3.7. (a) Determine bases for the range and null space of the matrices

$$A_1 = [1\,0\,1], \quad A_2 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \text{and} \quad A_3 = \begin{bmatrix} 3 & 2 & 1 \\ 3 & 2 & 1 \\ 3 & 2 & 1 \end{bmatrix}.$$

(b) Characterize all solutions of $A_1 x = 1$ (see Subsection A.3.1).

3.8. Show that $e^{(A_1+A_2)t} = e^{A_1 t} e^{A_2 t}$ if $A_1 A_2 = A_2 A_1$.

3.9. Show that there exists a similarity transformation matrix P such that

$$PAP^{-1} = A_c = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & -\alpha_2 & \cdots & -\alpha_{n-1} \end{bmatrix}$$

if and only if there exists a vector $b \in R^n$ such that the rank of $[b, Ab, \ldots, A^{n-1}b]$ is n; i.e., $\rho[b, Ab, \ldots, A^{n-1}b] = n$.

3.10. Show that if λ_i is an eigenvalue of the companion matrix A_c given in Exercise 3.9, then a corresponding eigenvector is $v^i = (1, \lambda_i, \ldots, \lambda_i^{n-1})^T$.

3.11. Let λ_i be an eigenvalue of a matrix A, and let v^i be a corresponding eigenvector. Let $f(\lambda) = \sum_{k=0}^{l} \alpha_k \lambda^k$ be a polynomial with real coefficients. Show that $f(\lambda_i)$ is an eigenvalue of the matrix function $f(A) = \sum_{k=0}^{l} \alpha_k A^k$. Determine an eigenvector corresponding to $f(\lambda_i)$.

3.12. For the matrices

$$A_1 = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

determine the matrices $A_1^{100}, A_2^{100}, e^{A_1 t}$, and $e^{A_2 t}, t \in R$.

3.13. For the system

$$\dot{x} = Ax + Bu, \tag{3.147}$$

where all symbols are as defined in (3.61a), derive the *variation of constants formula* (3.11), using the change of variables $z(t) = \Phi(t_0, t)x(t)$.

3.14. Show that $\frac{\partial}{\partial \tau}\Phi(t, \tau) = -\Phi(t, \tau)A$ for all $t, \tau \in R$.

3.15. The *adjoint equation* of (3.1) is given by

$$\dot{z} = -A^T z. \tag{3.148}$$

Let $\Phi(t, t_0)$ and $\Phi_a(t, t_0)$ denote the state transition matrices of (3.1) and its adjoint equation, respectively. Show that $\Phi_a(t, t_0) = [\Phi(t_0, t)]^T$.

3.16. Consider the system described by

$$\dot{x} = Ax + Bu, \quad y = Cx, \tag{3.149}$$

where all symbols are as in (3.61) with $D = 0$, and consider the *adjoint equation* of (3.149), given by

$$\dot{z} = -A^T z + C^T v, \quad w = B^T z. \tag{3.150}$$

(a) Let $H(t, \tau)$ and $H_a(t, \tau)$ denote the impulse response matrices of (3.149) and (3.150), respectively. Show that at the times when the impulse responses are nonzero, they satisfy $H(t, \tau) = H_a(\tau, t)^T$.
(b) Show that $H(s) = -H_a(-s)^T$, where $H(s)$ and $H_a(s)$ are the transfer matrices of (3.149) and (3.150), respectively.

3.17. Compute e^{At} for

$$A = \begin{bmatrix} 1 & 4 & 10 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

3.18. Given is the matrix

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

(a) Determine e^{At}, using the different methods covered in this text. Discuss the advantages and disadvantages of these methods.
(b) For system (3.1) let A be as given. Plot the components of the solution $\phi(t, t_0, x_0)$ when $x_0 = x(0) = (1, 1, 1)^T$ and $x_0 = x(0) = (2/3, 1, 0)^T$. Discuss the differences in these plots, if any.

3.19. Show that for $A = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$, we have $e^{At} = e^{at} \begin{bmatrix} \cos bt & \sin bt \\ -\sin bt & \cos bt \end{bmatrix}$.

3.20. Given is the system of equations

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u$$

with $x(0) = (1, 0)^T$ and

$$u(t) = p(t) = \begin{cases} 1, & t \geq 0, \\ 0, & \text{elsewhere.} \end{cases}$$

Plot the components of the solution of ϕ. For different initial conditions $x(0) = (a, b)^T$, investigate the changes in the asymptotic behavior of the solutions.

3.21. The system (3.1) with $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ is called the *harmonic oscillator* (refer to Chapter 1) because it has periodic solutions $\phi(t) = (\phi_1(t), \phi_2(t))^T$. Simultaneously, for the same values of t, plot $\phi_1(t)$ along the horizontal axis and $\phi_2(t)$ along the vertical axis in the x_1-x_2 plane to obtain a *trajectory* for this system for the specific initial condition $x(0) = x_0 = (x_1(0), x_2(0))^T = (1, 1)^T$. In plotting such trajectories, time t is viewed as a parameter, and arrows are used to indicate increasing time. When the horizontal axis corresponds to position and the vertical axis corresponds to velocity, the x_1-x_2 plane is called the *phase plane* and ϕ_1, ϕ_2 (resp. x_1, x_2) are called *phase variables.*

3.22. First, determine the solution ϕ of $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ with $x(0) = (1, 1)^T$. Next, determine the solution ϕ of the above system for $x(0) = \alpha(1, -1)^T, \alpha \in R, \alpha \neq 0$, and discuss the properties of the two solutions.

3.23. In Subsection 3.3.3 it is shown that when the n eigenvalues λ_i of a real $n \times n$ matrix A are distinct, then $e^{At} = \sum_{i=1}^n A_i e^{\lambda_i t}$ where $A_i = \lim_{s \to \lambda_i} [(s - \lambda_i)(sI - A)^{-1}] = v_i \tilde{v}_i$ [refer to (3.53), (3.54), and (3.57)], where $v_i, \tilde{v}_i$ are the right and left eigenvectors of A, respectively, corresponding to the eigenvalue λ_i. Show that (a) $\sum_{i=1}^n A_i = I$, where I denotes the $n \times n$ identity matrix, (b) $AA_i = \lambda_i A_i$, (c) $A_i A = \lambda_i A_i$, (d) $A_i A_j = \delta_{ij} A_i$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ when $i \neq j$.

3.24. Consider the system

$$\dot{x} = Ax + Bu, \quad y = Cx, \tag{3.151}$$

where all symbols are defined as in (3.61) with $D = 0$. Let

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 2 \\ 0 & 0 & 0 & 1 \\ 0 & -2 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = [1,\ 0,\ 1,\ 0]. \tag{3.152}$$

(a) Find equivalent representations for system (3.151), (3.152), given by

$$\dot{\tilde{x}} = \widetilde{A}\tilde{x} + \widetilde{B}u, \quad y = \widetilde{C}\tilde{x}, \tag{3.153}$$

where $\tilde{x} = Px$, when $\widetilde{A}$ is in (i) the Jordan canonical (or diagonal) form and (ii) the companion form.

(b) Determine the transfer function matrix for this system.

3.25. Consider the system (3.61) with $B = 0$.

(a) Let

$$A = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix} \quad \text{and} \quad C = [1,\ 1,\ 1].$$

If possible, select $x(0)$ in such a manner so that $y(t) = te^{-t}, t \geq 0$.

(b) Determine conditions under which it is possible to specify $y(t), t \geq 0$, using only the initial data $x(0)$.

3.26. Consider the system given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ -1/2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1/2 \end{bmatrix} u, \quad y = [1, 0] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

(a) Determine $x(0)$ so that for $u(t) = e^{-4t}, y(t) = ke^{-4t}$, where k is a real constant. Determine k for the present case. Notice that $y(t)$ does not have any transient components.
(b) Let $u(t) = e^{\alpha t}$. Determine $x(0)$ that will result in $y(t) = ke^{\alpha t}$. Determine the conditions on α for this to be true. What is k in this case?

3.27. Consider the system (3.61) with

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 3 & 0 & -3 & 1 \\ -1 & 1 & 4 & -1 \\ 1 & 0 & -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

(a) For $x(0) = [1, 1, 1, 1]^T$ and $u(t) = [1, 1]^T, t \geq 0$, determine the solution $\phi(t, 0, x(0))$ and the output $y(t)$ for this system and plot the components $\phi_i(t, 0, x(0)), i = 1, 2, 3, 4$ and $y_i(t), i = 1, 2$.
(b) Determine the transfer function matrix $H(s)$ for this system.

3.28. Consider the system

$$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k), \tag{3.154}$$

where all symbols are defined as in (3.91) with $D = 0$. Let

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad C = [1\, 1],$$

and let $x(0) = 0$ and $u(k) = 1, k \geq 0$.

(a) Determine $\{y(k)\}, k \geq 0$, by working in the (i) time domain and (ii) z-transform domain, using the transfer function $H(z)$.
(b) If it is known that when $u(k) = 0$, then $y(0) = y(1) = 1$, can $x(0)$ be uniquely determined? If your answer is affirmative, determine $x(0)$.

3.29. Consider $\hat{y}(z) = H(z)\hat{u}(z)$ with transfer function $H(z) = 1/(z + 0.5)$.

(a) Determine and plot the unit pulse response $\{h(k)\}$.
(b) Determine and plot the unit step response.

(c) If

$$u(k) = \begin{cases} 1, & k = 1, 2, \\ 0, & \text{elsewhere,} \end{cases}$$

determine $\{y(k)\}$ for $k = 0, 1, 2, 3$, and 4 via (i) convolution and (ii) the z-transform. Plot your answer.

(d) For $u(k)$ given in (c), determine $y(k)$ as $k \to \infty$.

3.30. Consider the system (3.91) with $x(0) = x_0$ and $k \geq 0$. Determine conditions under which there exists a sequence of inputs so that the state remains at x_0, i.e., so that $x(k) = x_0$ for all $k \geq 0$. How is this input sequence determined? Apply your method to the specific case

$$A = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad x_0 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

3.31. For system (3.92) with $x(0) = x_0$ and $k \geq 0$, it is desired to have the state go to the zero state for any initial condition x_0 in at most n steps; i.e., we desire that $x(k) = 0$ for any $x_0 = x(0)$ and for all $k \geq n$.

(a) Derive conditions in terms of the eigenvalues of A under which the above is true. Determine the minimum number of steps under which the above behavior will be true.

(b) For part (a), consider the specific cases

$$A_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hint: Use the Jordan canonical form for A. Results of this type are important in *deadbeat control*, where it is desired that a system variable attains some desired value and settles at that value in a finite number of time steps.

3.32. Consider a continuous-time system described by the transfer function $H(s) = 4/(s^2 + 2s + 2)$; i.e., $\hat{y}(s) = H(s)\hat{u}(s)$.

(a) Assume that the system is at rest, and assume a unit step input; i.e., $u(t) = 1, t \geq 0, u(t) = 0, t < 0$. Determine and plot $y(t)$ for $t \geq 0$.

(b) Obtain a discrete-time approximation for the above system by following these steps: (i) Determine a *realization* of the form (3.61) of $H(s)$ (see Exercise 3.33); (ii) assuming a sampler and a zero-order hold with sampling period T, use (3.151) to obtain a discrete-time system representation

$$\bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}\bar{u}(k), \quad \bar{y}(k) = \bar{C}\bar{x}(k) + \bar{D}\bar{u}(k) \tag{3.155}$$

and determine $\bar{A}, \bar{B}$, and $\bar{C}$ in terms of T.

(c) For the unit step input, $u(k) = 1$ for $k \geq 0$ and $u(k) = 0$ for $k < 0$, determine and plot $\bar{y}(k), k \geq 0$, for different values of T, assuming the system is at rest. Compare $\bar{y}(k)$ with $y(t)$ obtained in part (a).

(d) Determine for (3.155) the transfer function $\bar{H}(z)$ in terms of T. Note that $\bar{H}(z) = \bar{C}(zI - \bar{A})^{-1}\bar{B} + \bar{D}$. It can be shown that $\bar{H}(z) = (1 - z^{-1})\mathcal{Z}\{\mathcal{L}^{-1}[H(s)/s]_{t=kT}\}$. Verify this for the given $H(s)$.

3.33. Given a proper rational transfer function matrix $H(s)$, the state-space representation $\{A, B, C, D\}$ is called a *realization of* $H(s)$ if $H(s) = C(sI - A)^{-1}B + D$. Thus, the system (3.61) is a realizations of $H(s)$ if its transfer function matrix is equal to $H(s)$. Realizations of $H(s)$ are studied at length in Chapter 8. When $H(s)$ is scalar, it is straightforward to derive certain realizations, and in the following, we consider one such realization.

Given a proper rational scalar transfer function $H(s)$, let $D \triangleq \lim_{s\to\infty} H(s)$ and let

$$H_{sp}(s) \triangleq H(s) - D = \frac{b_{n-1}s^{n-1} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0},$$

a strictly proper rational function.

(a) Let

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{3.156}$$

$$C = [b_0 \; b_1 \cdots b_{n-1}],$$

and show that $\{A, B, C, D\}$ is indeed a realization of $H(s)$. Also, show that $\{\widetilde{A} = A^T, \widetilde{B} = C^T, \widetilde{C} = B^T, \widetilde{D} = D\}$ is a realization of $H(s)$ as well. These two state-space representations are said to be in *controller (companion) form* and in *observer (companion) form*, respectively (refer to Chapter 6).

(b) In particular find realizations in controller and observer form for (i) $H(s) = 1/s^2$, (ii) $H(s) = \omega_n^2/(s^2 + 2\zeta\omega_n s + \omega_n^2)$, and (iii) $H(s) = (s+1)^2/(s-1)^2$.

3.34. Assume that $H(s)$ is a $p \times m$ proper rational transfer function matrix. Expand $H(s)$ in a Laurent series about the origin to obtain

$$H(s) = H_0 + H_1 s^{-1} + \cdots + H_k s^{-k} + \cdots = \sum_{k=0}^{\infty} H_k s^{-k}. \tag{3.157}$$

The elements of the sequence $\{H_0, H_1, \ldots, H_k, \ldots\}$ are called the *Markov parameters* of the system. These parameters provide an alternative representation of the transfer function matrix $H(s)$, and they are useful in Realization Theory (refer to Chapter 8).

(a) Show that the impulse response $H(t, 0)$ can be expressed as

$$H(t, 0) = H_0\delta(t) + \sum_{k=1}^{\infty} H_k(t^{k-1}/(k-1)!). \tag{3.158}$$

In the following discussion, we assume that the system in question is described by (3.61).

(b) Show that

$$H(s) = D + C(sI - A)^{-1}B = D + \sum_{k=1}^{\infty}[CA^{k-1}B]s^{-k}, \tag{3.159}$$

which shows that the elements of the sequence $\{D, CB, CAB, \ldots, CA^{k-1}B, \ldots\}$ are the Markov parameters of the system; i.e., $H_0 = D$ and $H_k = CA^{k-1}B$, $k = 1, 2, \ldots$.

(c) Show that

$$H(s) = D + \frac{1}{\alpha(s)}C[R_{n-1}s^{n-1} + \cdots + R_1 s + R_0]B, \tag{3.160}$$

where $\alpha(s) = s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0 = \det(sI - A)$, the characteristic polynomial of A, and $R_{n-1} = I, R_{n-2} = AR_{n-1} + a_{n-1}I = A + a_{n-1}I, \ldots, R_0 = A^{n-1} + a_{n-1}A^{n-2} + \cdots + a_1 I$.

Hint: Write $(sI - A)^{-1} = \dfrac{1}{\alpha(s)}[\text{adjoint}(sI - A)] = \dfrac{1}{\alpha(s)}[R_{n-1}s^{n-1} + \cdots + R_1 s + R_0]$, and equate the coefficients of equal powers of s in the expression

$$\alpha(s)I = (sI - A)[R_{n-1}s^{n-1} + \cdots + R_1 s + R_0]. \tag{3.161}$$

3.35. The *frequency response matrix* of a system described by its $p \times m$ transfer function matrix evaluated at $s = j\omega$,

$$H(\omega) \triangleq \widehat{H}(s)|_{s=j\omega},$$

is a very useful means of characterizing a system, since typically it can be determined experimentally, and since control system specifications are frequently expressed in terms of the frequency responses of transfer functions. When the poles of $\widehat{H}(s)$ have negative real parts, the system turns out to be bounded-input/bounded-output (BIBO) stable (refer to Chapter 4). Under these conditions, the frequency response $H(\omega)$ has a clear physical meaning, and this fact can be used to determine $H(\omega)$ experimentally.

(a) Consider a stable SISO system given by $\hat{y}(s) = \widehat{H}(s)\hat{u}(s)$. Show that if $u(t) = k\sin(\omega_0 t + \phi)$ with k constant, then $y(t)$ at steady-state (i.e., after all transients have died out) is given by

$$y_{ss}(t) = k|H(\omega_0)|\sin(\omega_0 t + \phi + \theta(\omega_0)),$$

where $|H(\omega)|$ denotes the magnitude of $H(\omega)$ and $\theta(\omega) = \arg H(\omega)$ is the argument or phase of the complex quantity $H(\omega)$.
From the above it follows that $H(\omega)$ completely characterizes the system response at steady state (of a stable system) to a sinusoidal input. Since $u(t)$ can be expressed in terms of a series of sinusoidal terms via a Fourier series (recall that $u(t)$ is piecewise continuous), $H(\omega)$ characterizes the steady-state response of a stable system to any bounded input $u(t)$. This physical interpretation does not apply when the system is not stable.

(b) For the $p \times m$ transfer function matrix $\widehat{H}(s)$, consider the frequency response matrix $H(\omega)$ and extend the discussion of part (a) above to MIMO systems to give a physical interpretation of $H(\omega)$.

3.36. (*Double integrator*)

(a) Plot the response of the double integrator of Example 3.33 to a unit step input.

(b) Consider the discrete-time state-space representation of the double integrator of Example 3.33 for $T = 0.5$, 1, 5 sec and plot the unit step responses. Compare with your results in (a).

3.37. *(Spring mass system)* Consider the spring mass system of Example 1.1. For $M_1 = 1$ kg, $M_2 = 1$ kg, $K = 0.091$ N/m, $K_1 = 0.1$ N/m, $K_2 = 0.1$ N/m, $B = 0.0036$ N sec/m, $B_1 = 0.05$ N sec/m, and $B_2 = 0.05$ N sec/m, the state-space representation of the system in (1.27) assumes the form

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -0.1910 & -0.0536 & 0.0910 & 0.0036 \\ 0 & 0 & 0 & 1 \\ 0.0910 & 0.0036 & -0.1910 & -0.0536 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

where $x_1 \triangleq y_1$, $x_2 \triangleq \dot{y}_1$, $x_3 \triangleq y_2$, and $x_4 \triangleq \dot{y}_2$.

(a) Determine the eigenvalues and eigenvectors of the matrix A of the system and express $x(t)$ in terms of the modes and the initial conditions $x(0)$ of the system, assuming that $f_1 = f_2 = 0$.

(b) For $x(0) = [1, 0, -0.5, 0]^T$ and $f_1 = f_2 = 0$, plot the states for $t \geq 0$.

(c) Let $y = Cx$ with $C = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \end{bmatrix}$ denote the output of the system. Determine the transfer function between y and $u \triangleq [f_1, f_2]^T$.

(d) For zero initial conditions, $f_1(t) = \delta(t)$ (the unit impulse), and $f_2(t) = 0$, plot the states for $t \geq 0$ and comment on your results.

(e) It is desirable to explore what happens when the mass ratio M_2/M_1 takes on different values. For this, let $M_2 = \alpha M_1$ with $M_1 = 1$ kg and $\alpha = 0.1$, 0.5, 2, 5. All other parameter values remain the same. Repeat (a) to (d) for the different values of α and discuss your results.

3.38. *(Automobile suspension system)* [M.L. James, G.M. Smith, and J.C. Wolford, *Applied Numerical Methods for Digital Computation*, Harper and Row, 1985, p. 667.] Consider the spring mass system in Figure 3.4, which describes part of the suspension system of an automobile. The data for this system are given as

$$\begin{aligned}
m_1 &= \tfrac{1}{4} \times \text{(mass of the automobile)} = 375 \text{ kg},\\
m_2 &= \text{ mass of one wheel} = 30 \text{ kg},\\
k_1 &= \text{ spring constant} = 1500 \text{ N/m},\\
k_2 &= \text{ linear spring constant of tire} = 6500 \text{ N/m},\\
c &= \text{ damping constant of dashpot} = 0,\ 375,\ 750, \text{ and } 1125 \text{ N sec/m},\\
x_1 &= \text{ displacement of automobile body from equilibrium position m},\\
x_3 &= \text{ displacement of wheel from equilibrium position m},\\
v &= \text{ velocity of car} = 9,\ 18,\ 27, \text{ or } 36 \text{ m/sec}.
\end{aligned}$$

A linear model $\dot{x} = Ax + Bu$ for this system is given by

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -\frac{k_1}{m_1} & -\frac{c}{m_1} & \frac{k_1}{m_1} & \frac{c}{m_1} \\ 0 & 0 & 0 & 1 \\ \frac{k_1}{m_2} & \frac{c}{m_2} & -\frac{k_1+k_2}{m_2} & -\frac{c}{m_2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{k_2}{m_2} \end{bmatrix} u(t),$$

where $u(t) = \frac{1}{6} \sin \frac{2\pi vt}{20}$ describes the profile of the roadway.

(a) Determine the eigenvalues of A for all of the above cases.
(b) Plot the states for $t \geq 0$ when the input $u(t) = \frac{1}{6} \sin \frac{2\pi vt}{20}$ and $x(0) = [0, 0, 0, 0]^T$ for all the above cases. Comment on your results.

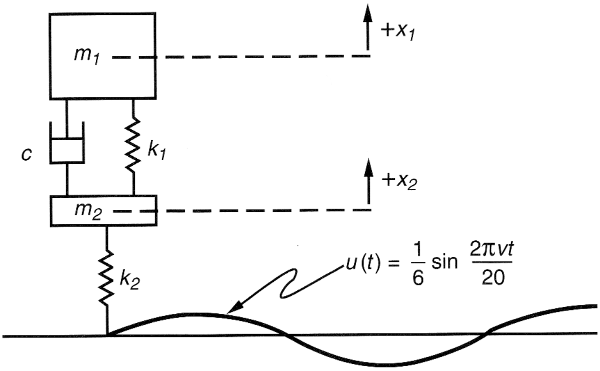


Figure 3.4. Model of an automobile suspension system

4

Stability

Dynamical systems, either occurring in nature or man made, usually function in some specified mode. The most common such modes are operating points that frequently turn out to be equilibria.

In this chapter we will concern ourselves primarily with the qualitative behavior of equilibria. Most of the time, we will be interested in the asymptotic stability of an equilibrium (operating point), which means that when the state of a given system is displaced (disturbed) from its desired operating point (equilibrium), the expectation is that the state will eventually return to the equilibrium. For example, in the case of an automobile under cruise control, traveling at the desired constant speed of 50 mph (which determines the operating point, or equilibrium condition), perturbations due to hill climbing (hill descending), will result in decreasing (increasing) speeds. In a properly designed cruise control system, it is expected that the car will return to its desired operating speed of 50 mph.

Another qualitative characterization of dynamical systems is the expectation that bounded system inputs will result in bounded system outputs, and that small changes in inputs will result in small changes in outputs. System properties of this type are referred to as input–output stability. Such properties are important for example in tracking systems, where the output of the system is expected to follow a desired input. Frequently, it is possible to establish a connection between the input–output stability properties and the Lyapunov stability properties of an equilibrium. In the case of linear systems, this connection is well understood. This will be addressed in Section 7.3.

4.1 Introduction

In this chapter we present a brief introduction to stability theory. We are concerned primarily with linear systems and systems that are a consequence of linearizations of nonlinear systems. As in the other chapters of this book, we consider finite-dimensional continuous-time systems and finite-dimensional discrete-time systems described by systems of first-order ordinary differential equations and systems of first-order ordinary difference equations, respectively.

In Section 4.2 we introduce the concept of equilibrium of dynamical systems described by systems of first-order ordinary differential equations, and in Section 4.3 we give definitions of various types of stability in the sense of Lyapunov (including stability, uniform stability, asymptotic stability, uniform asymptotic stability, exponential stability, and instability).

In Section 4.4 we establish conditions for the various Lyapunov stability and instability types enumerated in Section 4.3 for linear systems $\dot{x} = Ax$. Most of these results are phrased in terms of the properties of the state transition matrix for such systems.

In Section 4.5 we introduce the Second Method of Lyapunov, also called the Direct Method of Lyapunov, to establish necessary and sufficient conditions for various Lyapunov stability types of an equilibrium for linear systems $\dot{x} = Ax$. These results, which are phrased in terms of the system parameters [coefficients of the matrix A], give rise to the Lyapunov matrix equation.

In Section 4.6 we use the Direct Method of Lyapunov in deducing the asymptotic stability and instability of an equilibrium of nonlinear autonomous systems $\dot{x} = Ax + F(x)$ from the stability properties of their linearizations $\dot{w} = Aw$.

In Section 4.7 we establish necessary and sufficient conditions for the input–output stability (more precisely, for the bounded input/bounded output stability) of continuous-time, linear, time-invariant systems. These results involve the system impulse response matrix.

The stability results presented in Sections 4.2 through and including Section 4.7 pertain to continuous-time systems. In Section 4.8 we present analogous stability results for discrete-time systems.

4.2 The Concept of an Equilibrium

In this section we concern ourselves with systems of first-order autonomous ordinary differential equations,

$$\dot{x} = f(x), \tag{4.1}$$

where $x \in R^n$. When discussing global results, we shall assume that $f : R^n \to R^n$, while when considering local results, we may assume that $f : B(h) \to R^n$ for some $h > 0$, where $B(h) = \{x \in R^n :\| x \|< h\}$ and $\| \cdot \|$ denotes a norm on R^n. Unless otherwise stated, we shall assume that for every $(t_0, x_0), t_0 \in R^+$, the initial-value problem

$$\dot{x} = f(x), \quad x(t_0) = x_0 \tag{4.2}$$

possesses a unique solution $\phi(t, t_0, x_0)$ that exists for all $t \geq t_0$ and that depends continuously on the initial data (t_0, x_0). Refer to Section 1.5 for conditions that ensure that (4.2) has these properties. Since (4.1) is time-invariant, we may assume without loss of generality that $t_0 = 0$ and we will denote the solutions of (4.1) by $\phi(t, x_0)$ (rather than $\phi(t, t_0, x_0)$) with $x(0) = x_0$.

Definition 4.1. *A point $x_e \in R^n$ is called an* equilibrium point *of (4.1), or simply an* equilibrium *of (4.1), if*

$$f(x_e) = 0.$$

■

We note that

$$\phi(t, x_e) = x_e \quad \text{for all } t \geq 0;$$

i.e., the equilibrium x_e is the unique solution of (4.1) with initial data given by $\phi(0, x_e) = x_e$.

We will usually assume that in a given discussion, unless otherwise stated, the equilibrium of interest is located at the origin of R^n. This assumption can be made without loss of generality by noting that if $x_e \neq 0$ is an equilibrium point of (4.1), i.e., $f(x_e) = 0$, then by letting $w = x - x_e$, we obtain the transformed system

$$\dot{w} = F(w) \tag{4.3}$$

with $F(0) = 0$, where

$$F(w) = f(w + x_e). \tag{4.4}$$

Since the above transformation establishes a one-to-one correspondence between the solutions of (4.1) and (4.3), we may assume henceforth that the equilibrium of interest for (4.1) is located at the origin. This equilibrium, $x = 0$, will be referred to as the *trivial solution* of (4.1).

Before concluding this section, it may be fruitful to consider some specific cases.

Example 4.2. In Example 1.4 we considered the simple pendulum given in Figure 1.7. Letting $x_1 = x$ and $x_2 = \dot{x}$ in (1.37), we obtain the system of equations

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -k \sin x_1, \end{aligned} \tag{4.5}$$

where $k > 0$ is a constant. *Physically*, the pendulum has two equilibrium points: one where the mass M is located vertically at the bottom of the figure (i.e., at 6 o'clock) and the other where the mass is located vertically at the top of the figure (i.e., at 12 o'clock). The *model* of this pendulum, however, described by (4.5), has countably infinitely many equilibrium points that are located in R^2 at the points $(\pi n, 0)^T, n = 0, \pm 1, \pm 2, \ldots$.

Example 4.3. The linear, autonomous, homogenous system of ordinary differential equations

$$\dot{x} = Ax \tag{4.6}$$

has a unique equilibrium that is at the origin if and only if A is nonsingular. Otherwise, (4.6) has nondenumerably many equilibria. [Refer to Chapter 1 for the definitions of symbols in (4.6).]

Example 4.4. Assume that for

$$\dot{x} = f(x), \tag{4.7}$$

f is continuously differentiable with respect to all of its arguments, and let

$$J(x_e) = \frac{\partial f}{\partial x}(x)\bigg|_{x=x_e},$$

where $\partial f/\partial x$ denotes the $n \times n$ *Jacobian matrix* defined by

$$\frac{\partial f}{\partial x} = \left[\frac{\partial f_i}{\partial x_j}\right].$$

If $f(x_e) = 0$ and $J(x_e)$ is nonsingular, then x_e is an equilibrium of (4.7).

Example 4.5. The system of ordinary differential equations given by

$$\begin{aligned}\dot{x}_1 &= k + \sin(x_1 + x_2) + x_1,\\ \dot{x}_2 &= k + \sin(x_1 + x_2) - x_1,\end{aligned}$$

with $k > 1$, has no equilibrium points at all.

4.3 Qualitative Characterizations of an Equilibrium

In this section we consider several qualitative characterizations that are of fundamental importance in systems theory. These characterizations are concerned with various types of stability properties of an equilibrium and are referred to in the literature as *Lyapunov stability*.

Throughout this section, we consider systems of equations

$$\dot{x} = f(x), \tag{4.8}$$

and we assume that (4.8) possesses an equilibrium at the origin. We thus have $f(0) = 0$.

Definition 4.6. *The equilibrium $x = 0$ of (4.8) is said to be* stable *if for every $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that*

$$\| \phi(t, x_0) \| < \epsilon \textit{ for all } t \geq 0 \tag{4.9}$$

whenever

$$\| x_0 \| < \delta(\epsilon). \tag{4.10}$$

■

In Definition 4.6, $\| \cdot \|$ denotes any one of the equivalent norms on R^n, and (as in Chapters 1 and 2) $\phi(t, x_0)$ denotes the solution of (4.8) with initial condition $x(0) = x_0$. The notation $\delta(\epsilon)$ indicates that δ depends on the choice of ϵ.

In words, Definition 4.6 states that by choosing the initial points in a sufficiently small spherical neighborhood, when the equilibrium $x = 0$ of (4.8) is stable, we can force the graph of the solution for $t \geq 0$ to lie entirely inside a given cylinder. This is depicted in Figure 4.1 for the case $x \in R^2$.

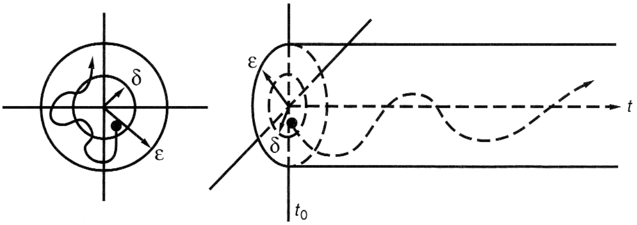


Figure 4.1. Stability of an equilibrium

Definition 4.7. *The equilibrium $x = 0$ of (4.8) is said to be* asymptotically stable *if*

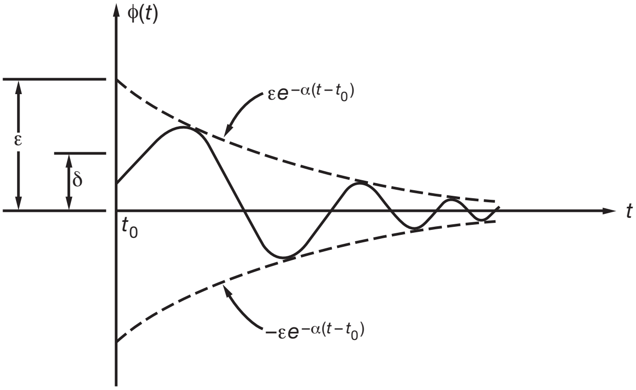
(i) it is stable,
(ii) there exists an $\eta > 0$ such that $\lim_{t \to \infty} \phi(t, x_0) = 0$ whenever $\| x_0 \| < \eta$. ■

The set of all $x_0 \in R^n$ such that $\phi(t, x_0) \to 0$ as $t \to \infty$ is called the *domain of attraction* of the equilibrium $x = 0$ of (4.8). Also, if for (4.8) condition (ii) is true, then the equilibrium $x = 0$ is said to be *attractive*.

Definition 4.8. *The equilibrium $x = 0$ of (4.8) is* exponentially stable *if there exists an $\alpha > 0$, and for every $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$, such that*

$$\| \phi(t, x_0) \| \leq \epsilon e^{\alpha t} \textit{ for all } t \geq 0$$

whenever $||x_0|| < \delta(\epsilon)$. ■



$\phi(t)$
$\varepsilon e^{-\alpha(t-t_0)}$
ε
δ
t_0
t
$-\varepsilon e^{-\alpha(t-t_0)}$

Figure 4.2. An exponentially stable equilibrium

Figure 4.2 shows the behavior of a solution in the vicinity of an exponentially stable equilibrium $x = 0$.

Definition 4.9. *The equilibrium $x = 0$ of (4.8) is* unstable *if it is not stable. In this case, there exists an $\epsilon > 0$, and a sequence $x_m \to 0$ of initial points and a sequence $\{t_m\}$ such that $\| \phi(t_m, x_m) \| \geq \epsilon$ for all $m, t_m \geq 0$.* ■

If $x = 0$ is an unstable equilibrium of (4.8), then it still can happen that all the solutions tend to zero with increasing t. This indicates that instability and attractivity of an equilibrium are compatible concepts. We note that the equilibrium $x = 0$ of (4.8) is necessarily unstable if every neighborhood of the origin contains initial conditions corresponding to unbounded solutions (i.e., solutions whose norm grows to infinity on a sequence $t_m \to \infty$). However, it can happen that a system (4.8) with unstable equilibrium $x = 0$ may have only bounded solutions.

The concepts that we have considered thus far pertain to *local* properties of an equilibrium. In the following discussion, we consider *global* characterizations of an equilibrium.

Definition 4.10. *The equilibrium $x = 0$ of (4.8) is* asymptotically stable in the large *if it is stable and if every solution of (4.8) tends to zero as $t \to \infty$.* ■

When the equilibrium $x = 0$ of (4.8) is asymptotically stable in the large, its domain of attraction is all of R^n. Note that in this case, $x = 0$ is the *only* equilibrium of (4.8).

Definition 4.11. *The equilibrium $x = 0$ of (4.8) is* exponentially stable in the large *if there exists $\alpha > 0$ and for any $\beta > 0$, there exists $k(\beta) > 0$ such that*

$$\| \phi(t, x_0) \| \leq k(\beta) \| x_0 \| e^{-\alpha t} \quad \textit{for all } t \geq 0$$

whenever $\| x_0 \| < \beta$. ■

We conclude this section with a few specific cases.

The scalar differential equation

$$\dot{x} = 0 \tag{4.11}$$

has for any initial condition $x(0) = x_0$ the solution $\phi(t, x_0) = x_0$; i.e., all solutions are equilibria of (4.11). The trivial solution is stable; however, it is not asymptotically stable.

The scalar differential equation

$$\dot{x} = ax \tag{4.12}$$

has for every $x(0) = x_0$ the solution $\phi(t, x_0) = x_0 e^{at}$, and $x = 0$ is the only equilibrium of (4.12). If $a > 0$, this equilibrium is unstable, and when $a < 0$, this equilibrium is exponentially stable in the large.

As mentioned earlier, a system

$$\dot{x} = f(x) \tag{4.13}$$

can have all solutions approaching an equilibrium, say, $x = 0$, without this equilibrium being asymptotically stable. An example of this type of behavior is given by the *nonlinear* system of equations

$$\begin{aligned} \dot{x}_1 &= \frac{x_1^2(x_2 - x_1) + x_2^5}{(x_1^2 + x_2^2)[1 + (x_1^2 + x_2^2)^2]}, \\ \dot{x}_2 &= \frac{x_2^2(x_2 - 2x_1)}{(x_1^2 + x_2^2)[1 + (x_1^2 + x_2^2)^2]}. \end{aligned}$$

For a detailed discussion of this system, refer to [6], pp. 191–194, cited at the end of this chapter.

Before proceeding any further, a few comments are in order concerning the reasons for considering equilibria and their stability properties as well as other types of stability that we will encounter. To this end we consider linear time-invariant systems given by

$$\dot{x} = Ax + Bu, \tag{4.14a}$$

$$y = Cx + Du, \tag{4.14b}$$

where all symbols in (4.14) are defined as in (2.7). The usual qualitative analysis of such systems involves two concepts, *internal stability* and *input–output stability*.

In the case of *internal stability*, the output equation (4.14b) plays no role whatsoever, the system input u is assumed to be identically zero, and the focus of the analysis is concerned with the qualitative behavior of the solutions of linear time-invariant systems

$$\dot{x} = Ax \tag{4.15}$$

near the equilibrium $x = 0$. This is accomplished by making use of the various types of Lyapunov stability concepts introduced in this section. In other words, internal stability of system (4.14) concerns the Lyapunov stability of the equilibrium $x = 0$ of system (4.15).

In the case of *input–output stability*, we view systems as operators determined by (4.14) that relate outputs y to inputs u and the focus of the analysis is concerned with qualitative relations between system inputs and system outputs. We will address this type of stability in Section 4.7.

4.4 Lyapunov Stability of Linear Systems

In this section we first study the stability properties of the equilibrium $x = 0$ of linear autonomous homogeneous systems

$$\dot{x} = Ax, \quad t \geq 0. \tag{4.16}$$

Recall that $x = 0$ is always an equilibrium of (4.16) and that $x = 0$ is the only equilibrium of (4.16) if A is nonsingular. Recall also that the solution of (4.16) for $x(0) = x_0$ is given by

$$\begin{aligned} \phi(t, x_0) &= \Phi(t, 0)x_0 = \Phi(t - 0, 0)x_0 \\ &\triangleq \Phi(t)x_0 = e^{At}x_0, \end{aligned}$$

where in the preceding equation, a slight abuse of notation has been used.

We first consider some of the basic properties of system (4.16).

Theorem 4.12. *The equilibrium $x = 0$ of (4.16) is* stable *if and only if the solutions of (4.16) are bounded, i.e., if and only if*

$$\sup_{t \geq t_0} \| \Phi(t) \| \triangleq k < \infty,$$

where $\| \Phi(t) \|$ denotes the matrix norm induced by the vector norm used on R^n and k denotes a constant.

Proof. Assume that the equilibrium $x = 0$ of (4.16) is stable. Then for $\epsilon = 1$ there is a $\delta = \delta(1) > 0$ such that $\| \phi(t, x_0) \| < 1$ for all $t \geq 0$ and all x_0 with $\| x_0 \| \leq \delta$. In this case

$$\| \phi(t, x_0) \| = \| \Phi(t)x_0 \| = \| [\Phi(t)(x_0\delta)/ \| x_0 \|] \| (\| x_0 \| /\delta) < \| x_0 \| /\delta$$

for all $x_0 \neq 0$ and all $t \geq 0$. Using the definition of matrix norm [refer to Section A.7], it follows that

$$\| \Phi(t) \| \leq \delta^{-1}, \quad t \geq 0.$$

We have proved that if the equilibrium $x = 0$ of (4.16) is stable, then the solutions of (4.16) are bounded.

Conversely, suppose that all solutions $\phi(t, x_0) = \Phi(t)x_0$ are bounded. Let $\{e_1, \dots, e_n\}$ denote the natural basis for n-space, and let $\| \phi(t, e_j) \| < \beta_j$ for all $t \geq 0$. Then for any vector $x_0 = \sum_{j=1}^n \alpha_j e_j$ we have that

$$\begin{aligned} \| \phi(t, x_0) \| = \| \sum_{j=1}^n \alpha_j \phi(t, e_j) \| &\leq \sum_{j=1}^n |\alpha_j| \beta_j \\ &\leq (\max_j \beta_j) \sum_{j=1}^n |\alpha_j| \leq k \| x_0 \| \end{aligned}$$

for some constant $k > 0$ for $t \geq 0$. For given $\epsilon > 0$, we choose $\delta = \epsilon / k$. Thus, if $\| x_0 \| < \delta$, then $\| \phi(t, x_0) \| < k \| x_0 \| < \epsilon$ for all $t \geq 0$. We have proved that if the solutions of (4.16) are bounded, then the equilibrium $x = 0$ of (4.16) is stable. ■

Theorem 4.13. *The following statements are equivalent.*

(i) *The equilibrium $x = 0$ of (4.16) is asymptotically stable.*
(ii) *The equilibrium $x = 0$ of (4.16) is asymptotically stable in the large.*
(iii) $\lim_{t \to \infty} \| \Phi(t) \| = 0$.

Proof. Assume that statement (i) is true. Then there is an $\eta > 0$ such that when $\| x_0 \| \leq \eta$, then $\phi(t, x_0) \to 0$ as $t \to \infty$. But then we have for any $x_0 \neq 0$ that

$$\phi(t, x_0) = \phi(t, \eta x_0 / \| x_0 \|)(\| x_0 \| / \eta) \to 0$$

as $t \to \infty$. It follows that statement (ii) is true.

Next, assume that statement (ii) is true. For any $\epsilon > 0$, there must exist a $T(\epsilon) > 0$ such that for all $t \geq T(\epsilon)$ we have that $\| \phi(t, x_0) \| = \| \Phi(t) x_0 \| < \epsilon$. To see this, let $\{e_1, \dots, e_n\}$ be the natural basis for R^n. Thus, for some fixed constant $k > 0$, if $x_0 = (\alpha_1, \dots, \alpha_n)^T$ and if $\| x_0 \| \leq 1$, then $x_0 = \sum_{j=1}^n \alpha_j e_j$ and $\sum_{j=1}^n |\alpha_j| \leq k$. For each j, there is a $T_j(\epsilon)$ such that $\| \Phi(t) e_j \| < \epsilon / k$ and $t \geq T_j(\epsilon)$. Define $T(\epsilon) = \max\{T_j(\epsilon) : j = 1, \dots, n\}$. For $\| x_0 \| \leq 1$ and $t \geq T(\epsilon)$, we have that

$$\| \Phi(t) x_0 \| = \| \sum_{j=1}^n \alpha_j \Phi(t) e_j \| \leq \sum_{j=1}^n |\alpha_j| (\epsilon / k) \leq \epsilon.$$

By the definition of the matrix norm [see the appendix], this means that $\| \Phi(t) \| \leq \epsilon$ for $t \geq T(\epsilon)$. Therefore, statement (iii) is true.

Finally, assume that statement (iii) is true. Then $\| \Phi(t) \|$ is bounded in t for all $t \geq 0$. By Theorem 4.12, the equilibrium $x = 0$ is stable. To prove asymptotic stability, fix $\epsilon > 0$. If $\| x_0 \| < \eta = 1$, then $\| \phi(t, x_0) \| \leq \| \Phi(t) \| \| x_0 \| \to 0$ as $t \to \infty$. Therefore, statement (i) is true. This completes the proof. ■

Theorem 4.14. *The equilibrium $x = 0$ of (4.16) is asymptotically stable if and only if it is exponentially stable.*

Proof. The exponential stability of the equilibrium $x = 0$ implies the asymptotic stability of the equilibrium $x = 0$ of systems (4.13) in general and, hence, for systems (4.16) in particular.

Conversely, assume that the equilibrium $x = 0$ of (4.16) is asymptotically stable. Then there is a $\delta > 0$ and a $T > 0$ such that if $\| x_0 \| \leq \delta$, then

$$\| \Phi(t+T)x_0 \| < \delta/2$$

for all $t \geq 0$. This implies that

$$\| \Phi(t+T) \| \leq \frac{1}{2} \text{ if } t \geq 0. \tag{4.17}$$

From Theorem 3.9 (iii) we have that $\Phi(t-\tau) = \Phi(t-\sigma)\Phi(\sigma-\tau)$ for any t, σ, and τ. Therefore,

$$\| \Phi(t+2T) \| = \| \Phi(t+2T-t-T)\Phi(t+T) \| \leq \frac{1}{4},$$

in view of (4.17). By induction, we obtain for $t \geq 0$ that

$$\| \Phi(t+nT) \| \leq 2^{-n}. \tag{4.18}$$

Now let $\alpha = (ln2)/T$. Then (4.18) implies that for $0 \leq t < T$ we have that

$$\begin{aligned} \| \phi(t+nT, x_0) \| &\leq 2 \| x_0 \| 2^{-(n+1)} = 2 \| x_0 \| e^{-\alpha(n+1)T} \\ &\leq 2 \| x_0 \| e^{-\alpha(t+nT)}, \end{aligned}$$

which proves the result. ■

Even though the preceding results require knowledge of the state transition matrix $\Phi(t)$ of (4.16), they are quite useful in the qualitative analysis of linear systems. In view of the above results, we can state the following equivalent definitions.

