1996

Mathematical systems and their stability

Marc Steven Knobel
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_theses

Recommended Citation

Knobel, Marc Steven, "Mathematical systems and their stability" (1996). *Master's Theses*. 1316.
DOI: <https://doi.org/10.31979/etd.455k-hk89>
https://scholarworks.sjsu.edu/etd_theses/1316

This Thesis is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Theses by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

MATHEMATICAL SYSTEMS AND THEIR STABILITY

A Thesis

Presented to

The Faculty of the Department of Mathematics

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Marc Steven Knobel

August 1996

UMI Number: 1381427

UMI Microform 1381427
Copyright 1996, by UMI Company. All rights reserved.

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

© 1996

Marc Steven Knobel

ALL RIGHTS RESERVED

APPROVED FOR THE DEPARTMENT OF MATHEMATICS

Dr. Jane Day

Dr. Igor Malyshev

Dr. Ho Kuen Ng

APPROVED FOR THE UNIVERSITY

ABSTRACT

MATHEMATICAL SYSTEMS AND THEIR STABILITY

by Marc Steven Knobel

The purpose of this thesis is to present the classical theory of dynamical systems and their stability in a precise mathematical way. Much of this classic theory was written for non-mathematicians, and, consequently, there is often a tendency to over-simplify notation and definitions, and not include proofs of important results.

The basic theory of continuous-time and discrete-time dynamical systems is developed, and linear dynamical systems are considered in detail; moreover, closed form solutions for such linear systems are given. Next, the theory of equilibrium and stability is presented. In addition, the work of Aleksandr Mikhaylovich Lyapunov is examined, which gives ways to study the stability of nonlinear systems. Finally, an in-depth stability analysis of a linear model for eye movement is performed, and the basic ideas of chaos theory are presented.

ACKNOWLEDGMENT

A special thanks to my thesis advisor, Dr. Jane Day, who gave generously of her time and resources, and whose suggestions have contributed to the clarity and exposition of this document. I would also like to thank Dr. Igor Malyshev and Dr. Ho Kuen Ng for their time in considering this thesis and for valuable suggestions. Finally I would like to thank my wife, Tracy, for her patience and support.

TABLE OF CONTENTS

1	BASIC THEORY OF DYNAMICAL SYSTEMS	
1.0	About This Chapter.....	1
1.1	Basic Theory of Continuous Dynamical Systems.....	3
1.1.1	An Existence and Uniqueness Theorem.....	7
1.1.2	A First Solution Method.....	15
1.1.3	Solving Continuous Linear Dynamical Systems.....	17
1.1.4	Continuous Dynamical Systems and Laplace Transform.....	30
1.1.5	Solving Continuous Linear Dynamical Systems: Examples.....	32
1.2	Basic Theory of Discrete Dynamical Systems.....	39
1.2.1	An Existence and Uniqueness Theorem.....	42
1.2.2	Solving Discrete Linear Dynamical Systems.....	43
1.2.3	Discrete Dynamical Systems and Z-transform.....	51
1.2.4	Solving Discrete Linear Dynamical Systems: Examples.....	54
1.3	Considering Higher-Order Equations.....	59
1.3.1	Transforming Higher-Order Differential Equations.....	59
1.3.2	Transforming Higher-Order Difference Equations.....	62
1.4	Generalized Definition of a Dynamical System.....	64
1.4.1	Time Invariant Deterministic Systems.....	67
1.4.2	Connecting Notions of Continuous-Time Dynamical Systems....	76
1.4.3	More General Examples of Deterministic Systems Described by Differential Equations.....	77
1.4.4	Connecting Notions of Discrete-Time Dynamical Systems.....	80

2 STABILITY OF DYNAMICAL SYSTEMS

2.0	About This Chapter.....	81
2.1	Equilibrium and Linearization.....	85
2.1.1	Linearization of Nonlinear Dynamical Systems with Known Equilibrium.....	86
2.1.2	Linearization with Unknown Equilibrium State.....	94
2.1.3	Example: The Satellite Problem.....	96
2.1.4	Asymptotic Stability Via Linearization.....	103
2.2	Types of Stability and Liapunov's Second Method.....	105
2.2.1	Types of Stability.....	107
2.2.2	Definition of a Liapunov Function.....	110
2.2.3	Liapunov Theorems for Discrete Dynamical Systems.....	112
2.2.4	Liapunov Theorems for Continuous Dynamical Systems.....	117
2.2.5	A Special Case: "Constants of Motion".....	123
2.3	Stability Analysis: Time Invariant Linear Systems.....	128
2.3.1	A Stability Representation of the Fundamental Matrix.....	131
2.3.2	Stability of Time Invariant Linear Dynamical Systems.....	136
2.4	BIBO Stability.....	147
2.4.1	BIBO Stability for Continuous Linear Dynamical Systems.....	148
2.4.2	BIBO Stability for Discrete Linear Dynamical Systems.....	151
2.4.3	BIBO Stability for Time Invariant Dynamical Systems.....	153
2.5	Generalizing the Concept of Stability.....	155

3 EXAMPLES FROM THE LITERATURE

3.0	About This Chapter.....	158
3.1	Stability Analysis: A Linear Model for Human Movement.....	160
3.1.1	Derivation of the Linear Model.....	160
3.1.2	Model of the Agonist Muscle.....	162
3.1.3	Model of the Antagonist Muscle.....	164
3.1.4	Extending the Model.....	165
3.1.5	A Discussion of Input to the Model.....	172
3.1.6	Stability Analysis of the Model.....	175
3.2	Chaos In Dynamical Systems.....	182
3.2.1	Definitions and Examples.....	185
3.2.2	Classification of Fixed and Periodic Points.....	196
3.2.3	The Orbit Diagram for the Quadratic Family.....	205
	LITERATURE CITED.....	212

LIST OF TABLES

1.1	Z-transforms of Common Functions.....	52
3.1	Values for Parameters Given in the Model.....	174
3.2	Examples of Definitions.....	186
3.3	Seed/c-value Summary of Eventual Orbit Behavior.....	209

LIST OF FIGURES

0.1	“Black Box” Description of a Dynamical System.....	xii
1.1	Solution of Example 1.3.....	6
1.2	Theorem 1.4 in One Dimension.....	8
1.3	The Translation Operator $z^{-\tau}$	68
2.1	Setup for the Satellite Problem.....	96
2.2	Forms of Stability.....	108
2.3	Geometric Visualization of Definition 2.7.....	109
2.4	Contours C_i in Two Dimensions.....	111
2.5	Regions in the Proof of Theorem 2.11A.....	113
2.6	Regions in the Proof of Theorem 2.15A.....	119
2.7	Contours of Function V in Example 2.19.....	127
3.1	A Simplified Mechanical Model of a Human Muscle.....	162
3.2	Eye Displacement Diagram for an Agonist Muscle Contraction.....	163
3.3	Agonist Muscle Model with Tension-Dampening Dashpot.....	166
3.4	The Linear Homeomorphic Model for Eye Movement.....	168
3.5	Input to the Eye Movement Model.....	173
3.6	Histogram for Example 3.5A.....	188
3.7	Histogram for Example 3.5B.....	189
3.8	Histogram for Example 3.5C.....	190
3.9	Graphical Analysis of the Orbit of $x_0 = 0$ for $F(x) = x^2 + 0.5$	197
3.10	The Graph of $F^2(x) = x^4 - 2x^2$ and an Attracting Fixed Point.....	202
3.11	The Orbit of $x_0 = 0.5$ for $F(x) = x^2 - 1$	203
3.12	Orbit Diagram for $Q_c(x) = x^2 + c$	206

INTRODUCTION

Webster's dictionary defines *dynamics* to be the branch of mechanics dealing with the motion and equilibrium of systems, often under the influence of external forces acting on the system. A *dynamical system* describes this motion as a series of events or *states* of the system, which occur in an ongoing process that evolves over time.

Dynamical systems are important because they apply—in varying degrees—to nearly all observed events. Specific examples occur in physical systems, social systems, and life systems:

- *physical systems*: a satellite in motion around the earth, a computer or electrical system such as a power distribution system, or a robotic control system;
- *social systems*: optimal policies for rental housing, theory of interaction in social groups, or national economic models;
- *life systems*: genetic systems, ecological decay, epidemic control, or population growth. ([Lue], p.1)

To study a particular dynamical system, we begin by creating a model that describes the important aspects of the system. While many dynamical systems can be understood and analyzed without the use of mathematics or a general theory, there are dynamical systems whose complexity requires the tools such a theory provides to gain an understanding of the dynamical system ([Lue], p. 1); in addition, even simple examples of dynamical systems will often give rise to nontrivial computations and analysis.

In Chapter 3, when we study eye movement, we will see that in practice there is typically a trade off between the desire to create a realistic model and the need

to end up with a tractable mathematical problem. In the process, one often must make simplifications to the model that jeopardize the ability of the model to accurately describe the dynamical system.

Henceforth, a reference to a “dynamical system” will mean the mathematical model that describes the system’s behavior.

Toward visualizing a dynamical system, we proceed as in Padulo and Arbib ([PA], pp. 19–20) and introduce the visual aid of a “black box” description of a dynamical system, which is given in Figure 0.1.

Figure 0.1 “Black box” description of a dynamical system

The input and output are usually taken to be vector valued functions $\mathbf{u}(t) = [u_1(t) \ \cdots \ u_m(t)]^T$ and $\mathbf{y}(t) = [y_1(t) \ \cdots \ y_p(t)]^T$, respectively. The inputs could be neural signals from the brain, voltages, population levels, market conditions, computer instructions, etc. The outputs could, correspondingly, be eye movement, electric motor speeds, world hunger levels, market value, screen outputs that create the visual display on a computer screen, etc.

Typically we are interested in the “internal” process that determines how inputs affect outputs. It is clear that in the case of eye movement, for example,

eye position is dependent not only upon brain input, but on the properties of the muscle systems that move the eye and the position of the eye in the previous instant. That is, an “internal description” of the dynamical system must be given if we are to explain the processes converting inputs to outputs. Hence, we must include in our model an additional feature that describes the “internal state” of the system, which is usually described by the vector function $\mathbf{x}(t) = [x_1(t) \cdots x_n(t)]^T$.

Since we are interested in describing some process as it evolves with respect to time, the next step in our development is to choose a time set T that is appropriate to the situation. In the case of neural signals or voltages, which may change at any instant in time, T is typically chosen to be the real numbers or some interval of real numbers, so that time unfolds in a smooth connected pattern, and we may associate a particular input, state, or output to any instant in time. Dynamical systems with time sets that form intervals of real numbers are referred to as “continuous-time” or “continuous” for short, if the context is clear.

In the case of population levels, market conditions, or computer programs, the inputs, states, and outputs occur in discrete time periods, i.e., daily, monthly, yearly, or in the case of computer programs, in milliseconds or nanoseconds. In this case, the integers or natural numbers are usually used for the time set T , and a dynamical system with such a time set is referred to as “discrete-time” or “discrete” for short, if the context is clear.

In practice the mathematical translation of a dynamical system usually takes one of two forms: (i) the time set is real and the dynamical system is modeled as a system of differential equations, or (ii) the time set is the natural numbers and the system is modeled by a system of difference equations.

If the dynamical system is continuous-time, then in general it is evolving instantaneously; our observation of this evolution gets translated into a differential equation describing the rate of change of the state with respect to time as it responds to inputs and determines corresponding outputs. To “solve” the dynamical system will be to solve the differential equation describing the “internal state” $x(t)$ of the dynamical system at any instant of time.

For the case of discrete time, dynamical behavior is usually described by equations relating the value of a dynamic variable at one time to the values at adjacent times. These equations are called *difference equations* ([Lue], p. 2). To “solve” the dynamical system here will be to solve the difference equation to obtain a mathematical description of the “internal state” $x(t)$ at any (discrete) time t —rather than in terms of the state at a previous time $t - 1$, say, given by the difference equation.

In both cases, a mathematical form of the state can then be used to study the behavior of the dynamical system as it evolves with respect to time, responding to various inputs and determining the corresponding outputs. Indeed, in using computers for the analysis of real-world continuous-time dynamical systems, we are forced to treat them actually as discrete dynamical systems, since we can sample only finitely many inputs and outputs of the dynamical system over any time interval ([PA], p. 20).