The equilibrium $x = 0$ of (4.16) is *stable* if and only if there exists a finite positive constant γ, such that for any x_0, the corresponding solution satisfies the inequality

$$\| \phi(t, x_0) \| \leq \gamma \| x_0 \|, \quad t \geq 0.$$

Furthermore, in view of the above results, if the equilibrium $x = 0$ of (4.16) is asymptotically stable, then in fact it must be globally asymptotically stable, and exponentially stable in the large. In this case there exist finite constants $\gamma \geq 1$ and $\lambda > 0$ such that

$$\| \phi(t, x_0) \| \leq \gamma e^{-\lambda t} \| x_0 \|$$

for $t \geq 0$ and $x_0 \in R^n$.

We now continue our investigation of system (4.16) by referring to the discussion in Subsection 3.3.2 [refer to (3.23) to (3.39)] concerning the use of the Jordan canonical form to compute $\exp(At)$. We let $J = P^{-1}AP$ and define $x = Py$. Then (4.16) yields

$$\dot{y} = P^{-1}APy = Jy. \tag{4.19}$$

It is easily verified (the reader is asked to do so in the Exercises section) that the equilibrium $x = 0$ of (4.16) is stable (resp., asymptotically stable or unstable) if and only if $y = 0$ of (4.19) is stable (resp., asymptotically stable or unstable). In view of this, we can assume without loss of generality that the matrix A in (4.16) is in Jordan canonical form, given by

$$A = \text{diag}[J_0, J_1, \dots, J_s],$$

where

$$J_0 = \text{diag}[\lambda_1, \dots, \lambda_k] \quad \text{and} \quad J_i = \lambda_{k+i} I_i + N_i$$

for the Jordan blocks $J_1, \dots, J_s$.

As in (3.33), (3.34), (3.38), and (3.39), we have

$$e^{At} = \begin{bmatrix} e^{J_0 t} & & & 0 \\ & e^{J_1 t} & & \\ & & \ddots & \\ 0 & & & e^{J_s t} \end{bmatrix},$$

where

$$e^{J_0 t} = \text{diag}[e^{\lambda_1 t}, \dots, e^{\lambda_k t}] \tag{4.20}$$

and

$$e^{J_i t} = e^{\lambda_{k+i} t} \begin{bmatrix} 1 & t & t^2/2 & \cdots & t^{n_i - 1}/(n_i - 1)! \\ 0 & 1 & t & \cdots & t^{n_i - 2}/(n_i - 2)! \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{4.21}$$

for $i = 1, \dots, s$.

Now suppose that $Re\lambda_i \leq \beta$ for all $i = 1, \dots, k$. Then it is clear that $\lim_{t\to\infty}(\| e^{J_0 t} \| / e^{\beta t}) < \infty$, where $\| e^{J_0 t} \|$ is the matrix norm induced by one of the equivalent vector norms defined on R^n. We write this as $\| e^{J_0 t} \| = \mathcal{O}(e^{\beta t})$. Similarly, if $\beta = Re\lambda_{k+i}$, then for any $\epsilon > 0$ we have that $\| e^{J_i t} \| = \mathcal{O}(t^{n_i - 1} e^{\beta t}) = \mathcal{O}(e^{(\beta + \epsilon)t})$.

From the foregoing it is now clear that $\| e^{At} \| \leq K$ for some $K > 0$ if and only if all eigenvalues of A have nonpositive real parts, and the eigenvalues with zero real part occur in the Jordan form only in J_0 and not in any of the Jordan blocks J_i, $1 \leq i \leq s$. Hence, by Theorem 4.12, the equilibrium $x = 0$ of (4.16) is under these conditions stable.

Now suppose that all eigenvalues of A have negative real parts. From the preceding discussion it is clear that there is a constant $K > 0$ and an $\alpha > 0$ such that $\| e^{At} \| \leq Ke^{-\alpha t}$, and therefore, $\| \phi(t, x_0) \| \leq Ke^{-\alpha t} \| x_0 \|$ for all $t \geq 0$ and for all $x_0 \in R^n$. It follows that the equilibrium $x = 0$ is asymptotically stable in the large, in fact exponentially stable in the large. Conversely, assume that there is an eigenvalue λ_i with a nonnegative real part. Then either one term in (4.20) does not tend to zero, or else a term in (4.21) is unbounded as $t \to \infty$. In either case, $e^{At}x(0)$ will not tend to zero when the initial condition $x(0) = x_0$ is properly chosen. Hence, the equilibrium $x = 0$ of (4.16) cannot be asymptotically stable (and, hence, it cannot be exponentially stable).

Summarizing the above, we have proved the following result.

Theorem 4.15. *The equilibrium $x = 0$ of (4.16) is* stable*, if and only if all eigenvalues of A have nonpositive real parts, and every eigenvalue with zero real part has an associated Jordan block of order one. The equilibrium $x = 0$ of (4.16) is* asymptotically stable in the large*, in fact* exponentially stable in the large*, if and only if all eigenvalues of A have negative real parts.* ■

A direct consequence of the above result is that the equilibrium $x = 0$ of (4.16) is *unstable* if and only if at least one of the eigenvalues of A has either positive real part or has zero real part that is associated with a Jordan block of order greater than one.

At this point, it may be appropriate to take note of certain conventions concerning matrices that are used in the literature. It should be noted that some of these are not entirely consistent with the terminology used in Theorem 4.15. Specifically, a real $n \times n$ matrix A is called *stable* or a *Hurwitz matrix* if all its eigenvalues have negative real parts. If at least one of the eigenvalues has a positive real part, then A is called *unstable*. A matrix A, which is neither stable nor unstable, is called *critical*, and the eigenvalues with zero real parts are called *critical eigenvalues*.

We conclude our discussion concerning the stability of (4.16) by noting that the results given above can also be obtained by directly using the facts established in Subsection 3.3.3, concerning modes and asymptotic behavior of time-invariant systems.

Example 4.16. We consider the system (4.16) with

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

The eigenvalues of A are $\lambda_1, \lambda_2 = \pm j$. According to Theorem 4.15, the equilibrium $x = 0$ of this system is stable. This can also be verified by computing the solution of this system for a given set of initial data $x(0)^T = (x_1(0), x_2(0))$,

$$\begin{aligned} \phi_1(t, x_0) &= x_1(0)\cos t + x_2(0)\sin t, \\ \phi_2(t, x_0) &= -x_1(0)\sin t + x_2(0)\cos t, \end{aligned}$$

$t \geq 0$, and then applying Definition 4.6.

Example 4.17. We consider the system (4.16) with

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

The eigenvalues of A are $\lambda_1 = 0, \lambda_2 = 0$. According to Theorem 4.15, the equilibrium $x = 0$ of this system is unstable. This can also be verified by computing the solution of this system for a given set of initial data $x(0)^T = (x_1(0), x_2(0))$,

$$\begin{aligned} \phi_1(t, x_0) &= x_1(0) + x_2(0)t, \\ \phi_2(t, x_0) &= x_2(0), \end{aligned}$$

$t \geq 0$, and then applying Definition 4.9. (Note that in this example, the entire x_1-axis consists of equilibria.)

Example 4.18. We consider the system (4.16) with

$$A = \begin{bmatrix} 2.8 & 9.6 \\ 9.6 & -2.8 \end{bmatrix}.$$

The eigenvalues of A are $\lambda_1, \lambda_2 = \pm 10$. According to Theorem 4.15, the equilibrium $x = 0$ of this system is unstable.

Example 4.19. We consider the system (4.16) with

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

The eigenvalues of A are $\lambda_1, \lambda_2 = -1, -2$. According to Theorem 4.15, the equilibrium $x = 0$ of this system is exponentially stable.

4.5 The Lyapunov Matrix Equation

In Section 4.4 we established a variety of stability results that require explicit knowledge of the solutions of (4.16). In this section we will develop stability criteria for (4.16) with *arbitrary* matrix A. In doing so, we will employ *Lyapunov's Second Method* (also called *Lyapunov's Direct Method*) for the case of linear systems (4.16). This method utilizes auxiliary real-valued functions

$v(x)$, called *Lyapunov functions*, that may be viewed as *generalized energy functions* or *generalized distance functions* (from the equilibrium $x = 0$), and the stability properties are then deduced directly from the properties of $v(x)$ and its time derivative $\dot{v}(x)$, evaluated along the solutions of (4.16).

A logical choice of Lyapunov function is $v(x) = x^T x = \| x \|^2$, which represents the square of the Euclidean distance of the state from the equilibrium $x = 0$ of (4.16). The stability properties of the equilibrium are then determined by examining the properties of $\dot{v}(x)$, the time derivative of $v(x)$ along the solutions of (4.16), which we repeat here,

$$\dot{x} = Ax. \tag{4.22}$$

This derivative can be determined *without explicitly solving for the solutions of (4.22)* by noting that

$$\begin{aligned}\dot{v}(x) &= \dot{x}^T x + x^T \dot{x} = (Ax)^T x + x^T (Ax)\\ &= x^T (A^T + A)x.\end{aligned}$$

If the matrix A is such that $\dot{v}(x)$ is negative for all $x \neq 0$, then it is reasonable to expect that the distance of the state of (4.22) from $x = 0$ will decrease with increasing time, and that the state will therefore tend to the equilibrium $x = 0$ of (4.22) with increasing time t.

It turns out that the Lyapunov function used in the above discussion is not sufficiently flexible. In the following discussion, we will employ as a "generalized distance function" the quadratic form given by

$$v(x) = x^T P x, \quad P = P^T, \tag{4.23}$$

where P is a real $n \times n$ matrix. The time derivative of $v(x)$ along the solutions of (4.22) is determined as

$$\begin{aligned}\dot{v}(x) &= \dot{x}^T P x + x^T P \dot{x} = x^T A^T P x + x^T P A x\\ &= x^T (A^T P + PA)x;\end{aligned}$$

i.e.,

$$\dot{v} = x^T C x, \tag{4.24}$$

where

$$C = A^T P + PA. \tag{4.25}$$

Note that C is real and $C^T = C$. The system of equations given in (4.25) is called the *Lyapunov Matrix Equation.*

We recall that since P is real and symmetric, all its eigenvalues are real. Also, we recall that P is said to be *positive definite* (resp., *positive semidefinite*) if all its eigenvalues are positive (resp., nonnegative), and it is called *indefinite* if P has eigenvalues of opposite sign. The concepts of *negative definite* and *negative semidefinite* (for P) are similarly defined. Furthermore, we recall that the *function* $v(x)$ given in (4.23) is said to be *positive definite, positive semidefinite, indefinite*, and so forth, if P has the corresponding definiteness properties.

Instead of solving for the eigenvalues of a real symmetric matrix to determine its definiteness properties, there are more efficient and direct methods of accomplishing this. We now digress to discuss some of these.

Let $G = [g_{ij}]$ be a real $n \times n$ matrix (not necessarily symmetric). Recall that the *minors* of G are the matrix itself and the matrix obtained by removing successively a row and a column. The *principal minors* of G are G itself and the matrices obtained by successively removing an ith row and an ith column, and the *leading principal minors* of G are G itself and the minors obtained by successively removing the last row and the last column. For example, if $G = [g_{ij}] \in R^{3\times 3}$, then the principal minors are

$$\begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ g_{31} & g_{32} & g_{33} \end{bmatrix}, \quad \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}, \quad [g_{11}],$$

$$\begin{bmatrix} g_{11} & g_{13} \\ g_{31} & g_{33} \end{bmatrix}, \quad \begin{bmatrix} g_{22} & g_{23} \\ g_{32} & g_{33} \end{bmatrix}, \quad [g_{22}], \quad [g_{33}].$$

The first three matrices above are the leading principal minors of G. On the other hand, the matrix

$$\begin{bmatrix} g_{21} & g_{22} \\ g_{31} & g_{32} \end{bmatrix}$$

is a minor but not a principal minor.

The following results, due to Sylvester, allow efficient determination of the definiteness properties of a *real, symmetric* matrix.

Proposition 4.20. *(i) A real symmetric matrix $P = [p_{ij}] \in R^{n\times n}$ is* positive definite *if and only if the determinants of its* leading principal minors *are positive, i.e., if and only if*

$$p_{11} > 0, \quad \det \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} > 0, \ldots, \det P > 0.$$

(ii) A real symmetric matrix P is positive semidefinite *if and only if the determinants of* all of its principal minors *are nonnegative.* ■

Still digressing, we consider next the quadratic form

$$v(w) = w^T G w, \quad G = G^T,$$

where $G \in R^{n\times n}$. Now recall that there exists an orthogonal matrix Q such that the matrix P defined by

$$P = Q^{-1} G Q = Q^T G Q$$

is diagonal. Therefore, if we let $w = Qx$, then

$$v(Qx) \triangleq v(x) = x^T Q^T G Q x = x^T P x,$$

where P is in the form given by

$$P = \mathrm{diag}[\Lambda_i] \quad i = 1, \ldots, p,$$

where $\Lambda_i = \mathrm{diag}\, \lambda_i$. From this, we immediately obtain the following useful result.

Proposition 4.21. *Let $P = P^T \in R^{n \times n}$, let $\lambda_M(P)$ and $\lambda_m(P)$ denote the largest and smallest eigenvalues of P, respectively, and let $\| \cdot \|$ denote the Euclidean norm. Then*

$$\lambda_m(P) \parallel x \parallel^2 \leq v(x) = x^T P x \leq \lambda_M(P) \parallel x \parallel^2 \tag{4.26}$$

for all $x \in R^n$ (refer to [1]). ■

Let $c_1 \triangleq \lambda_m(P)$ and $c_2 = \lambda_M(P)$. Clearly, $v(x)$ is positive definite if and only if $c_2 \geq c_1 > 0, v(x)$ is positive semidefinite if and only if $c_2 \geq c_1 \geq 0, v(x)$ is indefinite if and only if $c_2 > 0, c_1 < 0$, and so forth.

We are now in a position to prove several results.

Theorem 4.22. *The equilibrium $x = 0$ of (4.22) is* stable *if there exists a real, symmetric, and positive definite $n \times n$ matrix P such that the matrix C given in (4.25) is negative semidefinite.*

Proof. Along any solution $\phi(t, x_0) \triangleq \phi(t)$ of (4.22) with $\phi(0, x_0) = \phi(0) = x_0$, we have

$$\phi(t)^T P \phi(t) = x_0^T P x_0 + \int_0^t \frac{d}{d\eta} \phi(\eta)^T P \phi(\eta) d\eta = x_0^T P x_0 + \int_0^t \phi(\eta)^T C \phi(\eta) d\eta$$

for all $t \geq 0$. Since P is positive definite and C is negative semidefinite, we have

$$\phi(t)^T P \phi(t) - x_0^T P x_0 \leq 0$$

for all $t \geq 0$, and there exist $c_2 \geq c_1 > 0$ such that

$$c_1 \parallel \phi(t) \parallel^2 \leq \phi(t)^T P \phi(t) \leq x_0^T P x_0 \leq c_2 \parallel x_0 \parallel^2$$

for all $t \geq 0$. It follows that

$$\parallel \phi(t) \parallel \leq (c_2/c_1)^{1/2} \parallel x_0 \parallel$$

for all $t \geq 0$ and for any $x_0 \in R^n$. Therefore, the equilibrium $x = 0$ of (4.22) is stable (refer to Theorem 4.12). ■

Example 4.23. For the system given in Example 4.16 we choose $P = I$, and we compute

$$C = A^T P + PA = A^T + A = 0.$$

According to Theorem 4.22, the equilibrium $x = 0$ of this system is stable (as expected from Example 4.16).

Theorem 4.24. *The equilibrium $x = 0$ of (4.22) is* exponentially stable in the large *if there exists a real, symmetric, and positive definite $n \times n$ matrix P such that the matrix C given in (4.25) is negative definite.*

Proof. We let $\phi(t, x_0) \triangleq \phi(t)$ denote an arbitrary solution of (4.22) with $\phi(0) = x_0$. In view of the hypotheses of the theorem, there exist constants $c_2 \geq c_1 > 0$ and $c_3 \geq c_4 > 0$ such that

$$c_1 \parallel \phi(t) \parallel^2 \leq v(\phi(t)) = \phi(t)^T P \phi(t) \leq c_2 \parallel \phi(t) \parallel^2$$

and

$$-c_3 \parallel \phi(t) \parallel^2 \leq \dot{v}(\phi(t)) = \phi(t)^T C \phi(t) \leq -c_4 \parallel \phi(t) \parallel^2$$

for all $t \geq 0$ and for any $x_0 \in R^n$. Then

$$\begin{aligned} \dot{v}(\phi(t)) &= \frac{d}{dt}[\phi(t)^T P \phi(t)] \leq (-c_4/c_2)\phi(t)^T P \phi(t) \\ &= (-c_4/c_2) v(\phi(t)) \end{aligned}$$

for all $t \geq t_0$. This implies, after multiplication by the appropriate integrating factor, and integrating from 0 to t, that

$$v(\phi(t)) = \phi(t)^T P \phi(t) \leq x_0^T P x_0 e^{-(c_4/c_2)t}$$

or

$$c_1 \parallel \phi(t) \parallel^2 \leq \phi(t)^T P \phi(t) \leq c_2 \parallel x_0 \parallel^2 e^{-(c_4/c_2)t}$$

or

$$\parallel \phi(t) \parallel \leq (c_2/c_1)^{1/2} \parallel x_0 \parallel e^{-\frac{1}{2}(c_4/c_2)t}, \quad t \geq 0.$$

This inequality holds for all $x_0 \in R^n$. Therefore, the equilibrium $x = 0$ of (4.22) is exponentially stable in the large (refer to Sections 4.3 and 4.4). ■

In Figure 4.3 we provide an interpretation of Theorem 4.24 for the two-dimensional case ($n = 2$). The curves C_i, called *level curves*, depict loci where $v(x)$ is constant; i.e., $C_i = \{x \in R^2 : v(x) = x^T P x = c_i\}, i = 0, 1, 2, 3, \ldots$. When the hypotheses of Theorem 4.24 are satisfied, trajectories determined by (4.22) penetrate level curves corresponding to decreasing values of c_i as t increases, tending to the origin as t becomes arbitrarily large.

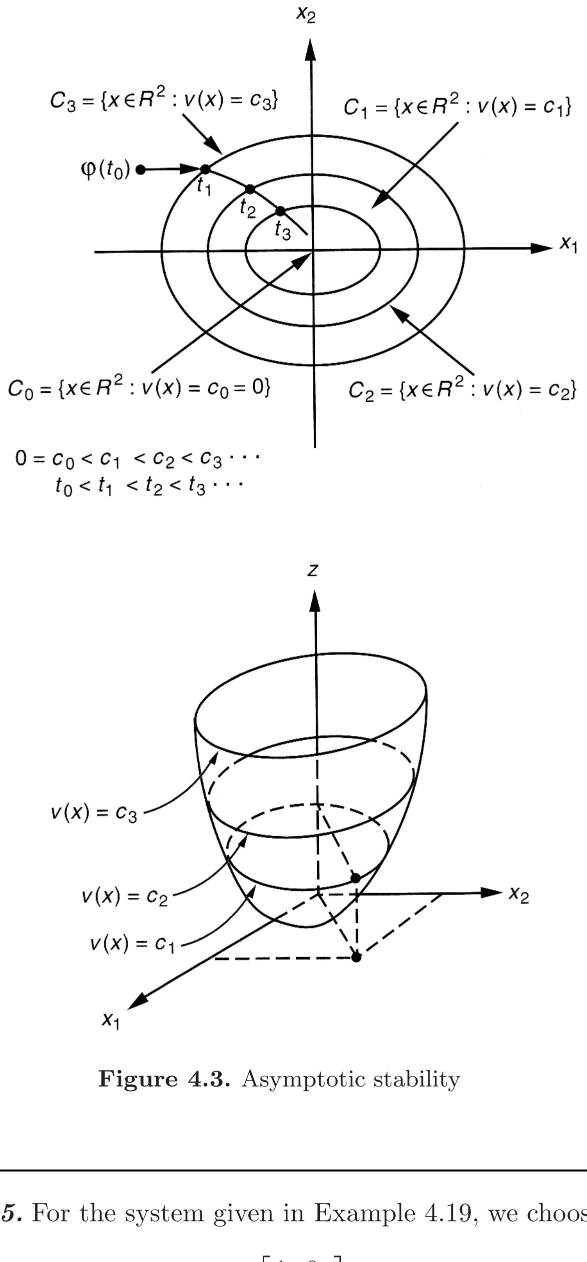

x_2
$C_3 = \{x \in R^2 : v(x) = c_3\}$
$C_1 = \{x \in R^2 : v(x) = c_1\}$
$\varphi(t_0)$
t_1
t_2
t_3
x_1
$C_0 = \{x \in R^2 : v(x) = c_0 = 0\}$
$C_2 = \{x \in R^2 : v(x) = c_2\}$
$0 = c_0 < c_1 < c_2 < c_3 \cdots$
$t_0 < t_1 < t_2 < t_3 \cdots$
z
$v(x) = c_3$
$v(x) = c_2$
$v(x) = c_1$
x_2
x_1

Figure 4.3. Asymptotic stability

Example 4.25. For the system given in Example 4.19, we choose

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix},$$

and we compute the matrix

$$C = A^T P + PA = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix}.$$

According to Theorem 4.24, the equilibrium $x = 0$ of this system is exponentially stable in the large (as expected from Example 4.19).

Theorem 4.26. *The equilibrium $x = 0$ of (4.22) is* unstable *if there exists a real, symmetric $n \times n$ matrix P that is either negative definite or indefinite such that the matrix C given in (4.25) is negative definite.*

Proof. We first assume that P is indefinite. Then P possesses eigenvalues of either sign, and every neighborhood of the origin contains points where the function

$$v(x) = x^T P x$$

is positive and negative. Consider the neighborhood

$$B(\epsilon) = \{x \in R^n : \| x \| < \epsilon\},$$

where $\| \cdot \|$ denotes the Euclidean norm, and let

$$G = \{x \in B(\epsilon) : v(x) < 0\}.$$

On the boundary of G we have either $\| x \| = \epsilon$ or $v(x) = 0$. In particular, note that the origin $x = 0$ is on the boundary of G. Now, since the matrix C is negative definite, there exist constants $c_3 > c_4 > 0$ such that

$$-c_3 \| x \|^2 \leq x^T C x = \dot{v}(x) \leq -c_4 \| x \|^2$$

for all $x \in R^n$. Let $\phi(t, x_0) \triangleq \phi(t)$ and let $x_0 = \phi(0) \in G$. Then $v(x_0) = -a < 0$. The solution $\phi(t)$ starting at x_0 must leave the set G. To see this, note that as long as $\phi(t) \in G, v(\phi(t)) \leq -a$ since $\dot{v}(x) < 0$ in G. Let $-c = \sup\{\dot{v}(x) : x \in G$ and $v(x) \leq -a\}$.

Then $c > 0$ and

$$\begin{aligned} v(\phi(t)) &= v(x_0) + \int_0^t \dot{v}(\phi(s)) ds \leq -a - \int_0^t c ds \\ &= -a - tc, t \geq t_0. \end{aligned}$$

This inequality shows that $\phi(t)$ must escape the set G (in finite time) because $v(x)$ is bounded from below on G. But $\phi(t)$ cannot leave G through the surface determined by $v(x) = 0$ since $v(\phi(t)) \leq -a$. Hence, it must leave G through the sphere determined by $\| x \| = \epsilon$. Since the above argument holds for arbitrarily small $\epsilon > 0$, it follows that the origin $x = 0$ of (4.22) is unstable.

Next, we assume that P is negative definite. Then G as defined is all of $B(\epsilon)$. The proof proceeds as above. ■

The proof of Theorem 4.26 shows that for $\epsilon > 0$ sufficiently small when P is negative definite, *all* solutions $\phi(t)$ of (4.22) with initial conditions $x_0 \in B(\epsilon)$ will tend away from the origin. This constitutes a severe case of instability, called *complete instability.*

Example 4.27. For the system given in Example 4.18, we choose

$$P = \begin{bmatrix} -0.28 & -0.96 \\ -0.96 & 0.28 \end{bmatrix},$$

and we compute the matrix

$$C = A^T P + PA = \begin{bmatrix} -20 & 0 \\ 0 & -20 \end{bmatrix}.$$

The eigenvalues of P are ±1. According to Theorem 4.26, the equilibrium $x = 0$ of this system is unstable (as expected from Example 4.18).

In applying the results derived thus far in this section, we start by choosing (guessing) a matrix P having certain desired properties. Next, we solve for the matrix C, using (4.25). If C possesses certain desired properties (i.e., it is negative definite), we draw appropriate conclusions by applying one of the preceding theorems of this section; if not, we need to choose another matrix P. This points to the principal shortcoming of Lyapunov's Direct Method, when applied to general systems. However, in the *special case* of linear systems described by (4.22), it is possible to *construct* Lyapunov functions of the form $v(x) = x^T Px$ in a *systematic* manner. In doing so, one first chooses the matrix C in (4.25) (having desired properties), and then one solves (4.25) for P. Conclusions are then drawn by applying the appropriate results of this section. In applying this construction procedure, we need to know conditions under which (4.25) possesses a (unique) solution P for a given C. We will address this topic next.

We consider the quadratic form

$$v(x) = x^T Px, \quad P = P^T, \tag{4.27}$$

and the time derivative of $v(x)$ along the solutions of (4.22), given by

$$\dot{v}(x) = x^T Cx, \quad C = C^T, \tag{4.28}$$

where

$$C = A^T P + PA \tag{4.29}$$

and where all symbols are as defined in (4.23) to (4.25). Our objective is to determine the as yet unknown matrix P in such a way that $\dot{v}(x)$ becomes a preassigned negative definite quadratic form, i.e., in such a way that C is a preassigned negative definite matrix.

Equation (4.29) constitutes a system of $n(n+1)/2$ linear equations. We need to determine under what conditions we can solve for the $n(n+1)/2$ elements, p_{ik}, given C and A. To this end, we choose a similarity transformation Q such that

$$QAQ^{-1} = \bar{A}, \tag{4.30}$$

or equivalently,

$$A = Q^{-1}\bar{A}Q, \tag{4.31}$$

where $\bar{A}$ is similar to A and Q is a real $n \times n$ nonsingular matrix. From (4.31) and (4.29) we obtain

$$(\bar{A})^T(Q^{-1})^T PQ^{-1} + (Q^{-1})^T PQ^{-1}\bar{A} = (Q^{-1})^T CQ^{-1} \tag{4.32}$$

or

$$(\bar{A})^T\bar{P} + \bar{P}\bar{A} = \bar{C}, \quad \bar{P} = (Q^{-1})^T PQ^{-1}, \quad \bar{C} = (Q^{-1})^T CQ^{-1}. \tag{4.33}$$

In (4.33), P and C are subjected to a congruence transformation and $\bar{P}$ and $\bar{C}$ have the same definiteness properties as P and C, respectively. Since every real $n \times n$ matrix can be triangularized (refer to [1]), we can choose Q in such a fashion that $\bar{A} = [\bar{a}_{ij}]$ is *triangular*; i.e., $\bar{a}_{ij} = 0$ for $i > j$. Note that in this case the eigenvalues of $A, \lambda_1, \dots, \lambda_n$, appear in the main diagonal of $\bar{A}$. To simplify our notation, we rewrite (4.33) in the form (4.29) by dropping the bars, i.e.,

$$A^T P + PA = C, \quad C = C^T, \tag{4.34}$$

and *we assume that* $A = [a_{ij}]$ *has been triangularized*; i.e., $a_{ij} = 0$ for $i > j$. Since the eigenvalues $\lambda_1, \dots, \lambda_n$ appear in the diagonal of A, we can rewrite (4.34) as

$$\begin{aligned} 2\lambda_1 p_{11} &= c_{11} \\ a_{12}p_{11} + (\lambda_1 + \lambda_2)p_{12} &= c_{12} \\ \dots\dots\dots\dots\dots\dots&\dots\dots \end{aligned} \tag{4.35}$$

Note that λ_1 may be a complex number; in which case, c_{11} will also be complex. Since this system of equations is triangular, and since its determinant is equal to

$$2^n \lambda_1 \dots \lambda_n \prod_{i<j}(\lambda_i + \lambda_j), \tag{4.36}$$

the matrix P can be determined (uniquely) if and only if this determinant is not zero. This is true when all eigenvalues of A are nonzero and no two of them are such that $\lambda_i + \lambda_j = 0$. This condition is not affected by a similarity transformation and is therefore also valid for the original system of equations (4.29).

We summarize the above discussion in the following lemma.

Lemma 4.28. *Let $A \in R^{n \times n}$ and let $\lambda_1, \dots, \lambda_n$ denote the (not necessarily distinct) eigenvalues of A. Then (4.34) has a unique solution for P corresponding to each $C \in R^{n \times n}$ if and only if*

$$\lambda_i \neq 0, \lambda_i + \lambda_j \neq 0 \text{ for all } i, j. \tag{4.37}$$

■

To construct $v(x)$, we must still check the definiteness of P. This can be done in a purely algebraic way; however, in the present case, it is much easier to apply the results of this section and argue as follows:

(a) If all the eigenvalues λ_i of A have negative real parts, then the equilibrium $x = 0$ of (4.22) is exponentially stable in the large, and if C in (4.29) is negative definite, then P must be positive definite. To prove this, we note that if P is not positive definite, then for $\delta > 0$ and sufficiently small, $(P - \delta I)$ has at least one negative eigenvalue, whereas the function $v(x) = x^T(P - \delta I)x$ has a negative definite derivative; i.e.,

$$v'_{(L)}(x) = x^T[C - \delta(A + A^T)]x < 0$$

for all $x \neq 0$. By Theorem 4.26, the equilibrium $x = 0$ of (4.22) is unstable. We have arrived at a contradiction. Therefore, P must be positive definite.

(b) If A has eigenvalues with positive real parts and no eigenvalues with zero real parts, we can use a similarity transformation $x = Qy$ in such a way that $Q^{-1}AQ$ is a block diagonal matrix of the form diag$[A_1,\ A_2]$, where all the eigenvalues of A_1 have positive real parts, whereas all eigenvalues of A_2 have negative real parts (refer to [1]). (If A does not have any eigenvalues with negative real parts, then we take $A = A_1$). By the result established in (a), noting that all eigenvalues of $-A_1$ have negative real parts, given any negative definite matrices C_1 and C_2, there exist positive definite matrices P_1 and P_2 such that

$$(-A_1^T)P_1 + P_1(-A_1) = C_1, \quad A_2^T P_2 + P_2 A_2 = C_2.$$

Then $w(y) = y^T P y$, with $P = \text{diag}[-P_1,\ P_2]$ is a Lyapunov function for the system $\dot{y} = Q^{-1}AQy$ (and, hence, for the system $\dot{x} = Ax$), which satisfies the hypotheses of Theorem 4.26. Therefore, the equilibrium $x = 0$ of system (4.22) is unstable. If A does not have any eigenvalues with negative real parts, then the equilibrium $x = 0$ of system (4.22) is *completely unstable*.]

In the above proof, we did not invoke Lemma 4.28. We note, however, that if additionally, (4.37) is true, then we can construct the Lyapunov function for (4.22) in a systematic manner.

Summarizing the above discussion, we can now state the main result of this subsection.

Theorem 4.29. *Assume that the matrix A [for system (4.22)] has no eigenvalues with real part equal to zero. If all the eigenvalues of A have negative real parts, or if at least one of the eigenvalues of A has a positive real part, then there exists a quadratic Lyapunov function*

$$v(x) = x^T P x, \quad P = P^T,$$

whose derivative along the solutions of (4.22) is definite (i.e., it is either negative definite or positive definite). ■

This result shows that when A is a stable matrix (i.e., all the eigenvalues of A have negative real parts), then for system (4.22) the conditions of Theorem 4.24 are also necessary conditions for exponential stability in the large. Moreover, in the case when the matrix A has at least one eigenvalue with positive real part and no eigenvalues on the imaginary axis, then the conditions of Theorem 4.26 are also necessary conditions for instability.

Example 4.30. We consider the system (4.22) with

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

The eigenvalues of A are $\lambda_1, \lambda_2 = \pm j$, and therefore condition (4.37) is violated. According to Lemma 4.28, the Lyapunov matrix equation

$$A^T P + PA = C$$

does not possess a unique solution for a given C. We now verify this for two specific cases.

(i) When $C = 0$, we obtain

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} -2p_{12} & p_{11} - p_{22} \\ p_{11} - p_{22} & 2p_{12} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

or $p_{12} = 0$ and $p_{11} = p_{22}$. Therefore, for any $a \in R$, the matrix $P = aI$ is a solution of the Lyapunov matrix equation. In other words, for $C = 0$, the Lyapunov matrix equation has in this example denumerably many solutions.

(ii) When $C = -2I$, we obtain

$$\begin{bmatrix} -2p_{12} & p_{11} - p_{22} \\ p_{11} - p_{22} & 2p_{12} \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & -2 \end{bmatrix},$$

or $p_{11} = p_{22}$ and $p_{12} = 1$ and $p_{12} = -1$, which is impossible. Therefore, for $C = -2I$, the Lyapunov matrix equation has in this example no solutions at all.

It turns out that if all the eigenvalues of matrix A have negative real parts, then we can compute P in (4.29) explicitly.

Theorem 4.31. *If all eigenvalues of a real $n \times n$ matrix A have negative real parts, then for each matrix $C \in R^{n\times n}$, the unique solution of (4.29) is given by*

$$P = \int_0^\infty e^{A^T t}(-C)e^{At}dt. \tag{4.38}$$

Proof. If all eigenvalues of A have negative real parts, then (4.37) is satisfied and therefore (4.29) has a unique solution for every $C \in R^{n\times n}$. To verify that (4.38) is indeed this solution, we first note that the right-hand side of (4.38) is well defined, since all eigenvalues of A have negative real parts. Substituting the right-hand side of (4.38) for P into (4.29), we obtain

$$\begin{aligned} A^T P + PA &= \int_0^\infty A^T e^{A^T t}(-C)e^{At}dt + \int_0^\infty e^{A^T t}(-C)e^{At}Adt \\ &= \int_0^\infty \frac{d}{dt}[e^{A^T t}(-C)e^{At}]dt \\ &= e^{A^T t}(-C)e^{At}\Big|_0^\infty = C, \end{aligned}$$

which proves the theorem. ■

4.6 Linearization

In this section we consider *nonlinear, finite-dimensional, continuous-time* dynamical systems described by equations of the form

$$\dot{w} = f(w), \tag{4.39}$$

where $f \in C^1(R^n, R^n)$. We assume that $w = 0$ is an equilibrium of (4.39). In accordance with Subsection 1.6.1, we linearize system (4.39) about the origin to obtain

$$\dot{x} = Ax + F(x), \tag{4.40}$$

$x \in R^n$, where $F \in C(R^n, R^n)$ and where A denotes the Jacobian of $f(w)$ evaluated at $w = 0$, given by

$$A = \frac{\partial f}{\partial w}(0), \tag{4.41}$$

and where

$$F(x) = o(\| x \|) \quad \text{as} \quad \| x \| \to 0. \tag{4.42}$$

Associated with (4.40) is the *linearization* of (4.39), given by

$$\dot{y} = Ay. \tag{4.43}$$

In the following discussion, we use the results of Section 4.5 to establish criteria that allow us to deduce the stability properties of the equilibrium $w = 0$ of the nonlinear system (4.39) from the stability properties of the equilibrium $y = 0$ of the linear system (4.43).

Theorem 4.32. *Let $A \in R^{n \times n}$ be a Hurwitz matrix (i.e., all of its eigevnalues have negative real parts), let $F \in C(R^n, R^n)$, and assume that (4.42) holds. Then the equilibrium $x = 0$ of (4.40) [and, hence, of (4.39)] is* exponentially stable*.*

Proof. Theorem 4.29 applies to (4.43) since all the eigenvalues of A have negative real parts. In view of that theorem (and the comments following Lemma 4.28), there exists a symmetric, real, positive definite $n \times n$ matrix P such that

$$PA + A^T P = C, \tag{4.44}$$

where C is negative definite. Consider the Lyapunov function

$$v(x) = x^T P x. \tag{4.45}$$

The derivative of v with respect to t along the solutions of (4.40) is given by

$$\begin{aligned} \dot{v}(x) &= \dot{x}^T P x + x^T P \dot{x} \\ &= (Ax + F(x))^T P x + x^T P (Ax + F(x)) \\ &= x^T C x + 2x^T P F(x). \end{aligned} \tag{4.46}$$

Now choose $\gamma < 0$ such that $x^T C x \leq 3\gamma \parallel x \parallel^2$ for all $x \in R^n$. Since it is assumed that (4.42) holds, there is a $\delta > 0$ such that if $\parallel x \parallel \leq \delta$, then $\parallel PF(x) \parallel \leq -\gamma \parallel x \parallel$ for all $x \in \overline{B(\delta)} = \{x \in R^n : \parallel x \parallel \leq \delta\}$. Therefore, for all $x \in \overline{B(\delta)}$, we obtain, in view of (4.46), the estimate

$$\dot{v}(x) \leq 3\gamma \parallel x \parallel^2 - 2\gamma \parallel x \parallel^2 = \gamma \parallel x \parallel^2 . \tag{4.47}$$

Now let $\alpha = \min_{\parallel x \parallel = \delta} v(x)$. Then $\alpha > 0$ (since P is positive definite). Take $\lambda \in (0, \alpha)$, and let

$$C_\lambda = \{x \in B(\delta) = \{x \in R^n : \parallel x \parallel < \delta\} : v(x) \leq \lambda\}. \tag{4.48}$$

Then $C_\lambda \subset B(\delta)$. [This can be shown by contradiction. Suppose that C_λ is not entirely inside $B(\delta)$. Then there is a point $\bar{x} \in C_\lambda$ that lies on the boundary of $B(\delta)$. At this point, $v(\bar{x}) \geq \alpha > \lambda$. We have thus arrived at a contradiction.] The set C_λ has the property that any solution of (4.40) starting in C_λ at $t = 0$ will stay in C_λ for all $t \geq 0$. To see this, we let $\phi(t, x_0) \triangleq \phi(t)$ and we recall that $\dot{v}(x) \leq \gamma \parallel x \parallel^2, \gamma < 0, x \in B(\delta) \supset C_\lambda$. Then $\dot{v}(\phi(t)) \leq 0$ implies that $v(\phi(t)) \leq v(x_0) \leq \lambda$ for all $t \geq t_0 \geq 0$. Therefore, $\phi(t) \in C_\lambda$ for all $t \geq t_0 \geq 0$.

We now proceed in a similar manner as in the proof of Theorem 4.24 to complete this proof. In doing so, we first obtain the estimate

$$\dot{v}(\phi(t)) \leq (\gamma/c_2)v(\phi(t)), \tag{4.49}$$

where γ is given in (4.47) and c_2 is determined by the relation

$$c_1 \parallel x \parallel^2 \leq v(x) = x^T Px \leq c_2 \parallel x \parallel^2 . \tag{4.50}$$

Following now in an identical manner as was done in the proof of Theorem 4.22, we have

$$\parallel \phi(t) \parallel \leq (c_2/c_1)^{\frac{1}{2}} \parallel x_0 \parallel e^{\frac{1}{2}(\gamma/c_2)t}, \quad t \geq 0, \tag{4.51}$$

whenever $x_0 \in B(r')$, where r' has been chosen sufficiently small so that $B(r') \subset C_\lambda$. This proves that the equilibrium $x = 0$ of (4.40) is exponentially stable. ■

It is important to recognize that Theorem 4.32 is a *local result* that yields sufficient conditions for the exponential stability of the equilibrium $x = 0$ of (4.40); it does not yield conditions for exponential stability in the large. The proof of Theorem 4.32, however, enables us to determine an estimate of the domain of attraction of the equilibrium $x = 0$ of (4.39), involving the following steps:

1. Determine an equilibrium, x_e, of (4.39) and transform (4.39) to a new system that translates x_e to the origin $x = 0$ (refer to Section 4.2).
2. Linearize (4.39) about the origin and determine $F(x), A$, and the eigenvalues of A.
3. If all eigenvalues of A have negative real parts, choose a negative definite matrix C and solve the Lyapunov matrix equation

$$C = A^T P + PA.$$

4. Determine the Lyapunov function

$$v(x) = x^T Px.$$

5. Compute the derivative of v along the solutions of (4.40), given by

$$\dot{v}(x) = x^T Cx + 2x^T PF(x).$$

6. Determine $\delta > 0$ such that $\dot{v}(x) < 0$ for all $x \in B(\delta) - \{0\}$.
7. Determine the largest $\lambda = \lambda_M$ such that $C_{\lambda_M} \subset B(\delta)$, where

$$C_\lambda = \{x \in R^n : v(x) < \lambda\}.$$

8. C_{λ_M} is a subset of the domain of attraction of the equilibrium $x = 0$ of (4.40) and, hence, of (4.39).

The above procedure may be repeated for different choices of matrix C given in step (3), resulting in different matrices P_i, which in turn may result in different estimates for the domain of attraction, $C^i_{\lambda_M}$, $i \in \Lambda$, where Λ is an index set. The union of the sets $C^i_{\lambda_M} \triangleq D_i, D = \cup_i D_i$, is also a subset of the domain of attraction of the equilibrium $x = 0$ of (4.39).

Theorem 4.33. *Assume that A is a real $n \times n$ matrix that has at least one eigenvalue with positive real part and no eigenvalue with zero real part. Let $F \in C(R^n, R^n)$, and assume that (4.42) holds. Then the equilibrium $x = 0$ of (4.40) [and, hence, of (4.39)] is* unstable*.*

Proof. We use Theorem 4.29 to choose a real, symmetric $n \times n$ matrix P such that the matrix $PA + A^T P = C$ is negative definite. The matrix P is not positive definite, or even positive semidefinite (refer to the comments following Lemma 4.28). Hence, the function $v(x) = x^T P x$ is negative at some points arbitrarily close to the origin. The derivative of $v(x)$ with respect to t along the solutions of (4.40) is given by (4.46). As in the proof of Theorem 4.32, we can choose a $\gamma < 0$ such that $x^T C x \leq 3\gamma \parallel x \parallel^2$ for all $x \in R^n$, and in view of (4.42) we can choose a $\delta > 0$ such that $\parallel PF(x) \parallel \leq -\gamma \parallel x \parallel$ for all $x \in B(\delta)$. Therefore, for all $x \in B(\delta)$, we obtain that

$$\dot{v}(x) \leq 3\gamma \parallel x \parallel^2 - 2\gamma \parallel x \parallel^2 = \gamma \parallel x \parallel^2 .$$

Now let

$$G = \{x \in B(\delta) : v(x) < 0\}.$$

The boundary of G is made up of points where $v(x) = 0$ and where $\parallel x \parallel = \delta$. Note in particular that the equilibrium $x = 0$ of (4.40) is in the boundary of G. Now following an identical procedure as in the proof of Theorem 4.26, we show that any solution $\phi(t)$ of (4.40) with $\phi(0) = x_0 \in G$ must escape G in finite time through the surface determined by $\parallel x \parallel = \delta$. Since the above argument holds for arbitrarily small $\delta > 0$, it follows that the origin $x = 0$ of (4.40) is unstable. ■

Before concluding this section, we consider a few specific cases.

Example 4.34. The *Lienard Equation* is given by

$$\ddot{w} + f(w)\dot{w} + w = 0, \tag{4.52}$$

where $f \in C^1(R, R)$ with $f(0) > 0$. Letting $x_1 = w$ and $x_2 = \dot{w}$, we obtain

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -x_1 - f(x_1)x_2. \end{aligned} \tag{4.53}$$

Let $x^T = (x_1, x_2)$, $f(x)^T = (f_1(x), f_2(x))$, and let

$$J(0) = A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(0) & \frac{\partial f_1}{\partial x_2}(0) \\ \frac{\partial f_2}{\partial x_1}(0) & \frac{\partial f_2}{\partial x_2}(0) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -f(0) \end{bmatrix}.$$

Then

$$\dot{x} = Ax + [f(x) - Ax] = Ax + F(x),$$

where

$$F(x) = \begin{bmatrix} 0 \\ [f(0) - f(x_1)]\, x_2 \end{bmatrix}.$$

The origin $x = 0$ is clearly an equilibrium of (4.52) and hence of (4.53). The eigenvalues of A are given by

$$\lambda_1, \lambda_2 = \frac{-f(0) \pm \sqrt{f(0)^2 - 4}}{2},$$

and therefore, A is a Hurwitz matrix. Also, (4.42) holds. Therefore, all the conditions of Theorem 4.32 are satisfied. We conclude that the equilibrium $x = 0$ of (4.53) is *exponentially stable.*

Example 4.35. We consider the system given by

$$\begin{aligned} \dot{x}_1 &= -x_1 + x_1(x_1^2 + x_2^2), \\ \dot{x}_2 &= -x_2 + x_2(x_1^2 + x_2^2). \end{aligned} \tag{4.54}$$

The origin is clearly an equilibrium of (4.54). Also, the system is already in the form (4.40) with

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad F(x) = \begin{bmatrix} x_1(x_1^2 + x_2^2) \\ x_2(x_1^2 + x_2^2) \end{bmatrix},$$

and condition (4.42) is clearly satisfied. The eigenvalues of A are $\lambda_1 = -1, \lambda_2 = -1$. Therefore, all conditions of Theorem 4.32 are satisfied and we conclude that the equilibrium $x^T = (x_1, x_2) = 0$ is *exponentially stable*; however, we cannot conclude that this equilibrium is exponentially stable in the large. Accordingly, we seek to determine an estimate for the domain of attraction of this equilibrium.

We choose $C = -I$ (where $I \in R^{2\times 2}$ denotes the identity matrix), and we solve the matrix equation $A^T P + PA = C$ to obtain $P = (1/2)I$, and therefore,

$$v(x_1, x_2) = x^T P x = \frac{1}{2}(x_1^2 + x_2^2).$$

Along the solutions of (4.54) we obtain

$$\begin{aligned} \dot{v}(x_1, x_2) &= x^T C x + 2x^T P F(x) \\ &= -(x_1^2 + x_2^2) + (x_1^2 + x_2^2)^2. \end{aligned}$$

Clearly, $\dot{v}(x_1, x_2) < 0$ when $(x_1, x_2) \neq (0,0)$ and $x_1^2 + x_2^2 < 1$. In the language of the proof of Theorem 4.32, we can therefore choose $\delta = 1$.

Now let

$$C_{1/2} = \{x \in R^2 : v(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2) < \frac{1}{2}\}.$$

Then clearly, $C_{1/2} \subset B(\delta), \delta = 1$, in fact $C_{1/2} = B(\delta)$. Therefore, the set $\{x \in R^2 : x_1^2 + x_2^2 < 1\}$ is a subset of the domain of attraction of the equilibrium $(x_1, x_2)^T = 0$ of system (4.54).

Example 4.36. The differential equation governing the motion of a pendulum is given by

$$\ddot{\theta} + a \sin\theta = 0, \tag{4.55}$$

where $a > 0$ is a constant (refer to Chapter 1). Letting $\theta = x_1$ and $\dot{\theta} = x_2$, we obtain the system description

$$\begin{aligned} \dot{x}_1 &= x_2, \\ \dot{x}_2 &= -a \sin x_1. \end{aligned} \tag{4.56}$$

The points $x_e^{(1)} = (0,0)^T$ and $x_e^{(2)} = (\pi, 0)^T$ are equilibria of (4.56).

(i) Linearizing (4.56) about the equilibrium $x_e^{(1)}$, we put (4.56) into the form (4.40) with

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

The eigenvalues of A are $\lambda_1, \lambda_2 = \pm j\sqrt{a}$. Therefore, the results of this section (Theorem 4.32 and 4.33) are not applicable in the present case.

(ii) In (4.56), we let $y_1 = x_1 - \pi$ and $y_2 = x_2$. Then (4.56) assumes the form

$$\begin{aligned} \dot{y}_1 &= y_2, \\ \dot{y}_2 &= -a \sin(y_1 + \pi). \end{aligned} \tag{4.57}$$

The point $(y_1, y_2)^T = (0,0)^T$ is clearly an equilibrium of system (4.57). Linearizing about this equilibrium, we put (4.57) into the form (4.40), where

$$A = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}, \quad F(y_1, y_2) = \begin{bmatrix} 0 \\ -a(\sin(y_1 + \pi) + y_1) \end{bmatrix}.$$

The eigenvalues of A are $\lambda_1, \lambda_2 = a, -a$. All conditions of Theorem 4.33 are satisfied, and we conclude that the equilibrium $x_e^{(2)} = (\pi, 0)^T$ of system (4.56) is *unstable.*

4.7 Input–Output Stability

We now turn our attention to systems described by the state equations

$$\begin{aligned}\dot{x} &= Ax + Bu,\\ y &= Cx + Du,\end{aligned} \tag{4.58}$$

where $A \in R^{n\times n}, B \in R^{n\times m}, C \in R^{p\times n}$, and $D \in R^{p\times m}$. In the preceding sections of this chapter we investigated the *internal stability properties* of system (4.58) by studying the Lyapunov stability of the trivial solution of the associated system

$$\dot{w} = Aw. \tag{4.59}$$

In this approach, system inputs and system outputs played no role. To account for these, we now consider the *external stability properties* of system (4.58), called *input–output stability*: Every bounded input of a system should produce a bounded output. More specifically, in the present context, we say that system (4.58) is *bounded-input/bounded-output (BIBO) stable*, if for zero initial conditions at $t = 0$, every bounded input defined on $[0, \infty)$ gives rise to a bounded response on $[0, \infty)$.

Matrix D does not affect the BIBO stability of (4.58). Accordingly, we will consider without any loss of generality the case where $D \equiv 0$; i.e., throughout this section we will concern ourselves with systems described by equations of the form

$$\begin{aligned}\dot{x} &= Ax + Bu,\\ y &= Cx.\end{aligned} \tag{4.60}$$

We will say that the system (4.60) is *BIBO stable* if there exists a constant $c > 0$ such that the conditions

$$\begin{aligned}x(0) &= 0,\\ ||u(t)|| &\leq 1, \quad t \geq 0,\end{aligned}$$

imply that $\| \, y(t) \, \| \leq c$ for all $t \geq 0$. (The symbol $\| \cdot \|$ denotes the Euclidean norm.)

Recall that for system (4.60) the impulse response matrix is given by

$$\begin{aligned}H(t) &= Ce^{At}B, \quad t \geq 0,\\ &= 0, \qquad\quad\;\; t < 0,\end{aligned} \tag{4.61}$$

and the transfer function matrix is given by

$$\widehat{H}(s) = C(sI - A)^{-1}B. \tag{4.62}$$

Theorem 4.37. *The system (4.60) is* BIBO stable *if and only if there exists a finite constant $L > 0$ such that for all t,*

$$\int_0^t \| \, H(t-\tau) \, \| \, d\tau \leq L. \tag{4.63}$$

Proof. The first part of the proof of Theorem 4.37 (sufficiency) is straightforward. Indeed, if $\| u(t) \| \leq 1$ for all $t \geq 0$ and if (4.63) is true, then we have for all $t \geq 0$ that

$$\begin{aligned} \| y(t) \| &= \| \int_0^t H(t-\tau)u(\tau)d\tau \| \\ &\leq \int_0^t \| H(t-\tau)u(\tau) \| d\tau \\ &\leq \int_0^t \| H(t-\tau) \| \| u(\tau) \| d\tau \\ &\leq \int_0^t \| H(t-\tau) \| d\tau \leq L. \end{aligned}$$

Therefore, system (4.60) is BIBO stable.

In proving the second part of Theorem 4.37 (necessity), we simplify matters by first considering in (4.60) the single-variable case ($n = 1$) with the input–output description given by

$$y(t) = \int_0^t h(t-\tau)u(\tau)d\tau. \tag{4.64}$$

For purposes of contradiction, we assume that the system is BIBO stable, but no finite L exists such that (4.63) is satisfied. Another way of stating this is that for *every finite* L, there exists $t_1 = t_1(L), t_1 > 0$, such that

$$\int_0^{t_1} |h(t_1, \tau)| d\tau > L.$$

We now choose in particular the input given by

$$u(t) = \begin{cases} +1 & \text{if } h(t-\tau) > 0, \\ 0 & \text{if } h(t-\tau) = 0, \\ -1 & \text{if } h(t-\tau) < 0, \end{cases} \tag{4.65}$$

$0 \leq t \leq t_1$. Clearly, $|u(t)| \leq 1$ for all $t \geq 0$. The output of the system at $t = t_1$ due to the above input, however, is

$$y(t_1) = \int_0^{t_1} h(t_1 - \tau)u(\tau)d\tau = \int_0^{t_1} |h(t_1 - \tau)| d\tau > L,$$

which contradicts the assumption that the system is BIBO stable.

The above can now be generalized to the multivariable case. In doing so, we apply the single-variable result to every possible pair of input and output vector components, we make use of the fact that the sum of a finite number of bounded sums will be bounded, and we recall that a vector is bounded if and only if each of its components is bounded. We leave the details to the reader. ■

In the preceding argument we made the tacit assumption that u is continuous, or piecewise continuous. However, our particular choice of u may involve nondenumerably many switchings (discontinuities) over a given finite-time interval. In such cases, u is no longer piecewise continuous; however, it is measurable (in the Lebesgue sense). This generalization can be handled, although in a broader mathematical setting that we do not wish to pursue here. The interested reader may want to refer, e.g., to the books by Desoer and Vidyasagar [5], Michel and Miller [13], and Vidyasagar [20] and the papers by Sandberg [17] to [19] and Zames [21], [22] for further details.

From Theorem 4.37 and from (4.61) it follows readily that a necessary and sufficient condition for the BIBO stability of system (4.60) is the condition

$$\int_0^\infty \| H(t) \| \, dt < \infty. \tag{4.66}$$

Corollary 4.38. *Assume that the equilibrium $w = 0$ of (4.59) is exponentially stable. Then system (4.60) is* BIBO stable.

Proof. Under the hypotheses of the corollary, we have

$$\begin{aligned} \| \int_0^t H(t-\tau) d\tau \| &\leq \int_0^t \| H(t-\tau) \| \, d\tau \\ &= \int_0^t \| C\Phi(t-\tau)B \| \, d\tau \leq \| C \| \| B \| \int_0^t \| \Phi(t-\tau) \| \, d\tau. \end{aligned}$$

Since the equilibrium $w = 0$ of (4.59) is exponentially stable, there exist $\delta > 0$, $\lambda > 0$ such that $\| \Phi(t,\tau) \| \leq \delta e^{-\lambda(t-\tau)}, t \geq \tau$. Therefore,

$$\begin{aligned} \int_0^t \| H(t-\tau) \| \, d\tau &\leq \int_0^t \| C \| \| B \| \, \delta e^{-\lambda(t-\tau)} d\tau \\ &\leq (\| C \| \| B \| \, \delta)/\lambda \triangleq L \end{aligned}$$

for all τ, t with $t \geq \tau$. It now follows from Theorem 4.37 that system (4.60) is BIBO stable. ■

In Section 7.3 we will establish a connection between the BIBO stability of (4.60) and the exponential stability of the trivial solution of (4.59).

Next, we recall that a complex number s_p is a *pole* of $\widehat{H}(s) = [\hat{h}_{ij}(s)]$ if for some pair (i,j), we have $|\hat{h}_{ij}(s_p)| = \infty$. If each entry of $\widehat{H}(s)$ has only poles with negative real values, then, as shown in Chapter 3, each entry of $H(t) = [h_{ij}(t)]$ has a sum of exponentials with exponents with real part negative. It follows that the integral

$$\int_0^\infty \| H(t) \| \, dt$$

is finite, and any realization of $\widehat{H}(s)$ will result in a system that is BIBO stable.

Now conversely, if

$$\int_0^\infty \parallel H(t) \parallel dt$$

is finite, then the exponential terms in any entry of $H(t)$ must have negative real parts. But then every entry of $\widehat{H}(s)$ has poles whose real parts are negative.

We have proved the following result.

Theorem 4.39. *The system (4.60) is BIBO stable if and only if all poles of the transfer function $\widehat{H}(s)$ given in (4.62) have only poles with negative real parts.* ■

Example 4.40. A system with $H(s) = 1/s$ is not BIBO stable. To see this consider a step input. The response is then given by $y(t) = t$, $t \geq 0$, which is not bounded.

4.8 Discrete-Time Systems

In this section we address the Lyapunov stability of an equilibrium of discrete-time systems (internal stability) and the input–output stability of discrete-time systems (external stability). We establish results for discrete-time systems that are analogous to practically all the stability results that we presented for continuous-time systems.

This section is organized into five subsections. In the first subsection we provide essential preliminary material. In the second and third subsections we establish results for the stability, instability, asymptotic stability, and exponential stability of an equilibrium and boundedness of solutions of systems described by linear autonomous ordinary difference equations. These results are used to develop Lyapunov stability results for linearizations of nonlinear systems described by ordinary difference equations in the fourth subsection. In the last subsection we present results for the input–output stability of linear time-invariant discrete-time systems.

4.8.1 Preliminaries

We concern ourselves here with finite-dimensional discrete-time systems described by difference equations of the form

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k), \\ y(k) &= Cx(k), \end{aligned} \tag{4.67}$$

where $A \in R^{n\times n}, B \in R^{n\times m}, C \in R^{p\times n}, k \geq k_0$, and $k, k_0 \in Z^+$. Since (4.67) is time-invariant, we will assume without loss of generality that $k_0 = 0$, and thus, $x : Z^+ \to R^n, y : Z^+ \to R^p$, and $u : Z^+ \to R^m$.

The internal dynamics of (4.67) under conditions of no input are described by equations of the form

$$x(k+1) = Ax(k). \tag{4.68}$$

Such equations may arise in the modeling process, or they may be the consequence of the linearization of nonlinear systems described by equations of the form

$$x(k+1) = g(x(k)), \tag{4.69}$$

where $g : R^n \to R^n$. For example, if $g \in C^1(R^n, R^n)$, then in linearizing (4.69) about, e.g., $x = 0$, we obtain

$$x(k+1) = Ax(k) + f(x(k)), \tag{4.70}$$

where $A = \frac{\partial f}{\partial x}(x)\Big|_{x=0}$ and where $f : R^n \to R^n$ is $o(||x||)$ as a norm of x (e.g., the Euclidean norm) approaches zero. Recall that this means that given $\epsilon > 0$, there is a $\delta > 0$ such that $\| f(x) \| < \epsilon \| x \|$ for all $\| x \| < \delta$.

As in Section 4.7, we will study the *external qualitative properties* of system (4.67) by means of the BIBO stability of such systems. Consistent with the definition of input–output stability of continuous-time systems, we will say that the system (4.67) is *BIBO stable* if there exists a constant $L > 0$ such that the conditions

$$\begin{aligned} x(0) &= 0, \\ \| u(k) \| &\leq 1, \quad k \geq 0, \end{aligned}$$

imply that $\| y(k) \| \leq L$ for all $k \geq 0$.

We will study the *internal qualitative properties* of system (4.67) by studying the *Lyapunov stability* properties of an equilibrium of (4.68).

Since system (4.69) is time-invariant, we will assume without loss of generality that $k_0 = 0$. As in Chapters 1 and 2, we will denote for a given set of initial data $x(0) = x_0$ the solution of (4.69) by $\phi(k, x_0)$. When x_0 is understood or of no importance, we will frequently write $\phi(k)$ in place of $\phi(k, x_0)$. Recall that for system (4.69) [as well as systems (4.67), (4.68), and (4.70)], there are no particular difficulties concerning the existence and uniqueness of solutions, and furthermore, as long as g in (4.69) is continuous, the solutions will be continuous with respect to initial data. Recall also that in contrast to systems described by ordinary differential equations, the solutions of systems described by ordinary difference equations [such as (4.69)] exist only in the forward direction of time ($k \geq 0$).

We say that $x_e \in R^n$ is an *equilibrium* of system (4.69) if $\phi(k, x_e) \equiv x_e$ for all $k \geq 0$, or equivalently,

$$g(x_e) = x_e. \tag{4.71}$$

As in the continuous-time case, we will assume without loss of generality that the equilibrium of interest will be the origin; i.e., $x_e = 0$. If this is not the case, then we can always transform (similarly as in the continuous-time case) system (4.69) into a system of equations that has an equilibrium at the origin.

Example 4.41. The system described by the equation

$$x(k+1) = x(k)[x(k) - 1]$$

has two equilibria, one at $x_{e1} = 0$ and another at $x_{e2} = 1$.

Example 4.42. The system described by the equations

$$\begin{aligned} x_1(k+1) &= x_2(k), \\ x_2(k+1) &= -x_1(k) \end{aligned}$$

has an equilibrium at $x_e^T = (0, 0)$.

Throughout this section we will assume that the function g in (4.69) is continuous, or if required, continuously differentiable. The various definitions of Lyapunov stability of the equilibrium $x = 0$ of system (4.69) are essentially identical to the corresponding definitions of Lyapunov stability of an equilibrium of continuous-time systems described by ordinary differential equations, replacing $t \in R^+$ by $k \in Z^+$. We will concern ourselves with stability, instability, asymptotic stability, and exponential stability of the equilibrium $x = 0$ of (4.69).

We say that the equilibrium $x = 0$ of (4.69) is *stable* if for every $\epsilon > 0$ there exists a $\delta = \delta(\epsilon) > 0$ such that $\| \phi(k, x_0) \| < \epsilon$ for all $k \geq 0$ whenever $\| x_0 \| < \delta$. If the equilibrium $x = 0$ of (4.69) is not stable, it is said to be *unstable*. We say that the equilibrium $x = 0$ of (4.69) is *asymptotically stable* if (i) it is stable and (ii) there exists an $\eta > 0$ such that if $\| x_0 \| < \eta$, then $\lim_{k\to\infty} \| \phi(k, x_0) \| = 0$. If the equilibrium $x = 0$ satisfies property (ii), it is said to be *attractive*, and we call the set of all $x_0 \in R^n$ for which $x = 0$ is attractive the *domain of attraction* of this equilibrium. If $x = 0$ is asymptotically stable and if its domain of attraction is all of R^n, then it is said to be *asymptotically stable in the large* or *globally asymptotically stable*. We say that the equililbrium $x = 0$ of (4.69) is *exponentially stable* if there exists an $\alpha > 0$ and for every $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$, such that $\| \phi(k, x_0) \| \leq \epsilon e^{-\alpha k}$ for all $k \geq 0$ whenever $\| x_0 \| < \delta(\epsilon)$. The equilibrium $x = 0$ of (4.69) is *exponentially stable in the large* if there exists $\alpha > 0$ and for any $\beta > 0$, there exists $k(\beta) > 0$ such that $\| \phi(t, x_0) \| \leq k(\beta) \| x_0 \| e^{-\alpha k}$ for all $k > 0$ whenever $\| x_0 \| < \beta$. Finally, we say that a solution of (4.69) through x_0 is *bounded* if there is a constant M such that $\| \phi(k, x_0) \| \leq M$ for all $k \geq 0$.

4.8.2 Linear Systems

In proving some of the results of this section, we require a result for system (4.68) that is analogous to Theorem 3.1. As in the proof of that theorem, we note that the linear combination of solutions of system (4.68) is also a solution of system (4.68), and hence, the set of solutions $\{\phi : Z^+ \times R^n \to R^n\}$ constitutes a vector space (over $F = R$ or $F = C$). The dimension of this vector space is n. To show this, we choose a set of linearly independent vectors $x_0^1, \dots, x_0^n$ in the n-dimensional x-space (R^n or C^n) and we show, in an identical manner as in the proof of Theorem 3.1, that the set of solutions $\phi(k, x_0^i), i = 1, \dots, n$, is linearly independent and spans the set of solutions of system (4.68). (We ask the reader in the Exercise section to provide the details of the proof of the above assertions.) This yields the following result.

Theorem 4.43. *The set of solutions of system (4.68) over the time interval Z^+ forms an n-dimensional vector space.* ■

Incidentally, if in particular we choose $\phi(k, e^i), i = 1, \dots, n$, where $e^i, i = 1, \dots, n$, denotes the natural basis for R^n, and if we let $\Phi(k, k_0 = 0) \triangleq \Phi(k) = [\phi(k, e^1), \dots, \phi(k, e^n)]$, then it is easily verified that the $n \times n$ matrix $\Phi(k)$ satisfies the matrix equation

$$\Phi(k+1) = A\Phi(k), \quad \Phi(0) = I,$$

and that $\Phi(k) = A^k, k \geq 0$ [i.e., $\Phi(k)$ is the state transition matrix for system (4.68)].

Theorem 4.44. *The equilibrium $x = 0$ of system (4.68) is stable if and only if the solutions of (4.68) are bounded.*

Proof. Assume that the equilibrium $x = 0$ of (4.68) is stable. Then for $\epsilon = 1$ there is a $\delta > 0$ such that $\| \phi(k, x_0) \| < 1$ for all $k \geq 0$ and all $\| x_0 \| \leq \delta$. In this case

$$\| \phi(k, x_0) \| = \| A^k x_0 \| = \| A^k x_0 \delta / \| x_0 \| \| (\| x_0 \| / \delta) < \| x_0 \| / \delta$$

for all $x_0 \neq 0$ and all $k \geq 0$. Using the definition of matrix norm [refer to Section A.7] it follows that $\| A^k \| \leq \delta^{-1}, k \geq 0$. We have proved that if the equilibrium $x = 0$ of (4.68) is stable, then the solutions of (4.68) are bounded.

Conversely, suppose that all solutions $\phi(k, x_0) = A^k x_0$ are bounded. Let $\{e^1, \dots, e^n\}$ denote the natural basis for n-space and let $\| \phi(k, e^j) \| < \beta_j$ for all $k \geq 0$. Then for any vector $x_0 = \sum_{j=1}^n \alpha_j e^j$ we have that

$$\begin{aligned} \| \phi(k, x_0) \| &= \| \sum_{j=1}^n \alpha_j \phi(k, e^j) \| \leq \sum_{j=1}^n |\alpha_j| \beta_j \leq (\max_j \beta_j) \sum_{j=1}^n |\alpha_j| \\ &\leq c \| x_0 \|, \quad k \geq 0, \end{aligned}$$

for some constant c. For given $\epsilon > 0$, we choose $\delta = \epsilon/c$. Then, if $\| x_0 \| < \delta$, we have $\| \phi(k, x_0) \| < c \| x_0 \| < \epsilon$ for all $k \geq 0$. We have proved that if the solutions of (4.68) are bounded, then the equilibrium $x = 0$ of (4.68) is stable. ■

Theorem 4.45. *The following statements are equivalent:*

(i) The equilibrium $x = 0$ of (4.68) is asymptotically stable,
(ii) The equilibrium $x = 0$ of (4.68) is asymptotically stable in the large,
(iii) $\lim_{k\to\infty} \| A^k \| = 0$.

Proof. Assume that statement (i) is true. Then there is an $\eta > 0$ such that when $\| x_0 \| \leq \eta$, then $\phi(k, x_0) \to 0$ as $k \to \infty$. But then we have for *any* $x_0 \neq 0$ that

$$\phi(k, x_0) = A^k x_0 = [A^k(\eta x_0 / \| x_0 \|)] \| x_0 \| /\eta \to 0 \text{ as } k \to \infty.$$

It follows that statement (ii) is true.

Next, assume that statement (ii) is true. Then for any $\epsilon > 0$ there must exist a $K = K(\epsilon)$ such that for all $k \geq K$ we have that $\| \phi(k, x_0) \| = \| A^k x_0 \| < \epsilon$. To see this, let $\{e^1, \ldots, e^n\}$ be the natural basis for R^n. Thus, for a fixed constant $c > 0$, if $x_0 = (\alpha_1, \ldots, \alpha_n)^T$ and if $\| x_0 \| \leq 1$, then $x_0 = \sum_{j=1}^n \alpha_j e^j$ and $\sum_{j=1}^n |\alpha_j| \leq c$. For each j there is a $K_j = K_j(\epsilon)$ such that $\| A^k e^j \| < \epsilon/c$ for $k \geq K_j$. Define $K = K(\epsilon) = \max\{K_j(\epsilon) : j = 1, \ldots, n\}$. For $\| x_0 \| \leq 1$ and $k \geq K$ we have that

$$\| A^k x_0 \| = \| \sum_{j=1}^n \alpha_j A^k e^j \| \leq \sum_{j=1}^n |\alpha_j| (\epsilon/c) \leq \epsilon.$$

By the definition of matrix norm [see Section A.7], this means that $\| A^k \| \leq \epsilon$ for $k > K$. Therefore, statement (iii) is true.

Finally, assume that statement (iii) is true. Then $\| A^k \|$ is bounded for all $k \geq 0$. By Theorem 4.44, the equilibrium $x = 0$ is stable. To prove asymptotic stability, fix $\epsilon > 0$. If $\| x_0 \| < \eta = 1$, then $\| \phi(k, x_0) \| \leq \| A^k \| \| x_0 \| \to 0$ as $k \to \infty$. Therefore, statement (i) is true. This completes the proof. ■

Theorem 4.46. *The equilibrium $x = 0$ of (4.68) is asymptotically stable if and only if it is exponentially stable.*

Proof. The exponential stability of the equilibrium $x = 0$ implies the asymptotic stability of the equilibrium $x = 0$ of systems (4.69) in general and, hence, for systems (4.68) in particular.

Conversely, assume that the equilibrium $x = 0$ of (4.68) is asymptotically stable. Then there is a $\delta > 0$ and a $K > 0$ such that if $\| x_0 \| \leq \delta$, then

$$\| \Phi(k + K) x_0 \| \leq \frac{\delta}{2}$$

for all $k \geq 0$. This implies that

$$\| \Phi(k+K) \| \leq \frac{1}{2} \quad \text{if } k \geq 0. \tag{4.72}$$

From Section 3.5.1 we have that $\Phi(k-l) = \Phi(k-s)\Phi(s-l)$ for any k, l, s. Therefore,

$$\| \Phi(k+2K) \| = \| \Phi[(k+2K)-(k+K)]\Phi(k+K) \| \leq \frac{1}{4}$$

in view of (4.72). By induction we obtain for $k \geq 0$ that

$$\| \Phi(k+nK) \| \leq 2^{-n}. \tag{4.73}$$

Let $\alpha = \frac{(\ln 2)}{K}$. Then (4.73) implies that for $0 \leq k < K$ we have that

$$\begin{aligned} \| (k+nK, x_0) \| &\leq 2 \| x_0 \| 2^{-(n+1)} \\ &= 2 \| x_0 \| e^{-\alpha(n+1)K} \\ &\leq 2 \| x_0 \| e^{-\alpha(k+nK)}, \end{aligned}$$

which proves the result. ■

To arrive at the next result, we make reference to the results of Subsection 3.5.5. Specifically, by inspecting the expressions for the modes of system (4.68) given in (3.131) and (3.132), or by utilizing the Jordan canonical form of A [refer to (3.135) and (3.136)], the following result is evident.

Theorem 4.47. *(i) The equilibrium $x = 0$ of system (4.68) is* asymptotically stable *if and only if all eigenvalues of A are within the unit circle of the complex plane (i.e., if $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of A, then $|\lambda_j| < 1, j = 1, \ldots, n$). In this case we say that* the matrix A is Schur stable*, or simply,* the matrix A is stable.

(ii) The equilibrium $x = 0$ of system (4.68) is stable *if and only if $|\lambda_j| \leq 1, j = 1, \ldots, n$, and for each eigenvalue with $|\lambda_j| = 1$ having multiplicity $n_j > 1$, it is true that*

$$\lim_{z \to \lambda_j} \left\{ \frac{d^{n_j-1-l}}{dz^{n_j-1-l}} [(z-\lambda_j)^{n_j} (zI-A)^{-1}] \right\} = 0, \quad l = 1, \ldots, n_j - 1.$$

(iii) The equilibrium $x = 0$ of system (4.68) is unstable *if and only if the conditions in (ii) above are not true.* ■

Alternatively, it is evident that the equilibrium $x = 0$ of system (4.68) is *stable* if and only if all eigenvalues of A are within or on the unit circle of the complex plane, and every eigenvalue that is on the unit circle has an associated Jordan block of order 1.

Example 4.48. (i) For the system in Example 4.42 we have

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

The eigenvalues of A are $\lambda_1, \lambda_2 = \pm\sqrt{-1}$. According to Theorem 4.47, the equilibrium $x = 0$ of the system is stable, and according to Theorem 4.44 the matrix A^k is bounded for all $k \geq 0$.

(ii) For system (4.68) let

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

The eigenvalues of A are $\lambda_1, \lambda_2 = \pm 1/\sqrt{2}$. According to Theorem 4.47, the equilibrium $x = 0$ of the system is asymptotically stable, and according to Theorem 4.45, $\lim_{k\to\infty} A^k = 0$.

(iii) For system (4.68) let

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

The eigenvalues of A are $\lambda_1, \lambda_2 = \pm\sqrt{3/2}$. According to Theorem 4.47, the equilibrium $x = 0$ of the system is unstable, and according to Theorem 4.44, the matrix A^k is not bounded with increasing k.

(iv) For system (4.68) let

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

The matrix A is a Jordan block of order 2 for the eigenvalue $\lambda = 1$. Accordingly, the equilibrium $x = 0$ of the system is unstable (refer to the remark following Theorem 4.47) and the matrix A^k is unbounded with increasing k.

4.8.3 The Lyapunov Matrix Equation

In this subsection we obtain another characterization of stable matrices by means of the *Lyapunov matrix equation.*

Returning to system (4.68) we choose as a Lyapunov function

$$v(x) = x^T Bx, B = B^T, \tag{4.74}$$

and we evaluate the first forward difference of v along the solutions of (4.68) as

$$\begin{aligned} Dv(x(k)) \triangleq v(x(k+1)) - v(x(k)) &= x(k+1)^T Bx(k+1) - x(k)^T Bx(k) \\ &= x(k)^T A^T BAx(k) - x(k)^T Bx(k) \\ &= x(k)^T (A^T BA - B)x(k), \end{aligned}$$

and therefore,

$$Dv(x) = x^T (A^T BA - B)x \triangleq -x^T Cx,$$

where

$$A^T BA - B = C, \quad C^T = C. \tag{4.75}$$

Theorem 4.49. *(i) The equilibrium $x = 0$ of system (4.68) is* stable *if there exists a real, symmetric, and positive definite matrix B such that the matrix C given in (4.75) is negative semidefinite.*

(ii) The equilibrium $x = 0$ of system (4.68) is asymptotically stable in the large *if there exists a real, symmetric, and positive definite matrix B such that the matrix C given in (4.75) is negative definite.*

(iii) The equilibrium $x = 0$ of system (4.68) is unstable *if there exists a real, symmetric matrix B that is either negative definite or indefinite such that the matrix C given in (4.75) is negative definite.* ■

In proving Theorem 4.49 one can follow a similar approach as in the proofs of Theorems 4.22, 4.24 and 4.26. We leave the details to the reader as an exercise.

In applying Theorem 4.49, we start by choosing (guessing) a matrix B having certain desired properties and we then solve for the matrix C, using equation (4.75). If C possesses certain desired properties (i.e., it is negative definite), we can draw appropriate conclusions by applying one of the results given in Theorem 4.49; if not, we need to choose another matrix B. This approach is not very satisfactory, and in the following we will derive results that will allow us (as in the case of continuous-time systems) to *construct* Lyapunov functions of the form $v(x) = x^T Bx$ in a systematic manner. In doing so, we first choose a matrix C in (4.75) that is either negative definite or positive definite, and then we solve (4.75) for B. Conclusions are then made by applying Theorem 4.49. In applying this construction procedure, we need to know conditions under which (4.75) possesses a (unique) solution B for *any* definite (i.e., positive or negative definite) matrix C. We will address this issue next.

We first show that if A is stable, i.e., if all eigenvalues of matrix A [in system (4.68)] are inside the unit circle of the complex plane, then we can compute B in (4.75) explicitly. To show this, we assume that in (4.75) C is a given matrix and that A is stable. Then

$$(A^T)^{k+1} BA^{k+1} - (A^T)^k BA^k = (A^T)^k CA^k,$$

and summing from $k = 0$ to l yields

$$A^TBA-B+(A^T)^2BA^2-A^TBA+\cdots+(A^T)^{l+1}BA^{l+1}-(A^T)^lBA^l=\sum_{k=0}^{l}(A^T)^kCA^k$$

or

$$(A^T)^{l+1}BA^{l+1}-B=\sum_{k=0}^{l}(A^T)^kCA^k.$$

Letting $l \to \infty$, we obtain

$$B=-\sum_{k=0}^{\infty}(A^T)^kCA^k. \tag{4.76}$$

It is easy to verify that (4.76) is a solution of (4.75). We have

$$-A^T\left[\sum_{k=0}^{\infty}(A^T)^kCA^k\right]A+\sum_{k=0}^{\infty}(A^T)^kCA^k=C$$

or

$$-A^TCA+C-(A^T)^2CA^2+A^TCA-(A^T)^3CA^3+(A^T)^2CA^2-\cdots=C.$$

Therefore (4.76) is a solution of (4.75). Furthermore, if C is negative definite, then B is positive definite.

Combining the above with Theorem 4.49(ii) we have the following result.

Theorem 4.50. *If there is a positive definite and symmetric matrix B and a negative definite and symmetric matrix C satisfying (4.75), then the matrix A is stable. Conversely, if A is stable, then, given* any *symmetric matrix C, (4.75) has a unique solution, and if C is negative definite, then B is positive definite.* ■

Next, we determine conditions under which the system of equations (4.75) has a (unique) solution $B=B^T\in R^{n\times n}$ for a given matrix $C=C^T\in R^{n\times n}$. To accomplish this, we consider the more general equation

$$A_1XA_2-X=C, \tag{4.77}$$

where $A_1\in R^{m\times m}, A_2\in R^{n\times n}$, and X and C are $m\times n$ matrices.

Lemma 4.51. *Let $A_1\in R^{m\times m}$ and $A_2\in R^{n\times n}$. Then (4.77) has a unique solution $X\in R^{m\times n}$ for a given $C\in R^{m\times n}$ if and only if no eigenvalue of A_1 is a reciprocal of an eigenvalue of A_2.*

Proof. We need to show that the condition on A_1 and A_2 is equivalent to the condition that $A_1XA_2=X$ implies $X=0$. Once we have proved that $A_1XA_2=X$ has the unique solution $X=0$, then it can be shown that (4.77) has a unique solution for every C, since (4.77) is a linear equation.

Assume first that the condition on A_1 and A_2 is satisfied. Now $A_1 X A_2 = X$ implies that $A_1^{k-j} X A_2^{k-j} = X$ and

$$A_1^j X = A_1^k X A_2^{k-j} \quad \text{for } k \geq j \geq 0.$$

Now for a polynomial of degree k,

$$p(\lambda) = \sum_{j=0}^{k} a_j \lambda^j,$$

we define the polynomial of degree k,

$$p^*(\lambda) = \sum_{j=0}^{k} a_j \lambda^{k-j} = \lambda^k p(1/\lambda),$$

from which it follows that

$$p(A_1)X = A_1^k X p^*(A_2).$$

Now let $\phi_i(\lambda)$ be the characteristic polynomial of $A_i, i = 1, 2$. Since $\phi_1(\lambda)$ and $\phi_2^*(\lambda)$ are relatively prime, there are polynomials $p(\lambda)$ and $q(\lambda)$ such that

$$p(\lambda)\phi_1(\lambda) + q(\lambda)\phi_2^*(\lambda) = 1.$$

Now define $\phi(\lambda) = q(\lambda)\phi_2^*(\lambda)$ and note that $\phi^*(\lambda) = q^*(\lambda)\phi_2(\lambda)$. It follows that $\phi^*(A_2) = 0$ and $\phi(A_1) = I$. From this it follows that $A_1 X A_2 = X$ implies $X = 0$.

To prove the converse, we assume that λ is an eigenvalue of A_1 and λ^{-1} is an eigenvalue of A_2 (and, hence, is also an eigenvalue of A_2^T). Let $A_1 x^1 = \lambda x^1$ and $A_2^T x^2 = \lambda^{-1} x^2, x^1 \neq 0$ and $x^2 \neq 0$. Define $X = (x_1^2 x^1, x_2^2 x^1, \ldots, x_n^2 x^1)$. Then $X \neq 0$ and $A_1 X A_2 = X$. ∎

To construct $v(x)$ by using Lemma 4.51, we must still check the definiteness of B. To accomplish this, we use Theorem 4.49.

1. If all eigenvalue of A [for system (4.68)] are inside the unit circle of the complex plane, then no reciprocal of an eigenvalue of A is an eigenvalue, and Lemma 4.51 gives another way of showing that (4.75) has a unique solution B for each C if A is stable. If C is negative definite, then B is positive definite. This can be shown as was done for the case of linear ordinary differential equations.
2. Suppose that at least one of the eigenvalues of A is outside the unit circle in the complex plane and that A has no eigenvalues on the unit circle. As in the case of linear differential equations (4.22) (Section 4.5), we use a similarity transformation $x = Qy$ in such a way that $Q^{-1}AQ = \text{diag}[A_1,\ A_2]$, where all eigenvalues of A_1 are outside the unit circle while all eigenvalues

of A_2 are within the unit circle. We then proceed identically as in the case of linear differential equations to show that under the present assumptions there exists for system (4.68) a Lyapunov function that satisfies the hypotheses of Theorem 4.49(iii). Therefore, the equilibrium $x = 0$ of system (4.68) is unstable. If A does not have any eigenvalues within the unit circle, then the equilibrium $x = 0$ of (4.68) is completely unstable. In this proof, Lemma 4.51 has not been invoked. If additionally, the hypotheses of Lemma 4.51 are true (i.e., no reciprocal of an eigenvalue of A is an eigenvalue of A), then we can construct the Lyapunov function for system (4.68) in a systematic manner.

Summarizing the above discussion, we have proved the following result.

Theorem 4.52. *Assume that the matrix A for system (4.68) has no eigenvalues on the unit circle in the complex plane. If all the eigenvalues of the matrix A are within the unit circle of the complex plane, or if at least one eigenvalue is outside the unit circle of the complex plane, then there exists a Lyapunov function of the form $v(x) = x^T Bx, B = B^T$, whose first forward difference along the solutions of system (4.68) is definite (i.e., it is either negative definite or positive definite).* ■

Theorem 4.52 shows that when all the eigenvalues of A are within the unit circle, then for system (4.68), the conditions of Theorem 4.49(ii) are also necessary conditions for exponential stability in the large. Furthermore, when at least one eigenvalue of A is outside the unit circle and no eigenvalues are on the unit circle, then the conditions of Theorem 4.49(iii) are also necessary conditions for instability.

We conclude this subsection with some specific examples.

Example 4.53. (i) For system (4.68), let

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

Let $B = I$, which is positive definite. From (4.75) we obtain

$$C = A^T A - I = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

It follows from Theorem 4.49(i) that the equilibrium $x = 0$ of this system is stable. This is the same conclusion that was made in Example 4.48.

(ii) For system (4.68), let

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

Choose

$$B = \begin{bmatrix} \frac{8}{3} & 0 \\ 0 & \frac{5}{3} \end{bmatrix},$$

which is positive definite. From (4.75) we obtain

$$C = A^T BA - B = \begin{bmatrix} 0 & -1 \\ -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{8}{3} & 0 \\ 0 & \frac{5}{3} \end{bmatrix} \begin{bmatrix} 0 & -\frac{1}{2} \\ -1 & 0 \end{bmatrix} - \begin{bmatrix} \frac{8}{3} & 0 \\ 0 & \frac{5}{3} \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$

which is negative definite. It follows from Theorem 4.49(ii) that the equilibrium $x = 0$ of this system is asymptotically stable in the large. This is the same conclusion that was made in Example 4.48(ii).

(iii) For system (4.68), let

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

Choose

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

which is negative definite. From (4.75) we obtain

$$C = A^T BA - B = \begin{bmatrix} 0 & -3 \\ -\frac{1}{2} & 0 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix} \begin{bmatrix} 0 & -\frac{1}{2} \\ -3 & 0 \end{bmatrix} - \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}$$

or

$$\begin{bmatrix} (9b_{22} - b_{11}) & \frac{1}{2}b_{12} \\ \frac{1}{2}b_{12} & (\frac{1}{4}b_{11} - b_{22}) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$

which yields

$$B = \begin{bmatrix} -8 & 0 \\ 0 & -1 \end{bmatrix},$$

which is also negative definite. It follows from Theorem 4.49(iii) that the equilibrium $x = 0$ of this system is unstable. This conclusion is consistent with the conclusion made in Example 4.48(iii).

(iv) For system (4.68), let

$$A = \begin{bmatrix} \frac{1}{3} & 1 \\ 0 & 3 \end{bmatrix}.$$

The eigenvalues of A are $\lambda_1 = \frac{1}{3}$ and $\lambda_2 = 3$. According to Lemma 4.51, for a given C, (4.77) does *not* have a unique solution in this case since $\lambda_1 = 1/\lambda_2$. For purposes of illustration, we choose $C = -I$. Then

$$-I = A^T BA - B = \begin{bmatrix} \frac{1}{3} & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 1 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ b_{12} & b_{22} \end{bmatrix}$$

or

$$\begin{bmatrix} -\frac{8}{9}b_{11} & \frac{1}{3}b_{11} \\ \frac{1}{3}b_{11} & b_{11} + 6b_{12} + 8b_{22} \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix},$$

which shows that for $C = -I$, (4.77) does not have any solution (for B) at all.

4.8.4 Linearization

In this subsection we determine conditions under which the stability properties of the equilibrium $w = 0$ of the linear system

$$w(k+1) = Aw(k) \tag{4.78}$$

determine the stability properties of the equilibrium $x = 0$ of the nonlinear system

$$x(k+1) = Ax(k) + f(x(k)), \tag{4.79}$$

under the assumption that $f(x) = o(\| x \|)$ as $\| x \| \to 0$ (i.e., given $\epsilon > 0$, there exists $\delta > 0$ such that $\| f(x(k)) \| < \epsilon \| x(k) \|$ for all $k \geq 0$ and all $\| x(k) \| < \delta$). [Refer to the discussion concerning (4.68) to (4.70) in Subsection 4.8.1.]

Theorem 4.54. *Assume that $f \in C(R^n, R^n)$ and that $f(x)$ is $o(\| x \|)$ as $\| x \| \to 0$. (i) If A is stable (i.e., all the eigenvalues of A are within the unit circle of the complex plane), then the equilibrium $x = 0$ of system (4.79) is* asymptotically stable*. (ii) If at least one eigenvalue of A is outside the unit circle of the complex plane and no eigenvalue is on the unit circle, then the equilibrium $x = 0$ of system (4.79) is* unstable. ■

In proving Theorem 4.54 one can follow a similar approach as in the proofs of Theorems 4.32 and 4.33. We leave the details to the reader as an exercise.

Before concluding this subsection, we consider some specific examples.

Example 4.55. (i) Consider the system

$$\begin{aligned} x_1(k+1) &= -\frac{1}{2}x_2(k) + x_1(k)^2 + x_2(k)^2, \\ x_2(k+1) &= -x_1(k) + x_1(k)^2 + x_2(k)^2. \end{aligned} \tag{4.80}$$

Using the notation of (4.79), we have

$$A = \begin{bmatrix} 0 & -\frac{1}{2} \\ -1 & 0 \end{bmatrix}, \quad f(x_1, x_2) = \begin{bmatrix} x_1^2 + x_2^2 \\ \\ x_1^2 + x_2^2 \end{bmatrix}.$$

The linearization of (4.80) is given by

$$w(k+1) = Aw(k). \tag{4.81}$$

From Example 4.48(ii) [and Example 4.53(ii)], it follows that the equilibrium $w = 0$ of (4.81) is asymptotically stable. Furthermore, in the present case $f(x) = o(\| x \|)$ as $\| x \| \to 0$. Therefore, in view of Theorem 4.54, the equilibrium $x = 0$ of system (4.80) is asymptotically stable.

(ii) Consider the system

$$\begin{aligned} x_1(k+1) &= -\frac{1}{2}x_2(k) + x_1(k)^3 + x_2(k)^2, \\ x_2(k+1) &= -3x_1(k) + x_1^4(k) - x_2(k)^5. \end{aligned} \tag{4.82}$$

Using the notation of (4.78) and (4.79), we have in the present case

$$A = \begin{bmatrix} 0 & -\frac{1}{2} \\ -3 & 0 \end{bmatrix}, \quad f(x_1, x_2) = \begin{bmatrix} x_1^3 + x_2^2 \\ \\ x_1^4 - x_2^5 \end{bmatrix}.$$

Since A is unstable [refer to Example 4.53(iii) and Example 4.48(iii)] and since $f(x) = o(\| x \|)$ as $\| x \| \to 0$, it follows from Theorem 4.54 that the equilibrium $x = 0$ of system (4.82) is unstable.

4.8.5 Input–Output Stability

We conclude this chapter by considering the input–output stability of discrete-time systems described by equations of the form

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k), \\ y(k) &= Cx(k), \end{aligned} \tag{4.83}$$

where all matrices and vectors are defined as in (4.67). Throughout this subsection we will assume that $k_0 = 0, x(0) = 0$, and $k \geq 0$.

As in the continuous-time case, we say that system (4.83) is *BIBO stable* if there exists a constant $c > 0$ such that the conditions

$$\begin{aligned} x(0) &= 0, \\ \| u(k) \| &\leq 1, \quad k \geq 0, \end{aligned}$$

imply that $\| y(k) \| \leq c$ for all $k \geq 0$.

The results that we will present involve the impulse response matrix of (4.83) given by

$$H(k) = \begin{cases} CA^{k-1}B, & k > 0, \\ 0, & k \leq 0, \end{cases} \tag{4.84}$$

and the transfer function matrix given by

$$\widehat{H}(z) = C(zI - A)^{-1}B. \tag{4.85}$$

Recall that

$$y(n) = \sum_{k=0}^{n} H(n-k)u(k). \tag{4.86}$$

Associated with system (4.83) is the free dynamical system described by the equation

$$p(k+1) = Ap(k). \tag{4.87}$$

Theorem 4.56. *The system (4.83) is* BIBO stable *if and only if there exists a constant* $L > 0$ *such that for all* $n \geq 0$,

$$\sum_{k=0}^{n} \parallel H(k) \parallel \leq L. \tag{4.88}$$

■

As in the continous-time case, the first part of the proof of Theorem 4.56 (sufficiency) is straightforward. Specifically, if $\parallel u(k) \parallel \leq 1$ for all $k \geq 0$ and if (4.88) is true, then we have for all $n \geq 0$,

$$\begin{aligned} \parallel y(n) \parallel &= \parallel \sum_{k=0}^{n} H(n-k)u(k) \parallel \leq \sum_{k=0}^{n} \parallel H(n-k)u(k) \parallel \\ &\leq \sum_{k=0}^{n} \parallel H(n-k) \parallel \parallel u(k) \parallel \leq \sum_{k=0}^{n} \parallel H(n-k) \parallel \leq L. \end{aligned}$$

Therefore, system (4.83) is BIBO stable.