Moreover, it is in this context of taking discrete “samplings” of the internal state $x(t)$ as it evolves with respect to (discrete) time, its “trajectory,” that has led to many unexpected surprises. During the 1950s and 60s, a number of ecologists studied first-order nonlinear difference equations of the form

$$N(t + 1) = R_0 N(t) e^{-a N(t)}$$

and observed regions of stable behavior but also witnessed “irregular behavior” that is now referred to as “chaos” [May]. Furthermore, there is much modern work that is dedicated to predicting long-term behavior of the state when discrete time sampling of trajectories appears chaotic. For a nice overview of current research, with many references to specific literature, see [May]. In Chapter 3, we will consider some simple examples of this chaotic behavior, as a prelude to study in this fascinating area.

However, the main focus of this thesis is to present the classical theory of dynamical systems and their stability in a precise mathematical way. Much of this classic theory was written for non-mathematicians, and consequently, there is often a tendency to over simplify notation and definitions, and not include proofs of important results—to the detriment of mathematical precision. Therefore in preparing this manuscript, much attention was given to the creation of precise definitions, elucidating examples, and complete proofs of all stated theorems.

In Chapter 1, we develop the basic theory of continuous-time and discrete-time dynamical systems, and in particular, consider linear dynamical systems in detail; moreover, we will give closed form solutions for such linear systems. Chapter 2 is dedicated to the theory of equilibrium and stability; in addition, the work of Aleksandr Mikhaylovich Lyapunov is included, which gives ways to study the stability of nonlinear systems. Chapter 3 presents an in depth stability analysis of a linear “homeomorphic” model for eye movement that appeared in a paper written in 1980, as well as an introduction to “chaos.”

CHAPTER 1

BASIC THEORY OF DYNAMICAL SYSTEMS

1.0 ABOUT THIS CHAPTER

In this first chapter, we develop the basic theory of dynamical systems that will be used extensively in Chapters 2–3 when we consider the stability of dynamical systems.

Section 1.1 develops the theory of first order continuous-time dynamical systems. The section opens with basic definitions, presents an important existence and uniqueness theorem for solutions of continuous-time systems, describes an iterative method to solve such systems, and then two methods for finding closed form solutions of a linear continuous-time dynamical system are presented. The first solution technique is based on computation of the *fundamental matrix* for continuous-time systems and is presented in Section 1.1.3. The second solution technique uses Laplace transform and is presented in Section 1.1.4.

In Section 1.2, first order discrete-time dynamical systems are considered, and the development parallels that given in Section 1.1, where we again present definitions and an important existence and uniqueness result for solutions of discrete-time dynamical systems. In the linear case, two solution methods are again given: one is based on computation of the fundamental matrix for discrete-time systems and is described in Section 1.2.2; the other solution method is based on Z-transform and is presented in Section 1.2.4. Similarities between the continuous-time and discrete-time solutions for linear systems will also be discussed.

Section 1.3 describes a technique by which higher-order differential and difference equations can be reduced to higher-dimensional first-order differential and difference equations. This technique will allow us then to apply the theory developed in Sections 1.1–1.2 to higher-order systems.

Finally, in Section 1.4 we extend the definitions and concepts given so far to allow for more general contexts than given in Sections 1.1–1.2. It is helpful to have seen the more simplified notation and contexts first. We will also show how the definitions of continuous-time and discrete-time dynamical systems given in Sections 1.1–1.2 are special cases of this more general definition.

1.1 BASIC THEORY OF CONTINUOUS DYNAMICAL SYSTEMS

In this section we develop the underlying mathematical theory of continuous dynamical systems that will be used in this paper. The working definition that will be used for the continuous case is typical of ones given in engineering literature. As we will see this definition simplifies notation; however, Kalman and Bertram point out that such a simplification can lead to confusion ([KB], p. 373). For our purposes the simpler form will be sufficient and will make the notation in this paper more familiar to readers of engineering literature. Toward understanding the concerns of Kalman and Bertram, the working definition given below is followed with clarifying comments.

Our working definition of a continuous dynamical system is a synthesis of the definitions given by Kalman and Bertram and Szidarovsky and Bahill ([SB], p. 84 and [KB], p. 373), and is as follows:

1.1 Definition (Dynamical system with continuous time scale) A dynamical system with continuous time scale and state-space X is given by

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \quad (1.1)$$

$$\mathbf{y}(t) = \mathbf{g}(t, \mathbf{x}(t)) \quad (1.2)$$

where $T \subseteq \mathcal{R}$, $\mathbf{x}(t) \in X \subseteq \mathcal{R}^n$, $\mathbf{u}(t) \in U \subseteq \mathcal{R}^m$, $\mathbf{f}: T \times X \times U \rightarrow \mathcal{R}^n$, and

$\mathbf{g}: T \times X \rightarrow \mathcal{R}^p$. Often $T = (0, \infty)$ or $[0, \infty)$. Observe that the notation in equation (1.1) is also given, equivalently, by the set of **n scalar differential equations**

$$\dot{x}_i = dx_i / dt = f_i(t, x_1(t), \dots, x_n(t), u_1(t), \dots, u_m(t)) \quad (1.3)$$

where $i = 1, \dots, n$, and f_i is the i th component of $\mathbf{f}$. Similarly, equation (1.2) is given, equivalently, by the set of **p scalar equations**

$$y_i = g_i(t, x_1(t), \dots, x_n(t)) \quad (1.4)$$

The sets X and U are called the *state space* and *input space*, respectively. The first equation, (1.1), is called the *state transition equation* and the second equation, (1.2), is referred to as the *output equation*. At a time t , the vector $\mathbf{x}(t)$ is called a *state* of the system (1.1)–(1.2), and its components x_i are called the *state variables*. The vector $\mathbf{u}(t)$ is called the *input* (or *forcing function*) of the system. ■

The function $\mathbf{f}$ is always assumed to be “sufficiently smooth” so that the system (1.1)–(1.2) has a unique solution for some initial state $\mathbf{x}_0$ and initial time t_0 . Below, this solution will be written as $\mathbf{x}(t)$, but first, we must make precise what we mean when we say a solution of system (1.1)–(1.2).

1.2 Definition (A solution of system (1.1)–(1.2)) A **solution** of the continuous system (1.1)–(1.2) is defined as follows. Begin by fixing $\mathbf{u}(t)$ and note that with $\mathbf{u}$ fixed we may view equation (1.1) as a function of $\mathbf{x}$ and t :

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)) \quad (1.6)$$

equation (1.2) is unchanged. Next, assume there exists a *unique* vector-valued function $\bar{\phi}: T \times X \times T \rightarrow X$ such that for any fixed t_0 and $\mathbf{x}_0$

$$(i) \quad \bar{\phi}(t_0; \mathbf{x}_0, t_0) = \mathbf{x}_0$$

$$(ii) \quad \bar{\phi} \text{ is differentiable in } t \text{ and } \dot{\bar{\phi}} = d\bar{\phi}(t; \mathbf{x}_0, t_0) / dt = \mathbf{f}(t, \bar{\phi}(t; \mathbf{x}_0, t_0))$$

in some open interval about t_0 . Then when t_0 and $\mathbf{x}_0$ are fixed, the function $\mathbf{x}(t) = \bar{\phi}(t; \mathbf{x}_0, t_0)$ is called a *solution* of the system (1.6), and it is also called the *transition function*. Observe finally that $\mathbf{x}(t_0) = \mathbf{x}_0$. ■

While the notation in (i)–(ii) above makes clear the notion of what a solution of a continuous dynamical system is, it is a bit cumbersome. Thus, when t_0 , $\mathbf{x}_0$, and $\mathbf{u}(t)$ have been fixed, we will let $\mathbf{x}(t)$ stand for both the variable and the solution of equation (1.6)—that is, a solution of (1.6), passing through the point

$\mathbf{x}(t_0)$ —while keeping in mind that this is an abuse of notation because $\bar{\phi}$ is dependent not only on t , but on $\mathbf{x}_0$ and t_0 as well.

Remark

In most cases, a reference to *the* state $\mathbf{x}(t)$ of a continuous dynamical system will mean the unique solution of differential equation (1.1)—as given by Definition 1.2—having initial state $\mathbf{x}_0 = \mathbf{x}(t_0)$. ■

To help visualize the ideas above, we first consider an example:

1.3 Example Consider the continuous dynamical system given by

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \\ &= \begin{bmatrix} f_1(t, x_1(t), x_2(t), u_1(t), u_2(t)) \\ f_2(t, x_1(t), x_2(t), u_1(t), u_2(t)) \end{bmatrix} \\ &= \begin{bmatrix} \frac{x_1(t)}{t} + u_1(t) \\ \frac{x_2(t)}{t} + u_2(t) \end{bmatrix}\end{aligned}\tag{1.7}$$

and

$$\begin{aligned}y(t) &= g(t, \mathbf{x}(t)) \\ &= g(t, x_1(t), x_2(t)) \\ &= x_1(t) + x_2(t)\end{aligned}\tag{1.8}$$

where $T = (0, \infty)$, $X = \mathbb{R}^2$, $U = \mathbb{R}^2$, and y is a scalar function of t . In Section 1.1.5 we will show that with $u_1(t) = u_2(t) \equiv 1$ for all $t > 0$, $t_0 = 1$, and $\mathbf{x}(t_0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

that the solution $\mathbf{x}$ of this dynamical system is given by

$$\mathbf{x}(t) = \begin{bmatrix} t(1 + \ln t) \\ t(1 + \ln t) \end{bmatrix}$$

whence $y(t) = 2t(1 + \ln t)$ by equation (1.8). This solution is sketched in Figure 1.1 below. ■

Finally, notice that we are assuming that every continuous dynamical system can be expressed in terms of a first order differential equation as given by equation (1.1), which may seem restrictive. We will see that a wide class of dynamical systems can be modeled and solved using this definition as a starting point. In Section 1.4 more general dynamical systems are considered, that are not defined in terms of a differential equation.

Figure 1.1 Solution of Example 1.3

1.1.1 An Existence and Uniqueness Theorem

As remarked above, we will always want our dynamical system to have a unique solution when t_0 , $\mathbf{x}_0 = \mathbf{x}(t_0)$, and input $\mathbf{u}(t)$ are given. In this first subsection we prove an important theorem that establishes sufficient conditions needed to guarantee this. Throughout this section, we will take equation (1.6) as the differential equation for our dynamical system (i.e., the input has already been fixed). The theorem and proof parallel those given in Szidarovsky and Bahill ([SB], pp. 46-50).

1.4 Theorem Suppose we are given the **initial value problem**:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)); \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.9)$$

where $t \in T = [t_0 - a, t_0 + a]$. Let $S = \{\mathbf{x} \in X : \|\mathbf{x} - \mathbf{x}_0\| \leq b\}$, where $a, b > 0$ are given, and $\|\bullet\|$ is some vector norm. Suppose further that $\mathbf{f}$ is continuous on $T \times S$ and that it satisfies the Lipschitz Condition, that is, there exists a constant $L > 0$ such that

$$\|\mathbf{f}(t, \mathbf{x}_1) - \mathbf{f}(t, \mathbf{x}_2)\| \leq L \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|$$

for all $t \in T$ and $\mathbf{x}_1, \mathbf{x}_2 \in S$. Let $Q = \max_{(t, \mathbf{x}) \in T \times S} \|\mathbf{f}(t, \mathbf{x})\|$. (Note that the continuity of $\mathbf{f}$ and compactness of $T \times S$ imply that Q exists.) Lastly, define $\alpha \equiv \min\{a, \frac{b}{Q}, \frac{1}{L} - \varepsilon\}$, where $\varepsilon > 0$ is small enough so that $\frac{1}{L} - \varepsilon > 0$. Then equation (1.9) has a unique solution on the interval $T_0 = [t_0 - \alpha, t_0 + \alpha] \subseteq T$. ■

Before we present a proof of this theorem, two examples are given next that show (i) that a solution of equation (1.9) may not exist, and (ii) that hypotheses of the theorem cannot be weakened to drop the Lipschitz Condition, that is, that continuity alone is not a strong enough assumption to guarantee uniqueness.