In proving the second part of Theorem 4.56 (necessity), we simplify matters by first considering in (4.83) the single-variable case ($n = 1$) with the system description given by

$$y(t) = \sum_{k=0}^{t} h(t-k)u(k), \quad t > 0. \tag{4.89}$$

For purposes of contradiction, we assume that the system is BIBO stable, but no finite L exists such that (4.88) is satisfied. Another way of expressing the last assumption is that for *any finite* L, there exists $t = k_1(L) \triangleq k_1$ such that

$$\sum_{k=0}^{k_1} |h(k_1 - k)| > L.$$

We now choose in particular the input u given by

$$u(k) = \begin{cases} +1 & \text{if } h(t-k) > 0, \\ 0 & \text{if } h(t-k) = 0, \\ -1 & \text{if } h(t-k) < 0, \end{cases}$$

$0 \leq k \leq k_1$. Clearly, $|u(k)| \leq 1$ for all $k \geq 0$. The output of the system at $t = k_1$ due to the above input, however, is

$$y(k_1) = \sum_{k=0}^{k_1} h(k_1 - k)u(k) = \sum_{k=0}^{k_1} |h(k_1 - k)| > L,$$

which contradicts the assumption that the system is BIBO stable.

The above can now be extended to the multivariable case. In doing so we apply the single-variable result to every possible pair of input and output vector components, we make use of the fact that the sum of a finite number of bounded sums will be bounded, and we note that a vector is bounded if and only if each of its components is bounded. We leave the details to the reader.

Next, as in the case of continuous-time systems, we note that the asymptotic stability of the equilibrium $p = 0$ of system (4.87) implies the BIBO stability of system (4.83) since the sum

$$\| \sum_{k=1}^{\infty} CA^{k-1}B \| \leq \sum_{k=1}^{\infty} \| C \| \, \| A^{k-1} \| \, \| B \|$$

is finite.

Next, we recall that a complex number z_p is a *pole* of $\widehat{H}(z) = [\hat{h}_{ij}(z)]$ if for some (i, j) we have $|\hat{h}_{ij}(z_p)| = \infty$. If each entry of $\widehat{H}(z)$ has only poles with modulus (magnitude) less than 1, then, as shown in Chapter 3, each entry of $H(k) = [h_{ij}(k)]$ consists of a sum of convergent terms. It follows that under these conditions the sum

$$\sum_{k=0}^{\infty} \| H(k) \|$$

is finite, and any realization of $\widehat{H}(z)$ will result in a system that is BIBO stable.

Conversely, if

$$\sum_{k=0}^{\infty} \| H(k) \|$$

is finite, then the terms in every entry of $H(k)$ must be convergent. But then every entry of $\widehat{H}(z)$ has poles whose modulus is within the unit circle of the complex plane. We have proved the final result of this section.

Theorem 4.57. *The time-invariant system (4.83) is BIBO stable if and only if the poles of the transfer function*

$$\widehat{H}(z) = C(zI - A)^{-1}B$$

are within the unit circle of the complex plane. ■

4.9 Summary and Highlights

In this chapter we first addressed the stability of an equilibrium of continuous-time finite-dimensional systems. In doing so, we first introduced the concept of equilibrium and defined several types of stability in the sense of Lyapunov (Sections 4.2 and 4.3). Next, we established several stability conditions of an equilibrium for linear systems $\dot{x} = Ax$, $t \geq 0$ in terms of the state transition matrix in Theorems 4.12–4.14 and in terms of eigenvalues in Theorem 4.15 (Section 4.4). Next, we established various stability conditions that are phrased in terms of the Lyapunov matrix equation (4.25) for system $\dot{x} = Ax$ (Section 4.5). The existence of Lyapunov functions for $\dot{x} = Ax$ of the form $x^T Px$ is established in Theorem 4.29. In Section 4.6 we established conditions under which the asymptotic stability and the instability of an equilibrium for a nonlinear time-invariant system can be deduced via linearization; see Theorems 4.32 and 4.33.

Next, we addressed the input–output stability of time-invariant linear, continuous-time, finite-dimensional systems (Section 4.7). For such systems we established several conditions for bounded input/bounded output stability (BIBO stability); see Theorems 4.37 and 4.39.

The chapter is concluded with Section 4.8, where we addressed the Lyapunov stability and the input–output stability of linear, time-invariant, discrete-time systems. For such systems, we established results that are analogous to the stability results of continuous-time systems. The stability of an equilibrium is expressed in terms of the state transition matrix in Theorem 4.45, in terms of the eigenvalues in Theorem 4.47, and in terms of the Lyapunov Matrix Equation in Theorems 4.49 and 4.50. The existence of Lyapunov functions of the form $x^T Px$ for $x(k + 1) = Ax(k)$ is established in Theorem 4.52. Stability results based on linearization are presented in Theorem 4.54 and for BIBO stability in Theorems 4.56 and 4.57.

4.10 Notes

The initial contributions to stability theory that took place toward the end of the nineteenth century are primarily due to physicists and mathematicians (Lyapunov [11]), whereas input–output stability is the brainchild of electrical engineers (Sandberg [17] to [19], Zames [21], [22]). Sources with extensive coverage of Lyapunov stability theory include, e.g., Hahn [6], Khalil [8], LaSalle [9], LaSalle and Lefschetz [10], Michel and Miller [13], Michel et al. [14], Miller and Michel [15], and Vidyasagar [20]. Input–output stability is addressed in great detail in Desoer and Vidyasagar [5], Vidyasagar [20], and Michel and Miller [13]. For a survey that traces many of the important developments of stability in feedback control, refer to Michel [12].

In the context of *linear systems*, sources on both Lyapunov stability and input–output stability can be found in numerous texts, including Antsaklis and Michel [1], Brockett [2], Chen [3], DeCarlo [4], Kailath [7], and Rugh [16]. In developing our presentation, we found the texts by Antsaklis and Michel [1], Brockett [2], Hahn [6], LaSalle [9], and Miller and Michel [15] especially helpful.

In this chapter, we addressed various types of Lyapunov stability and bounded input/bounded output stability of time-invariant systems. In the various stability concepts for such systems, the initial time t_0 (resp., k_0) plays no significant role, and for this reason, we chose without loss of generality $t_0 = 0$ (resp., $k_0 = 0$). In the case of time-varying systems, this is in general not true, and in defining the various Lyapunov stability concepts and the concept of bounded input/bounded output stability, one has to take into account the effects of initial time. In doing so, we have to distinguish between *uniformity* and *nonuniformity* when defining the various types of Lyapunov stability of an equilibrium and the BIBO stability of a system. For a treatment of the Lyapunov stability and the BIBO stablity of the time-varying counterparts of systems (4.14), (4.15) and (4.67), (4.68), we refer the reader to Chapter 6 in Antsaklis and Michel [1].

We conclude by noting that there are graphical criteria (i.e., frequency domain criteria), such as, the Leonhard–Mikhailov criterion, and algebraic criteria, such as the Routh–Hurwitz criterion and the Schur–Cohn criterion, which yield necessary and sufficient conditions for the asymptotic stability of the equilibrium $x = 0$ for system (4.15) and (4.68). For a presentation of these results, the reader should consult Chapter 6 in Antsaklis and Michel [1] and Michel [12].

References

1. P.J. Antsaklis and A.N. Michel, *Linear Systems*, Birkhäuser, Boston, MA, 2006.
2. R.W. Brockett, *Finite Dimensional Linear Systems*, Wiley, New York, NY, 1970.
3. C.T. Chen, *Linear System Theory and Design*, Holt, Rinehart and Winston, New York, NY, 1984.
4. R.A. DeCarlo, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
5. C.A. Desoer and M. Vidyasagar, *Feedback Systems: Input–Output Properties*, Academic Press, New York, NY, 1975.
6. W. Hahn, *Stability of Motion*, Springer-Verlag, New York, NY, 1967.
7. T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
8. H.K. Khalil, *Nonlinear Systems*, Macmillan, New York, NY, 1992.
9. J.P. LaSalle, *The Stability and Control of Discrete Processes*, Springer-Verlag, New York, NY, 1986.
10. J.P. LaSalle and S. Lefschetz, *Stability by Liapunov's Direct Method*, Academic Press, New York, NY, 1961.
11. M.A. Liapounoff, "Problème générale de la stabilité de mouvement," *Ann. Fac. Sci. Toulouse*, Vol. 9, 1907, pp. 203–474. (Translation of a paper published in *Comm. Soc. Math.* Kharkow 1893, reprinted in *Ann. Math. Studies*, Vol. 17, 1949, Princeton, NJ).
12. A.N. Michel, "Stability: the common thread in the evolution of feedback control," *IEEE Control Systems*, Vol. 16, 1996, pp. 50–60.
13. A.N. Michel and R.K. Miller, *Qualitative Analysis of Large Scale Dynamical Systems*, Academic Press, New York, NY, 1977.
14. A.N. Michel, K. Wang, and B. Hu, *Qualitative Theory of Dynamical Systems, Second Edition*, Marcel Dekker, New York, NY, 2001.

15. R.K. Miller and A.N. Michel, *Ordinary Differential Equations*, Academic Press, New York, NY, 1982.
16. W.J. Rugh, *Linear System Theory, Second Edition*, Prentice-Hall, Englewood Cliffs, NJ, 1999.
17. I.W. Sandberg, "On the L_2-boundedness of solutions of nonlinear functional equations," *Bell Syst. Tech. J.*, Vol. 43, 1964, pp. 1581–1599.
18. I.W. Sandberg, "A frequency-domain condition for stability of feedback systems containing a single time-varying nonlinear element," *Bell Syst. Tech. J.*, Vol. 43, 1964, pp. 1601–1608.
19. I.W. Sandberg, "Some results on the theory of physical systems governed by nonlinear functional equations," *Bell Syst. Tech. J.*, Vol. 44, 1965, pp. 871–898.
20. M. Vidyasagar, *Nonlinear Systems Analysis*, 2d edition, Prentice Hall, Englewood Cliffs, NJ, 1993.
21. G. Zames, "On the input–output stability of time-varying nonlinear feedback systems, Part I," *IEEE Trans. on Automat. Contr.*, Vol. 11, 1966, pp. 228–238.
22. G. Zames, "On the input–output stability of time-varying nonlinear feedback systems, Part II," *IEEE Trans. on Automat. Contr.*, Vol. 11, 1966, pp. 465–476.

Exercises

4.1. Determine the set of equilibrium points of a system described by the differential equations

$$\begin{aligned}\dot{x}_1 &= x_1 - x_2 + x_3,\\ \dot{x}_2 &= 2x_1 + 3x_2 + x_3,\\ \dot{x}_3 &= 3x_1 + 2x_2 + 2x_3.\end{aligned}$$

4.2. Determine the set of equilibria of a system described by the differential equations

$$\begin{aligned}\dot{x}_1 &= x_2,\\ \dot{x}_2 &= \begin{cases} x_1 \sin(1/x_1), & \text{when } x_1 \neq 0,\\ 0, & \text{when } x_1 = 0.\end{cases}\end{aligned}$$

4.3. Determine the equilibrium points and their stability properties of a system described by the ordinary differential equation

$$\dot{x} = x(x-1) \tag{4.90}$$

by solving (4.90) and then applying the definitions of stability, asymptotic stability, etc.

4.4. Prove that the equilibrium $x = 0$ of (4.16) is stable (resp., asymptotically stable or unstable) if and only if $y = 0$ of (4.19) is stable (resp., asymptotically stable or unstable).

4.5. Apply Proposition 4.20 to determine the definiteness properties of the matrix A given by

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 5 & -1 \\ 1 & -1 & 10 \end{bmatrix}.$$

4.6. Use Theorem 4.26 to prove that the trivial solution of the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

is unstable.

4.7. Determine the equilibrium points of a system described by the differential equation

$$\dot{x} = -x + x^2,$$

and determine the stability properties of the equilibrium points, if applicable, by using Theorem 4.32 or 4.33.

4.8. The system described by the differential equations

$$\begin{aligned} \dot{x}_1 &= x_2 + x_1(x_1^2 + x_2^2), \\ \dot{x}_2 &= -x_1 + x_2(x_1^2 + x_2^2) \end{aligned} \tag{4.91}$$

has an equilibrium at the origin $x^T = (x_1, x_2) = (0, 0)$. Show that the trivial solution of the *linearization* of system (4.91) is stable. Prove that the equilibrium $x = 0$ of system (4.91) is unstable. (This example shows that the assumptions on the matrix A in Theorems 4.32 and 4.33 are absolutely essential.)

4.9. Use Corollary 4.38 to analyze the stability properties of the system given by

$$\begin{aligned} \dot{x} &= Ax + Bu, \\ y &= Cx, \\ A &= \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad C = [0, 1]. \end{aligned}$$

4.10. Determine all equilibrium points for the discrete-time systems given by

(a)

$$\begin{aligned} x_1(k+1) &= x_2(k) + |x_1(k)|, \\ x_2(k+1) &= -x_1(k) + |x_2(k)|. \end{aligned}$$

(b)

$$\begin{aligned} x_1(k+1) &= x_1(k)x_2(k) - 1, \\ x_2(k+1) &= 2x_1(k)x_2(k) + 1. \end{aligned}$$

4.11. Prove Theorem 4.43.

4.12. Determine the stability properties of the trivial solution of the discrete-time system given by the equations

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}$$

with θ fixed.

4.13. Analyze the stability of the equilibrium $x = 0$ of the system described by the scalar-valued difference equation

$$x(k+1) = \sin[x(k)].$$

4.14. Analyze the stability of the equilibrium $x = 0$ of the system described by the difference equations

$$\begin{aligned} x_1(k+1) &= x_1(k) + x_2(k)[x_1(k)^2 + x_2(k)^2], \\ x_2(k+1) &= x_2(k) - x_1(k)[x_1(k)^2 + x_2(k)^2]. \end{aligned}$$

4.15. Determine a basis of the solution space of the system

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -6 & 5 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}.$$

Use your answer in analyzing the stability of the trivial solution of this system.

4.16. Let $A \in R^{n\times n}$. Prove that part (iii) of Theorem 4.45 is equivalent to the statement that all eigenvalues of A have modulus less than 1; i.e.,

$$\lim_{k\to\infty} \parallel A^k \parallel = 0$$

if and only if for any eigenvalue λ of A, it is true that $|\lambda| < 1$.

4.17. Use Theorem 4.44 to show that the equilibrium $x = 0$ of the system

$$x(k+1) = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} x(k)$$

is unstable.

4.18. (a) Use Theorem 4.47 to determine the stability of the equilibrium $x = 0$ of the system

$$x(k+1) = \begin{bmatrix} 1 & 1 & -2 \\ 0 & 1 & 3 \\ 0 & 9 & -1 \end{bmatrix} x(k).$$

(b) Use Theorem 4.47 to determine the stability of the equilibrium $x = 0$ of the system

$$x(k+1) = \begin{bmatrix} 1 & 0 & -2 \\ 0 & 1 & 3 \\ 0 & 9 & -1 \end{bmatrix} x(k).$$

4.19. Apply Theorems 4.24 and 4.49 to show that if the equilibrium $x = 0$ ($x \in R^n$) of the system

$$x(k+1) = e^A x(k)$$

is asymptotically stable, then the equilibrium $x = 0$ of the system

$$\dot{x} = Ax$$

is also asymptotically stable.

4.20. Apply Theorem 4.49 to show that the trivial solution of the system given by

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 2 & 0 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix}$$

is unstable.

4.21. Determine the stability of the equilibrium $x = 0$ of the scalar-valued system given by

$$x(k+1) = \frac{1}{2}x(k) + \frac{2}{3}\sin x(k).$$

4.22. Analyze the stability properties of the discrete-time system given by

$$\begin{aligned} x(k+1) &= x(k) + \frac{1}{2}u(k) \\ y(k) &= \frac{1}{2}x(k) \end{aligned}$$

where x, y, and u are scalar-valued variables. Is this system BIBO stable?

4.23. Prove Theorem 4.47.

4.24. Prove Theorem 4.49 by following a similar approach as was used in the proofs of Theorems 4.22, 4.24, and 4.26.

4.25. Prove Theorem 4.54 by following a similar approach as was used in the proofs of Theorems 4.32 and 4.33.

5

Controllability and Observability: Fundamental Results

5.1 Introduction

The principal goals of this chapter are to introduce the system properties of controllability and observability (and of reachability and constructibility), which play a central role in the study of state feedback controllers and state observers, and in establishing the relations between internal and external system representations, topics that will be studied in Chapters 7, 8, and 9. State controllability refers to the ability to manipulate the state by applying appropriate inputs (in particular, by steering the state vector from one vector value to any other vector value in finite time). Such is the case, for example, in satellite attitude control, where the satellite must change its orientation. State observability refers to the ability to determine the state vector of the system from knowledge of the input and the corresponding output over some finite time interval. Since it is frequently difficult or impossible to measure the state of a system directly (for example, internal temperatures and pressures in an internal combustion engine), it is very desirable to determine such states by observing the inputs and outputs of the system over some finite time interval.

In Section 5.2, the concepts of reachability and controllability and observability and constructibility are introduced, using discrete-time time-invariant systems. Discrete-time systems are selected for this exposition because the mathematical development is much simpler in this case. In subsection 5.2.3 the concept of duality is also introduced. Reachability and controllability are treated in detail in Section 5.3 and observability and constructibility in Section 5.4 for both continuous-time and discrete-time time-invariant systems.

5.2 A Brief Introduction to Reachability and Observability

Reachability and controllability are introduced first, followed by observability and constructibility. These important system concepts are more easily explained in the discrete-time case, and this is the approach taken in this section. Duality is also discussed at the end of the section.

5.2.1 Reachability and Controllability

The concepts of *state reachability* (or *controllability-from-the-origin*) and *controllability* (or *controllability-to-the-origin*) are introduced here and are discussed at length in Section 5.3.

In the case of time-invariant systems, a state x_1 is called *reachable* if there exists an input that transfers the state of the system $x(t)$ from the zero state to x_1 in some finite time T. The definition of reachability for the discrete-time case is completely analogous. Figure 5.1 shows that different control inputs $u_1(t)$ and $u_2(t)$ may force the state of a continuous-time system to reach the value x_1 from the origin at different finite times T_1 and T_2, following different paths. Note that reachability refers to the ability of the system to reach x_1 from the origin in some finite time; it specifies neither the exact time it takes to achieve this nor the trajectory to be followed.

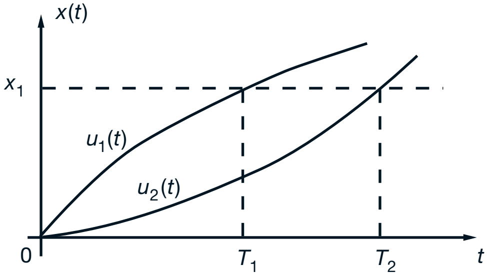


Figure 5.1. A reachable state x_1

A state x_0 is called *controllable* if there exists an input that transfers the state from x_0 to the zero state in some finite time T. See Figure 5.2. The definition of controllability for the discrete-time case is completely analogous. Similar to reachability, controllability specifies neither the time it takes to achieve the transfer nor the trajectory to be followed.

We note that when particular types of trajectories to be followed are of interest, then one seeks particular control inputs that will achieve such transfers. This leads to various control problem formulations, including the Linear Quadratic (Optimal) Regulator (LQR). The LQR problem is discussed in Chapter 9.

Section 5.3 shows that reachability always implies controllability, but controllability implies reachability only when the state transition matrix Φ of the system is nonsingular. This is always true for continuous-time systems, but is

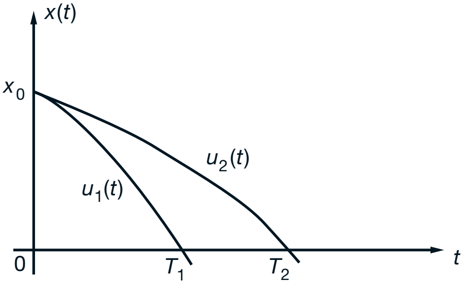


Figure 5.2. A controllable state x_0

true for discrete-time systems only when the matrix A of the system is nonsingular. If the system is state reachable, then there always exists an input that transfers any state x_0 to any other state x_1 in finite time.

In the time-invariant case, a system is *reachable* (or *controllable-from-the-origin*) if and only if its *controllability matrix* $\mathcal{C}$,

$$\mathcal{C} \triangleq [B, AB, \ldots, A^{n-1}B] \in R^{n \times mn}, \tag{5.1}$$

has full row rank n; that is, $\operatorname{rank} \mathcal{C} = n$. The matrices $A \in R^{n \times n}$ and $B \in R^{n \times m}$ come from either the continuous-time state equations

$$\dot{x} = Ax + Bu \tag{5.2}$$

or the discrete-time state equations

$$x(k+1) = Ax(k) + Bu(k), \tag{5.3}$$

$k \geq k_0 = 0$. Alternatively, we say that the pair (A, B) is reachable. The matrix $\mathcal{C}$ should perhaps more appropriately be called the "reachability matrix" or the "controllability-from-the-origin matrix." The term "controllability matrix," however, has been in use for some time and is expected to stay in use. Therefore, we shall call $\mathcal{C}$ the "controllability matrix," having in mind the "controllability-from-the-origin matrix."

We shall now discuss reachability and controllability for discrete-time time-invariant systems (5.3).

If the state $x(k)$ in (5.3) is expressed in terms of the initial vector $x(0)$, then (see Subsection 3.5.1)

$$x(k) = A^k x(0) + \sum_{i=0}^{k-1} A^{k-(i+1)} Bu(i) \tag{5.4}$$

for $k > 0$. Rewriting the summation in terms of matrix-vector multiplication, it follows that it is possible to transfer the state from some value $x(0) = x_0$ to some x_1 in n steps, that is, $x(n) = x_1$, if there exists an n-step input sequence $\{u(0), u(1), \ldots, u(n-1)\}$ that satisfies the equation

$$x_1 - A^n x_0 = \mathcal{C}_n U_n, \tag{5.5}$$

where $\mathcal{C}_n \triangleq [B, AB, \dots, A^{n-1}B] = \mathcal{C}$ [see (5.1)] and

$$U_n \triangleq [u^T(n-1), u^T(n-2), \dots, u^T(0)]^T. \tag{5.6}$$

From the theory of linear algebraic equations, (5.5) has a solution U_n if and only if

$$x_1 - A^n x_0 \in \mathcal{R}(\mathcal{C}), \tag{5.7}$$

where $\mathcal{R}(\mathcal{C}) = \text{range}(\mathcal{C})$. Note that it is not necessary to take more than n steps in the control sequence, since if this transfer cannot be accomplished in n steps, it cannot be accomplished at all. This follows from the Cayley–Hamilton Theorem, in view of which it can be shown that $\mathcal{R}(\mathcal{C}_n) = \mathcal{R}(\mathcal{C}_k)$ for $k \geq n$. Also note that $\mathcal{R}(\mathcal{C}_n)$ includes $\mathcal{R}(\mathcal{C}_k)$ for $k < n$ [i.e., $\mathcal{R}(\mathcal{C}_n) \supset \mathcal{R}(\mathcal{C}_k), k < n$]. (See Exercise 5.1.)

It is now easy to see that the system (5.3) or the pair (A, B) is *reachable* (*controllable-from-the-origin*), if and only if $\text{rank}\, \mathcal{C} = n$, since in this case $\mathcal{R}(\mathcal{C}) = R^n$, the entire state space. Note that $x_1 \in \mathcal{R}(\mathcal{C})$ is the condition for a particular state x_1 to be reachable from the zero state. Since $\mathcal{R}(\mathcal{C})$ contains all such states, it is called the *reachable subspace* of the system. It is also clear from (5.5) that if the system is reachable, any state x_0 can be transferred to any other state x_1 in n steps. In addition, the input that accomplishes this transfer is any solution U_n of (5.5). Note that, depending on x_1 and x_0, this transfer may be accomplished in fewer than n steps (see Section 5.3).

Example 5.1. Consider $x(k+1) = Ax(k) + Bu(k)$, where $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Here the controllability (-from-the-origin) matrix $\mathcal{C}$ is $\mathcal{C} = [B, AB] = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ with $\text{rank}\, \mathcal{C} = 2$. Therefore, the system [or the pair (A, B)] is reachable, meaning that any state x_1 can be reached from the zero state in a finite number of steps by applying at most n inputs $\{u(0), u(1), \dots, u(n-1)\}$ (presently, $n = 2$). To see this, let $x_1 = \begin{bmatrix} a \\ b \end{bmatrix}$. Then (5.5) implies that $\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u(1) \\ u(0) \end{bmatrix}$ or $\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} b-a \\ a \end{bmatrix}$. Thus, the control $u(0) = a, u(1) = b - a$ will transfer the state from the origin at $k = 0$ to the state $\begin{bmatrix} a \\ b \end{bmatrix}$ at $k = 2$. To verify this, we observe that $x(1) = Ax(0) + Bu(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} a = \begin{bmatrix} 0 \\ a \end{bmatrix}$ and $x(2) = Ax(1) + Bu(1) = \begin{bmatrix} a \\ a \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} (b-a) = \begin{bmatrix} a \\ b \end{bmatrix}$.

Reachability of the system also implies that a state x_1 can be reached from any other state x_0 in at most $n = 2$ steps. To illustrate this, let

$x(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Then (5.5) implies that $x_1 - A^2 x_0 = \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} a-2 \\ b-3 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u(1) \\ u(0) \end{bmatrix}$. Solving, $\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} b-a-1 \\ a-2 \end{bmatrix}$, which will drive the state from $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ at $k=0$ to $\begin{bmatrix} a \\ b \end{bmatrix}$ at $k=2$.

Notice that in general the solution U_n of (5.5) is not unique; i.e., many inputs can accomplish the transfer from $x(0) = x_0$ to $x(n) = x_1$, each corresponding to a particular state trajectory. In control problems, particular inputs are frequently selected that, in addition to transferring the state, satisfy additional criteria, such as, e.g., minimization of an appropriate performance index (optimal control).

A system [or the pair (A, B)] is *controllable, or controllable-to-the-origin*, when any state x_0 can be driven to the zero state in a finite number of steps. From (5.5) we see that a system is controllable when $A^n x_0 \in \mathcal{R}(\mathcal{C})$ for any x_0. If $\operatorname{rank} A = n$, a system is controllable when $\operatorname{rank} \mathcal{C} = n$, i.e., when the reachability condition is satisfied. In this case the $n \times mn$ matrix

$$A^{-n}\mathcal{C} = [A^{-n}B, \dots, A^{-1}B] \tag{5.8}$$

is of interest and the system is controllable if and only if $\operatorname{rank}(A^{-n}\mathcal{C}) = \operatorname{rank} \mathcal{C} = n$. If, however, $\operatorname{rank} A < n$, then controllability does not imply reachability (see Section 5.3).

Example 5.2. The system in Example 5.1 is controllable (-to-the-origin). To see this, we let, $x_1 = 0$ in (5.5) and write $-A^2 x_0 = -\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = [B, AB] \begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u(1) \\ u(0) \end{bmatrix}$, where $x_0 = \begin{bmatrix} a \\ b \end{bmatrix}$. From this we obtain $\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = -\begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -b \\ -a-b \end{bmatrix}$, which is the input that will drive the state from $\begin{bmatrix} a \\ b \end{bmatrix}$ at $k=0$ to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ at $k=2$.

Example 5.3. The system $x(k+1) = 0$ is controllable since any state, say, $x(0) = \left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]$, can be transferred to the zero state in one step. In this system, however, the input u does not affect the state at all! This example shows that reachability is a more useful concept than controllability for discrete-time systems.

It should be pointed out that nothing has been said up to now about maintaining the desired system state after reaching it [refer to (5.5)]. Zeroing the input for $k \geq n$, i.e., letting $u(k) = 0$ for $k \geq n$, will not typically work, unless $Ax_1 = x_1$. In general a state starting at x_1, will remain at x_1 for all $k \geq n$ if and only if there exists an input $u(k), k \geq n$, such that

$$x_1 = Ax_1 + Bu(k), \tag{5.9}$$

that is, if and only if $(I - A)x_1 \in \mathcal{R}(B)$. Clearly, there are states for which this condition may not be satisfied.

5.2.2 Observability and Constructibility

In Section 5.4, definitions for state *observability* and *constructibility* are given, and appropriate tests for these concepts are derived. It is shown that observability always implies constructibility, whereas constructibility implies observability only when the state transition matrix Φ of the system is nonsingular. Whereas this is always true for continuous-time systems, it is true for discrete-time systems only when the matrix A of the system is nonsingular. If a system is state observable, then its present state can be determined from knowledge of the present and future outputs and inputs. Constructibility refers to the ability of determining the present state from present and past outputs and inputs, and as such, it is of greater interest in applications.

In the time-invariant case a system [or a pair (A, C)] is observable if and only if its *observability matrix* $\mathcal{O}$, where

$$\mathcal{O} \triangleq \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \in R^{pn \times n}, \tag{5.10}$$

has full column rank; i.e., rank $\mathcal{O} = n$. The matrices $A \in R^{n \times n}$ and $C \in R^{p \times n}$ are given by the system description

$$\dot{x} = Ax + Bu, \quad y = Cx + Du \tag{5.11}$$

in the continuous-time case, and by the system description

$$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k), \tag{5.12}$$

with $k \geq k_0 = 0$, in the discrete-time case.

We shall now briefly discuss observability and constructibility for the discrete-time time-invariant case. As in the case of reachability and controllability, this discussion will provide insight into the underlying concepts and will clarify what these imply for a system.

If the output in (5.12) is expressed in terms of the initial vector $x(0)$, then

$$y(k) = CA^k x(0) + \sum_{i=0}^{k-1} CA^{k-(i+1)} Bu(i) + Du(k) \tag{5.13}$$

for $k > 0$ (see Section 3.5). This implies that

$$\tilde{y}(k) = CA^k x_0 \tag{5.14}$$

for $k \geq 0$, where

$$\tilde{y}(k) \triangleq y(k) - \left[\sum_{i=0}^{k-1} CA^{k-(i+1)} Bu(i) + Du(k) \right]$$

for $k > 0, \tilde{y}(0) \triangleq y(0) - Du(0)$, and $x_0 = x(0)$. In (5.14) x_0 is to be determined assuming that the system parameters are given and the inputs and outputs are measured. Note that if $u(k) = 0$ for $k \geq 0$, then the problem is simplified, since $\tilde{y}(k) = y(k)$ and since the output is generated only by the initial condition x_0. It is clear that the ability of determining x_0 from output and input measurements depends only on the matrices A and C, since the left-hand side of (5.14) is a known quantity. Now if $x(0) = x_0$ is known, then all $x(k), k \geq 0$, can be determined by means of (5.12). To determine x_0, we apply (5.14) for $k = 0, \ldots, n-1$. Then

$$\widetilde{Y}_{0,n-1} = \mathcal{O}_n x_0, \tag{5.15}$$

where $\mathcal{O}_n \triangleq [C^T, (CA)^T, \ldots, (CA^{n-1})^T]^T = \mathcal{O}$ [as in (5.10)] and

$$\widetilde{Y}_{0,n-1} \triangleq [\tilde{y}^T(0), \ldots, \tilde{y}^T(n-1)]^T.$$

Now (5.15) always has a solution x_0, by construction. A system is observable if the solution x_0 is unique, i.e., if it is the only initial condition that, together with the given input sequence, can generate the observed output sequence. From the theory of linear systems of equations, (5.15) has a unique solution x_0 if and only if the null space of $\mathcal{O}$ consists of only the zero vector, i.e., $\text{Null}(\mathcal{O}) = \mathcal{N}(\mathcal{O}) = \{0\}$, or equivalently, if and only if the only $x \in R^n$ that satisfies

$$\mathcal{O}x = 0 \tag{5.16}$$

is the zero vector. This is true if and only if $\operatorname{rank} \mathcal{O} = n$. Thus, a system is observable if and only if $\operatorname{rank} \mathcal{O} = n$. Any nonzero state vector $x \in R^n$ that satisfies (5.16) is said to be an unobservable state, and $\mathcal{N}(\mathcal{O})$ is said to be the *unobservable subspace*. Note that any such x satisfies $CA^k x = 0$ for $k = 0, 1, \ldots, n-1$. If $\operatorname{rank} \mathcal{O} < n$, then all vectors x_0 that satisfy (5.15) are given by $x_0 = x_{0p} + x_{0h}$, where x_{0p} is a particular solution and x_{0h} is any vector in $\mathcal{N}(\mathcal{O})$. Any of these state vectors, together with the given inputs, could have generated the measured outputs.

To determine x_0 from (5.15) it is not necessary to use more than n values for $\tilde{y}(k), k = 0, \ldots, n-1$, or to observe $y(k)$ for more than n steps in the future. This is true because, in view of the Cayley–Hamilton Theorem, it can be shown that $\mathcal{N}(\mathcal{O}_n) = \mathcal{N}(\mathcal{O}_k)$ for $k \geq n$. Note also that $\mathcal{N}(\mathcal{O}_n)$ is included in $\mathcal{N}(\mathcal{O}_k)$ $(\mathcal{N}(\mathcal{O}_n) \subset \mathcal{N}(\mathcal{O}_k))$ for $k < n$. Therefore, in general, one has to observe the output for n steps (see Exercise 5.1).

Example 5.4. Consider the system $x(k+1) = Ax(k), y(k) = Cx(k)$, where $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ and $C = [0\ 1]$. Here, $\mathcal{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ with $\operatorname{rank} \mathcal{O} = 2$. Therefore, the system [or the pair (A, C)] is observable. This means that $x(0)$ can uniquely be determined from $n = 2$ output measurements (in the present cases, the input is zero). In fact, in view of (5.15), $\begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$ or $\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} y(1) - y(0) \\ y(0) \end{bmatrix}$.

Example 5.5. Consider the system $x(k+1) = Ax(k), y(k) = Cx(k)$, where $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and $C = [1\ 0]$. Here, $\mathcal{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$ with $\operatorname{rank} \mathcal{O} = 1$. Therefore, the system is not observable. Note that a basis for $\mathcal{N}(\mathcal{O})$ is $\left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$, which in view of (5.16) implies that all state vectors of the form $\begin{bmatrix} 0 \\ c \end{bmatrix}, c \in R$, are unobservable. Relation (5.15) implies that $\begin{bmatrix} y(0) \\ y(1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$. For a solution $x(0)$ to exist, as it must, we have that $y(0) = y(1) = a$. Thus, this system will generate an identical output for $k \geq 0$. Accordingly, all $x(0)$ that satisfy (5.15) and can generate this output are given by $\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} a \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ c \end{bmatrix} = \begin{bmatrix} a \\ c \end{bmatrix}$, where $c \in R$.

In general, a system (5.12) [or a pair (A, C)] is *constructible* if the only vector x that satisfies $x = A^k \hat{x}$ with $C\hat{x} = 0$ for every $k \geq 0$ is the zero vector. When A is nonsingular, this condition can be stated more simply, namely, that the system is constructible if the only vector x that satisfies $CA^{-k}x = 0$ for every $k \geq 0$ is the zero vector. Compare this with the condition $CA^k x = 0, k \geq 0$, for x to be an unobservable state; or with the condition that a system is observable if the only vector x that satisfies $CA^k x = 0$ for every $k \geq 0$ is the zero vector. In view of (5.14), the above condition for a system to be constructible is the condition for the existence of a unique solution x_0 when past outputs and inputs are used. This, of course, makes sense since constructibility refers to determining the present state from knowledge of past outputs and inputs. Therefore, when A is nonsingular, the system is constructible if and only if the $pn \times n$ matrix

$$\mathcal{O}A^{-n} = \begin{bmatrix} CA^{-n} \\ \vdots \\ CA^{-1} \end{bmatrix} \tag{5.17}$$

has full rank, since in this case the only x that satisfies $CA^{-k}x = 0$ for every $k \geq 0$ is $x = 0$. Note that if the system is observable, then it is also constructible; however, if it is constructible, then it is also observable only when A is nonsingular (see Section 5.3).

Example 5.6. Consider the (unobservable) system in Example 5.5. Since A is nonsingular, $\mathcal{O}A^{-2} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -2 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. Since $\operatorname{rank} \mathcal{O}A^{-2} = 1 < 2$, the system [or the pair (A, C)] is not constructible. This can also be seen from the relation $CA^{-k}x = 0$, $k \geq 0$, that has nonzero solutions x, since $C = [1,\ 0] = CA^{-1} = CA^{-2} = \cdots = CA^{-k}$ for $k \geq 0$, which implies that any $x = \begin{bmatrix} 0 \\ c \end{bmatrix}$, $c \in R$, is a solution.

5.2.3 Dual Systems

Consider the system described by

$$\dot{x} = Ax + Bu, \quad y = Cx + Du, \tag{5.18}$$

where $A \in R^{n \times n}, B \in R^{n \times m}, C \in R^{p \times n}$, and $D \in R^{p \times m}$. The *dual system* of (5.18) is defined as the system

$$\dot{x}_D = A_D x_D + B_D u_D, \quad y_D = C_D x_D + D_D u_D, \tag{5.19}$$

where $A_D = A^T, B_D = C^T, C_D = B^T$, and $D_D = D^T$.

Lemma 5.7. *System (5.18), denoted by $\{A, B, C, D\}$, is reachable (controllable) if and only if its dual $\{A_D, B_D, C_D, D_D\}$ in (5.19) is observable (constructible), and vice versa.*

Proof. System $\{A, B, C, D\}$ is reachable if and only if $\mathcal{C} \triangleq [B, AB, \ldots, A^{n-1}B]$ has full rank n, and its dual is observable if and only if

$$\mathcal{O}_D \triangleq \begin{bmatrix} B^T \\ B^T A^T \\ \vdots \\ B^T (A^T)^{n-1} \end{bmatrix}$$

has full rank n. Since $\mathcal{O}_D^T = \mathcal{C}, \{A, B, C, D\}$ is reachable if and only if $\{A_D, B_D, C_D, D_D\}$ is observable. Similarly, $\{A, B, C, D\}$ is observable if and only if $\{A_D, B_D, C_D, D_D\}$ is reachable. Now $\{A, B, C, D\}$ is controllable if and only if its dual is constructible, and vice versa, since it is shown in Sections 5.3 and 5.4, that a continuous-time system is controllable if and only if it is reachable; it is constructible if and only if it is observable. ■

For the discrete-time time-invariant case, the dual system is again defined as $A_D = A^T, B_D = C^T, C_D = B^T$, and $D_D = D^T$. That such a system is reachable if and only if its dual is observable can be shown in exactly the same way as in the proof of Lemma 5.7. That such a system is controllable if and only if its dual is constructible in the case when A is nonsingular is because in this case the system is reachable if and only if it is controllable; and the same holds for observability and constructibility. The proof for the case when A is singular involves the controllable and unconstructible subspaces of a system and its dual. We omit the details. The reader is encouraged to complete this proof after studying Sections 5.3 and 5.4.

Figure 5.3 summarizes the relationships between reachability (observability) and controllability (constructibility) for continuous- and discrete-time systems.

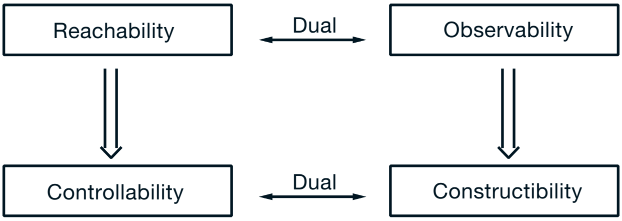


Figure 5.3. In continuous-time systems, reachability (observability) always implies and is implied by controllability (constructibility). In discrete-time systems, reachability (observability) always implies but in general is not implied by controllability (constructibility).

5.3 Reachability and Controllability

The objective here is to study the important properties of state controllability and reachability when a system is described by a state-space representation. In the previous section, a brief introduction to these concepts was given for discrete-time systems, and it was shown that a system is completely reachable if and only if the controllability (-from-the-origin) matrix $\mathcal{C}$ in (5.1) has full rank n ($\operatorname{rank} \mathcal{C} = n$). Furthermore, it was shown that the input sequence necessary to accomplish the transfer can be determined directly from $\mathcal{C}$ by solving a system of linear algebraic equations (5.5). In a similar manner, we would like to derive tests for reachability and controllability and determine the necessary system inputs to accomplish the state transfer for the continuous-time case. We note, however, that whereas the test for reachability can be derived by a number of methods, the appropriate sequence of system inputs to use cannot easily be determined directly from $\mathcal{C}$, as was the case for discrete-time systems. For this reason, we use an approach that utilizes ranges of maps, in particular, the range of an important $n \times n$ matrix—the reachability Gramian. The inputs that accomplish the desired state transfer can be determined directly from this matrix.

5.3.1 Continuous-Time Time-Invariant Systems

We consider the state equation

$$\dot{x} = Ax + Bu, \tag{5.20}$$

where $A \in R^{n\times n}, B \in R^{n\times m}$, and $u(t) \in R^m$ is (piecewise) continuous. The state at time t is given by

$$x(t) = \Phi(t, t_0)x(t_0) + \int_{t_0}^{t} \Phi(t, \tau)Bu(\tau)d\tau, \tag{5.21}$$

where $\Phi(t, \tau)$ is the state transition matrix of the system, and $x(t_0) = x_0$ denotes the state at initial time.

Here

$$\Phi(t, \tau) = \Phi(t - \tau, 0) = \exp[(t - \tau)A] = e^{A(t-\tau)}. \tag{5.22}$$

We are interested in using the input to transfer the state from x_0 to some other value x_1 at some finite time $t_1 > t_0$, [i.e., $x(t_1) = x_1$ in (5.21)]. Because of time invariance, the difference $t_1 - t_0 = T$, rather than the individual times t_0 and t_1, is important. Accordingly, we can always take $t_0 = 0$ and $t_1 = T$. Equation (5.21) assumes the form

$$x_1 - e^{AT}x_0 = \int_0^T e^{A(T-\tau)}Bu(\tau)d\tau, \tag{5.23}$$

and clearly, there exists $u(t)$, $t \in [0, T]$, that satisfies (5.23) if and only if such transfer of the state is possible. Letting $\hat{x}_1 \triangleq x_1 - e^{AT}x_0$, we note that the $u(t)$ that transfers the state from x_0 at time 0 to x_1 at time T will also cause the state to reach $\hat{x}_1$ at T, starting from the origin at 0 (i.e., $x(0) = 0$).

For the time-invariant system (5.20), we introduce the following concepts.

Definition 5.8. *(i) A* state x_1 *is* reachable *if there exists an input* $u(t)$, $t \in [0, T]$*, that transfers the state* $x(t)$ *from the origin at* $t = 0$ *to* x_1 *in some finite time* T*.*

(ii) The set of all reachable states R_r is the reachable subspace *of the system $\dot{x} = Ax + Bu$, or of the pair (A, B).*

(iii) The system $\dot{x} = Ax + Bu$, or the pair (A, B) is (completely state) reachable *if every state is reachable, i.e., if $R_r = R^n$.* ■

Regarding (ii), note that the set of all reachable states x_1 contains the origin and constitutes a linear subspace of the state space (R^n, R).

A reachable state is sometimes also called *controllable-from-the-origin*. Additionally, there are also states defined to be *controllable-to-the-origin* or simply *controllable*; see the definition later in this section.

Definition 5.9. *The $n \times n$* reachability Gramian *of the time-invariant system $\dot{x} = Ax + Bu$ is*

$$W_r(0,T) \triangleq \int_0^T e^{(T-\tau)A} BB^T e^{(T-\tau)A^T} d\tau. \tag{5.24}$$

■

Note that W_r is symmetric and positive semidefinite for every $T > 0$; i.e., $W_r = W_r^T$ and $W_r \geq 0$ (show this).

It can now be shown in [1, p. 230, Lemma 3.2.1] that the reachable subspace of the system (5.20) is exactly the range of the reachability Gramian W_r in (5.24). Let the $n \times mn$ *controllability (-from-the-origin) matrix* be

$$\mathcal{C} \triangleq [B, AB, \ldots, A^{n-1}B]. \tag{5.25}$$

The range of $W_r(0,T)$, denoted by $\mathcal{R}(W_r(0,T))$, is independent of T; i.e., it is the same for any finite $T(> 0)$, and in particular, it is equal to the range of the controllability matrix $\mathcal{C}$. Thus, the reachable subspace R_r of system (5.20), which is the set of all states that can be reached from the origin in finite time, is given by the range of $\mathcal{C}, \mathcal{R}(\mathcal{C})$, or the range of $W_r(0,T), \mathcal{R}(W_r(0,T))$, for some finite (and therefore for any) $T > 0$. This is stated as Lemma 5.10 below; for the proof, see [1, p. 236, Lemma 3.2.10].

Lemma 5.10. $\mathcal{R}(W_r(0,T)) = \mathcal{R}(\mathcal{C})$ *for every $T > 0$.* ■

Example 5.11. For the system $\dot{x} = Ax + Bu$ with $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, we have $e^{At} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$ and $e^{At}B = \begin{bmatrix} t \\ 1 \end{bmatrix}$. The reachability Gramian is $W_r(0,T) = \int_0^T \begin{bmatrix} T-\tau \\ 1 \end{bmatrix} [T-\tau, 1] d\tau = \int_0^T \begin{bmatrix} (T-\tau)^2 & T-\tau \\ T-\tau & 1 \end{bmatrix} d\tau = \begin{bmatrix} \frac{1}{3}T^3 & \frac{1}{2}T^2 \\ \frac{1}{2}T^2 & T \end{bmatrix}$. Since $\det W_r(0,T) = \frac{1}{12}T^4 \neq 0$ for any $T > 0$, $\operatorname{rank} W_r(0,T) = n$ and (A, B) is reachable. Note that $\mathcal{C} = [B, AB] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and that $\mathcal{R}(W_r(0,T)) = \mathcal{R}(\mathcal{C}) = R^2$, as expected (Lemma 5.10).

If $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, instead of $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, then $\mathcal{C} = [B, AB] = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ and (A, B) is not reachable. In this case $e^{At}B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and the reachability matrix is $W_r(0,T) = \int_0^T \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} d\tau = \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix}$. Notice again that $\mathcal{R}(\mathcal{C}) = \mathcal{R}(W_r(0,T))$ for every $T > 0$.

The following theorems and corollaries 5.12 to 5.15 contain the main reachability results. Their proofs may be found in [1, p. 237, Chapter 3], starting with Theorem 2.11.

Theorem 5.12. *Consider the system $\dot{x} = Ax + Bu$, and let $x(0) = 0$. There exists an input u that transfers the state to x_1 in finite time if and only if $x_1 \in \mathcal{R}(\mathcal{C})$, or equivalently, if and only if $x_1 \in \mathcal{R}(W_r(0,T))$ for some finite (and therefore for any) T. Thus, the reachable subspace $R_r = \mathcal{R}(\mathcal{C}) = \mathcal{R}(W_r(0,T))$. Furthermore, an appropriate u that will accomplish this transfer in time T is given by*

$$u(t) = B^T e^{A^T(T-t)} \eta_1 \tag{5.26}$$

with η_1 such that $W_r(0,T)\eta_1 = x_1$ and $t \in [0,T]$. ■

Note that in (5.26) no restrictions are imposed on time T, other than T be finite. T can be as small as we wish; i.e., the transfer can be accomplished in a very short time indeed.

Corollary 5.13. *The system $\dot{x} = Ax + Bu$, or the pair (A, B), is (completely state) reachable, if and only if*

$$\operatorname{rank} \mathcal{C} = n, \tag{5.27}$$

or equivalently, if and only if

$$\operatorname{rank} W_r(0,T) = n \tag{5.28}$$

for some finite (and therefore for any) T. ■

Theorem 5.14. *There exists an input u that transfers the state of the system $\dot{x} = Ax + Bu$ from x_0 to x_1 in some finite time T if and only if*

$$x_1 - e^{AT}x_0 \in \mathcal{R}(\mathcal{C}), \tag{5.29}$$

or equivalently, if and only if

$$x_1 - e^{AT}x_0 \in \mathcal{R}(W_r(0,T)). \tag{5.30}$$

Such an input is given by

$$u(t) = B^T e^{A^T(T-t)} \eta_1 \tag{5.31}$$

with $t \in [0, T]$, where η_1 is a solution of

$$W_r(0, T)\eta_1 = x_1 - e^{AT} x_0. \tag{5.32}$$

■

The above theorem leads to the next result, which establishes the importance of reachability in determining an input u to transfer the state from any x_0 to any x_1 in finite time.

Corollary 5.15. *Let the system $\dot{x} = Ax + Bu$ be (completely state) reachable, or the pair (A, B) be reachable. Then there exists an input that will transfer any state x_0 to any other state x_1 in some finite time T. Such input is given by*

$$u(t) = B^T e^{A^T(T-t)} W_r^{-1}(0, T)[x_1 - e^{AT} x_0] \tag{5.33}$$

for $t \in [0, T]$. ■

There are many different control inputs u that can accomplish the transfer from x_0 to x_1 in time T. It can be shown that the input u given by (5.33) accomplishes this transfer while expending a minimum amount of energy; in fact, u minimizes the cost functional $\int_0^T \| u(\tau) \|^2 \, d\tau$, where $\| u(t) \| \triangleq [u^T(t)u(t)]^{1/2}$ denotes the Euclidean norm of $u(t)$.

Example 5.16. The system $\dot{x} = Ax + Bu$ with $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is reachable (see Example 5.11). A control input $u(t)$ that will transfer any state x_0 to any other state x_1 in some finite time T is given by (see Corollary 5.15 and Example 5.11)

$$\begin{aligned} u(t) &= B^T e^{A^T(T-t)} W_r^{-1}(0, T)[x_1 - e^{AT} x_0] \\ &= [T - t, 1] \begin{bmatrix} 12/T^3 & -6/T^2 \\ -6/T^2 & 4/T \end{bmatrix} \left[x_1 - \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} x_0 \right]. \end{aligned}$$

Example 5.17. For the (scalar) system $\dot{x} = -ax + bu$, determine $u(t)$ that will transfer the state from $x(0) = x_0$ to the origin in T sec; i.e., $x(T) = 0$.

We shall apply Corollary 5.15. The reachability Gramian is $W_r(0, T) = \int_0^T e^{-(T-\tau)a} b b e^{-(T-\tau)a} d\tau = e^{-2aT} b^2 \int_0^T e^{2a\tau} d\tau = e^{-2aT} b^2 \frac{1}{2a}[e^{2aT} - 1] = \frac{b^2}{2a}[1 - e^{-2aT}]$. (Note [see (5.36) below] that the controllability Gramian is $W_c(0, T) = \frac{b^2}{2a}[e^{2aT} - 1]$.) Now in view of (5.33), we have

$$\begin{aligned} u(t) &= b e^{-(T-t)a} \frac{2a}{b^2} \frac{1}{1 - e^{-2aT}} [-e^{-aT} x_0] \\ &= -\frac{2a}{b} \frac{e^{-2aT}}{1 - e^{-2aT}} e^{aT} x_0 = -\frac{2a}{b} \frac{1}{e^{2aT} - 1} e^{at} x_0. \end{aligned}$$

To verify that this $u(t)$ accomplishes the desired transfer, we compute $x(t) = e^{At}x_0 + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau = e^{-at}x_0 + \int_0^t e^{-at}e^{a\tau}bu(\tau)d\tau = e^{-at}[x_0 + \int_0^t e^{a\tau}b \times \left(-\frac{2a}{b}\frac{1}{e^{2aT}-1} \times e^{a\tau}\right) d\tau = e^{-at}\left[1 - \frac{e^{2at}-1}{e^{2aT}-1}\right] x_0$. Note that $x(T) = 0$, as desired, and also that $x(0) = x_0$. The above expression shows also that for $t > T$, the state does not remain at the origin. An important point to notice here is that as $T \to 0$, the control magnitude $|u| \to \infty$. Thus, although it is (theoretically) possible to accomplish the desired transfer instantaneously, this will require infinite control magnitude. In general the faster the transfer, the larger the control magnitude required.

We now introduce the concept of a controllable state.

Definition 5.18. *(i) A* state x_0 *is* controllable *if there exists an input* $u(t)$, $t \in [0, T]$, *which transfers the state* $x(t)$ *from* x_0 *at* $t = 0$ *to the origin in some finite time* T.

(ii) The set of all controllable states R_c, *is the* controllable subspace *of the system* $\dot{x} = Ax + Bu$, *or of the pair* (A, B).

(iii) The system $\dot{x} = Ax + Bu$, *or the pair* (A, B), *is* (completely state) controllable *if every state is controllable, i.e., if* $R_c = R^n$. ■

We shall now establish the relationship between reachability and controllability for the continuous-time time-invariant systems (5.20).

In view of (5.23), x_0 is controllable when there exists $u(t), t \in [0, T]$, so that

$$-e^{AT}x_0 = \int_0^T e^{A(T-\tau)}Bu(\tau)d\tau$$

or when $e^{AT}x_0 \in \mathcal{R}(W_r(0,T))$ [1, p. 230, Lemma 3.2.1], or equivalently, in view of Lemma 5.10, when

$$e^{AT}x_0 \in \mathcal{R}(\mathcal{C}) \tag{5.34}$$

for some finite T. Recall that x_1 is reachable when

$$x_1 \in \mathcal{R}(\mathcal{C}). \tag{5.35}$$

We require the following result.

Lemma 5.19. *If* $x \in \mathcal{R}(\mathcal{C})$, *then* $Ax \in \mathcal{R}(\mathcal{C})$; *i.e., the reachable subspace* $R_r = \mathcal{R}(\mathcal{C})$ *is an* A-invariant *subspace.*

Proof. If $x \in \mathcal{R}(\mathcal{C})$, this means that there exists a vector α such that $[B, AB, \ldots, A^{n-1}B]\alpha = x$. Then $Ax = [AB, A^2B, \ldots, A^nB]\alpha$. In view of the Cayley–Hamilton Theorem, A^n can be expressed as a linear combination of $A^{n-1}, \ldots, A, I$, which implies that $Ax = \mathcal{C}\beta$ for some appropriate vector β. Therefore, $Ax \in \mathcal{R}(\mathcal{C})$. ■

Theorem 5.20. *Consider the system $\dot{x} = Ax + Bu$.*

(i) A state x is reachable if and only if it is controllable.
(ii) $R_c = R_r$.
(iii) The system (2.3), or the pair (A, B), is (completely state) reachable if and only if it is (completely state) controllable.

Proof. (i) Let x be reachable, that is, $x \in \mathcal{R}(\mathcal{C})$. Premultiply x by $e^{AT} = \sum_{k=0}^{\infty}(T^k/k!)A^k$ and notice that, in view of Lemma 5.19, $Ax, A^2x, \dots, A^kx \in \mathcal{R}(\mathcal{C})$. Therefore, $e^{AT}x \in \mathcal{R}(\mathcal{C})$ for any T that, in view of (5.34), implies that x is also controllable. If now x is controllable, i.e., $e^{AT}x \in \mathcal{R}(\mathcal{C})$, then premultiplying by e^{-AT}, the vector $e^{-AT}\left(e^{AT}x\right) = x$ will also be in $\mathcal{R}(\mathcal{C})$. Therefore, x is also reachable. Note that the second part of (i), that controllability implies reachability, is true because the inverse $(e^{AT})^{-1} = e^{-AT}$ does exist. This is in contrast to the discrete-time case where the state transition matrix $\Phi(k, 0)$ is nonsingular if and only if A is nonsingular [nonreversibility of time in discrete-time systems].

Parts (ii) and (iii) of the theorem follow directly from (i). ■

The reachability Gramian for the time-invariant case, $W_r(0, T)$, was defined in (5.24). For completeness the controllability Gramian is defined below.

Definition 5.21. *The* controllability Gramian *in the time-invariant case is the $n \times n$ matrix*

$$W_c(0, T) \triangleq \int_0^T e^{-A\tau}BB^Te^{-A^T\tau}d\tau. \tag{5.36}$$

■

We note that

$$W_r(0, T) = e^{AT}W_c(0, T)e^{A^TT},$$

which can be verified directly.

Additional Criteria for Reachability and Controllability

We first recall the definition of a set of linearly independent functions of time and consider in particular n complex-valued functions $f_i(t)$, $i = 1, \dots, n$, where $f_i^T(t) \in C^m$. Recall that the set of functions f_i, $i = 1, \dots, n$, is *linearly dependent* on a time interval $[t_1, t_2]$ over the field of complex numbers C if there exist complex numbers a_i, $i = 1, \dots, n$, not all zero, such that

$$a_1f_1(t) + \cdots + a_nf_n(t) = 0 \quad \text{for all } t \text{ in } [t_1, t_2];$$

otherwise, the set of functions is said to be *linearly independent* on $[t_1, t_2]$ over the field of complex numbers.

It is possible to test linear independence using the *Gram matrix of the functions f_i*.

Lemma 5.22. *Let $F(t) \in C^{n\times m}$ be a matrix with $f_i(t) \in C^{1\times m}$ in its ith row. Define the* Gram matrix *of $f_i(t)$, $i = 1, \ldots, n$, by*

$$W(t_1, t_2) \triangleq \int_{t_1}^{t_2} F(t)F^*(t)dt, \tag{5.37}$$

where $(\cdot)^*$ denotes the complex conjugate transpose. The set $f_i(t)$, $i = 1, \ldots, n$, is linearly independent on $[t_1, t_2]$ over the field of complex numbers if and only if the Gram matrix $W(t_1, t_2)$ is nonsingular, or equivalently, if and only if the Gram determinant *$\det W(t_1, t_2) \neq 0$.*

Proof. (*Necessity*) Assume the set $f_i, i = 1, \ldots, n$, is linearly independent but $W(t_1, t_2)$ is singular. Then there exists some nonzero $\alpha \in C^{1\times n}$ so that $\alpha W(t_1, t_2) = 0$, from which $\alpha W(t_1, t_2)\alpha^* = \int_{t_1}^{t_2}(\alpha F(t))(\alpha F(t))^* dt = 0$. Since $(\alpha F(t))(\alpha F(t))^* \geq 0$ for all t, this implies that $\alpha F(t) = 0$ for all t in $[t_1, t_2]$, which is a contradiction. Therefore, $W(t_1, t_2)$ is nonsingular.

(*Sufficiency*) Assume that $W(t_1, t_2)$ is nonsingular but the set $f_i, i = 1, \ldots, n$, is linearly dependent. Then there exists some nonzero $\alpha \in C^{1\times n}$ so that $\alpha F(t) = 0$. Then $\alpha W(t_1, t_2) = \int_{t_1}^{t_2} \alpha F(t)F^*(t)dt = 0$, which is a contradiction. Therefore, the set f_i, $i = 1, \ldots, n$, is linearly independent. ■

We now introduce a number of additional tests for reachability and controllability of time-invariant systems. Some earlier results are also repeated here for convenience.

Theorem 5.23. *The system $\dot{x} = Ax + Bu$ is reachable (controllable-from-the-origin)*

(i) *if and only if*

$$\operatorname{rank} W_r(0, T) = n \quad \textit{for some finite } T > 0,$$

where

$$W_r(0, T) \triangleq \int_0^T e^{(T-\tau)A}BB^T e^{(T-\tau)A^T} d\tau, \tag{5.38}$$

the reachability Gramian; or

(ii) *if and only if the n rows of*

$$e^{At}B \tag{5.39}$$

are linearly independent on $[0, \infty)$ over the field of complex numbers; or alternatively, if and only if the n rows of

$$(sI - A)^{-1}B \tag{5.40}$$

are linearly independent over the field of complex numbers; or

(iii) *if and only if*

$$\operatorname{rank} \mathcal{C} = n, \tag{5.41}$$

where $\mathcal{C} \triangleq [B, A, B, \ldots, A^{n-1}B]$, the controllability matrix; or *(iv) if and only if*

$$\operatorname{rank}[s_i I - A, B] = n \tag{5.42}$$

for all complex numbers s_i; or alternatively, for s_i, $i = 1, \ldots, n$, the eigenvalues of A.

Proof. Parts (i) and (ii) were proved in Corollary 5.13.

In part (ii), $\operatorname{rank} W_r(0,T) = n$ implies and is implied by the linear independence of the n rows of $e^{(T-t)A}B$ on $[0,T]$ over the field of complex numbers, in view of Lemma 5.22, or by the linear independence of the n rows of $e^{\hat{t}A}B$, where $\hat{t} \triangleq T - t$, on $[0,T]$. Therefore, the system is reachable if and only if the n rows of $e^{At}B$ are linearly independent on $[0,\infty)$ over the field of complex numbers. Note that the time interval can be taken to be $[0,\infty)$ since in $[0,T]$, T can be taken to be any finite positive real number. To prove the second part of (ii), recall that $\mathcal{L}(e^{At}B) = (sI - A)^{-1}B$ and that the Laplace transform is a one-to-one linear operator.

Part (iv) will be proved later in Section 6.3. ■

Since reachability implies and is implied by controllability, the criteria developed in the theorem for reachability are typically used to test the controllability of a system as well.

Example 5.24. For the system $\dot{x} = Ax + Bu$, where $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ (as in Example 5.11), we shall verify Theorem 5.23. The system is reachable since

(i) the reachability Gramian $W_r(0,T) = \begin{bmatrix} \frac{1}{3}T^3 & \frac{1}{2}T^2 \\ \frac{1}{2}T^2 & T \end{bmatrix}$ has $\operatorname{rank} W_r(0,T) = 2 = n$ for any $T > 0$, or since

(ii) $e^{At}B = \begin{bmatrix} t \\ 1 \end{bmatrix}$ has rows that are linearly independent on $[0,\infty)$ over the field of complex numbers (since $a_1 \times t + a_2 \times 1 = 0$, where a_1 and a_2 are complex numbers implies that $a_1 = a_2 = 0$). Similarly, the rows of $(sI - A)^{-1}B = \begin{bmatrix} 1/s^2 \\ 1/s \end{bmatrix}$ are linearly independent over the field of complex numbers. Also, since

(iii) $\operatorname{rank} \mathcal{C} = \operatorname{rank}[B, AB] = \operatorname{rank} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = 2 = n$, or

(iv) $\operatorname{rank}[s_i I - A, B] = \operatorname{rank} \begin{bmatrix} s_i & -1 & 0 \\ 0 & s_i & 1 \end{bmatrix} = 2 = n$ for $s_i = 0$, $i = 1, 2$, the eigenvalues of A.

If $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ in place of $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, then

(i) $W_r(0,T) = \begin{bmatrix} T & 0 \\ 0 & 0 \end{bmatrix}$ (see Example 5.11) with $\operatorname{rank} W_r(0,T) = 1 < 2 = n$, and

(ii) $e^{At}B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $(sI - A)^{-1}B = \begin{bmatrix} 1/s \\ 0 \end{bmatrix}$, neither of which has rows that are linearly independent over the complex numbers. Also,

(iii) $\operatorname{rank} \mathcal{C} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = 1 < 2 = n$, and

(iv) $\operatorname{rank}[s_iI - A, B] = \operatorname{rank} \begin{bmatrix} s_i & -1 & 1 \\ 0 & s_i & 0 \end{bmatrix} = 1 < 2 = n$ for $s_i = 0$.

Based on any of the above tests, it is concluded that the system is not reachable.

5.3.2 Discrete-Time Systems

The response of discrete-time systems was studied in Section 3.5. We consider systems described by equations of the form

$$x(k+1) = Ax(k) + Bu(k), \quad k \geq k_0, \tag{5.43}$$

where $A \in R^{n \times n}$ and $B \in R^{n \times m}$. The state $x(k)$ is given by

$$x(k) = \Phi(k, k_0)x(k_0) + \sum_{i=k_0}^{k-1} \Phi(k, i+1)Bu(i), \tag{5.44}$$

where the state transition matrix is

$$\Phi(k, k_0) = A^{k-k_0}, \quad k \geq k_0. \tag{5.45}$$

Let the state at time k_0 be x_0. For the state at some time $k_1 > k_0$ to assume the value x_1, an input u must exist that satisfies $x(k_1) = x_1$ in (5.44).

For the time-invariant system the elapsed time $k_1 - k_0$ is of interest, and we therefore take $k_0 = 0$ and $k_1 = K$. Recalling that $\Phi(k, 0) = A^k$, for the state x_1 to be reached from $x(0) = x_0$ in K steps, i.e., $x(K) = x_1$, an input u must exist that satisfies

$$x_1 = A^K x_0 + \sum_{i=0}^{K-1} A^{K-(i+1)}Bu(i), \tag{5.46}$$

when $K > 0$, or

$$x_1 = A^K x_0 + \mathcal{C}_K U_K, \tag{5.47}$$

where

$$\mathcal{C}_K \triangleq [B, AB, \ldots, A^{K-1}B] \tag{5.48}$$

and

$$U_K \triangleq [u^T(K-1), u^T(K-2), \ldots, u^T(0)]^T. \tag{5.49}$$

The definitions of *reachable state* x_1, *reachable subspace* R_r, and a *system being (completely state) reachable*, or *the pair (A,B) being reachable*, are the same as in the continuous-time case (see Definition 5.8, and use integer K in place of real time T).

To determine the finite input sequence for discrete-time systems that will accomplish a desired state transfer, if such a sequence exists, one does not have to define matrices comparable with the reachability Gramian W_r, as in the case for continuous-time systems, but we can work directly with the controllability matrix $\mathcal{C}_n = \mathcal{C}$; see also the introductory discussion in Section 5.2.1. In particular, we have the following result.

Theorem 5.25. *Consider the system $x(k+1) = Ax(k) + Bu(k)$ given in (5.43), and let $x(0) = 0$. There exists an input u that transfers the state to x_1 in finite time if and only if*

$$x_1 \in \mathcal{R}(\mathcal{C}).$$

In this case, x_1 is reachable and $R_r = \mathcal{R}(\mathcal{C})$. An appropriate input sequence $\{u(k)\}$, $k = 0, \ldots, n-1$, that accomplishes this transfer in n steps is determined by $U_n \triangleq [u^T(n-1), u^T(n-2), \ldots, u^T(0)]^T$, which is a solution to the equation

$$\mathcal{C}U_n = x_1. \tag{5.50}$$

Henceforth, with an abuse of language, we will refer to U_n as a control sequence, when in fact we actually have in mind $\{u(k)\}$.

Proof. In view of (5.47), x_1 can be reached from the origin in K steps if and only if $x_1 = \mathcal{C}_K U_K$ has a solution U_K, or if and only if $x_1 \in \mathcal{R}(\mathcal{C}_K)$. Furthermore, all input sequences that accomplish this are solutions to the equation $x_1 = \mathcal{C}_K U_K$. For x_1 to be reachable we must have $x_1 \in \mathcal{R}(\mathcal{C}_K)$ for some finite K. This range, however, cannot increase beyond the range of $\mathcal{C}_n = \mathcal{C}$; i.e., $\mathcal{R}(\mathcal{C}_K) = \mathcal{R}(\mathcal{C}_n)$ for $K \geq n$ [see Exercise 5.1]. This follows from the Cayley–Hamilton Theorem, which implies that any vector x in $\mathcal{R}(\mathcal{C}_K)$, $K \geq n$, can be expressed as a linear combination of $B, AB, \ldots, A^{n-1}B$. Therefore, $x \in \mathcal{R}(\mathcal{C}_n)$. It is of course possible to have $x_1 \in \mathcal{R}(\mathcal{C}_K)$ with $K < n$, for a particular x_1; however, in this case $x_1 \in \mathcal{R}(\mathcal{C}_n)$, since $\mathcal{C}_K$ is a subset of $\mathcal{C}_n$. Thus, x_1 is reachable if and only if it is in the range of $\mathcal{C}_n = \mathcal{C}$. Clearly, any U_n that accomplishes the transfer satisfies (5.50). ■

As pointed out in the above proof, for given x_1 we may have $x_1 \in \mathcal{R}(\mathcal{C}_K)$ for some $K < n$. In this case the transfer can be accomplished in fewer than n steps, and appropriate inputs are obtained by solving the equation $\mathcal{C}_K U_K = x_1$.

Corollary 5.26. *The system $x(k+1) = Ax(k) + Bu(k)$ in (5.43) is* (completely state) reachable*, or the pair (A, B) is reachable, if and only if*

$$\operatorname{rank} \mathcal{C} = n. \tag{5.51}$$

Proof. Apply Theorem 5.25, noting that $\mathcal{R}(\mathcal{C}) = R_r = R^n$ if and only if $\operatorname{rank} \mathcal{C} = n$. ■

Theorem 5.27. *There exists an input u that transfers the state of the system $x(k+1) = Ax(k) + Bu(k)$ in (5.43) from x_0 to x_1 in some finite number of steps K, if and only if*

$$x_1 - A^K x_0 \in \mathcal{R}(\mathcal{C}_K). \tag{5.52}$$

Such an input sequence $U_K \triangleq [u^T(K-1), u^T(K-2), \ldots, u^T(0)]^T$ is determined by solving the equation

$$\mathcal{C}_K U_K = x_1 - A^K x_0. \tag{5.53}$$

Proof. The proof follows directly from (5.47). ■

The above theorem leads to the following result that establishes the importance of reachability in determining u to transfer the state from any x_0 to any x_1 in a finite number of steps.

Corollary 5.28. *Let the system $x(k+1) = Ax(k) + Bu(k)$ given in (5.43) be (completely state) reachable or the pair (A, B) be reachable. Then there exists an input sequence that transfers the state from any x_0 to any x_1 in a finite number of steps. Such input is determined by solving Eq. (5.54).*

Proof. Consider (5.47). Since (A, B) is reachable, $\operatorname{rank} \mathcal{C}_n = \operatorname{rank} \mathcal{C} = n$ and $\mathcal{R}(\mathcal{C}) = R^n$. Then

$$\mathcal{C} U_n = x_1 - A^n x_0 \tag{5.54}$$

always has a solution $U_n = [u^T(n-1), \ldots, u^T(0)]^T$ for any x_0 and x_1. This input sequence transfers the state from x_0 to x_1 in n steps. ■

Note that, in view of Theorem 5.27, for particular x_0 and x_1, the state transfer may be accomplished in $K < n$ steps, using (5.53).

Example 5.29. Consider the system in Example 5.1, namely, $x(k+1) = Ax(k) + Bu(k)$, where $A = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$. Since $\operatorname{rank} \mathcal{C} = \operatorname{rank}[B, AB] = \operatorname{rank} \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right] = 2 = n$, the system is reachable and any state x_0 can be transferred to any other state x_1 in two steps. Let $x_1 = \begin{bmatrix} a \\ b \end{bmatrix}$, $x_0 = \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}$. Then (5.54) implies that $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}$ or $\begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} = \begin{bmatrix} b - 1 - b_0 \\ a - a_0 - b_0 \end{bmatrix}$. This agrees with the results obtained in Example 5.1. In view of (5.53), if x_1 and x_0 are chosen so that $x_1 - Ax_0 = \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} = \begin{bmatrix} a - b_0 \\ b - a_0 - b_0 \end{bmatrix}$ is in the $\mathcal{R}(\mathcal{C}_1) = \mathcal{R}(B) = \text{span}\left\{\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right\}$, then the state transfer can be achieved in one step. For example, if $x_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ and $x_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, then $Bu(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(0) = x_1 - Ax_0 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$ implies that the transfer from x_0 to x_1 can be accomplished in this case in $1 < 2 = n$ steps with $u(0) = 2$.

Example 5.30. Consider the system $x(k+1) = Ax(k) + Bu(k)$ with $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Since $\mathcal{C} = [B, AB] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ has full rank, there exists an input sequence that will transfer the state from any $x(0) = x_0$ to any $x(n) = x_1$ (in n steps), given by (5.54), $U_2 = \begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \mathcal{C}^{-1}(x_1 - A^2 x_0) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} (x_1 - x_0)$. Compare this with Example 5.16, where the continuous-time system had the same system parameters A and B.

Additional Criteria for Reachability. Note that completely analogous results to Theorem 5.23(ii)–(iv) exist for the discrete-time case.

We now turn to the concept of controllability. The definitions of *controllable state* x_0, *controllable subspace* R_c, and a *system* being *(completely state) controllable*, or *the pair (A,B) being controllable* are similar to the corresponding concepts given in Definition 5.18 for the case of continuous-time systems.

We shall now establish the relationship between reachability and controllability for the discrete-time time-invariant systems $x(k+1) = Ax(k) + Bu(k)$ in (5.43).

Consider (5.46). The state x_0 is controllable if it can be steered to the origin $x_1 = 0$ in a finite number of steps K. That is, x_0 is controllable if and only if

$$-A^K x_0 = \mathcal{C}_K U_K \tag{5.55}$$

for some finite positive integer K, or when

$$A^K x_0 \in \mathcal{R}(\mathcal{C}_K) \tag{5.56}$$

for some K. Recall that x_1 is reachable when

$$x_1 \in \mathcal{R}(\mathcal{C}). \tag{5.57}$$

Theorem 5.31. *Consider the system $x(k+1) = Ax(k) + Bu(k)$ in (5.43).*

(i) If state x is reachable, then it is controllable.

(ii) $R_r \subset R_c$.

(iii) If the system is (completely state) reachable, or the pair (A, B) *is reachable, then the system is also (completely state) controllable, or the pair* (A, B) *is controllable.*

Furthermore, if A *is nonsingular, then relations (i) and (iii) become if and only if statements, since controllability also implies reachability, and relation (ii) becomes an equality; i.e.,* $R_c = R_r$.

Proof. (i) If x is reachable, then $x \in \mathcal{R}(\mathcal{C})$. In view of Lemma 5.19, $\mathcal{R}(\mathcal{C})$ is an A-invariant subspace and so $A^n x \in \mathcal{R}(\mathcal{C})$, which in view of (5.56), implies that x is also controllable. Since x is an arbitrary vector in R_r, this implies (ii). If $\mathcal{R}(\mathcal{C}) = R^n$, the whole state space, then $A^n x$ for any x is in $\mathcal{R}(\mathcal{C})$ and so any vector x is also controllable. Thus, reachability implies controllability. Now, if A is nonsingular, then A^{-n} exists. If x is controllable, i.e., $A^n x \in \mathcal{R}(\mathcal{C})$, then $x \in \mathcal{R}(\mathcal{C})$, i.e., x is also reachable. This can be seen by noting that A^{-n} can be written as a power series in terms of A, which in view of Lemma 5.19, implies that $A^{-n}(A^n x) = x$ is also in $\mathcal{R}(\mathcal{C})$. ■

Matrix A being nonsingular is the necessary and sufficient condition for the state transition matrix $\Phi(k, k_0)$ to be nonsingular, which in turn is the condition for *time reversibility* in discrete-time systems. Recall that reversibility in time may not be present in such systems since $\Phi(k, k_0)$ may be singular. In contrast to this, in continuous-time systems, $\Phi(t, t_0)$ is always nonsingular. This causes differences in behavior between continuous- and discrete-time systems and implies that in discrete-time systems controllability may not imply reachability (see Theorem 5.31). Note that, in view of Theorem 5.20, in the case of continuous-time systems, it is not only reachability that always implies controllability, but also vice versa, controllability always implies reachability.

When A is nonsingular, the input that will transfer the state from x_0 at $k = 0$ to $x_1 = 0$ in n steps can be determined using (5.54). In particular, one needs to solve

$$[A^{-n}\mathcal{C}]U_n = [A^{-n}B, \ldots, A^{-1}B]U_n = -x_0 \tag{5.58}$$

for $U_n = [u^T(n-1), \ldots, u^T(0)]^T$. Note that x_0 is controllable if and only if $-A^n x_0 \in \mathcal{R}(\mathcal{C})$, or if and only if $x_0 \in \mathcal{R}(A^{-n}\mathcal{C})$ for A nonsingular.

Clearly, in the case of controllability (and under the assumption that A is nonsingular), the matrix $A^{-n}\mathcal{C}$ is of interest, instead of $\mathcal{C}$ [see also (5.8)]. In particular, a system is controllable if and only if $\text{rank}(A^{-n}\mathcal{C}) = \text{rank}\,\mathcal{C} = n$.

Example 5.32. Consider the system $x(k+1) = Ax(k) + Bu(k)$, where $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Since $\text{rank}\,\mathcal{C} = \text{rank}[B, AB] = \text{rank} \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = 1 < 2 = n$, this system is not (completely) reachable (controllable-from-the-origin). All reachable states are of the form $\alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, where $\alpha \in R$ since $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$ is a basis for the $\mathcal{R}(\mathcal{C}) = R_r$, the reachability subspace.

In view of (5.56) and the Cayley–Hamilton Theorem, all controllable states x_0 satisfy $A^2 x_0 \in \mathcal{R}(\mathcal{C})$; i.e., all controllable states are of the form $\alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, where $\alpha \in R$. This verifies Theorem 5.31 for the case when A is nonsingular. Note that presently $R_r = R_c$.

Example 5.33. Consider the system $x(k+1) = Ax(k) + Bu(k)$, where $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Since $\operatorname{rank} \mathcal{C} = \operatorname{rank}[B, AB] = \operatorname{rank} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = 1 < 2 = n$, the system is not (completely) reachable. All reachable states are of the form $\alpha \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, where $\alpha \in R$ since $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$ is a basis for $\mathcal{R}(\mathcal{C}) = R_r$, the reachability subspace.

To determine the controllable subspace R_c, consider (5.56) for $K = n$, in view of the Cayley–Hamilton Theorem. Note that $A^{-1}\mathcal{C}$ cannot be used in the present case, since A is singular. Since $A^2 x_0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} x_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \mathcal{R}(\mathcal{C})$, any state x_0 will be a controllable state; i.e., the system is (completely) controllable and $R_c = R^n$. This verifies Theorem 5.31 and illustrates that controllability does not in general imply reachability.

Note that (5.54) can be used to determine the control sequence that will drive any state x_0 to the origin ($x_1 = 0$). In particular,

$$\mathcal{C} U_n = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u(1) \\ u(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = -A^2 x_0 .$$

Therefore, $u(0) = \alpha$ and $u(1) = 0$, where $\alpha \in R$ will drive any state to the origin. To verify this, we consider $x(1) = Ax(0) + Bu(0) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \alpha = \begin{bmatrix} x_{02} + \alpha \\ 0 \end{bmatrix}$ and $x(2) = Ax(1) + Bu(1) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{02} + \alpha \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} 0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

5.4 Observability and Constructibility

In applications, the state of a system is frequently required but not accessible. Under such conditions, the question arises whether it is possible to determine the state by observing the response of the system to some input over some period of time. It turns out that the answer to this question is affirmative if the system is observable. *Observability* refers to the ability of determining the present state $x(t_0)$ from knowledge of current and future system outputs, $y(t)$, and system inputs, $u(t), t \geq t_0$. *Constructibility* refers to the ability of determining the present state $x(t_0)$ from knowledge of current and past system outputs, $y(t)$, and system inputs, $u(t), t \leq t_0$. Observability was briefly addressed in Section 5.2. In this section this concept is formally defined and the (present) state is explicitly determined from input and output measurements.

5.4.1 Continuous-Time Time-Invariant Systems

We shall now study observability and constructibility for time-invariant systems described by equations of the form

$$\dot{x} = Ax + Bu, \quad y = Cx + Du, \tag{5.59}$$

where $A \in R^{n\times n}, B \in R^{n\times m}, C \in R^{p\times n}, D \in R^{p\times m}$, and $u(t) \in R^m$ is (piecewise) continuous. As was shown in Section 3.3, the output of this system is given by

$$y(t) = Ce^{At}x(0) + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t). \tag{5.60}$$

We recall that the initial time can always be taken to be $t_0 = 0$. We will find it convenient to rewrite (5.60) as

$$\tilde{y}(t) = Ce^{At}x_0, \tag{5.61}$$

where $\tilde{y}(t) \triangleq y(t) - \left[\int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t)\right]$ and $x_0 = x(0)$.

Definition 5.34. *A* state x *is* unobservable *if the zero-input response of the system (5.59) is zero for every $t \geq 0$, i.e., if*

$$Ce^{At}x = 0 \quad \textit{for every } t \geq 0. \tag{5.62}$$

The set of all unobservable states $x, R_{\bar{o}}$, is called the unobservable subspace *of (5.59).* System *(5.59) is* (completely state) observable*, or the pair (A, C) is observable, if the only state $x \in R^n$ that is unobservable is $x = 0$, i.e., if $R_{\bar{o}} = \{0\}$.* ■

Definition 5.34 states that a state is unobservable precisely when it cannot be distinguished as an initial condition at time 0 from the initial condition $x(0) = 0$. This is because in this case the output is the same as if the initial condition were the zero vector. Note that the set of all unobservable states contains the zero vector and it can be shown to be a linear subspace. We now define the observability Gramian.

Definition 5.35. *The* observability Gramian *of system (5.59) is the* $n \times n$ *matrix*

$$W_o(0,T) \triangleq \int_0^T e^{A^T\tau} C^T C e^{A\tau} d\tau. \tag{5.63}$$

■

We note that W_o is symmetric and positive semidefinite for every $T > 0$; i.e., $W_o = W_o^T$ and $W_o \geq 0$ (show this). Recall that the $pn \times n$ *observability matrix*

$$\mathcal{O} \triangleq \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{5.64}$$

was defined in Section 5.2.

We now show that the null space of $W_o(0,T)$, denoted by $\mathcal{N}(W_o(0,T))$, is independent of T; i.e., it is the same for any $T > 0$, and in particular, it is equal to the null space of the observability matrix $\mathcal{O}$. Thus, the unobservable subspace $R_{\bar{o}}$ of the system is given by the null space of $\mathcal{O}, \mathcal{N}(\mathcal{O})$, or the null space of $W_o(0,T), \mathcal{N}(W_o(0,T))$ for some finite (and therefore for all) $T > 0$.

Lemma 5.36. $\mathcal{N}(\mathcal{O}) = \mathcal{N}(W_o(0,T))$ *for every* $T > 0$.

Proof. If $x \in \mathcal{N}(\mathcal{O})$, then $\mathcal{O}x = 0$. Thus, $CA^k x = 0$ for all $0 \leq k \leq n-1$, which is also true for every $k > n-1$, in view of the Cayley–Hamilton Theorem. Then $Ce^{At}x = C[\Sigma_{k=0}^{\infty}(t^k/k!)A^i]x = 0$ for every finite t. Therefore, in view of (5.63), $W_o(0,T)x = 0$ for every $T > 0$; i.e., $x \in \mathcal{N}(W_o(0,T))$ for every $T > 0$. Now let $x \in \mathcal{N}(W_o(0,T))$ for some $T > 0$, so that $x^T W(0,T)x = \int_0^T \| Ce^{A\tau}x \|^2 \, d\tau = 0$, or $Ce^{At}x = 0$ for every $t \in [0,T]$. Taking derivatives of the last equation with respect to t and evaluating at $t = 0$, we obtain $Cx = CAx = \cdots = CA^k x = 0$ for every $k > 0$. Therefore, $CA^k x = 0$ for every $k \geq 0$, i.e., $\mathcal{O}x = 0$ or $x \in \mathcal{N}(\mathcal{O})$. ■

Theorem 5.37. *A state* x *is unobservable if and only if*

$$x \in \mathcal{N}(\mathcal{O}), \tag{5.65}$$

or equivalently, if and only if

$$x \in \mathcal{N}(W_o(0,T)) \tag{5.66}$$

for some finite (and therefore for all) $T > 0$. *Thus, the unobservable subspace* $R_{\bar{0}} = \mathcal{N}(\mathcal{O}) = \mathcal{N}(W_o(0,T))$ *for some* $T > 0$.

Proof. If x is unobservable, (5.62) is satisfied. Taking derivatives with respect to t and evaluating at $t = 0$, we obtain $Cx = CAx = \cdots = CA^k x = 0$ for $k > 0$ or $CA^k x = 0$ for every $k \geq 0$. Therefore, $\mathcal{O}x = 0$ and (5.65) is satisfied.

Assume now that $\mathcal{O}x = 0$; i.e., $CA^k x = 0$ for $0 \leq k \leq n-1$, which is also true for every $k > n-1$, in view of the Cayley–Hamilton Theorem. Then $Ce^{At}x = C[\Sigma_{k=0}^{\infty}(t^k/k!)A^i]x = 0$ for every finite t; i.e., (5.62) is satisfied and x is unobservable. Therefore, x is unobservable if and only if (5.65) is satisfied. In view of Lemma 5.36, (5.66) follows. ■

Clearly, x is observable if and only if $\mathcal{O}x \neq 0$ or $W_o(0,T)x \neq 0$ for some $T > 0$.

Corollary 5.38. *The system (5.59) is (completely state) observable, or the pair (A, C) is observable, if and only if*

$$\operatorname{rank} \mathcal{O} = n, \tag{5.67}$$

or equivalently, if and only if

$$\operatorname{rank} W_o(0,T) = n \tag{5.68}$$

for some finite (and therefore for all) $T > 0$. If the system is observable, the state x_0 at $t = 0$ is given by

$$x_0 = W_o^{-1}(0,T)\left[\int_0^T e^{A^T\tau}C^T\tilde{y}(\tau)d\tau\right]. \tag{5.69}$$

Proof. The system is observable if and only if the only vector that satisfies (5.62) or (5.65) is the zero vector. This is true if and only if the null space is empty, i.e., if and only if (5.67) or (5.68) are true. To determine the state x_0 at $t = 0$, given the output and input values over some interval $[0, T]$, we premultiply (5.61) by $e^{A^T\tau}C^T$ and integrate over $[0, T]$ to obtain

$$W_o(0,T)x_0 = \int_0^T e^{A^T\tau}C^T\tilde{y}(\tau)d\tau, \tag{5.70}$$

in view of (5.63). When the system is observable, (5.70) has the unique solution (5.69). ■

Note that $T > 0$, the time span over which the input and output are observed, is arbitrary. Intuitively, one would expect in practice to have difficulties in evaluating x_0 accurately when T is small, using any numerical method. Note that for very small T, $||W_o(0,T)||$ can be very small, which can lead to numerical difficulties in solving (5.70). Compare this with the analogous case for reachability, where small T leads in general to large values in control action.

It is clear that if the state at some time t_0 is determined, then the state $x(t)$ at any subsequent time is easily determined, given $u(t), t \geq t_0$.

Alternative methods to (5.69) to determine the state of the system when the system is observable are provided in Section 9.3 on state observers.

Example 5.39. (i) Consider the system $\dot{x} = Ax, y = Cx$, where $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $C = [1,\ 0]$. Here $e^{At} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$ and $Ce^{At} = [1,\ t]$. The observability Gramian is then $W_o(0,T) = \int_0^T \begin{bmatrix} 1 \\ \tau \end{bmatrix} [1\ \tau] d\tau = \int_0^T \begin{bmatrix} 1 & \tau \\ \tau & \tau^2 \end{bmatrix} d\tau = \begin{bmatrix} T & \frac{1}{2}T^2 \\ \frac{1}{2}T^2 & \frac{1}{3}T^3 \end{bmatrix}$. Notice that $\det W_o(0,T) = \frac{1}{12}T^4 \neq 0$ for any $T > 0$, i.e., $\operatorname{rank} W_o(0,T) = 2 = n$ for any $T > 0$, and therefore (Corollary 5.38), the system is observable. Alternatively, note that the observability matrix $\mathcal{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ and $\operatorname{rank} \mathcal{O} = 2 = n$. Clearly, in this case $\mathcal{N}(\mathcal{O}) = \mathcal{N}(W_o(0,T)) = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$, which verifies Lemma 5.36.

(ii) If $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, as before, but $C = [0,\ 1]$, in place of $[1,\ 0]$, then $Ce^{At} = [0,\ 1]$ and the observability Gramian is $W_o(0,T) = \int_0^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} [0,\ 1] d\tau = \begin{bmatrix} 0 & 0 \\ 0 & T \end{bmatrix}$. We have $\operatorname{rank} W_o(0,T) = 1 < 2 = n$, and the system is not completely observable. In view of Theorem 5.37, all unobservable states $x \in \mathcal{N}(W_o(0,T))$ and are therefore of the form $\begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \alpha \in R$. Alternatively, the observability matrix $\mathcal{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Note that $\mathcal{N}(\mathcal{O}) = \mathcal{N}(W_0(0,T)) = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$.

Observability utilizes future output measurements to determine the present state. In (re)constructibility, past output measurements are used. Constructibility is defined in the following, and its relation to observability is determined.

Definition 5.40. *A* state x *is* unconstructible *if the zero-input response of the system (5.59) is zero for all* $t \leq 0$*; i.e.,*

$$Ce^{At}x = 0 \quad \textit{for every } t \leq 0. \tag{5.71}$$

The set of all unconstructible states $x, R_{\overline{cn}}$*, is called the* unconstructible subspace *of (5.59). The system (5.59) is (completely state)* (re)constructible*, or the pair* (A, C) *is* (re)constructible*, if the only state* $x \in R^n$ *that is unconstructible is* $x = 0$*; i.e.,* $R_{\overline{cn}} = \{0\}$*.*

We shall now establish a relationship between observability and constructibility for the continuous-time time-invariant systems (5.59). Recall that x is unobservable if and only if

$$Ce^{At}x = 0 \quad \text{for every } t \geq 0. \tag{5.72}$$

Theorem 5.41. *Consider the system $\dot{x} = Ax + Bu, y = Cx + Du$ given in (5.59).*

(i) A state x is unobservable if and only if it is unconstructible.
(ii) $R_{\bar{o}} = R_{\overline{cn}}$.
(iii) The system, or the pair (A, C), is (completely state) observable if and only if it is (completely state) (re)constructible.

Proof. (i) If x is unobservable, then $Ce^{At}x = 0$ for every $t \geq 0$. Taking derivatives with respect to t and evaluating at $t = 0$, we obtain $Cx = CAx = \cdots = CA^k x = 0$ for $k > 0$ or $CA^k x = 0$ for every $k \geq 0$. This, in view of $Ce^{At}x = \sum_{k=0}^{\infty}(t^k/k!)CA^k x$, implies that $Ce^{At}x = 0$ for every $t \leq 0$; i.e., x is unconstructible. The converse is proved in a similar manner. Parts (ii) and (iii) of the theorem follow directly from (i). ■

The observability Gramian for the time-invariant case, $W_o(0, T)$, was defined in (5.63). The constructibility Gramian is now defined.

Definition 5.42. *The* constructibility Gramian *of system (5.59) is the $n \times n$ matrix*

$$W_{cn}(0, T) \triangleq \int_0^T e^{A^T(\tau - T)} C^T C e^{A(\tau - T)} d\tau. \tag{5.73}$$

■

Note that

$$W_o(0, T) = e^{A^T T} W_{cn}(0, T) e^{AT}, \tag{5.74}$$

as can be verified directly.

Additional Criteria for Observability and Constructibility

We shall now use Lemma 5.22 to develop additional tests for observability and constructibility. These are analogous to the corresponding results established for reachability and controllability in Theorem 5.23.

Theorem 5.43. *The system $\dot{x} = Ax + Bu, y = Cx + Du$ is observable*

(i) if and only if

$$\operatorname{rank} W_o(0, T) = n \tag{5.75}$$

for some finite $T > 0$, where $W_0(0, T) \triangleq \int_0^T e^{A^T\tau} C^T C e^{A\tau} d\tau$, the observability Gramian, or *(ii) if and only if the n columns of*

$$Ce^{At} \tag{5.76}$$

are linearly independent on $[0,\infty)$ over the field of complex numbers, or alternatively, if and only if the n columns of

$$C(sI-A)^{-1} \tag{5.77}$$

are linearly independent over the field of complex numbers, or

(iii) if and only if

$$\operatorname{rank}\mathcal{O}=n, \tag{5.78}$$

where $\mathcal{O}\triangleq\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1}\end{bmatrix}$*, the observability matrix, or*

(iv) if and only if

$$\operatorname{rank}\begin{bmatrix} s_iI-A \\ C\end{bmatrix}=n \tag{5.79}$$

for all complex numbers s_i, or alternatively, for all eigenvalues of A.