Figure 1.2 Theorem 1.4 in one dimension

1.5 Example Let $T = \mathbb{R} = X$, $n = 1$, and

$$f(t, x(t)) = g(t) = \begin{cases} 0, & \text{if } t \text{ is rational} \\ 1, & \text{if } t \text{ is irrational} \end{cases}$$

This function is often referred to as the “Dirichlet function” in books on real analysis. One can easily show that $\dot{x}(t) = g(t)$ has no solution by using the denseness of the real number line to take integration partitions of $[a, b] \subseteq T$ with rational endpoints and then with irrational endpoints. Then in the limit we obtain 0 and $(b-a)$, respectively, for the two Riemann sums, whence $g(t)$ is not integrable. ■

In Example 1.5, we violated the continuity condition of Theorem 1.4. In Example 1.6, we will see that adding a continuity condition does not guarantee uniqueness.

1.6 Example Consider the following one-dimensional initial-value problem

$$\dot{x} = \sqrt{|x|}; \quad x(1) = 0$$

where $T = X = \mathbb{R}$. Note that the zero function $x_1(t) \equiv 0$ satisfies the differential equation and the initial condition, and hence is a solution of the initial-value problem. Now consider the function

$$x_2(t) = \begin{cases} 0, & \text{if } t \leq 1 \\ \left(\frac{t-1}{2}\right)^2, & \text{if } t > 1 \end{cases}$$

Thus, $x_2(1) = 0$ and

$$\dot{x}_2(t) = \begin{cases} 0, & (= \sqrt{|x_2|}), \text{ if } t \leq 1 \\ \left(\frac{t-1}{2}\right), & (= \sqrt{|x_2|}) \text{ if } t > 1 \end{cases}$$

So x_2 also satisfies the initial-value problem; therefore, since $x_1 \neq x_2$, we see that the solution of this initial-value problem is not unique. ■

The remainder of this section is devoted to proving Theorem 1.4. The proof will consist of two main steps:

- Step 1: We will show that the initial-value problem (1.9) is equivalent to a certain fixed-point problem in a metric space.
- Step 2: The fixed point problem in Step 1 will be shown to satisfy the conditions of another theorem regarding mappings in metric spaces—which will complete the proof.

Our first task is to prove the metric space theorem in Step 2. Toward this end, we begin with a definition of a contraction mapping in a metric space.

1.7 Definition Let (M, ρ) be a metric space and $D \subseteq M$. A mapping $A: D \rightarrow D$ is a **contraction** if and only if there exists K , $0 \leq K < 1$, such that for all $x, y \in D$,

$$\rho(A(x), A(y)) \leq K \cdot \rho(x, y)$$

Note that a contraction is continuous. ■

We are now ready to prove the key metric space theorem to be used in Theorem 1.4.

1.8 Theorem (Fixed Point Theorem) Suppose (M, ρ) is a complete metric space, D is a closed subset of M , and $A: D \rightarrow D$ is a contraction, that is, $\|A(x)\| < \|x\|$ for all $x \in D$. Then there exists a unique $x^* \in D$ such that $A(x^*) = x^*$. Furthermore, x^* is the limit of the iteration sequence

$$x_1 = A(x_0), x_2 = A(x_1), \dots \quad (1.10)$$

no matter what $x_0 \in D$ is chosen as the initial point.

Proof

The proof is essentially that given in Szidarovsky and Bahill ([SB], pp. 11–13). There are several steps.

(i) To see the iteration sequence (1.10) is a Cauchy sequence: By repeated application of the triangle inequality, for $m > n$,

$$\rho(x_n, x_m) \leq \rho(x_n, x_{n+1}) + \rho(x_{n+1}, x_{n+2}) + \dots + \rho(x_{m-1}, x_m) \quad (1.11)$$

From the definition of a contractive mapping, we have for all $k \geq 1$,

$$\begin{aligned} \rho(x_k, x_{k+1}) &= \rho(A(x_{k-1}), A(x_k)) \\ &\leq K \cdot \rho(x_{k-1}, x_k) \\ &= K \cdot \rho(A(x_{k-2}), A(x_{k-1})) \\ &\leq K^2 \cdot \rho(x_{k-2}, A(x_{k-1})) \\ &\leq \dots \\ &\leq K^k \cdot \rho(x_0, x_1) \end{aligned}$$

It now follows, using the last inequality above and substituting into inequality (1.11) that

$$\begin{aligned}
\rho(x_n, x_m) &\leq \rho(x_0, x_1)[K^n + K^{n+1} + \dots + K^{m-1}] \\
&\leq \rho(x_0, x_1)[K^n + K^{n+1} + \dots] \\
&= \rho(x_0, x_1)K^n \sum_{j=0}^{\infty} K^j \\
&= \rho(x_0, x_1)K^n \left(\frac{1}{1-K} \right) \\
&= \rho(x_0, x_1) \left(\frac{K^n}{1-K} \right)
\end{aligned}$$

Finally, since $0 \leq K < 1$, $\frac{K^n}{1-K} \rightarrow 0$ as $n \rightarrow \infty$, which implies that $\rho(x_n, x_m) \rightarrow 0$ as long as $m > n$ and $n \rightarrow \infty$. Thus the iteration sequence (1.11) is a Cauchy sequence, as desired. Since (M, ρ) is complete, it follows that this sequence converges to an element $x^* \in M$.

(ii) To see $x^* \in D$: Since D is closed by hypothesis, we have that $M - D$ is open so every point of $M - D$ is an interior point. For a contradiction, suppose $x^* \notin D$. Then $x^* \in M - D$ and hence is an interior point. But $x_n \in D$, for all $n \geq 1$ by hypothesis; hence, $x_n \notin M - D$ for every $n \geq 1$. Finally, choose any open neighborhood about x^* of radius $r > 0$. Since x^* is the limit of the sequence $\{x_n\}$, we can find some x_α inside this neighborhood. Since the size of this neighborhood is arbitrary, it now follows that x^* is a boundary point of M —which contradicts the fact that x^* is an interior point.

(iii) To see x^* is a fixed point of A : Since mapping A is a contraction, A is bounded and hence continuous. Let $n \rightarrow \infty$ in the iteration sequence $x_{n+1} = A(x_n)$. Then $x_{n+1} \rightarrow x^*$, $A(x_n) \rightarrow A(x^*)$ (recall A is continuous), and we have that the iteration sequence (1.11) converges to x^* and $A(x^*)$. The uniqueness of the limit of a convergent sequence then implies that $x^* = A(x^*)$. Therefore, x^* is a fixed point of A .

(iv) To see x^* is the unique solution of the given fixed-point problem: For a contradiction, suppose not. Suppose x^* and x^{**} and distinct solutions to the fixed-point problem $x = A(x)$. Since A is a contraction mapping,

$$\begin{aligned}\rho(x^*, x^{**}) &= \rho(A(x^*), A(x^{**})) \\ &\leq K \cdot \rho(x^*, x^{**})\end{aligned}$$

Since $x^* \neq x^{**}$, we can divide both sides of the inequality above by the positive quantity $\rho(x^*, x^{**})$ to get $1 \leq K$, which contradicts the hypothesis ($0 \leq K < 1$). Therefore, x^* is the unique solution of the fixed-point problem, completing Step (iv). ■

Finally, a proof of Theorem 1.4 can be given, which will complete our work in this subsection.

Proof of Theorem 1.4 Again, the proof consists of several steps.

(i) First, we prove the initial problem of the hypothesis is equivalent to the fixed point problem

$$x(t) = x_0 + \int_{t_0}^t f(\tau, x(\tau)) d\tau \quad (1.12)$$

Proof of (i): ($\Rightarrow$) Begin by assuming that x is a solution of equation (1.9). Integrating on both sides of (1.9) on the interval $[t_0, t]$ we obtain the following equation:

$$\int_{t_0}^t \dot{x} = \int_{t_0}^t f(\tau, x(\tau)) d\tau$$

Since x is a solution of the initial value problem we can rewrite the equation above as follows:

$$x(t) - x(t_0) = \int_{t_0}^t f(\tau, x(\tau)) d\tau \quad (1.13)$$

The initial condition of equation (1.9), $x(t_0) = x_0$, implies that equation (1.13) can be rewritten as follows:

$$x(t) = x_0 + \int_{t_0}^t f(\tau, x(\tau)) d\tau \quad (1.14)$$

which is equation (1.12).

($\Leftarrow$) Differentiate both sides of Equation (1.12) and use the fact that $f(\tau, x(\tau))$ is continuous in τ (so the Fundamental Theorem of Calculus applies) to obtain

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

Let $t = t_0$ and we have from equation (1.12) that $x(t_0) = x_0 + 0 = x_0$.

(ii) Next we show that the fixed-point problem of equation (1.12) satisfies all the conditions of Theorem 1.8, which implies that there exists a unique solution.

Define M as the set of continuous functions defined on T_0 , with metric

$$\rho(x_1, x_2) = \max_{t \in T_0} \|x_1(t) - x_2(t)\|$$

Select $D = \{x(t) : x \in M, x(t_0) = x_0, \|x(t) - x_0\| \leq b \text{ for all } t \in T_0\}$. Define a mapping A by

$$A(x)(t) = x_0 + \int_{t_0}^t f(\tau, x(\tau)) d\tau \quad (1.15)$$

where the right side of equation (1.15) gives the value of function $A(x)$ at t . It is true that (M, ρ) is a metric space and that D is a closed subset of M (see [Mun]).

Furthermore, we claim that A maps D to D and is a contraction.

To see that A maps D to D : Let $x \in D$. Since f and x are continuous, the function

$$g(t) = \int_{t_0}^t f(\tau, x(\tau)) d\tau$$

is continuous and hence differentiable with derivative $f(t, x(t))$ —by the Fundamental Theorem of Calculus. Since the derivative of a constant is zero, we see that the right-hand side of equation (1.15) is differentiable with derivative $f(t, x(t))$. Since differentiability implies continuity, function $A(x)$ is continuous. Clearly $A(x)(t_0) = x_0$. Furthermore, for all $t_0 \in T_0$

$$\begin{aligned}
\|A(x)(t) - x_0\| &= \left\| \int_{t_0}^t f(\tau, x(\tau)) d\tau \right\| \\
&\leq \int_{t_0}^t \|f(\tau, x(\tau))\| d\tau \\
&\leq Q \cdot |t - t_0| \\
&\leq Q \cdot \alpha \\
&\leq b
\end{aligned}$$

because $\alpha \leq \frac{b}{Q}$. Therefore $A(x) \in D$.

To see mapping $A: D \rightarrow D$ is a contraction: Suppose $x_1, x_2 \in D$. Then

$$\begin{aligned}
\rho(A(x_1(t)), A(x_2(t))) &= \max_{t \in T_0} \left\| x_0 + \int_{t_0}^t f(\tau, x_1(\tau)) d\tau - x_0 - \int_{t_0}^t f(\tau, x_2(\tau)) d\tau \right\| \\
&= \max_{t \in T_0} \left\| \int_{t_0}^t [f(\tau, x_1(\tau)) - f(\tau, x_2(\tau))] d\tau \right\| \\
&\leq \max_{t \in T_0} \left| \int_{t_0}^t \|f(\tau, x_1(\tau)) - f(\tau, x_2(\tau))\| d\tau \right| \\
&\leq \left| \int_{t_0}^t L \cdot \|x_1(\tau) - x_2(\tau)\| d\tau \right| \quad \{\text{Lipschitz condition}\} \\
&\leq \max_{t \in T_0} \left| \int_{t_0}^t L \cdot \max_{\tau \in T_0} \|x_1(\tau) - x_2(\tau)\| d\tau \right| \\
&= \max_{t \in T_0} \left| \int_{t_0}^t L \cdot \rho(x_1, x_2) d\tau \right| \\
&= L \cdot \rho(x_1, x_2) \max_{t \in T_0} \left| \int_{t_0}^t 1 d\tau \right| \\
&= L \cdot \rho(x_1, x_2) \cdot \alpha \quad \{\text{max occurs when } t = t_0 + \alpha\} \\
&= L\alpha \cdot \rho(x_1, x_2)
\end{aligned}$$

Finally, note that $L\alpha \leq L\left(\frac{1}{L} - \varepsilon\right) < 1$, whence A is a contraction. ■

1.1.2 A First Solution Method

Szidarovsky and Bahill ([SB], p. 50) remark that the proofs of the theorem above and of Theorem 1.8 suggest an iteration method for solving the initial problem given by equation (1.9). Begin by selecting an initial approximation for the solution as $x_0(t) \equiv x_0$. The constant function x_0 is clearly an element of the set D of the proof above. Use the recursion

$$x_{k+1}(t) = x_0 + \int_{t_0}^t f(\tau, x_k(\tau)) d\tau \quad (1.16)$$

to construct the iteration sequence of the proof, which by Theorem 1.2 converges uniformly on T_0 to the unique solution of the initial value problem. This method is demonstrated in the following example.