Proof. The proof of this theorem is completely analogous to the (dual) results on reachability (Theorem 5.23) and is omitted. ■

Since it was shown (in Theorem 5.41) that observability implies and is implied by constructibility, the tests developed in the theorem for observability are typically also used to test for constructibility.

Example 5.44. Consider the system $\dot{x}=Ax, y=Cx$, where $A=\begin{bmatrix}0 & 1\\ 0 & 0\end{bmatrix}$ and $C=[1,\ 0]$, as in Example 5.39(i). We shall verify (i) to (iv) of Theorem 5.43 for this case.

(i) For the observability Gramian, $W_o(0,T)=\begin{bmatrix} T & \frac{1}{2}T^2 \\ \frac{1}{2}T^2 & \frac{1}{3}T^3\end{bmatrix}$, we have $\operatorname{rank}W_o(0,T)=2=n$ for any $T>0$.

(ii) The columns of $Ce^{At}=[1,\ t]$ are linearly independent on $[0,\infty)$ over the field of complex numbers, since $a_1\times 1+a_2\times t=0$ implies that the complex numbers a_1 and a_2 must both be zero. Similarly, the columns of $C(sI-A)^{-1}=\left[\frac{1}{s},\frac{1}{s^2}\right]$ are linearly independent over the field of complex numbers.

(iii) $\operatorname{rank}\mathcal{O}=\operatorname{rank}\begin{bmatrix} C \\ CA\end{bmatrix}=\operatorname{rank}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}=2=n.$

(iv) $\operatorname{rank}\begin{bmatrix} s_iI-A \\ C\end{bmatrix}=\operatorname{rank}\begin{bmatrix} s_i & -1 \\ 0 & s_i \\ 1 & 0\end{bmatrix}=2=n$ for $s_i=0$, $i=1,2$, the eigenvalues of A.

Consider again $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ but $C = [0,\ 1]$ [in place of $[1,\ 0]$, as in Example 5.39(ii)].

The system is not observable for the reasons given below.

(i) $W_o(0,T) = \begin{bmatrix} 0 & 0 \\ 0 & T \end{bmatrix}$ with $\operatorname{rank} W_o(0,T) = 1 < 2 = n$.

(ii) $Ce^{At} = [0,\ 1]$ and its columns are not linearly independent. Similarly, the columns of $C(sI-A)^{-1} = [0,\ \frac{1}{s}]$ are not linearly independent.

(iii) $\operatorname{rank} \mathcal{O} = \operatorname{rank} \begin{bmatrix} C \\ CA \end{bmatrix} = \operatorname{rank} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = 1 < 2 = n$.

(iv) $\operatorname{rank} \begin{bmatrix} s_i I - A \\ C \end{bmatrix} = \operatorname{rank} \begin{bmatrix} s_i & -1 \\ 0 & s_i \\ 0 & 1 \end{bmatrix} = 1 < 2 = n$ for $s_i = 0$ an eigenvalue of A.

5.4.2 Discrete-Time Time-Invariant Systems

We consider systems described by equations of the form

$$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k), \quad k \geq k_0, \tag{5.80}$$

where $A \in R^{n\times n}, C \in R^{n\times m}, C \in R^{p\times n}, D \in R^{p\times m}$. The output $y(k)$ for $k > k_0$ is given by

$$y(k) = C(k)\Phi(k,k_0)x(k_0) + \sum_{i=k_0}^{k-1} C(k)\Phi(k,i+1)B(i)u(i) + D(k)u(k), \tag{5.81}$$

where the state transition matrix $\Phi(k,k_0)$ is given by

$$\Phi(k,k_0) = A^{k-k_0}, \quad k \geq k_0. \tag{5.82}$$

Observability and (re)constructibility for discrete-time systems are defined as in the continuous-time case. Observability refers to the ability to uniquely determine the state from knowledge of current and future outputs and inputs, whereas constructibility refers to the ability to determine the state from knowledge of current and past outputs and inputs. Without loss of generality, we take $k_0 = 0$. Then

$$y(k) = CA^k x(0) + \sum_{i=0}^{k-1} CA^{k-(i+1)}Bu(i) + Du(k) \tag{5.83}$$

for $k > 0$ and $y(0) = Cx(0) + Du(0)$. Rewrite as

$$\tilde{y}(k) = CA^k x_0 \tag{5.84}$$

for $k \geq 0$, where $\tilde{y}(k) \triangleq y(k) - \left[\sum_{i=0}^{k-1} CA^{k-(i+1)}Bu(i) + Du(k)\right]$ for $k > 0$ and $\tilde{y}(0) \triangleq y(0)$, and $x_0 = x(0)$.

Definition 5.45. *A state x is* unobservable *if the zero-input response of system (5.80) is zero for all $k \geq 0$, i.e., if*

$$CA^k x = 0 \quad \textit{for every } k \geq 0. \tag{5.85}$$

The set of all unobservable states x, $R_{\bar{o}}$, is called the unobservable subspace *of (5.80). The* system *(5.80) is* (completely state) observable*, or the pair (A, C) is observable, if the only state $x \in R^n$ that is unobservable is $x = 0$, i.e., if $R_{\bar{o}} = \{0\}$.* ■

The $pn \times n$ *observability matrix* $\mathcal{O}$ was defined in (5.64). Let $\mathcal{N}(\mathcal{O})$ denote the null space of $\mathcal{O}$.

Theorem 5.46. *A state x is unobservable if and only if*

$$x \in \mathcal{N}(\mathcal{O}); \tag{5.86}$$

i.e., the unobservable subspace $R_{\bar{o}} = \mathcal{N}(\mathcal{O})$.

Proof. If $x \in \mathcal{N}(\mathcal{O})$, then $\mathcal{O}x = 0$ or $CA^k x = 0$ for $0 \leq k \leq n-1$. This statement is also true for $k > n-1$, in view of the Cayley–Hamilton Theorem. Therefore, (5.85) is satisfied and x is unobservable. Conversely, if x is unobservable, then (5.85) is satisfied and $\mathcal{O}x = 0$. ■

Clearly, x is observable if and only if $\mathcal{O}x \neq 0$.

Corollary 5.47. *The system (5.80) is (completely state) observable, or the pair (A, C) is observable, if and only if*

$$\operatorname{rank} \mathcal{O} = n. \tag{5.87}$$

If the system is observable, the state x_0 at $k = 0$ can be determined as the unique solution of

$$[Y_{0,n-1} - M_n U_{0,n-1}] = \mathcal{O} x_0, \tag{5.88}$$

where

$$\begin{aligned} Y_{0,n-1} &\triangleq [y^T(0), y^T(1), \ldots, y^T(n-1)]^T \ \textit{ is a } pn \times 1 \textit{ matrix}, \\ U_{0,n-1} &\triangleq [u^T(0), u^T(1), \ldots, u^T(n-1)]^T \ \textit{ is an } mn \times 1 \textit{ matrix}, \end{aligned}$$

and M_n is the $pn \times mn$ matrix given by

$$M_n \triangleq \begin{bmatrix} D & 0 & \cdots & 0 & 0 \\ CB & D & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ CA^{n-2}B & CA^{n-3}B & \cdots & D & \\ CA^{n-1}B & CA^{n-2}B & \cdots & CB & D \end{bmatrix}.$$

Proof. The system is observable if and only if the only vector that satisfies (5.85) is the zero vector. This is true if and only if $\mathcal{N}(\mathcal{O}) = \{0\}$, or if (5.87) is true. To determine the state x_0, apply (5.83) for $k = 0, 1, \ldots, n-1$, and rearrange in a form of a system of linear equations to obtain (5.88). ■

The matrix M_n defined above has the special structure of a *Toeplitz* matrix. Note that a matrix T is Toeplitz if its (i, j)th entry depends on the value $i - j$; that is, T is "constant along the diagonals."

Additional Criteria for Observability. Note that completely analogous results to Theorem 5.43(ii)–(iv) exist for the discrete-time case.

Constructibility refers to the ability to determine uniquely the state $x(0)$ from knowledge of current and past outputs and inputs. This is in contrast to observability, which utilizes future outputs and inputs. The easiest way to define constructibility is by the use of (5.84), where $x(0) = x_0$ is to be determined from past data $\tilde{y}(k)$, $k \leq 0$. Note, however, that for $k \leq 0$, A^k may not exist; in fact, it exists only when A is nonsingular. To avoid making restrictive assumptions, we shall define unconstructible states in a slightly different way than anticipated. Unfortunately, this definition is not very transparent. It turns out that by using this definition, an unconstructible state can be related to an unobservable state in a manner analogous to the way a controllable state was related to a reachable state in Section 5.3 (see also the discussion of duality in Section 5.2).

Definition 5.48. *A state x is* unconstructible *if for every $k \geq 0$, there exists $\hat{x} \in R^n$ such that*

$$x = A^k \hat{x}, \quad C\hat{x} = 0. \tag{5.89}$$

The set of all unconstructible states, $R_{\overline{cn}}$, is called the unconstructible subspace*. The system (5.80) is* (completely state) constructible*, or the pair (A, C) is constructible, if the only state $x \in R^n$ that is unconstructible is $x = 0$, i.e., if $R_{\overline{cn}} = \{0\}$.* ■

Note that if A is nonsingular, then (5.89) simply states that x is unconstructible if $CA^{-k}x = 0$ for every $k \geq 0$ (compare this with Definition 5.45 of an unobservable state).

The results that can be derived for constructibility are simply dual to the results on controllability. They are presented briefly below, but first, a technical result must be established.

Lemma 5.49. *If $x \in \mathcal{N}(\mathcal{O})$, then $Ax \in \mathcal{N}(\mathcal{O})$; i.e., the unobservable subspace $R_{\bar{o}} = \mathcal{N}(\mathcal{O})$ is an* A-invariant subspace*.*

Proof. Let $x \in \mathcal{N}(\mathcal{O})$, so that $\mathcal{O}x = 0$. Then $CA^k x = 0$ for $0 \leq k \leq n-1$. This statement is also true for $k > n-1$, in view of the Cayley–Hamilton Theorem. Therefore, $\mathcal{O}Ax = 0$; i.e., $Ax \in \mathcal{N}(\mathcal{O})$. ■

Theorem 5.50. *Consider the system $x(k+1) = Ax(k) + Bu(k)$, $y(k) = Cx(k) + Du(k)$ given in (5.80).*

(i) If a state x is unconstructible, then it is unobservable.
(ii) $R_{\overline{cn}} \subset R_{\bar{o}}$.
(iii) If the system is (completely state) observable, or the pair (A, C) is observable, then the system is also (completely state) constructible, or the pair (A, C) is constructible.

If A is nonsingular, then relations (i) and (iii) are if and only if statements. In this case, constructibility also implies observability. Furthermore, in this case, (ii) becomes an equality; i.e., $R_{\overline{cn}} = R_{\bar{o}}$.

Proof. This theorem is dual to Theorem 5.31, which relates reachability and controllability in the discrete-time case. To verify (i), assume that x satisfies (5.89) and premultiply by C to obtain $Cx = CA^k\hat{x}$ for every $k \geq 0$. Note that $Cx = 0$ since for $k = 0$, $x = \hat{x}$, and $C\hat{x} = 0$. Therefore, $CA^k\hat{x} = 0$ for every $k \geq 0$; i.e., $\hat{x} \in \mathcal{N}(\mathcal{O})$. In view of Lemma 5.49, $x = A^k\hat{x} \in \mathcal{N}(\mathcal{O})$, and thus, x is unobservable. Since x is arbitrary, we have also verified (ii). When the system is observable, $R_{\bar{o}}$ is empty, which in view of (ii), implies that $R_{\bar{c}n} = \{0\}$ or that the system is constructible. This proves (iii). Alternatively, one could also prove this directly: Assume that the system is observable but not constructible. Then there exist $x, \hat{x} \neq 0$, which satisfy (5.89). As above, this implies that $\hat{x} \in \mathcal{N}(\mathcal{O})$, which is a contradiction since the system is observable.

Consider now the case when A is nonsingular and let x be unobservable. Then, in view of Lemma 5.49, $\hat{x} \triangleq A^{-k}x$ is also in $\mathcal{N}(\mathcal{O})$; i.e., $C\hat{x} = 0$. Therefore, $x = A^k\hat{x}$ is unconstructible, in view of Definition 5.48. This implies also that $R_{\bar{o}} \subset R_{\overline{cn}}$, and therefore, $R_{\bar{o}} = R_{\overline{cn}}$, which proves that in the present case constructibility also implies observability. ■

Example 5.51. Consider the system in Example 5.5, $x(k+1) = Ax(k), y(k) = Cx(k)$, where $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and $C = [1,\ 0]$. As shown, $\operatorname{rank} \mathcal{O} = \operatorname{rank} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = 1 < 2 = n$; i.e., the system is not observable. All unobservable states are of the form $\alpha \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, where $\alpha \in R$ since $\left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ is a basis for $\mathcal{N}(\mathcal{O}) = R_{\bar{o}}$, the unobservable subspace.

In Example 5.6 it was shown that all the states x that satisfy $CA^{-k}x = 0$ for every $k \geq 0$, i.e., all the unconstructible states, are given by $\alpha \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \alpha \in R$. This verifies Theorem 5.50(i) and (ii) for the case when A is nonsingular.

Example 5.52. Consider the system $x(k+1) = Ax(k)$, $y(k) = Cx(k)$, where $A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$ and $C = [1,\ 0]$. The observability matrix $\mathcal{O} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ is of rank 1, and therefore, the system is not observable. In fact, all states of the form $\alpha \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ are unobservable states since $\left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ is a basis for $\mathcal{N}(\mathcal{O})$.

To check constructibility, the defining relations (5.89) must be used since A is singular. $C\hat{x} = [1,\ 0]\hat{x} = 0$ implies $\hat{x} = \begin{bmatrix} 0 \\ \beta \end{bmatrix}$. Substituting into $x = A^k \hat{x}$, we obtain for $k = 0$, $x = \hat{x}$, and $x = 0$ for $k \geq 1$. Therefore, the only unconstructible state is $x = 0$, which implies that the system is constructible (although it is unobservable). This means that the initial state $x(0)$ can be uniquely determined from past measurements. In fact, from $x(k+1) = Ax(k)$ and $y(k) = Cx(k)$, we obtain $x(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(-1) \\ x_2(-1) \end{bmatrix} = \begin{bmatrix} 0 \\ x_1(-1) \end{bmatrix}$ and $y(-1) = Cx(-1) = [1,\ 0] \begin{bmatrix} x_1(-1) \\ x_2(-1) \end{bmatrix} = x_1(-1)$. Therefore, $x(0) = \begin{bmatrix} 0 \\ y(-1) \end{bmatrix}$.

When A is nonsingular, the state x_0 at $k = 0$ can be determined from past outputs and inputs in the following manner. We consider (5.84) and note that in this case

$$\tilde{y}(k) = CA^k x_0$$

is valid for $k \leq 0$ as well. This implies that

$$\widetilde{Y}_{-1,-n} = \mathcal{O}A^{-n}x_0 = \begin{bmatrix} CA^{-n} \\ \vdots \\ CA^{-1} \end{bmatrix} x_0 \tag{5.90}$$

with $\widetilde{Y}_{-1,-n} \triangleq [\tilde{y}^T(-n), \ldots, \tilde{y}^T(-1)]^T$. Equation (5.90) must be solved for x_0. Clearly, in the case of constructibility (and under the assumption that A is nonsingular), the matrix $\mathcal{O}A^{-n}$ is of interest instead of $\mathcal{O}$ [compare this with the dual results in (5.58)]. In particular, the system is constructible if and only if $\operatorname{rank}(\mathcal{O}A^{-n}) = \operatorname{rank} \mathcal{O} = n$.

Example 5.53. Consider the system in Example 5.4, namely, $x(k+1) = Ax(k), y(k) = Cx(k)$, where $A = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ and $C = [0,\ 1]$. Since A is nonsingular, to check constructibility we consider $\mathcal{O}A^{-2} = \left[\begin{smallmatrix} CA^{-2} \\ CA^{-1} \end{smallmatrix}\right] = \left[\begin{smallmatrix} -1 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, which has full rank. Therefore, the system is constructible (as expected), since it is observable. To determine $x(0)$, in view of (5.90), we note that $\begin{bmatrix} y(-1) \\ y(-2) \end{bmatrix} = \mathcal{O}A^{-2}x(0) = \begin{bmatrix} -1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix}$, from which $\begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} y(-1) \\ y(-2) \end{bmatrix} = \begin{bmatrix} y(-2) \\ y(-1) + y(-2) \end{bmatrix}$.

5.5 Summary and Highlights

Reachability and Controllability

- In continuous-time systems, reachability always implies and is implied by controllability. In discrete-time systems, reachability always implies controllability, but controllability implies reachability only when A is nonsingular. See Definitions 5.8 and 5.18 and Theorems 5.20 and 5.31.
- When a discrete-time system $x(k+1) = Ax(k)+Bu(k)$ [denoted by (A, B)] is completely reachable (controllable-from-the-origin), the input sequence $\{u(i)\}$, $i = 0, \ldots, K-1$ that transfers the state from any $x_0(= x(0))$ to any x_1 in some finite time K $(x_1 = x(K),\ K > 0)$ is determined by solving

$$x_1 = A^K x_0 + \sum_{i=0}^{K-1} A^{K-(i+1)} Bu(i) \quad \text{or}$$

$$x_1 - A^K x_0 = [B, AB, \ldots, A^{K-1}]\ [u^T(K-1), \ldots, u^T(0)]^T.$$

  A solution for this always exists when $K = n$. See Theorem 5.27.

- $$\mathcal{C} = [B, AB, \ldots, A^{n-1}B]\ (n \times mn) \tag{5.25}$$

  is the controllability matrix for both discrete- and continuous-time time-invariant systems, and it has full (row) rank when the system, denoted by (A, B), is (completely) reachable (controllable-from-the-origin).

- When a continuous-time system $\dot{x} = Ax + Bu$ [denoted by (A, B)] is controllable, an input that transfers any state $x_0(= x(0))$ to any other state x_1 in some finite time T $(x_1 = x(T)$ is

$$u(t) = B^T e^{A^T(T-t)} W_r^{-1}(0, T)[x_1 - e^{AT} x_0] \quad t \in [0, T], \tag{5.33}$$

  where

$$W_r(0, T) = \int_0^T e^{(T-\tau)A} BB^T e^{(T-\tau)A^T} d\tau \tag{5.24}$$

  is the reachability Gramian of the system.

- (A, B) is reachable if and only if

$$\text{rank}[s_i I - A, B] = n \tag{5.42}$$

  for s_i, $i = 1, \ldots, n$, all the eigenvalues of A.

Observability and Constructibility

- In continuous-time systems, observability always implies and is implied by constructibility. In discrete-time systems, observability always implies constructibility, but constructibility implies observability only when A is nonsingular. See Definitions 5.34 and 5.40 and Theorems 5.41 and 5.50.

- When a discrete-time system $x(k+1) = Ax(k) + Bu(k)$, $y(k) = Cx(k) + Du(k)$ [denoted by (A, C)] is completely observable, any initial state $x(0) = x_0$ can be uniquely determined by observing the input and output over some finite period of time, and using the relation
$$\tilde{y}(k) = CA^k x_0 \quad k = 0, 1, \ldots, n-1, \tag{5.84}$$
where $\tilde{y}(k) = y(k) - \left[\sum_{i=0}^{k-1} CA^{k-(i+1)} Bu(i) + D(k)u(k)\right]$. To determine x_0, solve
$$\begin{bmatrix} \tilde{y}(0) \\ \tilde{y}(1) \\ \vdots \\ \tilde{y}(n-1) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} x_0 .$$
See (5.88).
- $$\mathcal{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \quad (pn \times n) \tag{5.64}$$
is the observability matrix for both discrete- and continuous-time, time-invariant systems and it has full (column) rank when the system is completely observable.
- Consider the continuous-time system $\dot{x} = Ax + Bu$, $y = Cx + Du$. When this system [denoted by (A, C)] is completely observable, any initial state $x_0 = x(0)$ can be uniquely determined by observing the input and output over some finite period of time T and using the relation
$$\tilde{y}(t) = Ce^{At} x_0,$$
where $\tilde{y}(t) = y(t) - \left[\int_0^t Ce^{A(t-\tau)} Bu(\tau) d\tau + Du(t)\right]$. The initial state x_0 may be determined from
$$x_0 = W_o^{-1}(0, T) \left[\int_0^T e^{A^T \tau} C^T \tilde{y}(\tau) d\tau\right], \tag{5.69}$$
where
$$W_o(0, T) = \int_0^T e^{A^T \tau} C^T C e^{A\tau} d\tau \tag{5.63}$$
is the observability Gramian of the system.
- (A, C) is observable if and only if
$$\operatorname{rank} \begin{bmatrix} s_i I - A \\ C \end{bmatrix} = n \tag{5.79}$$
for s_i, $i = 1, \ldots, n$, all the eigenvalues of A.

Dual Systems

- $(A_D = A^T, B_D = C^T, C_D = B^T, D_D = D^T)$ is the dual of (A, B, C, D). Reachability is dual to observability. If a system is reachable (observable), its dual is observable (reachable).

5.6 Notes

The concept of controllability was first encountered as a technical condition in certain optimal control problems and also in the so-called finite-settling-time design problem for discrete-time systems (see Kalman [4]). In the latter, an input must be determined that returns the state x_0 to the origin as quickly as possible. Manipulating the input to assign particular values to the initial state in (analog-computer) simulations was not an issue since the individual capacitors could initially be charged independently. Also, observability was not an issue in simulations due to the particular system structures that were used (corresponding, e.g., to observer forms). The current definitions for controllability and observability and the recognition of the duality between them were worked out by Kalman in 1959–1960 (see Kalman [7] for historical comments) and were presented by Kalman in [5]. The significance of realizations that were both controllable and observable (see Chapter 5) was established later in Gilbert [2], Kalman [6], and Popov [8]. For further information regarding these historical issues, consult Kailath [3] and the original sources. Note that [3] has extensive references up to the late seventies with emphasis on the time-invariant case and a rather complete set of original references together with historical remarks for the period when the foundations of the state-space system theory were set, in the late fifties and sixties.

References

1. P.J. Antsaklis and A.N. Michel, *Linear Systems*, Birkhäuser, Boston, MA, 2006.
2. E. Gilbert, “Controllability and observability in multivariable control systems,” *SIAM J. Control*, Vol. 1, pp. 128–151, 1963.
3. T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
4. R.E. Kalman, “Optimal nonlinear control of saturating systems by intermittent control,” IRE WESCON Rec., Sec. IV, pp. 130–135, 1957.
5. R.E. Kalman, “On the general theory of control systems,” in *Proc. of the First Intern. Congress on Automatic Control*, pp. 481–493, Butterworth, London, 1960.
6. R.E. Kalman, “Mathematical descriptions of linear systems,” *SIAM J. Control*, Vol. 1, pp. 152–192, 1963.
7. R.E. Kalman, *Lectures on Controllability and Observability*, C.I.M.E., Bologna, 1968.
8. V.M. Popov, “On a new problem of stability for control systems,” *Autom. Remote Control*, Vol. 24, No. 1, pp. 1–23, 1963.

Exercises

5.1. (a) Let $\mathcal{C}_k \triangleq [B, AB, \ldots, A^{k-1}B]$, where $A \in R^{n\times n}, B \in R^{n\times m}$. Show that

$$\mathcal{R}\ (\mathcal{C}_k) = \mathcal{R}\ (\mathcal{C}_n) \text{ for } k \geq n, \quad \text{and} \quad \mathcal{R}(\mathcal{C}_k) \subset \mathcal{R}(\mathcal{C}_n) \text{ for } k < n.$$

(b) Let $\mathcal{O}_k \triangleq [C^T, (CA)^T, \ldots, (CA^{k-1})^T]^T$, where $A \in R^{n\times n}, C \in R^{p\times n}$. Show that

$$\mathcal{N}(\mathcal{O}_k) = \mathcal{N}(\mathcal{O}_n) \text{ for } k \geq n, \text{ and } \mathcal{N}(\mathcal{O}_k) \supset \mathcal{N}(\mathcal{O}_n) \text{ for } k < n.$$

5.2. Consider the state equation $\dot{x} = Ax + Bu$, where

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3w^2 & 0 & 0 & 2w \\ 0 & 0 & 0 & 1 \\ 0 & -2w & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix},$$

which was obtained by linearizing the nonlinear equations of motion of an orbiting satellite about a steady-state solution. In the state $x = [x_1, x_2, x_3, x_4]^T$, x_1 is the differential radius, whereas x_3 is the differential angle. In the input vector $u = [u_1, u_2]^T$, u_1 is the radial thrust and u_2 is the tangential thrust.

(a) Is this system controllable from u? If $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_3 \end{bmatrix}$, is the system observable from y?

(b) Can the system be controlled if the radial thruster fails? What if the tangential thruster fails?

(c) Is the system observable from y_1 only? From y_2 only?

5.3. Consider the state equation $\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1/2 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1/2 \\ 1 \end{bmatrix} u.$

(a) If $x(0) = \begin{bmatrix} a \\ b \end{bmatrix}$, derive an input that will drive the state to $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ in T sec.

(b) For $x(0) = \begin{bmatrix} 5 \\ -5 \end{bmatrix}$, plot $u(t), x_1(t), x_2(t)$ for $T = 1, 2,$ and 5 sec. Comment on the magnitude of the input in your results.

5.4. Consider the state equation $x(k+1) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} u(k), y(k) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} x(k).$

(a) Is $x^1 = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}$ reachable? If yes, what is the minimum number of steps required to transfer the state from the zero state to x^1? What inputs do you need?
(b) Determine all states that are reachable.
(c) Determine all states that are unobservable.
(d) If $\dot{x} = Ax + Bu$ is given with A, B as in (a), what is the minimum time required to transfer the state from the zero state to x^1? What is an appropriate $u(t)$?

5.5. *Output reachability (controllability)* can be defined in a manner analogous to state reachability (controllability). In particular, a system will be called output reachable if there exists an input that transfers the output from some y_0 to any y_1 in finite time.

Consider now a discrete-time time-invariant system $x(k+1) = Ax(k) + Bu(k), y(k) = Cx(k) + Du(k)$ with $A \in R^{n\times n}, B \in R^{n\times m}, C \in R^{p\times n}$ and $D \in R^{p\times m}$. Recall that

$$y(k) = CA^k x(0) + \sum_{i=0}^{k-1} CA^{k-(i+1)} Bu(i) + Du(k).$$

(a) Show that the system $\{A, B, C, D\}$ is output reachable if and only if

$$\operatorname{rank}[D, CB, CAB, \dots, CA^{n-1}B] = p.$$

Note that this rank condition is also the condition for output reachability for continuous-time time-invariant systems $\dot{x} = Ax + Bu, y = Cx + Du$. It should be noted that, in general, state reachability is neither necessary nor sufficient for output reachability. Notice for example that if $\operatorname{rank} D = p$, then the system is output reachable.
(b) Let $D = 0$. Show that if (A, B) is (state) reachable, then $\{A, B, C, D\}$ is output reachable if and only if $\operatorname{rank} C = p$.
(c) Let $A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & -1 \end{bmatrix}$, $B = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$, $C = [1,\ 1,\ 0]$, and $D = 0$.
  (i) Is the system output reachable? Is it state reachable?
  (ii) Let $x(0) = 0$. Determine an appropriate input sequence to transfer the output to $y_1 = 3$ in minimum time. Repeat for $x(0) = [1, -1, 2]^T$.

5.6. (a) Given $\dot{x} = Ax + Bu, y = Cx + Du$, show that this system is output reachable if and only if the rows of the $p \times m$ transfer matrix $H(s)$ are linearly independent over the field of complex numbers. In view of this result, is the system $H(s) = \begin{bmatrix} \frac{1}{s+2} \\ \frac{s}{s+1} \end{bmatrix}$ output reachable?

(b) Similarly, for discrete-time systems, the system is output reachable if and only if the rows of the transfer function matrix $H(z)$ are linearly independent over the field of complex numbers. Consider now the system of Exercise 5.5 and determine whether it is output reachable.

5.7. Show that the circuit depicted in Figure 5.4 with input u and output y is neither state reachable nor observable but is output reachable.

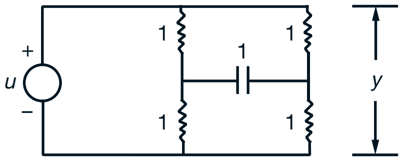


Figure 5.4. Circuit for Exercise 5.7

5.8. A system $\dot{x} = Ax + Bu, y = Cx + Du$ is called *output function controllable* if there exists an input $u(t)$, $t \in [0, \infty)$, that will cause the output $y(t)$ to follow a prescribed trajectory for $0 \leq t < \infty$, assuming that the system is at rest at $t = 0$. It is easiest to derive a test for output function controllability in terms of the $p \times m$ transfer function matrix $H(s)$, and this is the approach taken in the following. We say that the $m \times p$ rational matrix $H_R(s)$ is a *right inverse of* $H(s)$ if

$$H(s)H_R(s) = I_p.$$

(a) Show that the right inverse $H_R(s)$ exists if and only if $\operatorname{rank} H(s) = p$. *Hint:* In the sufficiency proof, select $H_R = H^T(HH^T)^{-1}$, the (right) pseudoinverse of H.

(b) Show that the system is output function controllable if and only if $H(s)$ has a right inverse $H_R(s)$. *Hint:* Consider $\hat{y} = H\hat{u}$. In the necessity proof, show that if $\operatorname{rank} H < p$, then the system may not be output function controllable.

Input function observability is the dual to output function controllablity. Here, the *left inverse of* $H(s)$, $H_L(s)$, is of interest and is defined by

$$H_L(s)H(s) = I_m.$$

(c) Show that the left inverse $H_L(s)$ of $H(s)$ exists if and only if $\operatorname{rank} H(s) = m$. *Hint:* This is the dual result to part (a).

(d) Let $H(s) = \left[\frac{s+1}{s}, \frac{1}{s}\right]$ and characterize all inputs $u(t)$ that will cause the system (at rest at $t = 0$) to exactly follow a step, $\hat{y}(s) = 1/s$.

Part (d) points to a variety of questions that may arise when inverses are considered, including: Is $H_R(s)$ proper? Is it unique? Is it stable? What is the minimum degree possible?

5.9. Consider the system $\dot{x} = Ax + Bu, y = Cx$. Show that output function controllability implies output controllability (-from-the-origin, or reachability).

5.10. Given $x(k+1) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} x(k) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(k)$, $y(k) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} x(k)$, and assume zero initial conditions.

(a) Is there a sequence of inputs $\{u(0), u(1), \ldots\}$ that transfers the output from $y(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ to $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ in finite time? If the answer is yes, determine such a sequence.

(b) Characterize all outputs that can be reached from the zero output ($y(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$), in one step.

5.11. Suppose that for system $x(k+1) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(k)$, $y(k) = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} x(k)$, it is known that $y(0) = y(1) = y(2) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Based on this information, what can be said about the initial condition $x(0)$?

5.12. (a) Consider the system $\dot{x} = Ax + Bu, y = Cx + Du$, where (A, C) is assumed to be observable. Express $x(t)$ as a function of $y(t), u(t)$ and their derivatives. *Hint:* Write $y(t), y^{(1)}(t), \ldots, y^{(n-1)}(t)$ in terms of $x(t)$ and $u(t), u^{(1)}(t), \ldots, u^{(n-1)}(t)$ ($x(t) \in R^n$).

(b) Given the system $\dot{x} = Ax + Bu, y = Cx + Du$ with (A, C) observable. Determine $x(0)$ in terms of $y(t), u(t)$ and their derivatives up to order $n - 1$. Note that in general this is not a practical way of determining $x(0)$, since this method requires differentiation of signals, which is very susceptible to measurement noise.

(c) Consider the system $x(k+1) = Ax(k) + Bu(k), y(k) = Cx(k) + Du(k)$, where (A, C) is observable. Express $x(k)$ as a function of $y(k), y(k+1), \ldots, y(k+n-1)$ and $u(k), u(k+1), \ldots, y(k+n-1)$. *Hint:* Express $y(k), \ldots, y(k+n-1)$ in terms of $x(k)$ and $u(k), u(k+1), \ldots, u(k+n-1)$ [$x(k) \in R^n$]. Note the relation to expression (5.88) in Section 5.4.

6

Controllability and Observability: Special Forms

6.1 Introduction

In this chapter, important special forms for the state-space description of time-invariant systems are presented. These forms are obtained by means of similarity transformations and are designed to reveal those features of a system that are related to the properties of controllability and observability. In Section 6.2, special state-space forms that separate the controllable (observable) from the uncontrollable (unobservable) part of a system are presented. These forms, referred to as the standard forms for uncontrollable and unobservable systems, are very useful in establishing a number of results. In particular, these forms are used in Section 6.3 to derive alternative tests for controllability and observability and in Section 7.2 to relate state-space and input–output descriptions. In Section 6.4 the controller and observer state-space forms are introduced. These are useful in the study of state-space realizations in Chapter 8 and state feedback and state estimators in Chapter 9.

6.2 Standard Forms for Uncontrollable and Unobservable Systems

We consider time-invariant systems described by equations of the form

$$\dot{x} = Ax + Bu, \quad y = Cx + Du, \tag{6.1}$$

where $A \in R^{n\times n}$, $B \in R^{n\times m}$, $C \in R^{p\times n}$, and $D \in R^{p\times m}$. It was shown in the previous chapter that this system is state reachable if and only if the $n \times mn$ controllability matrix

$$\mathcal{C} \triangleq [B, AB, \ldots, A^{n-1}B] \tag{6.2}$$

has full row rank n; i.e., $\operatorname{rank}\mathcal{C} = n$. If the system is reachable (or controllable-from-the-origin), then it is also controllable (or controllable-to-the-origin), and vice versa (see Section 5.3.1).

It was also shown earlier that system (6.1) is state observable if and only if the $pn \times n$ observability matrix

$$\mathcal{O} \triangleq \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{6.3}$$

has full column rank; i.e., rank $\mathcal{O} = n$. If the system is observable, then it is also constructible, and vice versa (see Section 5.4.1).

Similar results were also derived for discrete-time time-invariant systems described by equations of the form

$$x(k+1) = Ax(k) + Bu(k), \quad y(k) = Cx(k) + Du(k). \tag{6.4}$$

Again, rank $\mathcal{C} = n$ and rank $\mathcal{O} = n$ are the necessary and sufficient conditions for state reachability and observability, respectively. Reachability always implies controllability and observability always implies constructibility, as in the continuous-time case. However, in the discrete-time case, controllability does not necessarily imply reachability and constructibility does not imply observability, unless A is nonsingular (see Sections 5.3.2 and 5.4.2).

Next, we will introduce standard forms for unreachable and unobservable systems both for the continuous-time and the discrete-time time-invariant cases. These forms will be referred to as standard forms for *uncontrollable systems*, rather than unreachable systems, and standard forms for *unobservable systems*, respectively.

6.2.1 Standard Form for Uncontrollable Systems

If the system (6.1) [or (6.4)] is not completely reachable or controllable-from-the-origin, then it is possible to "separate" the controllable part of the system by means of an appropriate similarity transformation. This amounts to changing the basis of the state space so that all the vectors in the reachable subspace R_r have a certain structure. In particular, let rank $\mathcal{C} = n_r < n$; i.e., the pair (A, B) is not controllable. This implies that the subspace $R_r = \mathcal{R}(\mathcal{C})$ has dimension n_r. Let $\{v_1, v_2, \ldots, v_{n_r}\}$ be a basis for R_r. These n_r vectors can be, for example, any n_r linearly independent columns of $\mathcal{C}$. Define the $n \times n$ similarity transformation matrix

$$Q \triangleq [v_1, v_2, \ldots, v_{n_r}, Q_{n-n_r}], \tag{6.5}$$

where the $n \times (n - n_r)$ matrix Q_{n-n_r} contains $n - n_r$ linearly independent vectors chosen so that Q is nonsingular. There are many such choices. We are now in a position to prove the following result.

Lemma 6.1. *For (A, B) uncontrollable, there exists a nonsingular matrix Q such that*

$$\widehat{A} = Q^{-1}AQ = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix} \quad \textit{and} \quad \widehat{B} = Q^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{6.6}$$

where $A_1 \in R^{n_r \times n_r}, B_1 \in R^{n_r \times m}$, and the pair (A_1, B_1) is controllable. The pair $(\widehat{A}, \widehat{B})$ is in the standard form for uncontrollable systems.

Proof. We need to show that

$$AQ = A[v_1, \ldots, v_{n_r}, Q_{n-n_r}] = [v_1, \ldots, v_{n_r}, Q_{n-n_r}] \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix} = Q\widehat{A}.$$

Since the subspace R_r is A-invariant (see Lemma 5.19), $Av_i \in R_r$, which can be written as a linear combination of only the n_r vectors in a basis of R_r. Thus, A_1 in $\widehat{A}$ is an $n_r \times n_r$ matrix, and the $(n - n_r) \times n_r$ matrix below it in $\widehat{A}$ is a zero matrix. Similarly, we also need to show that

$$B = [v_1, \ldots, v_{n_r}, Q_{n-n_r}] \begin{bmatrix} B_1 \\ 0 \end{bmatrix} = Q\widehat{B}.$$

But this is true for similar reasons: The columns of B are in the range of $\mathcal{C}$ or in R_r. ■

The $n \times nm$ controllability matrix $\widehat{\mathcal{C}}$ of $(\widehat{A}, \widehat{B})$ is

$$\widehat{\mathcal{C}} = [\widehat{B}, \widehat{A}\widehat{B}, \ldots, \widehat{A}^{n-1}\widehat{B}] = \begin{bmatrix} B_1 & A_1B_1 & \cdots & A_1^{n-1}B_1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}, \tag{6.7}$$

which clearly has $\operatorname{rank} \widehat{\mathcal{C}} = \operatorname{rank}[B_1, A_1B_1, \ldots, A_1^{n_r-1}B_1, \ldots, A_1^{n-1}B_1] = n_r$. Note that

$$\widehat{\mathcal{C}} = Q^{-1}\mathcal{C}. \tag{6.8}$$

The range of $\widehat{\mathcal{C}}$ is the controllable subspace of $(\widehat{A}, \widehat{B})$. It contains vectors only of the form $[\alpha^T, 0]^T$, where $\alpha \in R^{n_r}$. Since $\dim \mathcal{R}(\widehat{\mathcal{C}}) = \operatorname{rank} \widehat{\mathcal{C}} = n_r$, every vector of the form $[\alpha^T, 0]^T$ is a controllable (state) vector. In other words, the similarity transformation has changed the basis of R^n in such a manner so that all controllable vectors, expressed in terms of this new basis, have this very particular structure with zeros in the last $n - n_r$ entries.

Given system (6.1) [or (6.4)], if a new state $\hat{x}(t)$ is taken to be $\hat{x}(t) = Q^{-1}x(t)$, then

$$\dot{\hat{x}} = \widehat{A}\hat{x} + \widehat{B}u, \quad y = \widehat{C}\hat{x} + \widehat{D}u, \tag{6.9}$$

where $\widehat{A} = Q^{-1}AQ$, $\widehat{B} = Q^{-1}B$, $\widehat{C} = CQ$, and $\widehat{D} = D$ constitutes an equivalent representation (see Section 3.4.3). For Q as in Lemma 6.1, we obtain

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \end{bmatrix} = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} u, y = [C_1, C_2] \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + Du, \tag{6.10}$$

where $\hat{x} = [\hat{x}_1^T, \hat{x}_2^T]^T$ with $\hat{x}_1 \in R^{n_r}$ and where (A_1, B_1) is controllable. The matrix $\widehat{C} = [C_1, C_2]$ does not have any particular structure. This representation is called a *standard form for the uncontrollable system.* The state equation can now be written as

$$\dot{\hat{x}}_1 = A_1\hat{x}_1 + B_1 u + A_{12}\hat{x}_2, \dot{\hat{x}}_2 = A_2\hat{x}_2, \tag{6.11}$$

which shows that the input u does not affect the trajectory component $\hat{x}_2(t)$ at all, and therefore, $\hat{x}_2(t)$ is determined only by the value of its initial vector. The input u certainly affects $\hat{x}_1(t)$. Note also that the trajectory component $\hat{x}_1(t)$ is also influenced by $\hat{x}_2(t)$. In fact,

$$\hat{x}_1(t) = e^{A_1 t}\hat{x}_1(0) + \int_0^t e^{A_1(t-\tau)} B_1 u(\tau) d\tau + \left[\int_0^t e^{A_1(t-\tau)} A_{12} e^{A_2 \tau} d\tau\right] \hat{x}_2(0). \tag{6.12}$$

The n_r eigenvalues of A_1 and the corresponding modes are the *controllable eigenvalues* and *controllable modes* of the pair (A, B) or of system (6.1) [or of (6.4)]. The $n - n_r$ eigenvalues of A_2 and the corresponding modes are the *uncontrollable eigenvalues* and *uncontrollable modes*, respectively.

It is interesting to observe that in the zero-state response of the system (zero initial conditions), the uncontrollable modes are completely absent. In particular, in the solution $x(t) = e^{At}x(0) + \int_0^t e^{A(t-\tau)} Bu(\tau) d\tau$ of $\dot{x} = Ax + Bu$, given $x(0)$, notice that

$$e^{A(t-\tau)}B = [Qe^{\widehat{A}(t-\tau)}Q^{-1}][Q\widehat{B}] = Q \begin{bmatrix} e^{A_1(t-\tau)}B_1 \\ 0 \end{bmatrix},$$

where A_1 [from (6.6)] contains only the controllable eigenvalues. Therefore, the input $u(t)$ cannot directly influence the uncontrollable modes. Note, however, that the uncontrollable modes do appear in the zero-input response $e^{At}x(0)$. The same observations can be made for discrete-time systems (6.4) where the quantity $A^k B$ is of interest.

Example 6.2. Given $A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$, we wish to reduce system (6.1) to the standard form (6.6). Here

$$\mathcal{C} = [B, AB, A^2B] = \left[\begin{array}{cc|cc|cc} 1 & 0 & 0 & 1 & 0 & -1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & -1 & 0 & 1 \end{array}\right]$$

and $\operatorname{rank} \mathcal{C} = n_r = 2 < 3 = n$. Thus, the subspace $R_r = \mathcal{R}(\mathcal{C})$ has dimension $n_r = 2$, and a basis $\{v_1, v_2\}$ can be found by taking two linearly independent columns of $\mathcal{C}$, say, the first two, to obtain

$$Q = [v_1, v_2, Q_1] = \left[\begin{array}{cc|c} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{array}\right].$$

The third column of Q was selected so that Q is nonsingular. Note that the first two columns of Q could have been the first and fourth columns of $\mathcal{C}$ instead, or any other two linearly independent vectors obtained as a linear combination of the columns in $\mathcal{C}$. For the above choice for Q, we have

$$\begin{aligned}
\widehat{A} = Q^{-1}AQ &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 0 & -1 & 1 \\ 1 & -1 & 0 \\ -2 & 4 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \\
&= \left[\begin{array}{cc|c} 0 & 1 & 1 \\ 0 & -1 & 0 \\ \hline 0 & 0 & -2 \end{array}\right] = \left[\begin{array}{c|c} A_1 & A_{12} \\ \hline 0 & A_2 \end{array}\right], \\
\widehat{B} = Q^{-1}B &= \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ \hline 0 & 0 \end{array}\right] = \left[\begin{array}{c} B_1 \\ \hline 0 \end{array}\right],
\end{aligned}$$

where (A_1, B_1) is controllable. The matrix A has three eigenvalues at 0, -1, and -2. It is clear from $(\widehat{A}, \widehat{B})$ that the eigenvalues $0, -1$ are controllable (in A_1), whereas -2 is an uncontrollable eigenvalue (in A_2).

6.2.2 Standard Form for Unobservable Systems

The standard form for an unobservable system can be derived in a similar way as the standard form of uncontrollable systems. If the system (6.1) [or (6.4)] is not completely state observable, then it is possible to "separate" the unobservable part of the system by means of a similarity transformation. This amounts to changing the basis of the state space so that all the vectors in the unobservable subspace $R_{\bar{o}}$ have a certain structure.

As in the preceding discussion concerning systems or pairs (A, B) that are not completely controllable, we shall select a similarity transformation Q to reduce a pair (A, C), which is not completely observable, to a particular form. This can be accomplished in two ways. The simplest way is to invoke duality and to work with the pair $(A_D = A^T, B_D = C^T)$, which is not controllable (refer to the discussion of dual systems in Section 5.2.3). If Lemma 6.1 is applied, then

$$\widehat{A}_D = Q_D^{-1} A_D Q_D = \begin{bmatrix} A_{D1} & A_{D12} \\ 0 & A_{D2} \end{bmatrix}, \quad \widehat{B}_D = Q_D^{-1} B_D = \begin{bmatrix} B_{D1} \\ 0 \end{bmatrix},$$

where (A_{D1}, B_{D1}) is controllable.