1.9 Example It is easy to verify that the unique solution of the one-dimensional initial value problem

$$\dot{x} = x, \quad x(0) = 1,$$

is $x(t) = e^t$. To find it by iteration observe that equation (1.16) becomes

$$x_{k+1}(t) = 1 + \int_0^t x_k(\tau) d\tau.$$

As suggested above, select $x_0(t) \equiv x_0 = 1$, whereby we get the following iteration sequence:

$$\begin{aligned} x_1(t) &= 1 + \int_0^t 1 d\tau = 1 + t, \\ x_2(t) &= 1 + \int_0^t (1 + \tau) d\tau = 1 + t + \frac{t^2}{2!}, \\ &\vdots \\ x_k(t) &= \sum_{i=0}^k \frac{t^i}{i!} \quad \{\text{i.e., by finite induction}\} \end{aligned}$$

We see that this last equation is the k th degree Taylor polynomial of e^t , which confirms that this sequence converges to e^t . ■

Remark

It should be noted that slow convergence or difficulty of calculating the integrals limit the usefulness of this method. Using computer methods, however, we can always use this technique to obtain a numerical approximation to the solution—that is, when other solution methods fail or are more difficult.

In what follows, various techniques will be applied to obtain a closed form solution when possible. Although we will develop only a limited number of closed form solutions, we will see that a surprising number of dynamical systems can be solved using them.

1.1.3 Solving Continuous Linear Dynamical Systems

In this subsection, a closed form solution for continuous *linear* systems of the form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x} + \mathbf{f}(t) \quad (1.17)$$

will be derived, where $\mathbf{A}(t)$ is an $n \times n$ matrix and $\mathbf{f}(t)$ is an n -element vector. Usually $\mathbf{f}(t) = \mathbf{B}(t)\mathbf{u}(t)$, where $\mathbf{B}(t)$ is an $n \times n$ matrix, and the input $\mathbf{u}$ is a function from $T \rightarrow \mathbb{R}^n$. The elements a_{ij} , b_{ij} of $\mathbf{A}$ and $\mathbf{B}$, respectively, are assumed to be continuous functions of t on a closed finite interval $T \subseteq \mathbb{R}$, and the elements u_i of $\mathbf{u}$ are assumed to be piecewise continuous on T .

First, note that the system in equation (1.17) fits the conditions of Definition 1.1, that is, fits our definition of a continuous dynamical system. Here, we will take $X = U = \mathbb{R}^n$. Moreover, under the assumptions above, the right-hand side of equation (1.17) is continuous. If we can show that the Lipschitz condition holds, then by Theorem 1.4 there exists a unique solution to equation (1.17) for each initial condition $\mathbf{x}_0 \in \mathbb{R}^n$. Theorem 1.12 below shows how to construct it.

1.10 Lemma The right-hand side of equation (1.17) satisfies the Lipschitz condition ([SB], p. 52).

Proof Choose a vector norm and a compatible matrix norm. Since $\|\mathbf{A}(t)\|$ is continuous and defined over a closed and bounded interval, it is bounded by some constant $L > 0$. Hence,

$$\begin{aligned} \|\mathbf{A}(t)\mathbf{x}_1 + \mathbf{f}(t) - (\mathbf{A}(t)\mathbf{x}_2 + \mathbf{f}(t))\| &= \|\mathbf{A}(t)(\mathbf{x}_1 - \mathbf{x}_2)\| \\ &\leq \|\mathbf{A}(t)\| \cdot \|\mathbf{x}_1 - \mathbf{x}_2\| \\ &\leq L \cdot \|\mathbf{x}_1 - \mathbf{x}_2\| \end{aligned}$$

for all $t \in I$, and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, where L is an upper bound on $\|\mathbf{A}(t)\|$. The first inequality is valid because the matrix and vector norms here are compatible. ■

For the remainder of this section we will be concerned with constructing a particular matrix, which we will call the fundamental (or transition) matrix of system (1.17). We will see that this matrix plays a central role in computing the solution of the system.

Begin by noting that by Lemma 1.10 and Theorem 1.4, the initial value problem

$$\dot{\mathbf{x}} = A(t)\mathbf{x} + \mathbf{f}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.18)$$

has a unique solution $\mathbf{x}(t)$ for any fixed $t_0 \in T$ and $\mathbf{x}_0 \in \mathbb{R}^n$. The first step in obtaining this solution will be to construct the matrix given in the following definition.

1.11 Definition (Fundamental, or transition matrix) Consider the homogeneous linear dynamical system

$$\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) \quad (1.19)$$

For $t_0 \in T$, and fixed k , $1 \leq k \leq n$, let

$$\mathbf{x}_0 = \mathbf{x}(t_0) = \mathbf{e}_k = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \{k^{\text{th}} \text{ position}\}$$

As discussed above, system (1.19) and each initial condition above has a unique solution associated with it, which we will denote as follows:

$$\mathbf{x}_k(t) = \tilde{\Phi}(t; \mathbf{e}_k, t_0); \quad k = 1, \dots, n$$

Create such $\mathbf{x}_k(t)$, for each k , $1 \leq k \leq n$, and construct the $n \times n$ matrix

$$\Phi(t, t_0) = [\mathbf{x}_1(t) \cdots \mathbf{x}_n(t)]$$

The matrix $\Phi(t, t_0)$ is called the **fundamental, or transition matrix**, of system (1.19). ■

We are now ready to present the first major result of this section.

1.12 Theorem Suppose $\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t)$, where each $a_{ij}(t)$ is continuous. Fix $t_0 \in T$ and define $\Phi(t, t_0)$ as above. Then

A) $\Phi(t, t_0)$ is invertible; and

B) the general solution of the dynamical system of the hypothesis is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{c} \quad (1.20)$$

where $\mathbf{c}$ is some constant vector ([SB], pp. 53–54).

Proof The proof is essentially that in [SB], and involves three main steps.

(i): We begin by establishing that equation (1.20) satisfies the homogeneous equation (1.19) for arbitrary $\mathbf{c} \in \mathbb{R}^n$. If we define $\mathbf{c} = (c_k)$, then it follows from equation (1.20) that

$$\begin{aligned} \mathbf{x}(t) &= [\mathbf{x}_1(t) \cdots \mathbf{x}_n(t)] \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \\ &= \sum_{k=1}^n c_k \mathbf{x}_k(t) \end{aligned}$$

which implies that

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \sum_{k=1}^n c_k \dot{\mathbf{x}}_k(t) \\ &= \sum_{k=1}^n A(t) \mathbf{x}_k(t) c_k \\ &= A(t) \sum_k c_k \mathbf{x}_k(t) \\ &= A(t) \mathbf{x}(t) \end{aligned}$$

Thus, $\mathbf{x}(t) = \Phi(t, t_0)\mathbf{c}$ satisfies equation (1.19) as desired.

(ii) To see $\Phi(t, t_0)$ is invertible: We show that the columns are linearly independent for all $t \in T$. For a contradiction, suppose not. Then there exists $t_1 \in T$ such that

$$\alpha_1 \mathbf{x}_1(t_1) + \cdots + \alpha_n \mathbf{x}_n(t_1) = \theta \quad (1.21)$$

where θ is the zero vector, and the constants α_i are not all zero. Define a function

$$\mathbf{z}(t) = \alpha_1 \mathbf{x}_1(t) + \cdots + \alpha_n \mathbf{x}_n(t)$$

and note that $\mathbf{z}(t) = \Phi(t, t_0)\mathbf{c}$, where $\mathbf{c} = (\alpha_i)$ whence, by (i), it satisfies equation (1.19). But notice that by (1.21) above, $\mathbf{z}$ satisfies the initial value problem

$$\dot{\mathbf{z}} = A(t)\mathbf{z}, \quad \mathbf{z}(t_1) = \theta \quad (1.22)$$

Note further that $\mathbf{z}$ is not the zero function on the interval T :

$$\begin{aligned} \mathbf{z}(t_0) &= \alpha_1 \mathbf{x}_1(t_0) + \cdots + \alpha_n \mathbf{x}_n(t_0) \\ &= \alpha_1 \mathbf{e}_1 + \cdots + \alpha_n \mathbf{e}_n \\ &\neq \theta \end{aligned}$$

because some $\alpha_j \neq 0$. Finally, note that the zero function $\theta(t) \equiv \theta$, all $t \in T$, satisfies the initial value problem (1.22), contradicting the uniqueness of the solution as remarked above Theorem 1.12.

(iii): Lastly, we show that any solution $\mathbf{x}(t)$ of the homogeneous equation (1.19) can be written in the form of equation (1.20). So suppose $\mathbf{x}(t)$ is a solution of equation (1.19) and suppose $\mathbf{x}(t_0) = \mathbf{x}_0$. Consider the function

$$\mathbf{z}(t) = \Phi(t, t_0)\mathbf{x}_0$$

which satisfies homogeneous equation (1.19) by (i); in addition,

$$\begin{aligned} \mathbf{z}(t_0) &= [\mathbf{x}_1(t_0) \cdots \mathbf{x}_n(t_0)] \mathbf{x}_0 \\ &= [\mathbf{e}_1 \cdots \mathbf{e}_n] \mathbf{x}_0 \\ &= I_n \mathbf{x}_0 \\ &= \mathbf{x}_0 \end{aligned}$$

Hence $\mathbf{x}$ and $\mathbf{z}$ satisfy the same initial value problem. By the uniqueness of the solution, we have $\mathbf{x}(t) = \mathbf{z}(t)$, for all $t \in T$, completing the proof. ■

1.13 Corollary From Step (iii) of the proof above, we conclude that the particular solution of the initial value problem

$$\dot{\mathbf{x}} = A(t)\mathbf{x}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.23)$$

is given as

$$\mathbf{x}(t) = \Phi(t, t_0) \mathbf{x}_0 \quad (1.24)$$

([SB], p. 54) ■

We turn next to the key properties of the fundamental matrix. They will first be used in deriving the solution of the initial value problem (1.18). Again, our presentation follows Szidarovsky and Bahill ([SB], pp. 54–55).

1.14 Theorem (Properties of the fundamental matrix) Let $\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t)$, where each $a_{ij}(t)$ is continuous. Then the fundamental matrix for this system has the following properties, for fixed $t_0 \in T$ and all $t, t_1 \in T$:

- i.) $\Phi(t_0, t_0) = I_n$;
- ii.) $\Phi(t, t_1)\Phi(t_1, t_0) = \Phi(t, t_0)$ {"transitivity" in t };
- iii.) $\Phi(t_1, t_0)^{-1} = \Phi(t_0, t_1)$;
- iv.) $\frac{\partial}{\partial t} \Phi(t, t_0) = A(t)\Phi(t, t_0)$; and
- v.) $\frac{\partial}{\partial t} \Phi(t_0, t) = -\Phi(t_0, t)A(t)$

Proof

(i): This is immediate from the construction of the fundamental matrix.

(ii): The basis of the proof here is a useful fact that for $n \times n$ matrices B and C , if $B\mathbf{x} = C\mathbf{x}$ for every $\mathbf{x} \in \mathbb{R}^n$, then $B = C$. Select an arbitrary (and fixed) initial vector $\mathbf{c} \in \mathbb{R}^n$. Note that the construction of the fundamental matrix is independent of the choice of initial vector. By Corollary 1.13, the solution of the initial value problem of equation (1.23) is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{c}, \quad \mathbf{x}(t_0) = \mathbf{c}$$

Next, denote $\mathbf{x}(t_1) = \Phi(t_1, t_0)\mathbf{c}$. Then $\mathbf{x}$ satisfies the (new) initial value problem

$$\dot{\mathbf{x}} = A(t)\mathbf{x}, \quad \mathbf{x}(t_1) = \mathbf{x}_1$$

Using equation (1.23), the solution of this new initial value problem is given by

$$\begin{aligned} \mathbf{y}(t) &= \Phi(t, t_1)\mathbf{x}_1 \\ &= \Phi(t, t_1)\Phi(t_1, t_0)\mathbf{c} \end{aligned}$$

But again the uniqueness of the solution implies $\mathbf{y} = \mathbf{x}$, that is

$$\Phi(t, t_1)\Phi(t_1, t_0)\mathbf{c} = \Phi(t, t_0)\mathbf{c}$$

Since $\mathbf{c}$ was arbitrary, we have the desired result.