Taking the dual again, we obtain the pair $(\widehat{A}, \widehat{C})$, which has the desired properties. In particular,

$$\begin{aligned} \widehat{A} &= \widehat{A}_D^T = Q_D^T A_D^T (Q_D^T)^{-1} = Q_D^T A (Q_D^T)^{-1} = \begin{bmatrix} A_{D1}^T & 0 \\ A_{D12}^T & A_{D2}^T \end{bmatrix}, \\ \widehat{C} &= \widehat{B}_D^T = B_D^T (Q_D^T)^{-1} = C (Q_D^T)^{-1} = [B_{D1}^T, 0], \end{aligned} \tag{6.13}$$

where (A_{D1}^T, B_{D1}^T) is completely observable by duality (see Lemma 5.7).

Example 6.3. Given $A = \begin{bmatrix} 0 & 1 & 0 \\ -1 & -2 & 1 \\ 1 & 1 & -1 \end{bmatrix}$ and $C = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}$, we wish to reduce system (6.1) to the standard form (6.13). To accomplish this, let $A_D = A^T$ and $B_D = C^T$. Notice that the pair (A_D, B_D) is precisely the pair (A, B) of Example 6.2.

A pair (A, C) can of course also be reduced directly to the standard form for unobservable systems. This is accomplished in the following.

Consider the system (6.1) [or (6.4)] and the observability matrix $\mathcal{O}$ in (6.3). Let $\operatorname{rank} \mathcal{O} = n_o < n$; i.e., the pair (A, C) is not completely observable. This implies that the unobservable subspace $R_{\bar{o}} = \mathcal{N}(\mathcal{O})$ has dimension $n - n_o$. Let $\{v_1, \dots, v_{n-n_o}\}$ be a basis for $R_{\bar{o}}$, and define an $n \times n$ similarity transformation matrix Q as

$$Q \triangleq [Q_{n_o}, v_1, \dots, v_{n-n_o}], \tag{6.14}$$

where the $n \times n_o$ matrix Q_{n_o} contains n_o linearly independent vectors chosen so that Q is nonsingular. Clearly, there are many such choices.

Lemma 6.4. *For (A, C) unobservable, there is a nonsingular matrix Q such that*

$$\widehat{A} = Q^{-1} A Q = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix} \quad \textit{and} \quad \widehat{C} = CQ = [C_1,\ 0], \tag{6.15}$$

where $A_1 \in R^{n_o \times n_o}, C_1 \in R^{p \times n_o}$, and the pair (A_1, C_1) is observable. The pair $(\widehat{A}, \widehat{C})$ is in the standard form for unobservable systems.

Proof. We need to show that

$$AQ = A[Q_{n_0}, v_1, \dots, v_{n-n_o}] = [Q_{n_o}, v_1, \dots, v_{n-n_o}] \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix} = Q\widehat{A}.$$

Since the unobservable subspace $R_{\bar{o}}$ is A-invariant (see Lemma 5.49), $Av_i \in R_{\bar{o}}$, which can be written as a linear combination of only the $n - n_o$ vectors in a basis of $R_{\bar{o}}$. Thus, A_2 in $\widehat{A}$ is an $(n - n_o) \times (n - n_o)$ matrix, and the $n_o \times (n - n_o)$ matrix above it in $\widehat{A}$ is a zero matrix. Similarly, we also need to show that

$$CQ = C[Q_{n_o}, v_1, \dots, v_{n-n_o}] = [C_1, 0] = \widehat{C}.$$

This is true since $Cv_i = 0$. ■

The $pn \times n$ observability matrix $\widehat{\mathcal{O}}$ of $(\widehat{A}, \widehat{C})$ is

$$\widehat{\mathcal{O}} = \begin{bmatrix} \widehat{C} \\ \widehat{C}\widehat{A} \\ \vdots \\ \widehat{C}\widehat{A}^{n-1} \end{bmatrix} = \begin{bmatrix} C_1 & 0 \\ C_1 A_1 & 0 \\ \vdots & \vdots \\ C_1 A_1^{n-1} & 0 \end{bmatrix}, \tag{6.16}$$

which clearly has

$$\operatorname{rank} \widehat{\mathcal{O}} = \operatorname{rank} \begin{bmatrix} C_1 \\ C_1 A_1 \\ \vdots \\ C_1 A_1^{n_o - 1} \\ \vdots \\ C_1 A_1^{n-1} \end{bmatrix} = n_o.$$

Note that

$$\widehat{\mathcal{O}} = \mathcal{O} Q. \tag{6.17}$$

The null space of $\widehat{\mathcal{O}}$ is the unobservable subspace of $(\widehat{A}, \widehat{C})$. It contains vectors only of the form $[0, \alpha^T]^T$, where $\alpha \in R^{n-n_0}$. Since $\dim \mathcal{N}(\widehat{\mathcal{O}}) = n - \operatorname{rank} \widehat{\mathcal{O}} = n - n_0$, every vector of the form $[0, \alpha^T]^T$ is an unobservable (state) vector. In other words, the similarity transformation has changed the basis of R^n in such a manner so that all unobservable vectors expressed in terms of this new basis have this very particular structure—zeros in the first n_o entries.

For Q chosen as in Lemma 6.4,

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u, y = [C_1, \ 0] \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + Du, \tag{6.18}$$

where $\hat{x} = [\hat{x}_1^T, \hat{x}_2^T]^T$ with $\hat{x}_1 \in R^{n_o}$ and (A_1, C_1) is observable. The matrix $\widehat{B} = [B_1^T, B_2^T]^T$ does not have any particular form. This representation is called a *standard form for the unobservable system.*

The n_o eigenvalues of A_1 and the corresponding modes are called *observable eigenvalues* and *observable modes* of the pair (A, C) or of the system (6.1) [or of (6.4)]. The $n - n_o$ eigenvalues of A_2 and the corresponding modes are called *unobservable eigenvalues* and *unobservable modes*, respectively.

Notice that the trajectory component $\hat{x}(t)$, which is observed via the output y, is not influenced at all by $\hat{x}_2$, the trajectory of which is determined primarily by the eigenvalues of A_2.

The unobservable modes of the system are completely absent from the output. In particular, given $\dot{x} = Ax + Bu, y = Cx$ with initial state $x(0)$, we have

$$y(t) = Ce^{At}x(0) + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau$$

and $Ce^{At} = [\widehat{C}Q^{-1}][Qe^{\widehat{A}t}Q^{-1}] = [C_1e^{A_1t}, 0]Q^{-1}$, where A_1 [from (6.15)] contains only the observable eigenvalues. Therefore, the unobservable modes cannot be seen by observing the output. The same observations can be made for discrete-time systems where the quantity CA^k is of interest.

Example 6.5. Given $A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$ and $C = [1, 1]$, we wish to reduce system (6.1) to the standard form (6.15). To accomplish this, we compute $\mathcal{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix}$, which has $\operatorname{rank} \mathcal{O} = n_o = 1 < 2 = n$. Therefore, the unobservable subspace $R_{\bar{o}} = \mathcal{N}(\mathcal{O})$ has dimension $n - n_o = 1$. In view of (6.14),

$$Q = [Q_1, v_1] = \left[\begin{array}{c|c} 0 & 1 \\ 1 & -1 \end{array}\right],$$

where $v_1 = [1, -1]^T$ is a basis for $R_{\bar{o}}$, and Q_1 was chosen so that Q is nonsingular. Then

$$\begin{aligned} \widehat{A} = Q^{-1}AQ &= \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} \\ &= \left[\begin{array}{c|c} -2 & 0 \\ \hline 1 & -1 \end{array}\right] = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}, \\ \widehat{C} = CQ = [1,\ 1] &\begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} = [1,\ 0] = [C_1,\ 0], \end{aligned}$$

where (A_1, C_1) is observable. The matrix A has two eigenvalues at $-1, -2$. It is clear from $(\widehat{A}, \widehat{C})$ that the eigenvalue -2 is observable (in A_1), whereas -1 is an unobservable eigenvalue (in A_2).

6.2.3 Kalman's Decomposition Theorem

Lemmas 6.1 and 6.4 can be combined to obtain an equivalent representation of (6.1) where the reachable and observable parts of this system can readily be identified. We consider system (6.9) and proceed, in the following, to construct the $n \times n$ required similarity transformation matrix Q.

As before, we let n_r denote the dimension of the controllable subspace R_r; i.e., $n_r = \dim R_r = \dim \mathcal{R}(\mathcal{C}) = \operatorname{rank} \mathcal{C}$. The dimension of the unobservable subspace $R_{\bar{o}} = \mathcal{N}(\mathcal{O})$ is given by $n_{\bar{o}} = n - \operatorname{rank} \mathcal{O} = n - n_o$. Let $n_{r\bar{o}}$ be the dimension of the subspace $R_{r\bar{o}} \triangleq R_r \cap R_{\bar{o}}$, which contains all the state vectors $x \in R^n$ that are controllable but unobservable. We choose

$$Q \triangleq [v_1, \dots, v_{n_r - n_{r\bar{o}}+1}, \dots, v_{n_r}, Q_N, \hat{v}_1, \dots, \hat{v}_{n_{\bar{o}} - n_{r\bar{o}}}], \tag{6.19}$$

where the n_r vectors in $\{v_1, \dots, v_{n_r}\}$ form a basis for R_r. The last $n_{r\bar{o}}$ vectors $\{v_{n_r - n_{r\bar{o}}+1}, \dots, v_{n_r}\}$ in the basis for R_r are chosen so that they form a basis for $R_{r\bar{o}} = R_r \cap R_{\bar{o}}$. The $n_{\bar{o}} - n_{r\bar{o}} = (n - n_o - n_{r\bar{o}})$ vectors $\{\hat{v}_1, \ldots, \hat{v}_{n_{\bar{o}}-n_{r\bar{o}}}\}$ are selected so that when taken together with the $n_{r\bar{o}}$ vectors $\{v_{n_r - n_{r\bar{o}}+1}, \ldots, v_{n_r}\}$ they form a basis for $R_{\bar{o}}$, the unobservable subspace. The remaining $N = n - (n_r + n_{\bar{o}} - n_{r\bar{o}})$ columns in Q_N are simply selected so that Q is nonsingular.

The following theorem is called the *Canonical Structure Theorem* or *Kalman's Decomposition Theorem.*

Theorem 6.6. *For (A, B) uncontrollable and (A, C) unobservable, there is a nonsingular matrix Q such that*

$$\widehat{A} = Q^{-1}AQ = \begin{bmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & A_{43} & A_{44} \end{bmatrix}, \quad \widehat{B} = Q^{-1}B = \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix}, \tag{6.20}$$
$$\widehat{C} = CQ = [C_1,\ 0,\ C_3,\ 0],$$

where

(i) (A_c, B_c) *with*

$$A_c \triangleq \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix} \quad \textit{and} \quad B_c \triangleq \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

is controllable, where $A_c \in R^{n_r \times n_r}, B_c \in R^{n_r \times m}$;

(ii) (A_o, C_o) *with*

$$A_o \triangleq \begin{bmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{bmatrix} \quad \textit{and} \quad C_o \triangleq [C_1, C_3]$$

is observable, where $A_o \in R^{n_o \times n_o}$ and $C_o \in R^{p \times n_o}$ and where the dimensions of the matrices A_{ij}, B_i, and C_j are as follows:

$$\begin{aligned} &A_{11} : (n_r - n_{r\bar{o}}) \times (n_r - n_{r\bar{o}}), && A_{22} : n_{r\bar{o}} \times n_{r_{\bar{o}}}, \\ &A_{33} : (n - (n_r + n_{\bar{o}} - n_{r_{\bar{o}}})) \times && A_{44} : (n_{\bar{o}} - n_{r\bar{o}}) \times (n_{\bar{o}} - n_{r\bar{o}}), \\ &\qquad (n - (n_r + n_{\bar{o}} - n_{r_{\bar{o}}})), && \\ &B_1 : (n_r - n_{r\bar{o}}) \times m, && B_2 : n_{r\bar{o}} \times m, \\ &C_1 : p \times (n_r - n_{r\bar{o}}), && C_3 : p \times (n - (n_r + n_{\bar{o}} - n_{r\bar{o}})); \end{aligned}$$

(iii) the triple (A_{11}, B_1, C_1) is such that (A_{11}, B_1) is controllable and (A_{11}, C_1) is observable.

Proof. For details of the proof, refer to [6] and to [7], where further clarifications to [6] and an updated method of selecting Q are given. ■

The similarity transformation (6.19) has altered the basis of the state space in such a manner that the vectors in the controllable subspace R_r, the vectors in the unobservable subspace $R_{\bar{o}}$, and the vectors in the subspace $R_{r\bar{o}} \cap R_{\bar{o}}$ all have specific forms. To see this, we construct the controllability matrix $\widehat{C} = [\widehat{B}, \ldots, \widehat{A}^{n-1}\widehat{B}]$ whose range is the controllable subspace and the observability matrix $\widehat{\mathcal{O}} = [\widehat{C}^T, \ldots, (\widehat{C}\widehat{A}^{n-1})^T]^T$, whose null space is the unobservable subspace. Then, all controllable states are of the form $[x_1^T, x_2^T, 0, 0]^T$, all the unobservable ones have the structure $[0, x_2^T, 0, x_4^T]^T$, and states of the form $[0, x_2^T, 0, 0]^T$ characterize $R_{r\bar{o}}$; i.e., they are controllable but unobservable.

Similarly to the previous two lemmas, the eigenvalues of $\widehat{A}$, or of A, are the eigenvalues of A_{11}, A_{22}, A_{33}, and A_{44}; i.e.,

$$|\lambda I - A| = |\lambda I - \widehat{A}| = |\lambda I - A_{11}||\lambda I - A_{22}||\lambda I - A_{33}||\lambda I - A_{44}|. \tag{6.21}$$

If we consider the representation $\{\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}\}$ given in (6.20), then

$$\begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \\ \dot{\hat{x}}_4 \end{bmatrix} = \begin{bmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix} u,$$
$$y = [C_1,\ 0,\ C_3,\ 0] \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} + Du. \tag{6.22}$$

This shows that the trajectory components corresponding to $\hat{x}_3$ and $\hat{x}_4$ are not affected by the input u. The modes associated with the eigenvalues of A_{33} and A_{44} determine the trajectory components for $\hat{x}_3$ and $\hat{x}_4$ (compare this with the results in Lemma 6.1). Similarly to Lemma 6.4, the trajectory components for $\hat{x}_2$ and $\hat{x}_4$ are not influenced by $\hat{x}_1$ and $\hat{x}_3$ (observed via y), and they are determined by the eigenvalues of A_{22} and A_{44}. The following is now apparent (see also Figure 6.1):

The eigenvalues of

A_{11} are controllable and observable,
A_{22} are controllable and unobservable,
A_{33} are uncontrollable and observable,
A_{44} are uncontrollable and unobservable.

Example 6.7. Given $A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$, and $C = [0,\ 1,\ 0]$, we wish to reduce system (6.1) to the canonical structure (or Kalman decomposition) form (6.20). The appropriate transformation matrix Q is given by (6.19). The matrix $\mathcal{C}$ was found in Example 6.2 and

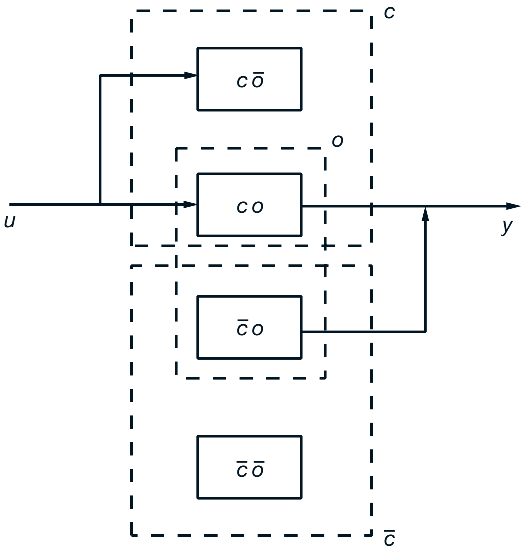


Figure 6.1. Canonical decomposition (c and $\bar{c}$ denote controllable and uncontrollable, respectively). The connections of the $c/\bar{c}$ and $o/\bar{o}$ parts of the system to the input and output are emphasized. Note that the impulse response (transfer function) of the system, which is an input–output description only, represents the part of the system that is both controllable and observable (see Chapter 7).

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -2 & 1 \\ -2 & 4 & -2 \end{bmatrix}.$$

A basis for $R_{\bar{o}} = \mathcal{N}(\mathcal{O})$ is $\{(1,\ 0,\ -1)^T\}$. Note that $n_r = 2, n_{\bar{o}} = 1$, and $n_{r\bar{o}} = 1$. Therefore,

$$Q = [v_1, v_2, Q_N] = \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{array}\right]$$

is an appropriate similarity matrix (check that $\det Q \neq 0$). We compute

$$\begin{aligned} \widehat{A} = Q^{-1}AQ &= \begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix} \\ &= \left[\begin{array}{c|c|c} 0 & 0 & 1 \\ \hline 0 & -1 & 0 \\ \hline 0 & 0 & -2 \end{array}\right] = \begin{bmatrix} A_{11} & 0 & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}, \\ \widehat{B} = Q^{-1}B = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & -2 & 1 \end{bmatrix} &\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = \left[\begin{array}{cc} 1 & 1 \\ \hline 0 & -1 \\ \hline 0 & 0 \end{array}\right] = \begin{bmatrix} B_1 \\ B_2 \\ 0 \end{bmatrix}, \end{aligned}$$

and

$$\widehat{C} = CQ = [0,\ 1,\ 0] \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \end{bmatrix} = [1, 0, 0] = [C_1,\ 0,\ C_3].$$

The eigenvalue 0 (in A_{11}) is controllable and observable, the eigenvalue -1 (in A_{22}) is controllable and unobservable and the eigenvalue -2 (in A_{33}) is uncontrollable and observable. There are no eigenvalues that are both uncontrollable and unobservable.

6.3 Eigenvalue/Eigenvector Tests for Controllability and Observability

There are tests for controllability and observability for both continuous- and discrete-time time-invariant systems that involve the eigenvalues and eigenvectors of A. Some of these criteria are called PBH tests, after the initials of the codiscoverers (Popov–Belevitch–Hautus) of these tests. These tests are useful in theoretical analysis, and in addition, they are also attractive as computational tools.

Theorem 6.8. *(i) The pair (A, B) is uncontrollable if and only if there exists a $1 \times n$ (in general) complex vector $\hat{v}_i \neq 0$ such that*

$$\hat{v}_i[\lambda_i I - A, B] = 0, \tag{6.23}$$

where λ_i is some complex scalar.

(ii) The pair (A, C) is unobservable if and only if there exists an $n \times 1$ (in general) complex vector $v_i \neq 0$ such that

$$\begin{bmatrix} \lambda_i I - A \\ C \end{bmatrix} v_i = 0, \tag{6.24}$$

where λ_i is some complex scalar.

Proof. Only part (i) will be considered since (ii) can be proved using a similar argument or, directly, by duality arguments.

(*Sufficiency*) Assume that (6.23) is satisfied. In view of $\hat{v}_i A = \lambda_i \hat{v}_i$ and $\hat{v}_i B = 0, \hat{v}_i AB = \lambda_i \hat{v}_i B = 0$ and $\hat{v}_i A^k B = 0 \quad k = 0, 1, 2, \ldots$. Therefore, $\hat{v}_i \mathcal{C} = \hat{v}_i [B, AB, \ldots, A^{n-1}B] = 0$, which shows that (A, B) is not completely controllable.

(*Necessity*) Let (A, B) be uncontrollable and assume without loss of generality the standard form for A and B given in Lemma 6.1. We will show that there exist λ_i and $\hat{v}_i$ so that (6.23) holds. Let λ_i be an uncontrollable eigenvalue, and let $\hat{v}_i = [0, \alpha], \alpha^T \in C^{n-n_r}$, where $\alpha(\lambda_i I - A_2) = 0$; i.e., α is a left eigenvector of A_2 corresponding to λ_i. Then $\hat{v}_i[\lambda_i I - A, B] = [0, \alpha(\lambda_i I - A_2), 0] = 0$; i.e., (6.23) is satisfied. ■

Corollary 6.9. *(i)* λ_i *is an uncontrollable eigenvalue of* (A, B) *if and only if there exists a* $1 \times n$ *(in general) complex vector* $\hat{v}_i \neq 0$ *that satisfies (6.23).*
(ii) λ_i *is an unobservable eigenvalue of* (A, C) *if and only if there exists an* $n \times 1$ *(in general) complex vector* $v_i \neq 0$ *that satisfies (6.24).*

Proof. See [1, p. 273, Corollary 4.6]. ■

Example 6.10. Given are $A = \begin{bmatrix} 0 & -1 & 1 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}$, $B = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$, and $C = [0,\ 1,\ 0]$, as in Example 6.7. The matrix A has three eigenvalues, $\lambda_1 = 0, \lambda_2 = -1$, and $\lambda_3 = -2$, with corresponding right eigenvectors $v_1 = [1,\ 1,\ 1]^T$, $v_2 = [1,\ 0,\ -1]^T$, $v_3 = [1,\ 1,\ -1]^T$ and with left eigenvectors $\hat{v}_1 = [1/2,\ 0,\ 1/2]$, $\hat{v}_2 = [1,\ -1,\ 0]$, and $\hat{v}_3 = [-1/2,\ 1,\ -1/2]$, respectively.

In view of Corollary 6.9, $\hat{v}_1 B = [1, 1] \neq 0$ implies that $\lambda_1 = 0$ is controllable. This is because $\hat{v}_1$ is the only nonzero vector (within a multiplication by a nonzero scalar) that satisfies $\hat{v}_1(\lambda_1 I - A) = 0$, and so $\hat{v}_1 B \neq 0$ implies that the only 1×3 vector α that satisfies $\alpha[\lambda_1 I - A, B] = 0$ is the zero vector, which in turn implies that λ_1 is controllable in view of (i) of Corollary 6.9. For similar reasons $Cv_1 = 1 \neq 0$ implies that $\lambda_1 = 0$ is observable; see (ii) of Corollary 6.9. Similarly, $\hat{v}_2 B = [0, -1] \neq 0$ implies that $\lambda_2 = -1$ is controllable, and $Cv_2 = 0$ implies that $\lambda_2 = -1$ is *unobservable*. Also, $\hat{v}_3 B = [0, 0]$ implies that $\lambda_3 = -2$ is *uncontrollable*, and $Cv_3 = 1 \neq 0$ implies that $\lambda_3 = -2$ is observable. These results agree with the results derived in Example 6.7.

Corollary 6.11. *(*Rank Tests*)*

(ia) The pair (A, B) *is controllable if and only if*

$$\operatorname{rank}[\lambda I - A, B] = n \tag{6.25}$$

for all complex numbers λ*, or for all* n *eigenvalues* λ_i *of* A*.*
(ib) λ_i *is an uncontrollable eigenvalue of* A *if and only if*

$$\operatorname{rank}[\lambda_i I - A, B] < n. \tag{6.26}$$

(iia) The pair (A, C) *is observable if and only if*

$$\operatorname{rank} \begin{bmatrix} \lambda I - A \\ C \end{bmatrix} = n \tag{6.27}$$

for all complex numbers λ*, or for all* n *eigenvalues* λ_i*.*
(iib) λ_i *is an unobservable eigenvalue of* A *if and only if*

$$\operatorname{rank} \begin{bmatrix} \lambda_i I - A \\ C \end{bmatrix} < n. \tag{6.28}$$

Proof. The proofs follow in a straightforward manner from Theorem 6.8. Notice that the only values of λ that can possibly reduce the rank of $[\lambda I - A, B]$ are the eigenvalues of A. ■

Example 6.12. If in Example 6.10 the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of A are known, but the corresponding eigenvectors are not, consider the system matrix

$$P(s) = \begin{bmatrix} sI - A & B \\ -C & 0 \end{bmatrix} = \left[\begin{array}{ccc:cc} s & 1 & -1 & 1 & 0 \\ -1 & s+2 & -1 & 1 & 1 \\ 0 & -1 & s+1 & 1 & 2 \\ \hdashline 0 & -1 & 0 & 0 & 0 \end{array}\right]$$

and determine $\operatorname{rank}[\lambda_i I - A, B]$ and $\operatorname{rank}\begin{bmatrix} \lambda_i I - A \\ C \end{bmatrix}$. Notice that

$$\operatorname{rank}\begin{bmatrix} sI - A \\ C \end{bmatrix}_{s=\lambda_2} = \operatorname{rank}\begin{bmatrix} -1 & 1 & -1 \\ -1 & 1 & -1 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix} = 2 < 3 = n$$

and

$$\operatorname{rank}[sI - A, B]_{s=\lambda_3} = \operatorname{rank}\begin{bmatrix} -2 & 2 & -1 & 1 & 0 \\ -1 & 0 & -1 & 1 & 1 \\ 0 & -1 & -1 & 1 & 2 \end{bmatrix} = 2 < 3 = n.$$

In view of Corollary 6.11, $\lambda_2 = -1$ is unobservable and $\lambda_3 = -2$ is uncontrollable.

6.4 Controller and Observer Forms

It has been seen several times in this book that equivalent representations of systems

$$\dot{x} = Ax + Bu, \quad y = Cx + Du, \tag{6.29}$$

given by the equations

$$\dot{\hat{x}} = \widehat{A}\hat{x} + \widehat{B}u, \quad y = \widehat{C}\hat{x} + \widehat{D}u, \tag{6.30}$$

where $\hat{x} = Px, \widehat{A} = PAP^{-1}$, $\widehat{B} = PB$, $\widehat{C} = CP^{-1}$, and $\widehat{D} = D$ may offer advantages over the original representation when P (or $Q = P^{-1}$) is chosen in an appropriate manner. This is the case when P (or Q) is such that the new basis of the state space provides a natural setting for the properties of interest. This section shows how to select Q when (A, B) is controllable [or (A, C) is observable] to obtain the controller and observer forms. These special forms are very useful, in realizations discussed in Chapter 8 and especially when studying state-feedback control (and state observers) discussed in Chapter 9. They are also very useful in establishing a convenient way to transition between state-space representations and another very useful class of equivalent internal representations, the polynomial matrix representations.

Controller forms are considered first. Observer forms can of course be obtained directly in a similar manner to the controller forms, or they may be obtained by duality. This is addressed in the latter part of this section.

6.4.1 Controller Forms

The controller form is a particular system representation where both matrices (A, B) have a certain special structure. Since in this case A is in the companion form, the controller form is sometimes also referred to as the *controllable companion form.* Consider the system

$$\dot{x} = Ax + Bu, \quad y = Cx + Du, \tag{6.31}$$

where $A \in R^{n\times n}$, $B \in R^{n\times m}$, $C \in R^{p\times n}$, and $D \in R^{p\times m}$ and let (A, B) be controllable. Then $\operatorname{rank} \mathcal{C} = n$, where

$$\mathcal{C} = [B, AB, \ldots, A^{n-1}B]. \tag{6.32}$$

Assume that

$$\operatorname{rank} B = m \leq n. \tag{6.33}$$

Under these assumptions, $\operatorname{rank} \mathcal{C} = n$ and $\operatorname{rank} B = m$. We will show how to obtain an equivalent pair $(\widehat{A}, \widehat{B})$ in controller form, first for the single-input case $(m = 1)$ and then for the multi-input case $(m > 1)$. Before this is accomplished, we discuss how to deal with two special cases that do not satisfy the above assumptions that $\operatorname{rank} B = m$ and that (A, B) is controllable.

1. If the m columns of B are not linearly independent ($\operatorname{rank} B = r < m$), then there exists an $m \times m$ nonsingular matrix K so that $BK = [B_r, 0]$, where the r columns of B_r are linearly independent ($\operatorname{rank} B_r = r$). Note that $\dot{x} = Ax + Bu = Ax + (BK)(K^{-1}u) = Ax + [B_r, 0] \begin{bmatrix} u_r \\ u_{m-r} \end{bmatrix} = Ax + B_r u_r$, which shows that when $\operatorname{rank} B = r < m$ the same input action to the system can be accomplished by only r inputs, instead of m inputs. The pair (A, B_r), which is controllable when (A, B) is controllable, can now be reduced to controller form, using the method described below.
2. When (A, B) is not completely controllable, then a two-step approach can be taken. First, the controllable part is isolated (see Subsection 6.2.1) and then is reduced to the controller form, using the methods of this section. In particular, consider the system $\dot{x} = Ax + Bu$ with $A \in R^{n\times n}, B \in R^{n\times m}$, and $\operatorname{rank} B = m$. Let $\operatorname{rank}[B, AB, \ldots, A^{n-1}B] = n_r < n$. Then

there exists a transformation P_1 such that $P_1AP_1^{-1} = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix}$ and $P_1B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$, where $A_1 \in R^{n_r \times n_r}, B_1 \in R^{n_r \times m}$, and (A_1, B_1) is controllable (Subsection 6.2.1). Since (A_1, B_1) is controllable, there exists a transformation P_2 such that $P_2A_1P_2^{-1} = A_{1c}$, and $P_2B_1 = B_{1c}$, where A_{1c}, B_{1c} is in controller form, defined below. Combining, we obtain

$$PAP^{-1} = \begin{bmatrix} A_{1c} & P_2A_{12} \\ 0 & A_2 \end{bmatrix}, \quad \text{and} \quad PB = \begin{bmatrix} B_{1c} \\ 0 \end{bmatrix} \tag{6.34}$$

[where $A_{1c} \in R^{n_r \times n_r}, B_{1c} \in R^{n_r \times m}$, and (A_{1c}, B_{1c}) is controllable], which is in controller form. Note that

$$P = \begin{bmatrix} P_2 & 0 \\ 0 & I \end{bmatrix} P_1. \tag{6.35}$$

Single-Input Case ($m = 1$)

The representation $\{A_c, B_c, C_c, D_c\}$ in controller form is given by $A_c \triangleq \widehat{A} = PAP^{-1}$ and $B_c \triangleq \widehat{B} = PB$ with

$$A_c = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-1} \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{6.36}$$

where the coefficients α_i are the coefficients of the characteristic polynomial $\alpha(s)$ of A; that is,

$$\alpha(s) \triangleq \det(sI - A) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0. \tag{6.37}$$

Note that $C_c \triangleq \widehat{C} = CP^{-1}$ and $D_c = D$ do not have any particular structure. The structure of (A_c, B_c) is very useful (in control problems), and the representation $\{A_c, B_c, C_c, D_c\}$ shall be referred to as the *controller form* of the system. The similarity transformation matrix P is obtained as follows. The controllability matrix $\mathcal{C} = [B, AB, \ldots, A^{n-1}B]$ is in this case an $n \times n$ nonsingular matrix. Let $\mathcal{C}^{-1} = \begin{bmatrix} \times \\ q \end{bmatrix}$, where q is the nth row of $\mathcal{C}^{-1}$ and × indicates the remaining entries of $\mathcal{C}^{-1}$. Then

$$P \triangleq \begin{bmatrix} q \\ qA \\ \cdots \\ qA^{n-1} \end{bmatrix}. \tag{6.38}$$

To show that $PAP^{-1} = A_c$ and $PB = B_c$ given in (6.36), note first that $qA^{i-1}B = 0 \; i = 1, \ldots, n-1$ and $qA^{n-1}\, B = 1$. This can be verified from the definition of q, which implies that $q\,\mathcal{C} = [0, 0, \ldots, 1]$. Now

$$P\mathcal{C} = P[B, AB, \ldots, A^{n-1}B] = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 1 \\ 0 & 0 & \cdots & 1 & \times \\ \vdots & 1 & & \vdots & \vdots \\ 1 & \times & \cdots & \times & \times \end{bmatrix} = \mathcal{C}_c, \tag{6.39}$$

which implies that $|P\mathcal{C}| = |P|\,|\mathcal{C}| \neq 0$ or that $|P| \neq 0$. Therefore, P qualifies as a similarity transformation matrix. In view of (6.39), $PB = [0, 0, \ldots, 1]^T = B_c$. Furthermore,

$$A_cP = \begin{bmatrix} qA \\ \vdots \\ qA^{n-1} \\ qA^n \end{bmatrix} = PA, \tag{6.40}$$

where in the last row of A_cP, the relation $-\sum_{i=0}^{n-1} \alpha_i A^i = A^n$ was used [which is the Cayley–Hamilton Theorem, namely, $\alpha(A) = 0$].

Example 6.13. Let $A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$. Since $n = 3$ and $|sI - A| = (s+1)(s-1)(s+2) = s^3 + 2s^2 - s - 2$, $\{A_c, B_c\}$ in controller form is given by

$$A_c = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2 & 1 & -2 \end{bmatrix} \text{ and } B_c = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The transformation matrix P that reduces (A, B) to $(A_c = PAP^{-1}, B_c = PB)$ is now derived. We have

$$\mathcal{C} = [B, AB, A^2B] = \begin{bmatrix} 1 & -1 & 1 \\ -1 & -1 & -1 \\ 1 & -2 & 4 \end{bmatrix} \quad \text{and} \quad \mathcal{C}^{-1} = \begin{bmatrix} 1 & -1/3 & -1/3 \\ -1/2 & -1/2 & 0 \\ -1/2 & -1/6 & 1/3 \end{bmatrix}.$$

The third (the nth) row of $\mathcal{C}^{-1}$ is $q = [-1/2, -1/6, 1/3]$, and therefore,

$$P \triangleq \begin{bmatrix} q \\ qA \\ qA^2 \end{bmatrix} = \begin{bmatrix} -1/2 & -1/6 & 1/3 \\ 1/2 & -1/6 & -2/3 \\ -1/2 & -1/6 & 4/3 \end{bmatrix}.$$

It can now easily be verified that $A_c = PAP^{-1}$, or

$$A_cP = \begin{bmatrix} 1/2 & -1/6 & -2/3 \\ -1/2 & -1/6 & -2/3 \\ 1/2 & -1/6 & -8/3 \end{bmatrix} = PA,$$

and that $B_c = PB$.

An alternative form to (6.36) is

$$A_{c1} = \begin{bmatrix} -\alpha_{n-1} & \cdots & -\alpha_1 & -\alpha_0 \\ 1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}, \quad B_{c1} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{6.41}$$

which is obtained if the similarity transformation matrix is taken to be

$$P_1 \triangleq \begin{bmatrix} qA^{n-1} \\ \vdots \\ qA \\ q \end{bmatrix}, \tag{6.42}$$

i.e., by reversing the order of the rows of P in (6.38). (See Exercise 6.5 and Example 6.14.)

In the above, A_c is a companion matrix of the form $\begin{bmatrix} 0 & I \\ \times & \times \end{bmatrix}$ or $\begin{bmatrix} \times & \times \\ I & 0 \end{bmatrix}$. It could also be of the form $\begin{bmatrix} 0 & \times \\ I & \times \end{bmatrix}$ or $\begin{bmatrix} \times & 0 \\ \times & I \end{bmatrix}$ with coefficients $-[\alpha_0, \ldots, \alpha_{n-1}]^T$ in the last or the first column. It is shown here, for completeness, how to determine controller forms where A_c are such companion matrices. In particular, if

$$Q_2 = P_2^{-1} = [B, AB, \ldots, A^{n-1}B] = \mathcal{C}, \tag{6.43}$$

then

$$A_{c2} = Q_2^{-1}AQ_2 = \begin{bmatrix} 0 & \cdots & 0 & -\alpha_0 \\ 1 & \cdots & 0 & -\alpha_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & -\alpha_{n-1} \end{bmatrix}, B_{c2} = Q_2^{-1}B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{6.44}$$

Also, if

$$Q_3 = P_3^{-1} = [A^{n-1}B, \ldots, B], \tag{6.45}$$

then

$$A_{c3} = Q_3^{-1}AQ_3 = \begin{bmatrix} -\alpha_{n-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\alpha_1 & 0 & \cdots & 1 \\ -\alpha_0 & 0 & \cdots & 0 \end{bmatrix}, B_{c3} = Q_3^{-1}B = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \tag{6.46}$$

(A_c, B_c) in (6.44) and (6.46) are also in controller canonical or controllable companion form. (See also Exercise 6.5 and Example 6.14.)

Example 6.14. Let $A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$, as in Example 6.13. Alternative controller forms can be derived for different P. In particular, if

(i) $P = P_1 = \begin{bmatrix} qA^2 \\ qA \\ q \end{bmatrix} = \begin{bmatrix} -1/2 & -1/6 & 4/3 \\ 1/2 & -1/6 & -2/3 \\ -1/2 & -1/6 & 1/3 \end{bmatrix}$, as in (6.42) ($\mathcal{C}, \mathcal{C}^{-1}$, and q were found in Example 6.13), then

$$A_{c1} = \begin{bmatrix} -2 & 1 & 2 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B_{c1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

as in (6.41). Note that in the present case $A_{c1}P_1 = \begin{bmatrix} 1/2 & -1/6 & -8/3 \\ -1/2 & -1/6 & 4/3 \\ 1/2 & -1/6 & -2/3 \end{bmatrix} = P_1 A, \quad B_{c1} = P_1 B.$

(ii) $Q_2 = \mathcal{C} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & -1 & -1 \\ 1 & -2 & 4 \end{bmatrix}$, as in (6.43). Then

$$A_{c2} = \begin{bmatrix} 0 & 0 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{bmatrix}, \quad B_{c2} = Q_2^{-1} B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

as in (6.44).

(iii) $Q_3 = [A^2 B, AB, B] = \begin{bmatrix} 1 & -1 & 1 \\ -1 & -1 & -1 \\ 4 & -2 & 1 \end{bmatrix}$, as in (6.45). Then

$$A_{c3} = \begin{bmatrix} -2 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & 0 & 0 \end{bmatrix}, \quad B_{c3} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

as in (6.46). Note that $Q_3 A_{c3} = \begin{bmatrix} -1 & 1 & -1 \\ -1 & -1 & -1 \\ -8 & 4 & -2 \end{bmatrix} = AQ_3, \quad Q_3 B_{c3} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = B.$

Multi-Input Case ($m > 1$)

In this case, the $n \times mn$ matrix $\mathcal{C}$ given in (6.32) is not square, and there are typically many sets of n columns of $\mathcal{C}$ that are linearly independent ($\operatorname{rank}\mathcal{C} = n$). Depending on which columns are chosen and in what order, different controller forms (controllable companion forms) are derived. Note that in the case when $m = 1$, four different controller forms were derived, even though there was only one set of n linearly independent columns. In the present case there are many more such choices. The form that will be used most often in the following is a generalization of (A_c, B_c) given in (6.36). Further discussion including derivation and alternative forms may be found in [1, Subsection 3.4D].

Let $\widehat{A} = PAP^{-1}$ and $\widehat{B} = PB$, where P is constructed as follows. Consider

$$\begin{aligned} \mathcal{C} &= [B, AB, \ldots, A^{n-1}B] \\ &= [b_1, \ldots, b_m, Ab_1, \ldots, Ab_m, \ldots, A^{n-1}b_1, \ldots, A^{n-1}b_m], \end{aligned} \tag{6.47}$$

where the $b_1, \ldots, b_m$ are the m columns of B. Select, starting from the left and moving to the right, the first n independent columns (rank $\mathcal{C} = n$). Reorder these columns by taking first b_1, Ab_1, A^2b_1, etc., until all columns involving b_1 have been taken; then take b_2, Ab_2, etc.; and lastly, take b_m, Ab_m, etc., to obtain

$$\bar{\mathcal{C}} \triangleq [b_1, Ab_1, \ldots, A^{\mu_1 - 1}b_1, \ldots, b_m, \ldots, A^{\mu_m - 1}b_m], \tag{6.48}$$

an $n \times n$ matrix. The integer μ_i denotes the number of columns involving b_i in the set of the first n linearly independent columns found in $\mathcal{C}$ when moving from left to right.

Definition 6.15. *The m integers μ_i, $i = 1, \ldots, m$, are the* controllability indices of the system, *and $\mu \triangleq \max \mu_i$ is called* the controllability index of the system. *Note that*

$$\sum_{i=1}^{m} \mu_i = n \quad \textit{and} \quad m\mu \geq n. \tag{6.49}$$

■

An alternative but equivalent definition for μ is that μ is the minimum integer k such that

$$\operatorname{rank}[B, AB, \ldots, A^{k-1}B] = n. \tag{6.50}$$

Notice that in (6.48) all columns of B are always present since rank $B = m$. This implies that $\mu_i \geq 1$ for all i. Notice further that if $A^k b_i$ is present, then $A^{k-1}b_i$ must also be present.

Now define

$$\sigma_k \triangleq \sum_{i=1}^{k} \mu_i, \quad k = 1, \ldots, m; \tag{6.51}$$

i.e., $\sigma_1 = \mu_1, \sigma_2 = \mu_1 + \mu_2, \dots, \sigma_m = \mu_1 + \cdots + \mu_m = n$. Also, consider $\bar{\mathcal{C}}^{-1}$ and let q_k, where $q_k^T \in R^n$, $k = 1, \dots, m$, denote its ${\sigma_k}^{th}$ row; i.e.,

$$\bar{\mathcal{C}}^{-1} = [\times, \dots, \times, q_1^T \vdots \cdots \vdots \times, \dots, \times, q_m^T]^T. \tag{6.52}$$

Next, define

$$P \triangleq \begin{bmatrix} q_1 \\ q_1 A \\ \vdots \\ q_1 A^{\mu_1 - 1} \\ \cdots \\ \vdots \\ \cdots \\ q_m \\ q_m A \\ \vdots \\ q_m A^{\mu_m - 1} \end{bmatrix}. \tag{6.53}$$

It can now be shown that $PAP^{-1} = A_c$ and $PB = B_c$ with

$$A_c = [A_{ij}], \qquad i, j = 1, \dots, m,$$

$$A_{ii} = \begin{bmatrix} 0 & \\ \vdots & I_{\mu_i - 1} \\ 0 & \\ \times & \times \cdots \times \end{bmatrix} \in R^{\mu_i \times \mu_i},\ i = j, \quad A_{ij} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \\ \times & \times \cdots & \times \end{bmatrix} \in R^{\mu_i \times \mu_j},\ i \neq j,$$

and

$$B_c = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_m \end{bmatrix}, \qquad B_i = \begin{bmatrix} 0 \cdots 0 & 0 & \cdots & 0 \\ \vdots \quad \vdots & \vdots & & \vdots \\ 0 \cdots 0 & 1 & \times \cdots & \times \end{bmatrix} \in R^{\mu_i \times m}, \tag{6.54}$$

where the 1 in the last row of B_i occurs at the ith column location, $i = 1, \dots, m$, and × denotes nonfixed entries. Note that $C_c = CP^{-1}$ does not have any particular structure. The expression (6.54) is a very useful form (in control problems) and shall be referred to as the *controller form* of the system. The derivation of this result is discussed in [1, Subsection 3.4D] .

Example 6.16. Given are $A \in R^{n \times n}$ and $B \in R^{n \times m}$ with (A, B) controllable and with $\operatorname{rank} B = m$. Let $n = 4$ and $m = 2$. Then there must be two controllability indices μ_1 and μ_2 such that $n = 4 = \sum_{i=1}^{2} \mu_i = \mu_1 + \mu_2$. Under these conditions, there are three possibilities:

(i) $\mu_1 = 2, \mu_2 = 2$,

$$A_c = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ \times & \times & \times & \times \\ \hline 0 & 0 & 0 & 1 \\ \times & \times & \times & \times \end{array}\right], \quad B_c = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \left[\begin{array}{cc} 0 & 0 \\ 1 & \times \\ \hline 0 & 0 \\ 0 & 1 \end{array}\right].$$

(ii) $\mu_1 = 1, \mu_2 = 3$,

$$A_c = \left[\begin{array}{cc|cc} \times & \times & \times & \times \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \times & \times & \times & \times \end{array}\right], \quad B_c = \left[\begin{array}{cc} 1 & \times \\ \hline 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{array}\right].$$

(iii) $\mu_1 = 3$, $\mu_2 = 1$,

$$A_c = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \times & \times & \times & \times \\ \hline \times & \times & \times & \times \end{array}\right], \quad B_c = \left[\begin{array}{cc} 0 & 0 \\ 0 & 0 \\ 1 & \times \\ \hline 0 & 1 \end{array}\right].$$

It is possible to write A_c, B_c in a systematic and perhaps more transparent way. In particular, notice that A_c, B_c in (6.54) can be expressed as

$$A_c = \bar{A}_c + \bar{B}_c A_m, \quad B_c = \bar{B}_c B_m, \tag{6.55}$$

where $\bar{A}_c = \text{block diag}[\bar{A}_{11}, \bar{A}_{22}, \ldots, \bar{A}_{mm}]$ with

$$\bar{A}_{ii} = \begin{bmatrix} 0 & \\ \vdots & I_{\mu_i - 1} \\ 0 & \\ 0 & 0 \cdots 0 \end{bmatrix} \in R^{\mu_i \times \mu_i}, \ \bar{B}_c = \text{block diag}\left(\begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in R^{\mu_i \times 1}, \quad i = 1, \ldots, m \right),$$

and $A_m \in R^{m \times n}$ and $B_m \in R^{m \times m}$ are some appropriate matrices with $\sum_{i=1}^{m} \mu_i = n$. Note that the matrices $\bar{A}_c, \bar{B}_c$ are completely determined by the m controllability indices μ_i, $i = 1, \ldots, m$. The matrices A_m and B_m consist of the σ_1th, σ_2th, $\ldots, \sigma_m$th rows of A_c (entries denoted by $\times$) and the same rows of B_c, respectively [see (6.57) and (6.58) below].