(iii): Let $t = t_0$ in (ii) to get

$$\begin{aligned} \Phi(t_0, t_1)\Phi(t_1, t_0) &= \Phi(t_0, t_0) \\ &= I_n \quad \{\text{by (i)}\} \end{aligned}$$

which implies the result.

(iv): A relatively straightforward computation gives

$$\begin{aligned} \frac{\partial}{\partial t} \Phi(t, t_0) &= \frac{d}{dt} [\mathbf{x}_1(t) \cdots \mathbf{x}_n(t)] \\ &= [\dot{\mathbf{x}}_1(t) \cdots \dot{\mathbf{x}}_n(t)] \\ &= [A(t)\mathbf{x}_1(t) \cdots A(t)\mathbf{x}_n(t)] \\ &= A(t)[\mathbf{x}_1(t) \cdots \mathbf{x}_n(t)] \\ &= A(t)\Phi(t, t_0) \end{aligned}$$

(v): By Property (iii) we have $\Phi(t_0, t)\Phi(t, t_0) = I_n$. Differentiating both sides of this equation with respect to t (and using the product rule) gives

$$\left[\frac{\partial}{\partial t} \Phi(t_0, t) \right] \Phi(t, t_0) + \Phi(t_0, t) \left[\frac{\partial}{\partial t} \Phi(t, t_0) \right] = O_n,$$

where O_n denotes the $n \times n$ zero matrix. This last equation implies

$$\begin{aligned} \frac{\partial}{\partial t} \Phi(t_0, t) &= -\Phi(t_0, t) \left[\frac{\partial}{\partial t} \Phi(t, t_0) \right] \Phi(t, t_0)^{-1} \\ &= -\Phi(t_0, t) [A(t)\Phi(t, t_0)] \Phi(t, t_0)^{-1} \quad \{\text{by (iv)}\} \\ &= -\Phi(t_0, t) A(t) \end{aligned}$$

■

Notice that the definition of the fundamental matrix requires solving n initial value problems in order to obtain its columns. The next lemma transforms the solution of these n initial value problems into the solution of a single matrix differential equation, and sometimes the matrix equation can be solved quite easily.

1.15 Lemma Assume that the $n \times n$ matrix $X(t)$ satisfies the **matrix** initial value problem

$$\dot{X}(t) = A(t)X(t); \quad X(t_0) = I_n \quad (1.25)$$

Then $X(t) = \Phi(t, t_0)$.

Proof This result is immediate if we restate this matrix initial value problem in terms of n column equations. A solution matrix X must then have columns that satisfy the same initial value problem as the vectors that make up the columns of $\Phi(t, t_0)$. But then the uniqueness of the solutions of these n vector equations implies that the columns of $X(t)$ are identical to the columns of $\Phi(t, t_0)$. ■

1.16 Example Consider the following initial value problem.

$$\dot{\mathbf{x}} = \begin{bmatrix} t & t \\ 0 & 2t \end{bmatrix} \mathbf{x}, \quad \mathbf{x}(0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

In this case $t_0 = 0$. We won't explain yet how we found it, but notice that the matrix

$$X(t) = \begin{bmatrix} e^{t^2/2} & -e^{-t^2/2} + e^{t^2} \\ 0 & e^{t^2} \end{bmatrix}$$

satisfies relations (1.25):

$$\begin{aligned} \frac{d}{dt} X(t) &= \begin{bmatrix} te^{t^2/2} & -te^{t^2/2} + 2te^{t^2} \\ 0 & 2te^{t^2} \end{bmatrix} \\ &= \begin{bmatrix} t & t \\ 0 & 2t \end{bmatrix} \begin{bmatrix} e^{t^2/2} & -e^{-t^2/2} + e^{t^2} \\ 0 & e^{t^2} \end{bmatrix} \\ &= A(t)X(t) \end{aligned}$$

and also

$$X(0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$$

Hence, $X(t) = \Phi(t, 0)$ by Lemma 1.15. Then by Corollary 1.13, we have that the particular solution to the initial value problem is given as

$$\begin{aligned} \mathbf{x}(t) &= \Phi(t, 0)\mathbf{x}_0 \\ &= \begin{bmatrix} e^{t^2/2} & -e^{-t^2/2} + e^{t^2} \\ 0 & e^{t^2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= e^{t^2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

■

At last we are ready to find a solution to linear systems of the form of equation (1.17), posed at the beginning of this subsection.

1.17 Theorem The general solution of the inhomogeneous equation

$$\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) + \mathbf{f}(t)$$

is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{k}_1 + \int_{t_0}^t \Phi(t, \tau)\mathbf{f}(\tau)d\tau \quad (1.26)$$

where $\mathbf{k}_1$ is an arbitrary constant vector. Furthermore, the particular solution of the initial value problem

$$\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) + \mathbf{f}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.27)$$

has the form

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^t \Phi(t, \tau)\mathbf{f}(\tau)d\tau \quad (1.28)$$

([SB], pp. 56–57).

Proof By the comments following the proof of Lemma 1.10, a solution $\mathbf{x}(t)$ exists.

Fix t_0 and define $\mathbf{k}(t) = \Phi(t, t_0)^{-1} \mathbf{x}(t)$, so we have

$$\mathbf{x}(t) = \Phi(t, t_0) \mathbf{k}(t) \quad (1.29)$$

(recall $\Phi(t, t_0)$ is invertible by Theorem 1.12). Substitution into the

inhomogeneous equation (1.27) above shows that $\mathbf{x}(t)$ (as given by (1.29)) is a solution if and only if

$$\left[\frac{\partial}{\partial t} \Phi(t, t_0) \right] \mathbf{k}(t) + \Phi(t, t_0) \dot{\mathbf{k}}(t) = A(t) \Phi(t, t_0) \mathbf{k}(t) + \mathbf{f}(t)$$

Using properties (iii)–(iv) of Theorem 1.14, it follows that

$$\begin{aligned} \Phi(t, t_0) \dot{\mathbf{k}}(t) &= A(t) \Phi(t, t_0) \mathbf{k}(t) + \mathbf{f}(t) - \frac{\partial}{\partial t} \Phi(t, t_0) \cdot \mathbf{k}(t) \\ &= \left[A(t) \Phi(t, t_0) - \frac{\partial}{\partial t} \Phi(t, t_0) \right] \mathbf{k}(t) + \mathbf{f}(t) \\ &= \left[\frac{\partial}{\partial t} \Phi(t, t_0) - \frac{\partial}{\partial t} \Phi(t, t_0) \right] \mathbf{k}(t) + \mathbf{f}(t) \quad \{\text{by (iv)}\} \\ &= \mathbf{f}(t) \end{aligned}$$

The invertibility of the fundamental matrix now implies that

$$\begin{aligned} \dot{\mathbf{k}}(t) &= \Phi(t, t_0)^{-1} \mathbf{f}(t) \\ &= \Phi(t_0, t) \mathbf{f}(t) \quad \{\text{by (iii)}\} \end{aligned}$$

If $t_0 \in T$ is an arbitrary point, then integrating on both sides gives

$$\mathbf{k}(t) = \int_{t_0}^t \Phi(t_0, \tau) \mathbf{f}(\tau) d\tau + \mathbf{k}_1 \quad (1.30)$$

where $\mathbf{k}_1$ is an arbitrary constant vector. Substituting equation (1.30) into equation (1.29) gives

$$\begin{aligned} \mathbf{x}(t) &= \Phi(t, t_0) \mathbf{k}_1 + \int_{t_0}^t \Phi(t, t_0) \Phi(t_0, \tau) \mathbf{f}(\tau) d\tau \\ &= \Phi(t, t_0) \mathbf{k}_1 + \int_{t_0}^t \Phi(t, \tau) \mathbf{f}(\tau) d\tau \quad \{\text{Theorem 1.14 (ii)}\} \end{aligned}$$

thus establishing the first result.

Next, we substitute the general solution just obtained into the initial condition of equation (1.27):

$$\begin{aligned} \mathbf{x}(t_0) &= \Phi(t_0, t_0)\mathbf{k}_1 + \int_{t_0}^{t_0} \Phi(t_0, \tau)\mathbf{f}(\tau)d\tau \\ &= I_n\mathbf{k}_1 + \mathbf{0} \\ &= \mathbf{k}_1 \end{aligned}$$

We see that the initial condition is satisfied if and only if $\mathbf{x}_0 = \mathbf{k}_1$. Hence the particular solution of the initial value problem given in equation (1.27) is obtained by substituting $\mathbf{x}_0$ into the general solution—which results in equation (1.28) and completes the proof. ■

As a corollary to this result, we establish the closed form solution of a specific continuous linear dynamical system that often occurs in the literature.

1.18 Corollary Suppose $A(t)$, $B(t)$, and $C(t)$ are matrices whose entries are continuous functions. Fix $t_0 \in \mathfrak{R}$. Then the solution of the dynamical system given by

$$\dot{\mathbf{x}}(t) = A(t)\mathbf{x}(t) + B(t)\mathbf{u}(t); \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.31)$$

$$\mathbf{y}(t) = C(t)\mathbf{x}(t) \quad (1.32)$$

where the vector $\mathbf{u}$ is the input to the system and vector $\mathbf{y}$ gives the output of the system, is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^t \Phi(t, \tau)B(\tau)\mathbf{u}(\tau)d\tau \quad (1.33)$$

whence it follows that

$$\mathbf{y}(t) = C(t)\Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^t C(t)\Phi(t, \tau)B(\tau)\mathbf{u}(\tau)d\tau \quad (1.34)$$

Proof This result is immediate from Theorem 1.17 where, in this case, $\mathbf{f}(t) = B(t)\mathbf{u}(t)$. ■

By now it should be apparent that computation of the fundamental matrix is central to obtaining the solution of a continuous linear dynamical system. In the special case that matrices A , B , and C in Corollary 1.18 are constant, the system is called “time invariant,” and we state this formally as a definition.

1.19 Definition (Time invariant continuous linear dynamical system) Suppose A , B , and C are $n \times n$ constant matrices. Then the continuous linear dynamical system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.35)$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} \quad (1.36)$$

where the vector $\mathbf{u}$ is the input to the system and vector $\mathbf{y}$ gives the system's output, is called **time invariant**, because of the time independence of the entries of matrices A , B , and C ([SB], p. 94). ■

We will see that the fundamental matrix for time invariant continuous linear dynamical systems can be expressed in terms of a matrix exponential, which we define next.

1.20 Definition (Matrix exponential e^A) The **matrix exponential** form of the $n \times n$ matrix A , e^A , is defined to be the power series

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k \quad \blacksquare$$

The matrix exponential is well defined for every square matrix A [HJ85].

1.21 Theorem Suppose A is an $n \times n$ **constant** matrix. Then

$$\begin{aligned} \Phi(t, t_0) &= e^{(t-t_0)A} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} (t-t_0)^k A^k \end{aligned}$$

([SB], p. 57).

Proof Note that it suffices to show that the matrix $e^{(t-t_0)A}$ satisfies the conditions of Lemma 1.15.

$$\begin{aligned}\frac{\partial}{\partial t}e^{(t-t_0)A} &= \frac{d}{dt}e^{tA} \cdot e^{-t_0A} \\ &= Ae^{tA} \cdot e^{-t_0A} \\ &= Ae^{(t-t_0)A}\end{aligned}$$

where the various equalities above follow from theorems in linear algebra which, in particular, establish that the usual exponent laws and chain rule work for matrix exponentials just as they do in regular one-variable calculus ([HJ85] and [HJ91]). Also,

$$e^{(t_0-t_0)A} = e^{0 \cdot A} = I_n \quad \blacksquare$$

1.22 Corollary Suppose A is an $n \times n$ constant matrix, which is diagonalizable.

Then there exists an invertible matrix S such that

$$\Phi(t, t_0) = SES^{-1}$$

where the matrix E is given by

$$E = \begin{bmatrix} e^{\lambda_1(t-t_0)} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2(t-t_0)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n(t-t_0)} \end{bmatrix}$$

and $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A in some order.

Proof Since A is diagonalizable, there exists an $n \times n$ invertible matrix S such that

$$S^{-1}AS = \Lambda$$

where

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

is a diagonal matrix containing the eigenvalues of A . It now follows from Theorem 1.21 that

$$\begin{aligned}
 \Phi(t, t_0) &= \sum_{k=0}^{\infty} \frac{1}{k!} (t - t_0)^k A^k \\
 &= \sum_{k=0}^{\infty} \frac{1}{k!} (t - t_0)^k (S \Lambda S^{-1})^k \\
 &= S \left\{ \sum_{k=0}^{\infty} \frac{1}{k!} (t - t_0)^k \Lambda^k \right\} S^{-1} \\
 &= S E S^{-1}
 \end{aligned}$$

where the last equality above follows from the Taylor series expansion of e^x and the fact that $\Lambda^k = \text{diag}(\lambda_1^k, \dots, \lambda_n^k)$. ■

Using the notation of the proof above, note that $E = e^{(t-t_0)\Lambda}$.

1.1.4 Continuous Dynamical Systems and Laplace Transform

One ostensible limitation to how we defined continuous dynamical systems is that the systems are equivalent to first-order scalar differential equations.

However, in Section 1.3 we show how higher-order scalar differential equations can be transformed into higher-dimensional continuous dynamical systems, which also satisfy Definition 1.1. Further, it is known that certain higher-order differential equations can be solved easily using Laplace transforms. This is especially true in the case of linear ordinary differential equations of the form

$$x^{(n)}(t) + a_{n-1}x^{(n-1)}(t) + \cdots + a_1\dot{x}(t) + a_0x(t) = f(t)$$

where the a_j are constants, or where the output $y(t)$ and input $u(t)$ of the

dynamical system are related by the differential equation

$$\begin{aligned} y^{(n)}(t) + p_{n-1}y^{(n-1)}(t) + \cdots + p_1\dot{y} + p_0y(t) \\ = q_{n-1}u^{(n-1)}(t) + \cdots + q_1\dot{u}(t) + q_0u(t) \end{aligned}$$

where the p_i and q_j are constants and $y(t)$ and $u(t)$ are real-valued functions.

Since our focus here is on continuous dynamical systems as given in Definition 1.1, the reader is referred to Szidarovsky and Bahill ([SB], pp. 58-69 and 98-111) or Rice and Strange [RS] for a review of Laplace transform and its application to the differential equations given above. Note especially that the Laplace operator is a linear operator.

In this subsection we show how Laplace transform can be used to obtain the solution of a continuous time invariant linear dynamical system. This method provides an alternative to methods based on computation of the fundamental matrix. To simplify notation, we will assume that $t \geq 0$ and $t_0 = 0$ throughout this subsection.

1.23 Theorem Suppose we are given the continuous linear dynamical system

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + \mathbf{f}(t), \quad \mathbf{x}(0) = \mathbf{x}_0 \quad (1.35)$$

where A is an $n \times n$ constant matrix, $T = [0, \infty)$, $\mathbf{f}: T \rightarrow \mathbb{R}^n$, and $t_0 = 0$. Then the Laplace transform of the solution $\mathbf{x}(t)$ of System (1.35) is given by

$$\mathbf{X}(s) = (sI - A)^{-1}(\mathbf{x}_0 + \mathbf{F}(s)) \quad (1.36)$$

where $\mathbf{X}(s)$ and $\mathbf{F}(s)$ are the Laplace transforms of $\mathbf{x}(t)$ and $\mathbf{f}(t)$, respectively and s is not an eigenvalue of A . The components of $\mathbf{x}(t)$ can then be recovered by applying inverse Laplace transform to the components of $\mathbf{X}(s)$ ([SB], p. 68).

Proof Consider the resulting equation of the j th entry of system (1.35):

$$\dot{x}_j = \mathbf{a}^{(j)}\mathbf{x} + f_j(t) \quad (1.37)$$

where $\mathbf{a}^{(j)}$ is the j th row vector of matrix A . Applying Laplace transform to both sides of (1.37), and using properties of the Laplace operator, gives

$$\begin{aligned} -x_j(0) + sX_j(s) &= \sum_{k=1}^n a_{jk}X_k(s) + F_j(s) \\ &= \mathbf{a}^{(j)}\mathbf{X}(s) + F_j(s) \end{aligned}$$

which implies

$$-\mathbf{x}(0) + s\mathbf{X}(s) = A\mathbf{X}(s) + \mathbf{F}(s)$$

Since s is not an eigenvalue of A , $(sI - A)$ is invertible and hence

$$\mathbf{X}(s) = (sI - A)^{-1}(\mathbf{x}_0 + \mathbf{F}(s)) \quad \blacksquare$$

An example of how this theorem can be used to solve continuous linear dynamical systems is presented in the next subsection.

1.1.5 Solving Continuous Linear Dynamical Systems: Examples

In this subsection we present examples that summarize the methodology developed so far to solve continuous linear dynamical systems. In the first example we solve the dynamical system that was presented in Example 1.3 at the beginning of Section 1.1.

1.24 Example Recall the continuous dynamical system of Example 1.3, where we take $u_1(t) = u_2(t) = u(t)$:

$$\begin{aligned}\dot{\mathbf{x}}(t) &= \begin{bmatrix} \frac{x_1(t)}{t} + u(t) \\ \frac{x_2(t)}{t} + u(t) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{t} & 0 \\ 0 & \frac{1}{t} \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t)\end{aligned}$$

and

$$\begin{aligned}y(t) &= x_1(t) + x_2(t) \\ &= [1 \quad 1] \mathbf{x}(t)\end{aligned}$$

Here $T = (0, \infty)$. This is a continuous linear dynamical system. In the notation of equations (1.31)–(1.32), we have

$$\begin{aligned}A(t) &= \begin{bmatrix} \frac{1}{t} & 0 \\ 0 & \frac{1}{t} \end{bmatrix} \\ B(t) &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ C(t) &= [1 \quad 1]\end{aligned}$$

To obtain the solution of the dynamical system above by using Corollary 1.18, we must first compute $\Phi(t, \tau)$. We do this following Definition 1.11. The first column of $\Phi(t, \tau)$, $\mathbf{x}_1(t)$, is obtained by solving the system

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) = \begin{bmatrix} \frac{x_1(t)}{t} \\ \frac{x_2(t)}{t} \end{bmatrix}, \mathbf{x}(t_0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

We must do this for arbitrary values of $t_0 \in T = (0, \infty)$, since $t_0 = \tau$ is a variable in Corollary 1.18. The corresponding differential equation in each component in the equation above is of the form $\frac{dy}{dt} = \frac{y}{t}$, which can be written, equivalently, as $\frac{dy}{y} = \frac{dt}{t}$. Integrating both sides of this last equation, and using the fact that $t, t_0 > 0$, gives $y(t) = kt$ for a solution. The initial condition above can now be rewritten

$$\begin{bmatrix} x_1(t_0) \\ x_2(t_0) \end{bmatrix} = \begin{bmatrix} k_1 t_0 \\ k_2 t_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

which implies that

$$\mathbf{x}_1(t) = \begin{bmatrix} \frac{t}{t_0} \\ 0 \end{bmatrix}$$

The computation for the second column $\mathbf{x}_2(t)$ of $\Phi(t, \tau)$ is identical to the computation above except that the initial condition is different: $\mathbf{x}_2(t_0) = [0 \ 1]^T$.

It follows that the fundamental matrix can be written

$$\Phi(t, \tau) = \begin{bmatrix} \frac{t}{\tau} & 0 \\ 0 & \frac{t}{\tau} \end{bmatrix}$$

The solution of the system can then be taken from equations (1.33)–(1.34). By equation (1.33), we have

$$\begin{aligned} \mathbf{x}(t) &= \Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^t \Phi(t, \tau)B(\tau)u(\tau)d\tau \\ &= \begin{bmatrix} \frac{t}{t_0} & 0 \\ 0 & \frac{t}{t_0} \end{bmatrix} \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix} + \int_{t_0}^t \begin{bmatrix} \frac{t}{\tau} & 0 \\ 0 & \frac{t}{\tau} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(\tau)d\tau \\ &= \frac{t}{t_0} \begin{bmatrix} x_{10} \\ x_{20} \end{bmatrix} + \int_{t_0}^t \frac{tu(\tau)}{\tau} d\tau \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

and from equation (1.34), we have

$$\begin{aligned}
 y(t) &= C(t)\Phi(t, t_0)x_0 + \int_{t_0}^t C(t)\Phi(t, \tau)B(\tau)u(\tau)d\tau \\
 &= (1, 1) \begin{bmatrix} \frac{t \cdot x_{10}}{t_0} \\ \frac{t \cdot x_{20}}{t_0} \end{bmatrix} + \int_{t_0}^t (1, 1) \begin{bmatrix} \frac{t}{\tau} \\ \frac{t}{\tau} \end{bmatrix} u(\tau)d\tau \\
 &= \frac{t}{t_0}(x_{10} + x_{20}) + \int_{t_0}^t \frac{2t \cdot u(\tau)}{\tau} d\tau
 \end{aligned}$$

For the special case $u(t) \equiv 1$ for all $t > 0$, $t_0 = 1$, and $x(t_0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ we have

$$\begin{aligned}
 x(t) &= t \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \left\{ t \int_1^t \frac{1}{\tau} d\tau \right\} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
 &= t \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \{t(\ln t - \ln 1)\} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\
 &= \begin{bmatrix} t(1 + \ln t) \\ t(1 + \ln t) \end{bmatrix}
 \end{aligned}$$

and

$$\begin{aligned}
 y(t) &= 2t + 2t \int_1^t \frac{1}{\tau} d\tau \\
 &= 2t + 2t(\ln t - \ln 1) \\
 &= 2t(1 + \ln t)
 \end{aligned}$$

Note that these solutions agree with those given in Example 1.3. ■

In the next example we present all of the methods presented thus far to solve a continuous linear dynamical system of the form given in equation (1.18), where the matrices $A(t)$, $B(t)$, and $C(t)$ are constant, that is, a time-invariant continuous linear dynamical system.

1.25 Example For the time-invariant continuous linear dynamical system

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u, \quad \mathbf{x}(t_0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- i) derive the fundamental matrix using Theorem 1.21 and Corollary 1.22;
- ii) find the state, $\mathbf{x}(t)$, of the system using equation (1.33) and for the particular case of $t_0 = 0$ and $u(t) \equiv 1$; and
- iii) find the state, $\mathbf{x}(t)$, of the system using Laplace transform, for the case $t_0 = 0$ and $u(t) \equiv 1$.

Solution.

- i) In contrast to Example 1.24, in this case, we can use Theorem 1.21 or Corollary 1.22 to compute the fundamental matrix, since the system is time invariant.