Example 6.17. Let $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & -1 \end{bmatrix}$ and $B = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}$. To determine the controller form (6.54), consider

$$\mathcal{C} = [B, AB, A^2 B] = [b_1, b_2, Ab_1, Ab_2, A^2 b_1, A^2 b_2] = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 2 & -2 & -2 \end{bmatrix},$$

where rank $\mathcal{C} = 3 = n$; i.e., (A, B) is controllable. Searching from left to right, the first three columns of $\mathcal{C}$ are selected since they are linearly independent. Then

$$\bar{\mathcal{C}} = [b_1, Ab_1, b_2] = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \end{bmatrix}$$

and the controllability indices are $\mu_1 = 2$ and $\mu_2 = 1$. Also, $\sigma_1 = \mu_1 = 2$ and $\sigma_2 = \mu_1 + \mu_2 = 3 = n$, and

$$\bar{\mathcal{C}}^{-1} = \begin{bmatrix} -1 & 1 & 1/2 \\ 0 & 0 & 1/2 \\ 1 & 0 & -1/2 \end{bmatrix}.$$

Notice that $q_1 = [0, 0, 1/2]$ and $q_2 = [1, 0, -1/2]$, the second and third rows of $\bar{\mathcal{C}}^{-1}$, respectively. In view of (6.53), $P = \left[\begin{array}{c} q_1 \\ q_1 A \\ \hdashline q_2 \end{array}\right] = \begin{bmatrix} 0 & 0 & 1/2 \\ 0 & 1 & -1/2 \\ 1 & 0 & -1/2 \end{bmatrix}$, $P^{-1} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 2 & 0 & 0 \end{bmatrix}$, and $A_c = PAP^{-1} = \left[\begin{array}{c:c} A_{11} & A_{12} \\ \hdashline A_{21} & A_{22} \end{array}\right] = \left[\begin{array}{cc:c} 0 & 1 & 0 \\ 2 & -1 & 0 \\ \hdashline 1 & 0 & 0 \end{array}\right]$, $B_c = PB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \left[\begin{array}{cc} 0 & 0 \\ 1 & 1 \\ \hdashline 0 & 1 \end{array}\right]$.

One can also verify (6.55) quite easily. We have

$$A_c = \left[\begin{array}{cc:c} 0 & 1 & 0 \\ 2 & -1 & 0 \\ \hdashline 1 & 0 & 0 \end{array}\right] = \bar{A}_c + \bar{B}_c A_m = \left[\begin{array}{cc:c} 0 & 1 & 0 \\ 0 & 0 & 0 \\ \hdashline 0 & 0 & 0 \end{array}\right] + \left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \\ \hdashline 0 & 1 \end{array}\right] \begin{bmatrix} 2 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

and

$$B_c = \left[\begin{array}{cc} 0 & 0 \\ 1 & 1 \\ \hdashline 0 & 1 \end{array}\right] = \bar{B}_c B_m = \left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \\ \hdashline 0 & 1 \end{array}\right] \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

It is interesting to note that in this example, the given pair (A, B) could have already been in controller form if B were different but A were the same. For example, consider the following three cases:

1. $A = \left[\begin{array}{c:cc} 0 & 1 & 0 \\ \hdashline 0 & 0 & 1 \\ 0 & 2 & -1 \end{array}\right]$, $\quad B = \left[\begin{array}{cc} 1 & \times \\ \hdashline 0 & 0 \\ 0 & 1 \end{array}\right]$, $\quad \mu_1 = 1, \mu_2 = 2$,
2. $A = \left[\begin{array}{cc:c} 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hdashline 0 & 2 & -1 \end{array}\right]$, $\quad B = \left[\begin{array}{cc} 0 & 0 \\ 1 & \times \\ \hdashline 0 & 1 \end{array}\right]$, $\quad \mu_1 = 2, \mu_1 = 1$,

3. $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mu_1 = 3 = n.$

Note that case 3 is the single-input case (6.36).

Remarks

(i) An important result involving the controllability indices of (A, B) is the following: Given (A, B) controllable, then $(P(A + BGF)P^{-1}, PBG)$ will have the same controllability indices, within reordering, for any P, F, and G ($|P| \neq 0, |G| \neq 0$) of appropriate dimensions. In other words, *the controllability indices are invariant under similarity and input transformations P and G, and state feedback F [or similarity transformation P and state feedback (F, G)].* (For further discussion, see [1, Subsection 3.4D].)

(ii) It is not difficult to derive explicit expressions for A_m and B_m in (6.55). Using

$$\begin{aligned} &q_i A^{k-1} b_j = 0 \quad k = 1, \ldots, \mu_j, \quad i \neq j, \\ &q_i A^{k-1} b_i = 0 \quad k = 1, \ldots, \mu_i - 1, \text{ and } q_i A^{\mu_i - 1} b_i = 1, \quad i = j, \end{aligned} \tag{6.56}$$

where $i = 1, \ldots, m$, and $j = 1, \ldots, m$, it can be shown that the m σ_1th, σ_2th, $\ldots, \sigma_m$th rows of A_c that are denoted by A_m in (6.55) are given by

$$A_m = \begin{bmatrix} q_1 A^{\mu_1} \\ \vdots \\ q_m A^{\mu_m} \end{bmatrix} P^{-1}. \tag{6.57}$$

Similarly

$$B_m = \begin{bmatrix} q_1 A^{\mu_1 - 1} \\ \vdots \\ q_m A^{\mu_m - 1} \end{bmatrix} B. \tag{6.58}$$

The matrix B_m is an upper triangular matrix with ones on the diagonal. (For details, see [1, Subsection 3.4D].)

Example 6.18. We wish to reduce $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & -1 \end{bmatrix}, B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$ to controller form. Note that A and B are almost the same as in Example 6.17; however, here $\mu_1 = 1 < 2 = \mu_2$, as will be seen. We have $\mathcal{C} = [B, AB, A^2B] = [b_1, b_2, Ab_1, Ab_2, \ldots] = \begin{bmatrix} 1 & 1 & 0 & 1 & \\ 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 2 & \end{bmatrix}$. Searching from left to right, the first three linearly independent columns are b_1, b_2, Ab_2, and $\bar{\mathcal{C}} = [b_1, b_2, Ab_2] = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$, from which we conclude that $\mu_1 = 1$, $\mu_2 = 2$, $\sigma_1 = 1$, and $\sigma_2 = 3$. We compute $\bar{\mathcal{C}}^{-1} = \begin{bmatrix} 1 & -1 & -1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$. Note that $q_1 = [1, -1, -1/2]$ and $q_2 = [0, 0, 1/2]$, the first and third rows of $\bar{\mathcal{C}}^{-1}$, respectively. Then $P = \begin{bmatrix} q_1 \\ q_2 \\ q_2 A \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1/2 \\ 0 & 0 & 1/2 \\ 0 & 1 & -1/2 \end{bmatrix}$, $P^{-1} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 0 & 2 & 0 \end{bmatrix}$, and

$$A_c = PAP^{-1} = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] = \left[\begin{array}{c|cc} 0 & -1 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 2 & -1 \end{array}\right],$$

$$B_c = PB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \left[\begin{array}{cc} 1 & 0 \\ \hline 0 & 0 \\ 0 & 1 \end{array}\right].$$

It is easy to verify relations (6.57) and (6.58).

Structure Theorem—Controllable Version

The transfer function matrix $H(s)$ of the system $\dot{x} = Ax + Bu$, $y = Cx + Du$ is given by $H(s) = C(sI - A)^{-1}B + D$. If (A, B) is in *controller form* (6.54), then $H(s)$ can alternatively be characterized by the Structure Theorem stated in Theorem 6.19 below. This result is very useful in the realization of systems, which is addressed in Chapter 8 and in the study of state feedback in Chapter 9.

Let $A = A_c = \bar{A}_c + \bar{B}_c A_m$ and $B = B_c = \bar{B}_c B_m$, as in (6.55), with $|B_m| \neq 0$, and let $C = C_c$ and $D = D_c$. Define

$$\Lambda(s) \triangleq \operatorname{diag}[s^{\mu_1}, s^{\mu_2}, \ldots, s^{\mu_m}], \tag{6.59}$$

$$S(s) \triangleq \text{block diag}([1, s, \ldots, s^{\mu_i - 1}]^T, \quad i = 1, \ldots, m). \tag{6.60}$$

Note that $S(s)$ is an $n \times m$ polynomial matrix ($n = \sum_{i=1}^m \mu_i$), i.e., a matrix with polynomials as entries. Now define the $m \times m$ polynomial matrix $D(s)$ and the $p \times m$ polynomial matrix $N(s)$ by

$$D(s) \triangleq B_m^{-1}[\Lambda(s) - A_m S(s)], N(s) \triangleq C_c S(s) + D_c D(s). \tag{6.61}$$

The following is the controllable version of the *Structure Theorem.*

Theorem 6.19. $H(s) = N(s)D^{-1}(s)$, *where* $N(s)$ *and* $D(s)$ *are defined in (6.61).*

Proof. First, note that

$$(sI - A_c)S(s) = B_c D(s). \tag{6.62}$$

To see this, we write $B_c D(s) = \bar{B}_c B_m B_m^{-1}[\Lambda(s) - A_m S(s)] = \bar{B}_c \Lambda(s) - \bar{B}_c A_m S(s)$ and $(sI - A_c)S(s) = sS(s) - (\bar{A}_c + \bar{B}_c A_m)S(s) = (sI - \bar{A}_c)S(s) - \bar{B}_c A_m S(s) = \bar{B}_c \Lambda(s) - \bar{B}_c A_m S(s)$, which proves (6.62). Now $H(s) = C_c(sI - A_c)^{-1}B_c + D_c = C_c S(s) D^{-1}(s) + D_c = [C_c S(s) + D_c D(s)]D^{-1}(s) = ND^{-1}$. ■

Example 6.20. Let $A_c = \begin{bmatrix} 0 & 1 & 0 \\ 2 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$, $B_c = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$, as in Example 6.17. Here $\mu_1 = 2, \mu_2 = 1$ and $A_m = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$, $B_m = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$. Then $\Lambda(s) = \begin{bmatrix} s^2 & 0 \\ 0 & s \end{bmatrix}$, $S(s) = \begin{bmatrix} 1 & 0 \\ s & 0 \\ 0 & 1 \end{bmatrix}$ and

$$\begin{aligned} D(s) = B_m^{-1}[\Lambda(s) - A_m S(s)] &= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \left[\begin{bmatrix} s^2 & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} -s+2 & 0 \\ 1 & 0 \end{bmatrix} \right] \\ &= \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} s^2+s-2 & 0 \\ -1 & s \end{bmatrix} = \begin{bmatrix} s^2+s-1 & -s \\ -1 & s \end{bmatrix}. \end{aligned}$$

Now $C_c = [0,\ 1,\ 1]$, and $D_c = [0,\ 0]$,

$$N(s) = C_c S(s) + D_c D(s) = [s,\ 1],$$

and

$$\begin{aligned} H(s) = [s,1] \begin{bmatrix} s^2+s-1 & -s \\ -1 & s \end{bmatrix}^{-1} &= [s,1] \begin{bmatrix} s & s \\ 1 & s^2+s-1 \end{bmatrix} \frac{1}{s(s^2+s-2)} \\ &= \frac{1}{s(s^2+s-2)}[s^2+1, 2s^2+s-1] \\ &= C_c(sI - A_c)^{-1}B_c + D_c. \end{aligned}$$

Example 6.21. Let $A_c = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 2 & 1 & -2 \end{bmatrix}$, $B_c = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, $C_c = [0,\ 1,\ 0]$, and $D_c = 0$ (see Example 6.13). In the present case, we have $A_m = [2, 1, -2]$, $B_m = 1$, $\Lambda(s) = s^3$, $S(s) = [1, s, s^2]^T$, and

$$D(s) = 1 \cdot [s^3 - [2,\ 1,\ -2][1,\ s,\ s^2]^T] = s^3 + 2s^2 - s - 2, \quad N(s) = s.$$

Then

$$H(s) = N(s)D^{-1}(s) = s/(s^3 + 2s^2 - s - 2) = C_c(sI - A_c)^{-1}B_c + D_c.$$

6.4.2 Observer Forms

Consider the system $\dot{x} = Ax + Bu$, $y = Cx + Du$ given in (6.1) and assume that (A, C) is observable; i.e., $\operatorname{rank} \mathcal{O} = n$, where

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}. \tag{6.63}$$

Also, assume that the $p \times n$ matrix C has a full row rank p; i.e.,

$$\operatorname{rank} C = p \leq n. \tag{6.64}$$

It is of interest to determine a transformation matrix P so that the equivalent system representation $\{A_o, B_o, C_o, D_o\}$ with

$$A_o = PAP^{-1}, \quad B_o = PB, \quad C_o = CP^{-1}, \quad D_o = D \tag{6.65}$$

will have (A_o, C_o) in an observer form (defined below). As will become clear in the following discussion, these forms are dual to the controller forms previously discussed and can be derived by taking advantage of this fact. In particular, let $\widetilde{A} \triangleq A^T$, $\widetilde{B} \triangleq C^T$ [$(\widetilde{A}, \widetilde{B})$ is controllable], and determine a nonsingular transformation $\widetilde{P}$ so that $\widetilde{A}_c = \widetilde{P}\widetilde{A}\widetilde{P}^{-1}$, $\widetilde{B}_c = \widetilde{P}\widetilde{B}$ are in controller form given in (6.54). Then $A_o = \widetilde{A}_c^T$ and $C_o = \widetilde{B}_c^T$ is in observer form.

It will be demonstrated in the following discussion how to obtain observer forms directly, in a way that parallels the approach described for controller forms. This is done for the sake of completeness and to define the observability indices. The approach of using duality just given can be used in each case to verify the results.

We first note that if $\operatorname{rank} C = r < p$, an approach analogous to the case when $\operatorname{rank} B < m$ can be followed, as in Subsection 6.4.1. The fact that the rows of C are not linearly independent means that the same information can be extracted from only r outputs, and therefore, the choice for the outputs should perhaps be reconsidered. Now if (A, C) is unobservable, one may use two steps to first isolate the observable part and then reduce it to the observer form, in an analogous way to the uncontrollable case previously given.

Single-Output Case ($p = 1$)

Let

$$P^{-1} = Q \triangleq [\tilde{q}, A\tilde{q}, \ldots, A^{n-1}\tilde{q}], \tag{6.66}$$

where $\tilde{q}$ is the nth column in $\mathcal{O}^{-1}$. Then

$$A_0 = \begin{bmatrix} 0 \cdots 0 & -\alpha_0 \\ 1 \cdots 0 & -\alpha_1 \\ \vdots \ \ddots \ \vdots & \vdots \\ 0 \cdots 1 & -\alpha_{n-1} \end{bmatrix}, \quad C_o = [0, \ldots, 0, 1], \tag{6.67}$$

where the α_i denote the coefficients of the characteristic polynomial $\alpha(s) \triangleq \det(sI - A) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$. Here $A_o = PAP^{-1} = Q^{-1}AQ, C_o = CP^{-1} = CQ$, and the desired result can be established by using a proof that is completely analogous to the proof in determining the (dual) controller form presented in Subsection 6.4.1. Note that $B_o = PB$ does not have any particular structure. The representation $\{A_o, B_o, C_o, D_o\}$ will be referred to as the *observer form* of the system.

Reversing the order of columns in P^{-1} given in (6.66) or selecting P to be exactly $\mathcal{O}$, or to be equal to the matrix obtained after the order of the columns in $\mathcal{O}$ has been reversed, leads to alternative observer forms in a manner analogous to the controller form case.

Example 6.22. Let $A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$ and $C = [1, -1, 1]$. To derive the observer form (6.67), we could use duality, by defining $\widetilde{A} = A^T, \widetilde{B} = C^T$, and deriving the controller form of $\widetilde{A}, \widetilde{B}$, i.e., by following the procedure outlined above. We note that the $\widetilde{A}, \widetilde{B}$ are exactly the matrices given in Examples 6.13 and 6.14. As an alternative approach, the observer form is now derived directly. In particular, we have

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & -1 & -2 \\ 1 & -1 & 4 \end{bmatrix}, \mathcal{O}^{-1} = \begin{bmatrix} 1 & -1/2 & -1/2 \\ -1/3 & -1/2 & -1/6 \\ -1/3 & 0 & 1/3 \end{bmatrix},$$

and in view of (6.66),

$$Q = P^{-1} = [\tilde{q}, A\tilde{q}, A^2\tilde{q}] = \begin{bmatrix} -1/2 & 1/2 & -1/2 \\ -1/6 & -1/6 & -1/6 \\ 1/3 & -2/3 & 4/3 \end{bmatrix}.$$

Note that $\tilde{q} = [-1/2, -1/6, 1/3]^T$, the last column of $\mathcal{O}^{-1}$. Then

$$A_o = Q^{-1}AQ = \begin{bmatrix} 0 & 0 & 2 \\ 1 & 0 & 1 \\ 0 & 1 & -2 \end{bmatrix}, \text{ and } C_o = CQ = [0, 0, 1],$$

where $|sI - A| = s^3 + 2s - s - 2 = s^3 + \alpha_2 s^2 + \alpha_1 s + \alpha_0$. Hence, $QA_o = \begin{bmatrix} 1/2 & -1/2 & 1/2 \\ -1/6 & -1/6 & -1/6 \\ -2/3 & 4/3 & -8/3 \end{bmatrix} = AQ$.

Multi-Output Case ($p > 1$)

Consider

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_p \\ c_1 A \\ \vdots \\ c_p A \\ \vdots \\ c_1 A^{n-1} \\ \vdots \\ c_p A^{n-1} \end{bmatrix}, \tag{6.68}$$

where $c_1, \ldots, c_p$ denote the p rows of C, and select the first n linearly independent rows in $\mathcal{O}$, moving from the top to bottom (rank $\mathcal{O} = n$). Next, reorder the selected rows by first taking all rows involving c_1, then c_2, etc., to obtain

$$\bar{\mathcal{O}} \triangleq \begin{bmatrix} c_1 \\ c_1 A \\ \vdots \\ c_1 A^{\nu_1 - 1} \\ \vdots \\ c_p \\ \vdots \\ c_p A^{\nu_p - 1} \end{bmatrix}, \tag{6.69}$$

an $n \times n$ matrix. The integer ν_i denotes the number of rows involving c_i in the set of the first n linearly independent rows found in $\mathcal{O}$ when moving from top to bottom.

Definition 6.23. *The p integers ν_i, $i = 1, \ldots, p$, are the* observability indices of the system, *and $\nu \triangleq \max \nu_i$ is called* the observability index of the system. *Note that*

$$\sum_{i=1}^{p} \nu_i = n \quad \textit{and} \quad p\nu \geq n. \tag{6.70}$$

■

When rank $C = p$, then $\nu_i \geq 1$. Now define

$$\tilde{\sigma}_k \triangleq \sum_{i=1}^{k} \nu_i \quad k = 1, \ldots, p; \tag{6.71}$$

i.e., $\tilde{\sigma}_1 = \nu_1, \tilde{\sigma}_2 = \nu_1 + \nu_2, \ldots, \tilde{\sigma}_p = \nu_1 + \cdots + \nu_p = n$. Consider $\bar{\mathcal{O}}^{-1}$ and let $\tilde{q}_k \in R^n$, $k = 1, \ldots, p$, represent its $\tilde{\sigma}_k$th column; i.e.,

$$\bar{\mathcal{O}}^{-1} = [\times \cdots \times \tilde{q}_1 | \times \cdots \times \tilde{q}_2 | \cdots | \times \cdots \times \tilde{q}_p]. \tag{6.72}$$

Define

$$P^{-1} = Q = [\tilde{q}_1, \ldots, A^{\nu_1 - 1}\tilde{q}_1, \ldots, \tilde{q}_p, \ldots, A^{\nu_p - 1}\tilde{q}_p]. \tag{6.73}$$

Then $A_o = PAP^{-1} = Q^{-1}AQ$ and $C_o = CP^{-1} = CQ$ are given by

$$A_o = [A_{ij}], \quad i, j = 1, \ldots, p,$$

$$A_{ii} = \begin{bmatrix} 0 \cdots 0 & \times \\ I_{\nu_i - 1} & \vdots \\ & \times \end{bmatrix} \in R^{\nu_i \times \nu_i}, \quad i = j, A_{ij} = \begin{bmatrix} 0 \cdots & 0 & \times \\ \vdots & \vdots & \vdots \\ 0 \cdots & 0 & \times \end{bmatrix} \in R^{\nu_i \times \nu_j}, \quad i \neq j,$$

and

$$C_o = [C_1, C_2, \ldots, C_p], C_i = \begin{bmatrix} 0 \cdots & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 \cdots & 0 & 0 \\ 0 \cdots & 0 & 1 \\ 0 \cdots & 0 & \times \\ \vdots & \vdots & \vdots \\ 0 \cdots & 0 & \times \end{bmatrix} \in R^{p \times \nu_i}, \tag{6.74}$$

where the 1 in the last column of C_i occurs at the ith row location ($i = 1, \ldots, p$) and × denotes nonfixed entries. Note that the matrix $B_o = PB = Q^{-1}B$ does not have any particular structure. Equation (6.74) is a very useful form (in the observer problem) and shall be referred to as the *observer form* of the system.

Analogous to (6.55), we express A_o and C_o as

$$A_o = \bar{A}_o + A_p \bar{C}_o, \quad C_o = C_p \bar{C}_o, \tag{6.75}$$

where $\bar{A}_o =$ block diag$[A_1, A_2, \ldots, A_p]$ with $A_i = \begin{bmatrix} 0 \cdots & 0 \\ I_{\nu_i - 1} & \vdots \\ & 0 \end{bmatrix} \in R^{\nu_i \times \nu_i}, \bar{C}_o =$ block diag($[0, \ldots, 0, 1]^T \in R^{\nu_i}$, $i = 1, \ldots, p$), and $A_p \in R^{n \times p}$, and $C_p \in R^{p \times p}$ are appropriate matrices ($\sum_{i=1}^{p} \nu_i = n$). Note that $\bar{A}_o, \bar{C}_o$ are completely determined by the p observability indices ν_i, $i = 1, \ldots, p$, and A_p and C_p contain this information in the $\tilde{\sigma}_1$th$, \ldots, \tilde{\sigma}_p$th columns of A_o and in the same columns of C_o, respectively.

Example 6.24. Given $A = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 2 \\ 0 & 1 & -1 \end{bmatrix}$ and $C = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$, we wish to reduce these to observer form. This can be accomplished using duality, i.e., by first reducing $\widetilde{A} \triangleq A^T, \widetilde{B} \triangleq C^T$ to controller form. Note that $\widetilde{A}, \widetilde{B}$ are the matrices used in Example 6.17, and therefore, the desired answer is easily obtained. Presently, we shall follow the direct algorithm described above. We have

$$\mathcal{O} = \begin{bmatrix} C \\ CA \\ CA^2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 2 \\ 1 & 0 & 2 \\ 0 & 2 & -2 \\ 0 & 2 & -2 \end{bmatrix}.$$

Searching from top to bottom, the first three linearly independent rows are $c_1, c_2, c_1 A$, and

$$\bar{\mathcal{O}} = \begin{bmatrix} c_1 \\ c_1 A \\ c_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 2 \\ 1 & 1 & 0 \end{bmatrix}.$$

Note that the observability indices are $\nu_1 = 2, \nu_2 = 1$ and $\tilde{\sigma}_1 = 2, \tilde{\sigma}_2 = 3$. We compute

$$\bar{\mathcal{O}}^{-1} = \begin{bmatrix} -1 & 0 & 1 \\ 1 & 0 & 0 \\ 1/2 & 1/2 & -1/2 \end{bmatrix} = \begin{bmatrix} \times & 0 & 1 \\ \times & 0 & 0 \\ \times & 1/2 & -1/2 \end{bmatrix}.$$

Then, $Q = [\tilde{q}_1, A\tilde{q}_1, \tilde{q}_2] = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1/2 & -1/2 & -1/2 \end{bmatrix}$ and $Q^{-1} = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. Therefore,

$$A_o = Q^{-1}AQ = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] = \left[\begin{array}{cc|c} 0 & 2 & 1 \\ 1 & -1 & 0 \\ \hline 0 & 0 & 0 \end{array}\right], C_o = CQ = [C_1 \vdots C_2] = \left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 1 & 1 \end{array}\right].$$

We can also verify (6.47), namely

$$A_o = \left[\begin{array}{cc|c} 0 & 2 & 1 \\ 1 & -1 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] = \bar{A}_o + A_p \bar{C}_o = \left[\begin{array}{cc|c} 0 & 0 & 0 \\ 1 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right] + \begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and

$$C_o = \left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & 1 & 1 \end{array}\right] = C_p \bar{C}_o = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Structure Theorem—Observable Version

The transfer function matrix $H(s)$ of system $\dot{x} = Ax + Bu$, $y = Cx + Du$ is given by $H(s) = C(sI - A)^{-1}B + D$. If (A, C) is in the *observer form*, given in (6.74), then $H(s)$ can alternatively be characterized by the Structure

Theorem stated in Theorem 6.25 below. This result will be very useful in the realization of systems, addressed in Chapter 8 and also in the study of observers in Chapter 9.

Let $A = A_o = \bar{A}_o + A_p\bar{C}_o$ and $C = C_o = C_p\bar{C}_o$ as in (6.75) with $|C_p| \neq 0$; let $B = B_o$ and $D = D_o$, and define

$$\widetilde{\Lambda}(s) \triangleq \text{diag}[s^{\nu_1}, s^{\nu_2}, \dots, s^{\nu_p}], \widetilde{S}(s) \triangleq \text{block diag}([1, s, \dots, s^{\nu_i - 1}], i = 1, \dots, p). \tag{6.76}$$

Note that $\widetilde{S}(s)$ is a $p \times n$ polynomial matrix, where $n = \sum_{i=1}^{p} \nu_i$. Now define the $p \times p$ polynomial matrix $\widetilde{D}(s)$ and the $p \times m$ polynomial matrix $\widetilde{N}(s)$ as

$$\widetilde{D}(s) \triangleq [\widetilde{\Lambda}(s) - \widetilde{S}(s)A_p]C_p^{-1}, \quad \widetilde{N}(s) \triangleq \widetilde{S}(s)B_o + \widetilde{D}(s)D_o. \tag{6.77}$$

The following result is the observable version of the *Structure Theorem*. It is the dual of Theorem 6.19 and can therefore be proved using duality arguments. The proof given is direct.

Theorem 6.25. *$H(s) = \widetilde{D}^{-1}(s)\widetilde{N}(s)$, where $\widetilde{N}(s), \widetilde{D}(s)$ are defined in (6.77).*

Proof. First we note that

$$\widetilde{D}(s)C_o = \widetilde{S}(s)(sI - A_o). \tag{6.78}$$

To see this, write $\widetilde{D}(s)C_o = [\widetilde{\Lambda}(s) - \widetilde{S}(s)A_p]C_p^{-1}C_p\bar{C}_o = \widetilde{\Lambda}(s)\bar{C}_o - \widetilde{S}(s)A_p\bar{C}_o$, and also, $\widetilde{S}(s)(sI - A_o) = \widetilde{S}(s)s - \widetilde{S}(s)(\bar{A}_o + A_p\bar{C}_o) = \widetilde{S}(s)(sI - \bar{A}_o) - \widetilde{S}(s)A_p\bar{C}_o = \widetilde{\Lambda}(s)\bar{C}_o - \widetilde{S}(s)A_p\bar{C}_o$, which proves (6.78). We now obtain $H(s) = C_o(sI - A_o)^{-1}B_o + D_o = \widetilde{D}^{-1}(s)\widetilde{S}(s)B_o + D_o = \widetilde{D}^{-1}(s)[\widetilde{S}(s)B_o + \widetilde{D}(s)D_o] = \widetilde{D}^{-1}(s)\widetilde{N}(s)$. ■

Example 6.26. Consider $A_o = \begin{bmatrix} 0 & 2 & 1 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $C_o = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$ of Example 6.24. Here $\nu_1 = 2, \nu_2 = 1$, $\widetilde{\Lambda}(s) = \begin{bmatrix} s^2 & 0 \\ 0 & s \end{bmatrix}$, and $\widetilde{S}(s) = \begin{bmatrix} 1 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Then $\widetilde{D}(s) = [\widetilde{\Lambda}(s) - \widetilde{S}(s)A_p]C_p^{-1} = \left[\begin{bmatrix} s^2 & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} 1 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 0 & 0 \end{bmatrix}\right] \cdot \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}^{-1} = \left[\begin{bmatrix} s^2 & 0 \\ 0 & s \end{bmatrix} - \begin{bmatrix} -s+2 & 1 \\ 0 & 0 \end{bmatrix}\right] \cdot \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} s^2+s-2, & -1 \\ 0 & s \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} s^2+s-1 & -1 \\ -s & s \end{bmatrix}$.

Now if $B_o = [0,\ 1,\ 1]^T, D_o = 0$, and $\widetilde{N}(s) = \widetilde{S}(s)B_o + \widetilde{D}(s)D_o = [s, 1]^T$, then $H(s) = \widetilde{D}^{-1}(s)\widetilde{N}(s) = \frac{1}{s(s^2+s-2)}[s^2+1, 2s^2+s-1]^T = C_o(sI - A_o)^{-1}B_o + D_o$.

6.5 Summary and Highlights

- The standard form for uncontrollable systems is

$$\widehat{A} = Q^{-1}AQ = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix}, \quad \widehat{B} = Q^{-1}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \tag{6.6}$$

where $A_1 \in R^{n_r \times n_r}$, $B_1 \in R^{n_r \times m}$, and (A_1, B_1) is controllable. $n_r < n$ is the rank of the controllability matrix $\mathcal{C} = [B, AB, \ldots, A^{n-1}B]$; i.e.,

$$\operatorname{rank} \mathcal{C} = n_r.$$

- The standard form for unobservable systems is

$$\widehat{A} = Q^{-1}AQ = \begin{bmatrix} A_1 & 0 \\ A_{21} & A_2 \end{bmatrix}, \quad \widehat{C} = CQ = \begin{bmatrix} C_1 \\ 0 \end{bmatrix}, \tag{6.15}$$

where $A_1 \in R^{n_o \times n_o}$, $C_1 \in R^{p \times n_o}$, and (A_1, C_1) is observable. $n_o < n$ is the rank of the observability matrix

$$\mathcal{O} = \begin{bmatrix} \widehat{C} \\ \widehat{C}\widehat{A} \\ \vdots \\ \widehat{C}\widehat{A}^{n-1} \end{bmatrix};$$

i.e.,

$$\operatorname{rank} \mathcal{O} = n_o.$$

- Kalman's Decomposition Theorem.

$$\begin{aligned} &\widehat{A} = Q^{-1}AQ = \begin{bmatrix} A_{11} & 0 & A_{13} & 0 \\ A_{21} & A_{22} & A_{23} & A_{24} \\ 0 & 0 & A_{33} & 0 \\ 0 & 0 & A_{43} & A_{44} \end{bmatrix}, \quad \widehat{B} = Q^{-1}B = \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix}, \\ &\widehat{C} = CQ = [C_1,\ 0,\ C_3,\ 0], \end{aligned} \tag{6.20}$$

where (A_{11}, B_1, C_1) is controllable and observable.

- λ_i is an uncontrollable eigenvalue if and only if

$$\hat{v}_i[\lambda_i I - A, B] = 0, \tag{6.23}$$

where $\hat{v}_i$ is the corresponding (left) eigenvector.

- λ_i is an unobservable eigenvalue if and only if

$$\begin{bmatrix} \lambda_i I - A \\ C \end{bmatrix} v_i = 0, \tag{6.24}$$

where v_i is the corresponding (right) eigenvector.

Controller Forms (for Controllable Systems)

- $m = 1$ case.

$$A_c = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-1} \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \tag{6.36}$$

where

$$\alpha(s) \triangleq \det(sI - A) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0. \tag{6.37}$$

- $m > 1$ case.

$$A_c = [A_{ij}], \qquad i, j = 1, \ldots, m,$$

$$A_{ii} = \begin{bmatrix} 0 & \\ \vdots & I_{\mu_i - 1} \\ 0 & \\ \times \; \times \cdots \times & \end{bmatrix} \in R^{\mu_i \times \mu_i},\ i = j,\ A_{ij} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \\ \times & \times \cdots & \times \end{bmatrix} \in R^{\mu_i \times \mu_j},\ i \neq j,$$

and

$$B_c = \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_m \end{bmatrix}, \qquad B_i = \begin{bmatrix} 0 \cdots 0 & 0 & \cdots & 0 \\ \vdots \quad\ \vdots & \vdots & & \vdots \\ 0 \cdots 0 & 1 & \times \cdots & \times \end{bmatrix} \in R^{\mu_i \times m}. \tag{6.54}$$

An example for $n = 4$, $m = 2$ and $\mu_1 = 2, \mu_2 = 2$ is

$$A_c = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ \times & \times & \times & \times \\ \hline 0 & 0 & 0 & 1 \\ \times & \times & \times & \times \end{array}\right], \quad B_c = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \left[\begin{array}{cc} 0 & 0 \\ 1 & \times \\ \hline 0 & 0 \\ 0 & 1 \end{array}\right].$$

- $$A_c = \bar{A}_c + \bar{B}_c A_m, \quad B_c = \bar{B}_c B_m. \tag{6.55}$$

- *Structure theorem—controllable version*
  $H(s) = N(s)D^{-1}(s)$, where

$$D(s) = B_m^{-1}[\Lambda(s) - A_m S(s)], N(s) = C_c S(s) + D_c D(s). \tag{6.61}$$

Note that

$$(sI - A_c)S(s) = B_c D(s). \tag{6.62}$$

Observer Forms (for Observable Systems)

- $p = 1$ case.

$$A_0 = \begin{bmatrix} 0 \cdots 0 & -\alpha_0 \\ 1 \cdots 0 & -\alpha_1 \\ \vdots \ddots \vdots & \vdots \\ 0 \cdots 1 & -\alpha_{n-1} \end{bmatrix}, \quad C_o = [0, \ldots, 0, 1]. \tag{6.67}$$

- $p > 1$.

$$A_o = [A_{ij}], \quad i, j = 1, \ldots, p,$$

$$A_{ii} = \begin{bmatrix} 0 \cdots 0 & \times \\ I_{\nu_i - 1} & \vdots \\ & \times \end{bmatrix} \in R^{\nu_i \times \nu_i},\ i = j,\ A_{ij} = \begin{bmatrix} 0 & \cdots & 0 & \times \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \times \end{bmatrix} \in R^{\nu_i \times \nu_j},\ i \neq j,$$

and

$$C_o = [C_1, C_2, \ldots, C_p], C_i = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & 1 \\ 0 & \cdots & 0 & \times \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \times \end{bmatrix} \in R^{p \times \nu_i}, \tag{6.74}$$

If (A_c, B_c) is in controller form, $(A_o = A_c^T, C_o = B_c^T)$ will be in observer form.

- $$A_o = \bar{A}_o + A_p \bar{C}_o, \quad C_o = C_p \bar{C}_o. \tag{6.75}$$

- *Structure theorem—observable version*
$H(s) = \widetilde{D}^{-1}(s)\widetilde{N}(s)$, where

$$\widetilde{D}(s) = [\widetilde{\Lambda}(s) - \widetilde{S}(s)A_p]C_p^{-1}, \quad \widetilde{N}(s) = \widetilde{S}(s)B_o + \widetilde{D}(s)D_o. \tag{6.77}$$

Note that

$$\widetilde{D}(s)C_o = \widetilde{S}(s)(sI - A_o). \tag{6.78}$$

6.6 Notes

Special state-space forms for controllable and observable systems obtained by similarity transformations are discussed at length in Kailath [5]. Wolovich [13] discusses the algorithms for controller and observer forms and introduces the Structure Theorems. The controller form is based on results by Luenberger [9] (see also Popov [10]). A detailed derivation of the controller form can also be found in Rugh [12].

Original sources for the Canonical Structure Theorem include Kalman [6] and Gilbert [3].

The eigenvector and rank tests for controllability and observability are called PBH tests in Kailath [5]. Original sources for these include Popov [10], Belevich [2], and Hautus [4]. Consult also Rosenbrock [11], and for the case when A can be diagonalized via a similarity transformation, see Gilbert [3]. Note that in the eigenvalue/eigenvector tests presented herein the uncontrollable (unobservable) eigenvalues are also explicitly identified, which represents a modification of the above original results.

The fact that the controllability indices appear in the work of Kronecker was recognized by Rosenbrock [11] and Kalman [8].

For an extensive introductory discussion and a formal definition of canonical forms, see Kailath [5].

References

1. P.J. Antsaklis and A.N. Michel, *Linear Systems*, Birkhäuser, Boston, MA, 2006.
2. V. Belevich, *Classical Network Theory*, Holden-Day, San Francisco, CA, 1968.
3. E. Gilbert, "Controllability and observability in multivariable control systems," *SIAM J. Control*, Vol. 1, pp. 128–151, 1963.
4. M.L.J. Hautus, "Controllability and observability conditions of linear autonomous systems," *Proc. Koninklijke Akademie van Wetenschappen, Serie A*, Vol. 72, pp. 443–448, 1969.
5. T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
6. R.E. Kalman, "Mathematical descriptions of linear systems," *SIAM J. Control*, Vol. 1, pp. 152–192, 1963.
7. R.E. Kalman, "On the computation of the reachable/observable canonical form," *SIAM J. Control Optimization*, Vol. 20, no. 2, pp. 258–260, 1982.
8. R.E. Kalman, "Kronecker invariants and feedback," in *Ordinary Differential Equations*, L. Weiss, ed., pp. 459–471, Academic Press, New York, NY, 1972.
9. D.G. Luenberger, "Canonical forms for linear multivariable systems," *IEEE Trans. Auto. Control*, Vol. 12, pp. 290–293, 1967.
10. V.M. Popov, "Invariant description of linear, time-invariant controllable systems," *SIAM J. Control Optimization*, Vol. 10, No. 2, pp. 252–264, 1972.
11. H.H. Rosenbrock, *State-Space and Multivariable Theory*, Wiley, New York, NY, 1970.
12. W.J. Rugh, *Linear System Theory*, Second Ed., Prentice-Hall, Englewood Cliffs, NJ, 1996.
13. W.A. Wolovich, *Linear Multivariable Systems*, Springer-Verlag, New York, NY, 1974.

Exercises

6.1. Write software programs to implement the algorithms of Section 6.2. In particular:

(a) Given the pair (A, B), where $A \in R^{n \times n}, B \in R^{n \times m}$ with

$$\operatorname{rank}[B, AB, \ldots, A^{n-1}B] = n_r < n,$$

reduce this pair to the standard uncontrollable form

$$\widehat{A} = PAP^{-1} = \begin{bmatrix} A_1 & A_{12} \\ 0 & A_2 \end{bmatrix}, \widehat{B} = PB = \begin{bmatrix} B_1 \\ 0 \end{bmatrix},$$

where (A_1, B_1) is controllable and $A_1 \in R^{n_r \times n_r}, B_1 \in R^{n_r \times m}$.

(b) Given the controllable pair (A, B), where $A \in R^{n \times n}, B \in R^{n \times m}$ with $\operatorname{rank} B = m$, reduce this pair to the controller form $A_c = PAP^{-1}, B_c = PB$.

6.2. Determine the uncontrollable modes of each pair (A, B) given below by

(a) Reducing (A, B), using a similarity transformation.
(b) Using eigenvalue/eigenvector criteria:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

6.3. Reduce the pair

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 3 & 0 & -3 & 1 \\ -1 & 1 & 4 & -1 \\ 1 & 0 & -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

into controller form $A_c = PAP^{-1}, B_c = PB$. What is the similarity transformation matrix in this case? What are the controllability indices?

6.4. Consider

$$A_c = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-1} \end{bmatrix}, \quad B_c = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Show that

$$\mathcal{C} = [B_c, A_cB_c, \ldots, A_c^{n-1}B_c] = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & c_1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 1 & \cdots & c_{n-3} \\ 0 & 1 & c_1 & \cdots & c_{n-2} \\ 1 & c_1 & c_2 & \cdots & c_{n-1} \end{bmatrix},$$

where $c_k = -\sum_{i=0}^{k-1} \alpha_{n-i-1} c_{k-i-1}$, $k = 1, \dots, n-1$, with $c_0 = 1$. Also, show that

$$\mathcal{C}^{-1} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{n-1} & 1 \\ \alpha_2 & \alpha_3 & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \alpha_{n-1} & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

6.5. Show that the matrices $A_c = PAP^{-1}, B_c = PB$ are as follows:

(a) Given by (6.41) if P is given by (6.42).
(b) Given by (6.44) if $Q(= P^{-1})$ is given by (6.43).
(c) Given by (6.46) if $Q(= P^{-1})$ is given by (6.45).

6.6. Consider the pair (A, b), where $A \in R^{n\times n}, b \in R^n$. Show that if more than one linearly independent eigenvector can be associated with a single eigenvalue, then (A, b) is uncontrollable. *Hint:* Use the eigenvector test. Let $\hat{v}_1, \hat{v}_2$ be linearly independent left eigenvectors associated with eigenvalue $\lambda_1 = \lambda_2 = \lambda$. Notice that if $\hat{v}_1 b = \alpha_1$ and $\hat{v}_2 b = \alpha_2$, then $(\alpha_1 \hat{v}_1 - \alpha_1 \hat{v}_2) b = 0$.

6.7. Show that if (A, B) is controllable, where $A \in R^{n\times n}$, and $B \in R^{n\times m}$, and $\operatorname{rank} B = m$, then $\operatorname{rank} A \geq n - m$.

6.8. Given $A \in R^{n\times n}$, and $B \in R^{n\times m}$, let $\operatorname{rank}\mathcal{C} = n$, where $\mathcal{C} = [B, AB, \dots, A^{n-1}B]$. Consider $\widehat{A} \in R^{n\times n}, \widehat{B} \in R^{n\times m}$ with $\operatorname{rank}\widehat{\mathcal{C}} = n$, where $\widehat{\mathcal{C}} = [\widehat{B}, \widehat{A}\widehat{B}, \dots, \widehat{A}^{n-1}\widehat{B}]$, and assume that $P \in R^{n\times n}$ with $\det P \neq 0$ exists such that

$$P[\mathcal{C}, A^n B] = [\widehat{\mathcal{C}}, \widehat{A}^n \widehat{B}].$$

Show that $\widehat{B} = PB$ and $\widehat{A} = PAP^{-1}$. *Hint:* Show that $(PA - \widehat{A}P)\mathcal{C} = 0$.

6.9. Let $A = \bar{A}_c + \bar{B}_c A_m$ and $B = \bar{B}_c B_m$, where the $\bar{A}_c, \bar{B}_c$ are as in (6.55) with $A_m \in R^{m\times n}, B_m \in R^{m\times m}$, and $|B_m| \neq 0$. Show that (A, B) is controllable with controllability indices μ_i. *Hint:* Use the eigenvalue test to show that (A, B) is controllable. Use state feedback to simplify (A, B) (see Exercise 6.11), and show that the μ_i are the controllability indices.

6.10. Show that the controllability indices of the state equation $\dot{x} = Ax + BGv$, where $|G| \neq 0$ and (A, B) is controllable, with $A \in R^{n\times n}$, $B \in R^{n\times m}$, are the same as the controllability indices of $\dot{x} = Ax + Bu$, within reordering. *Hint:* Write $\bar{\mathcal{C}}_k = [BG, ABG, \dots, A^{k-1}BG] = [B, AB, \dots, A^{k-1}B] \cdot [\text{block diag}\, G] = \mathcal{C}_k \cdot [\text{block diag}\, G]$ and show that the number of linearly dependent columns in $A^k BG$ that occur while searching from left to right in $\bar{\mathcal{C}}_n$ is the same as the corresponding number in $\mathcal{C}_n$.

6.11. Consider the state equation $\dot{x} = Ax + Bu$, where $A \in R^{n\times n}, B \in R^{n\times m}$ with (A, B) controllable. Let the linear state-feedback control law be $u = Fx + Gv, F \in R^{m\times n}, G \in R^{m\times m}$ with $|G| \neq 0$. Show that

(a) $(A + BF, BG)$ is controllable.
(b) The controllability indices of $(A+BF, B)$ are identical to those of (A, B).
(c) The controllability indices of $(A+BF, BG)$ are equal to the controllability indices of (A, B) within reordering. *Hint:* Use the eigenvalue test to show (a). To show (b), use the controller forms in Section 6.4.

6.12. For the system $\dot{x} = Ax + Bu, y = Cx$, consider the corresponding sampled-data system $\bar{x}(k+1) = \bar{A}\bar{x}(k) + \bar{B}\bar{u}(k), \bar{y}(k) = \bar{C}\bar{x}(k)$, where

$$\bar{A} = e^{AT}, \bar{B} = [\int_0^T e^{A\tau} d\tau]B, \quad \text{and} \quad \bar{C} = C.$$

(a) Let the continuous-time system $\{A, B, C\}$ be controllable (observable), and assume it is a SISO system. Show that $\{\bar{A}, \bar{B}, \bar{C}\}$ is controllable (observable) if and only if the sampling period T is such that

$$Im\ (\lambda_i - \lambda_j) \neq \frac{2\pi k}{T}, \text{ where } k = \pm 1, \pm 2, \ldots \text{ whenever } Re\ (\lambda_i - \lambda_j) = 0,$$

where $\{\lambda_i\}$ are the eigenvalues of A. *Hint:* Use the PBH test.
(b) Apply the results of (a) to the double integrator (Example 3.33 in Chapter 3), where $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, and $C = [1, 0]$, and also to $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $C = [1, 0]$. Determine the values of T that preserve controllability (observability).

6.13. (**Spring mass system**) Consider the spring mass given in Exercise 3.37.

(a) Is the system controllable from $[f_1, f_2]^T$? If yes, reduce (A, B) to controller form.
(b) Is the system controllable from input f_1 only? Is it controllable from f_2 only? Discuss your answers.
(c) Let $y = Cx$ with $C = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$. Is the system observable from y? If yes, reduce (A, C) to observer form.