- i.1) Using Theorem 1.21, we have

$$\begin{aligned} \Phi(t, t_0) &= e^{(t-t_0)A} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} (t-t_0)^k A^k \end{aligned}$$

In this case, we would like a closed form for the powers of A. Note that

$$\begin{aligned} A^2 &= \begin{bmatrix} 3 & 3 \\ 6 & 6 \end{bmatrix} = 3 \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \\ A^3 &= 3 \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = 3^2 \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \end{aligned}$$

and a straightforward induction gives

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

for all k such that $1 \leq k$. Hence

$$\begin{aligned}
\Phi(t, t_0) &= I + \sum_{k=1}^{\infty} \frac{3^{k-1}(t-t_0)}{k!} \cdot A \\
&= I + \frac{1}{3} \sum_{k \geq 1} \frac{[3(t-t_0)]^k}{k!} \cdot A \\
&= I - \frac{1}{3}A + \frac{1}{3}e^{3(t-t_0)} \cdot A \\
&= \frac{1}{3} \begin{bmatrix} 2+e^{3(t-t_0)} & -1+e^{3(t-t_0)} \\ -2+2e^{3(t-t_0)} & 1+2e^{3(t-t_0)} \end{bmatrix}
\end{aligned}$$

Observe finally that $\Phi(t_0, t_0) = I_2$.

i.2) First, note that matrix A has distinct eigenvalues (namely, 0 and 3), and that

$$\begin{aligned}
A &= SAS^{-1} \\
&= \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix}
\end{aligned}$$

Then by Corollary 1.22, we have

$$\begin{aligned}
\Phi(t, t_0) &= SES^{-1} \\
&= \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} e^{3(t-t_0)} & 0 \\ 0 & e^{0(t-t_0)} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \\
&= \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} e^{3(t-t_0)} & e^{3(t-t_0)} \\ 2 & -1 \end{bmatrix} \\
&= \frac{1}{3} \begin{bmatrix} 2+e^{3(t-t_0)} & -1+e^{3(t-t_0)} \\ -2+2e^{3(t-t_0)} & 1+2e^{3(t-t_0)} \end{bmatrix}
\end{aligned}$$

which agrees with the result in Step (i.1) above.

ii) By equation (1.33) of Corollary 1.18, the state of the dynamical system above is given by

$$\begin{aligned} \mathbf{x}(t) &= \Phi(t, t_0) \mathbf{x}_0 + \int_{t_0}^t \Phi(t, \tau) B(\tau) u(\tau) d\tau \\ &= \Phi(t, t_0) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \int_{t_0}^t \Phi(t, \tau) \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(\tau) d\tau \\ &= \frac{1}{3} \begin{bmatrix} 1 + 2e^{3(t-t_0)} \\ -1 + 4e^{3(t-t_0)} \end{bmatrix} + \frac{1}{3} \int_{t_0}^t u(\tau) \begin{bmatrix} 2 + e^{3(t-\tau)} \\ -2 + 2e^{3(t-\tau)} \end{bmatrix} d\tau \end{aligned}$$

ii.1) For the special case $t_0 = 0$ and $u(t) \equiv 1$, we have

$$\begin{aligned} \mathbf{x}(t) &= \frac{1}{3} \begin{bmatrix} 1 + 2e^{3t} \\ -1 + 4e^{3t} \end{bmatrix} + \frac{1}{3} \int_0^t \begin{bmatrix} 2 + e^{3(t-\tau)} \\ -2 + 2e^{3(t-\tau)} \end{bmatrix} d\tau \\ &= \frac{1}{3} \begin{bmatrix} 1 + 2e^{3t} \\ -1 + 4e^{3t} \end{bmatrix} + \frac{1}{3} \begin{bmatrix} \int_0^t 2 + e^{3(t-\tau)} d\tau \\ \int_0^t -2 + 2e^{3(t-\tau)} d\tau \end{bmatrix} \\ &= \frac{1}{3} \begin{bmatrix} 1 + 2e^{3t} \\ -1 + 4e^{3t} \end{bmatrix} + \frac{1}{3} \begin{bmatrix} 2(t-0) - \frac{1}{3}(1 - e^{3t}) \\ -2(t-0) - \frac{2}{3}(1 - e^{3t}) \end{bmatrix} \end{aligned}$$

that is,

$$\mathbf{x}(t) = \frac{1}{3} \begin{bmatrix} \frac{2}{3} + 2t + \frac{7}{3}e^{3t} \\ -\frac{5}{3} - 2t + \frac{14}{3}e^{3t} \end{bmatrix}$$

iii) The differential equation given by equation (1.35) of Theorem 1.23,

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{f}(t)$$

takes the particular form

$$\dot{\mathbf{x}} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} u(t) \\ 0 \end{bmatrix}$$

So by Theorem 1.23, we have

$$\begin{aligned}
 \mathbf{X}(s) &= (sI - A)^{-1}(\mathbf{x}_0 + \mathbf{F}(s)) \\
 &= \begin{bmatrix} s-1 & -1 \\ -2 & s-2 \end{bmatrix}^{-1} \begin{bmatrix} 1+U(s) \\ 1 \end{bmatrix} \\
 &= \frac{1}{s(s-3)} \begin{bmatrix} s-2 & 1 \\ 2 & s-1 \end{bmatrix} \begin{bmatrix} 1+U(s) \\ 1 \end{bmatrix} \\
 &= \frac{1}{s(s-3)} \begin{bmatrix} s-1+sU(s)-2U(s) \\ 1+2U(s)+s \end{bmatrix}
 \end{aligned}$$

where $\mathbf{X}(s)$ and $U(s)$ are the Laplace transforms of $\mathbf{x}(t)$ and $u(t)$, respectively. For the particular case that $u(t) \equiv 1$ (and recall $t_0 = 0$) we have $U(s) = 1/s$, whence it follows that

$$\begin{aligned}
 \mathbf{X}(s) &= \frac{1}{s(s-3)} \begin{bmatrix} s - \frac{2}{s} \\ s + \frac{2}{s} + 1 \end{bmatrix} \\
 &= \begin{bmatrix} \frac{s^2 - 2}{s^2(s-3)} \\ \frac{s^2 + s + 2}{s^2(s-3)} \end{bmatrix}
 \end{aligned}$$

Next, we compute partial fractions decompositions of the components of $\mathbf{X}(s)$:

$$\frac{s^2 - 2}{s^2(s-3)} = \frac{2}{9} \cdot \frac{1}{s} + \frac{2}{3} \cdot \frac{1}{s^2} + \frac{7}{9} \cdot \frac{1}{s-3}$$

and

$$\frac{s^2 + s + 2}{s^2(s-3)} = -\frac{5}{9} \cdot \frac{1}{s} - \frac{2}{3} \cdot \frac{1}{s^2} + \frac{14}{9} \cdot \frac{1}{s-3}$$

Finally, using the linearity of the inverse Laplace transform operator and a standard table of Laplace transforms of common functions, we have

$$\mathbf{x}(t) = \begin{bmatrix} \frac{2}{9} + \frac{2}{3}t + \frac{7}{9}e^{3t} \\ -\frac{5}{9} - \frac{2}{3}t + \frac{14}{9}e^{3t} \end{bmatrix}$$

which agrees with the result obtained in Step (ii.1). ■

1.2 BASIC THEORY OF DISCRETE DYNAMICAL SYSTEMS

We now turn to dynamical systems for which the time scale is discrete, that is defined for discrete values of time. The development here parallels that in Section 1.1, except the time scale is no longer a continuum. We begin with the basic definition and commit similar notational abuses as in the development of continuous systems.

Our working definition of a dynamical system with discrete time scale is, again, a synthesis of those given by Kalman and Bertram and Szidarovsky and Bahill ([KB], p. 395 and [SB], p. 69).

1.26 Definition (Dynamical system with discrete time scale) A dynamical system with discrete time scale and state-space X is given by the vector difference equations

$$\mathbf{x}(t_{k+1}) = \mathbf{f}(t_k, \mathbf{x}(t_k), \mathbf{u}(t_k)) \quad (1.38)$$

$$\mathbf{y}(t_{k+1}) = \mathbf{g}(t_k, \mathbf{x}(t_k)) \quad (1.39)$$

where the t_k indicate discrete values of time (k an integer) and

$$-\infty < \dots, t_{k-1} < t_k < t_{k+1}, \dots < \infty; \quad t_k \rightarrow \infty \text{ when } k \rightarrow \infty \quad (1.40)$$

The variable t_k is an independent variable, analogous to the variable t in the continuous case. Here $\mathbf{f}: T \times X \times U \rightarrow \mathbb{R}^n$ and $\mathbf{g}: T \times X \rightarrow \mathbb{R}^p$, where T is composed of the (discrete) values of t_k given in equation (1.40). Often, $T = \mathbf{N} = \{0, 1, 2, \dots\}$, the set of natural numbers, or $T = \mathbf{Z}$, the integers. Further, it is assumed for all $t_k \in T$ that $\mathbf{x}(t_k) \in X \subseteq \mathbb{R}^n$ and $\mathbf{u}(t_k) \in U \subseteq \mathbb{R}^m$. The sets X and U are called the *state space* and *input space*, respectively. The first equation, (1.38), is called the *state transition equation* and is equivalent to the set of n scalar difference equations

$$x_i(t_{k+1}) = f_i(t_k, x_1(t_k), \dots, x_n(t_k), u_1(t_k), \dots, u_m(t_k))$$

$$i = 1, \dots, n$$

The second equation, (1.39), is referred to as the *output equation* and is equivalent to the set of p scalar difference equations

$$\begin{aligned} y_i(t_{k+1}) &= g_i(t_k, x_1(t_k), \dots, x_n(t_k)) \\ i &= 1, \dots, p \end{aligned}$$

The vector $\mathbf{x}$ is called the *state* of system (1.38)–(1.39), and its components x_i are called the *state variables*. The vector function $\mathbf{u}(t_k)$ is the *input* (or *forcing function* or *control function*) of the system, and the variables u_i are the *control variables*. ■

A number of simplifications can be made if one assumes that the time values come from the integers or natural numbers, and since a more generalized definition is given later, we will assume that $T = \mathbb{N}$ or $\mathbb{Z}$ throughout this section. The justification for this simplification is given in Section 1.4. For now, we present definitions based on the assumptions above.

Also assume, as we did in the continuous case, that we have fixed an input function $\mathbf{u}: \mathbb{N} \rightarrow \mathbb{R}^m$ (resp. $\mathbf{u}: \mathbb{Z} \rightarrow \mathbb{R}^m$) and fixed initial time $t_0 = 0$ (resp. $t_0 \in \mathbb{Z}$). With fixed input, we may view the discrete dynamical system as a function of t and $\mathbf{x}(t)$:

$$\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t)) \text{ for } t \geq 0 \text{ (resp. } t \geq t_0) \quad (1.41)$$

where $\mathbf{x}: \mathbb{N} \rightarrow \mathbb{R}^n$ (resp. $\mathbf{x}: \mathbb{Z} \rightarrow \mathbb{R}^n$) is an unknown vector-valued function with domain $\mathbb{N}$ (resp. $\mathbb{Z}$) and $\mathbf{f}: \mathbb{N} \times X \rightarrow \mathbb{R}^n$ (resp. $\mathbf{f}: \mathbb{Z} \times X \rightarrow \mathbb{R}^n$), where $X \subseteq \mathbb{R}^n$.

1.27 Definition (A solution of system (1.41)) Suppose $\bar{\phi}: \mathbb{N} \times X \rightarrow \mathbb{R}^n$ (resp. $\bar{\phi}: \mathbb{Z} \times X \times \mathbb{Z} \rightarrow \mathbb{R}^n$) is a vector-valued function, and $t_0 = 0$ (resp. $t_0 \in \mathbb{Z}$ is fixed). If

- (i) for any fixed $\mathbf{x}_0 \in X$, $\bar{\phi}(0; \mathbf{x}_0) = \mathbf{x}_0$ (resp. $\bar{\phi}(t_0; \mathbf{x}_0, t_0) = \mathbf{x}_0$); and
- (ii) it's possible to define recursively, $\bar{\phi}(t+1; \mathbf{x}_0) = \mathbf{f}(t, \bar{\phi}(t; \mathbf{x}_0))$ (resp. $\bar{\phi}(t+1; \mathbf{x}_0, t_0) = \mathbf{f}(t, \bar{\phi}(t; \mathbf{x}_0, t_0))$) for $t \geq t_0$, then we say the function $\mathbf{x}(t) = \bar{\phi}(t; \mathbf{x}_0)$

(resp. $\mathbf{x}(t) = \bar{\phi}(t; \mathbf{x}_0, t_0)$) is called a *solution* of the system (1.41), and it is also called the *transition function*. ■

As in the continuous case, with fixed initial time t_0 , input $\mathbf{u}(t)$, and initial state $\mathbf{x}_0$, we will let $\mathbf{x}(t)$ stand for the variable and the solution of equation (1.41)—while keeping in mind that this is an abuse of notation because $\bar{\phi}$ is dependent not only on t , but on $\mathbf{x}_0$ and t_0 as well. Also, if the range of $\mathbf{f}$ is a subset of the state space X , then condition (ii) of Definition 1.27 will imply that the solution is unique, by Theorem 1.30 below.

Remark

As in the continuous case, a reference to *the* state $\mathbf{x}(t)$ of a discrete dynamical system will mean the unique solution of difference equation (1.41)—as given by Definition 1.27—having initial state $\mathbf{x}_0 = \mathbf{x}(t_0)$. ■

1.28 Example Consider the discrete dynamical system given by

$$\begin{aligned}\mathbf{x}(t+1) &= \mathbf{f}(t, \mathbf{x}(t)) \\ &= \begin{bmatrix} f_1(t, x_1(t), x_2(t)) \\ f_2(t, x_1(t), x_2(t)) \end{bmatrix} \\ &= \begin{bmatrix} x_1(t) + x_2(t) \\ 2x_1(t) + 2x_2(t) + 1 \end{bmatrix}\end{aligned}$$

where $X = \mathbb{R}^2$ in this case. This system is linear and can be rewritten in matrix form as

$$\begin{aligned}\mathbf{x}(t+1) &= \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\end{aligned}$$

The solution of this system will be derived in Subsection 1.2.4. ■

1.2.1 An Existence and Uniqueness Theorem

As in the continuous case, we require the existence and uniqueness of a solution to a given discrete dynamical system. The following example shows why this is necessary.

1.29 Example Consider the one-dimensional discrete dynamical system

$$x(t+1) = |\ln(1-|x(t)|)| + 1, \quad t_0 = 0$$

In this case, $f(t, x) = |\ln(1-|x|)| + 1$. Observe that $f(t, x)$ is defined only for $|x| < 1$.

Hence, for f to be defined, we must have $|x(t)| < 1$ for all $t \geq 0$. But then function x does not satisfy the given difference equation, because even if $|x(0)| = 0$, then

$$\begin{aligned} x(1) &= |\ln(1-|x(0)|)| + 1 \\ &= 1 \end{aligned}$$

and if $0 < |x(0)| < 1$, then $|x(1)| > 1$. Hence, no solution to this discrete dynamical system exists. ■

1.30 Theorem Suppose that for every $x \in X$ and $t \in \mathbb{N}$ (resp. $t \in \mathbb{Z}$), $f(t, x)$ is defined and in X . Then there exists a family of solutions to the discrete dynamical system of equation (1.41). Further, once an initial value $x(0) = x_0$ (resp. $x(t_0) = x_0$ with t_0 fixed) is chosen, the solution is unique. ([SB], p. 70)

Proof We prove the more general case when $T = \mathbb{Z}$ only. Define for arbitrary (fixed) $t_0 \in \mathbb{Z}$ and $x_0 \in X$ the function $\varphi: T \times X \times T \rightarrow \mathbb{R}^n$ by $\varphi(t_0; x_0, t_0) = x_0$ and observe that, by induction, it is possible to define recursively

$$\varphi(t+1; x_0, t_0) = f(t, \varphi(t; x_0, t_0))$$

The base case is satisfied by definition of φ and the fact that $f(t_0, x_0) \in X$ by hypothesis, and $\varphi(t+1; x_0, t_0) = f(t, \varphi(t; x_0, t_0)) \in X$, since $\varphi(t; x_0, t_0) \in X$ by the inductive assumption and $f(t, \varphi(t; x_0, t_0)) \in X$ by hypothesis. Thus function φ is a solution to system (1.41) by Definition 1.27. It is obvious that once the particular initial state is chosen, φ will be uniquely determined by the recursion above. ■

1.2.2 Solving Discrete Linear Dynamical Systems

A discrete linear dynamical system is one of the form

$$\mathbf{x}(t+1) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{f}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.42)$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) \quad (1.43)$$

where $\mathbf{A}(t)$ and $\mathbf{C}(t)$ are $n \times n$ matrices, $\mathbf{f}(t)$ is an n -element vector, and the elements a_{ij} , c_{ij} , and f_i of $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{f}$, respectively, are real-valued functions defined on $\mathbf{N} = \{0, 1, 2, \dots\}$ or the integers $\mathbf{Z}$. As in the continuous case, we usually will have $\mathbf{f}(t) = \mathbf{B}(t)\mathbf{u}(t)$, where $\mathbf{B}(t)$ is an $n \times n$ matrix, and input to the system is given by the function $\mathbf{u}: \mathbf{N} \rightarrow \mathcal{R}^n$ or $\mathbf{u}: \mathbf{Z} \rightarrow \mathcal{R}^n$.

Clearly such a system satisfies conditions (1.38)–(1.39) of Definition 1.26, that is, equation (1.42) qualifies as a state transition equation of a discrete dynamical system, and equation (1.43) is a valid output equation. For the remainder of this section we will take $X = U = \mathcal{R}^n$; then, by Theorem 1.30, there exists a unique solution to equation (1.42) for each initial state $\mathbf{x}_0 = \mathbf{x}(t_0)$ in $\mathcal{R}^n$.

In contrast to the continuous case, a closed form solution of equation (1.42) is easily obtained. Again, we first define the fundamental matrix for a discrete linear dynamical system, which is computed as a first step toward obtaining the solution.

1.31 Definition (Fundamental, or transition matrix) For the discrete dynamical system given by $\mathbf{x}(t+1) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{f}(t)$, define the **fundamental** or **transition matrix**, $\Phi(t, \tau)$, for this system as follows:

$$\Phi(t, \tau) = \begin{cases} \mathbf{A}(t-1)\mathbf{A}(t-2)\cdots\mathbf{A}(\tau), & \text{if } t > \tau \geq t_0 \\ \mathbf{I}_n, & \text{if } t = \tau \geq t_0 \end{cases}$$

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

1.32 Theorem Under the assumptions given at the beginning of this subsection, the general solution of equation (1.42) is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{c} + \sum_{\tau=t_0}^{t-1} \Phi(t, \tau+1)\mathbf{f}(\tau) \quad (1.44)$$

where $t > t_0$, $\mathbf{c} \in \mathbb{R}^n$ is arbitrary, and $\Phi(t, \tau)$ is defined above ([SB], pp. 71–72).

When $t = t_0$, define $\mathbf{x}(t_0) = \mathbf{c}$.

Proof First let $t_0 = 0$. Beginning at $t = 0$, repeated application of equation (1.42) gives

$$\begin{aligned} \mathbf{x}(1) &= A(0)\mathbf{x}(0) + \mathbf{f}(0) \\ \mathbf{x}(2) &= A(1)\mathbf{x}(1) + \mathbf{f}(1) \\ &= A(1)A(0)\mathbf{x}(0) + A(1)\mathbf{f}(0) + \mathbf{f}(1) \\ \mathbf{x}(3) &= A(2)\mathbf{x}(2) + \mathbf{f}(2) \\ &= A(2)A(1)A(0)\mathbf{x}(0) + A(2)A(1)\mathbf{f}(0) + A(2)\mathbf{f}(1) + \mathbf{f}(2) \end{aligned}$$

and so on. Using finite induction we obtain the general form

$$\begin{aligned} \mathbf{x}(t) &= A(t-1)A(t-2)\cdots A(0)\mathbf{x}(0) + \\ &\quad \sum_{\tau=0}^{t-2} A(t-1)A(t-2)\cdots A(\tau+1)\mathbf{f}(\tau) + \mathbf{f}(t-1) \end{aligned} \quad (1.45)$$

whenever $t \geq 2$. Define $\Phi(t, \tau)$ as in Definition 1.31:

$$\Phi(t, \tau) \equiv \begin{cases} A(t-1)A(t-2)\cdots A(\tau), & \text{if } t > \tau \geq 0 \\ I_n, & \text{if } t = \tau \geq 0 \end{cases}$$

We claim $\mathbf{x}(t)$ equals the expression in equation (1.44), for $t \geq 1$ and $\mathbf{c} = \mathbf{x}(0)$. To

see this note that

$$\begin{aligned} \mathbf{x}(1) &= \Phi(1, 0)\mathbf{x}(0) + \Phi(1, 1)\mathbf{f}(1) \\ &= A(0)\mathbf{x}(0) + I_n\mathbf{f}(1) \\ &= A(0)\mathbf{x}(0) + \mathbf{f}(1) \end{aligned}$$

and for $t \geq 2$ we have

$$\begin{aligned}
\mathbf{x}(t) &= \Phi(t, 0)\mathbf{x}(0) + \sum_{\tau=0}^{t-1} \Phi(t, \tau+1)\mathbf{f}(\tau) \\
&= A(t-1)A(t-2)\cdots A(0)\mathbf{x}(0) + \sum_{\tau=0}^{t-2} \Phi(t, \tau+1)\mathbf{f}(\tau) + \mathbf{f}(t-1) \\
&= A(t-1)A(t-2)\cdots A(0)\mathbf{x}(0) + \sum_{\tau=0}^{t-2} A(t-1)A(t-2)\cdots A(\tau+1)\mathbf{f}(\tau) + \mathbf{f}(t-1)
\end{aligned}$$

where we have used the fact that $\Phi(t, t) = I_n$ to simplify the summation in the first line above. Hence, equation (1.44) agrees with solution formula (1.45) for all values of $t \geq 2$, and we verified the case $t = 1$ separately. That is, when $t_0 = 0$ the general solution of equation (1.42) is given by equation (1.44). It is straightforward to substitute $t_0, t_0 + 1, t_0 + 2, \dots$ for $0, 1, 2, \dots$ in the above derivation, thus obtaining equation (1.44) for any initial value t_0 . ■

1.33 Corollary Once the initial state $\mathbf{x}_0 = \mathbf{x}(t_0)$ is determined, the unique solution of system (1.42) is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}_0 + \sum_{\tau=t_0}^{t-1} \Phi(t, \tau+1)\mathbf{f}(\tau) \text{ for } t > t_0 \quad (1.46)$$

In addition, the particular solution of the homogeneous form of Equation (1.42), where $\mathbf{f}(t) \equiv \theta$, is given by

$$\mathbf{x}(t) = \Phi(t, t_0)\mathbf{x}_0 \quad (1.47)$$

Notice that if we define $\sum_{\tau=t_0}^{t_0-1} \Phi(t, \tau+1)\mathbf{f}(\tau) = \theta$, then

$$\mathbf{x}(t_0) = \Phi(t_0, t_0)\mathbf{x}_0 = I_n\mathbf{x}_0 = \mathbf{x}_0$$

under the assumptions of the proof above, so equation (1.46) makes sense for $t = t_0$ also.

Proof Equation (1.46) follows immediately from the end of the proof of Theorem 1.32 and from Theorem 1.30, if we take $\mathbf{x}(t_0) = \mathbf{c} = \mathbf{x}_0$. Equation (1.47) follows immediately from equation (1.46) if we substitute the zero vector $\mathbf{0}$ for $\mathbf{f}(t)$. ■

Notice the similarities between the forms of the solutions given in this section, namely equations (1.46)–(1.47) and those of the continuous case, which are given by equations (1.24) and (1.28). The forms are identical if we replace integration in the continuous case by summation in the discrete case. However, note carefully that the forms of the matrix $\Phi(t, \tau)$, in the continuous and discrete cases, are quite different.

The next corollary is an immediate consequence of the preceding corollary; it is stated for reference purposes, as it applies to a particular dynamical system that occurs frequently in applications.

1.34 Corollary Suppose $A(t)$, $B(t)$, and $C(t)$ are matrices whose entries are real valued functions defined on $\mathbf{N} = \{0, 1, 2, \dots\}$ and $t_0 = 0$. Then the solution of the discrete dynamical system

$$\begin{aligned}\mathbf{x}(t+1) &= A(t)\mathbf{x}(t) + B(t)\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{x}_0 \\ \mathbf{y}(t) &= C(t)\mathbf{x}(t)\end{aligned}$$

where the vector function $\mathbf{u}(t)$ is input to the system, and the function $\mathbf{y}(t)$ is the system's output, is given by

$$\mathbf{x}(t) = \Phi(t, 0)\mathbf{x}_0 + \sum_{\tau=0}^{t-1} \Phi(t, \tau+1)B(\tau)\mathbf{u}(\tau) \quad (1.48)$$

and

$$\mathbf{y}(t) = C(t)\Phi(t, 0)\mathbf{x}_0 + \sum_{\tau=0}^{t-1} C(t)\Phi(t, \tau+1)B(\tau)\mathbf{u}(\tau) \quad (1.49)$$

■

Recall that in the continuous case, we first defined the fundamental matrix and derived special properties of it, which led to the solution of the dynamical system. However, in the discrete case, we began with the solution and used it to define the fundamental matrix. In the continuous case it was shown that the fundamental matrix is invertible. In the discrete case, this need not be true, but if it is invertible, we obtain a theorem analogous to Theorem 1.14.

1.35 Theorem (Properties of the fundamental matrix—discrete case) Suppose $A(t)$ is invertible for all $t \in \mathbb{Z}$ (resp. $t \geq 0$), and for $t \geq \tau$ (resp. $t \geq \tau \geq 0$), $\Phi(t, \tau)$ is defined as above. Then $\Phi(t, \tau)$ is invertible when $t \geq \tau$ (resp. $t \geq \tau \geq 0$). For $t < \tau$ (resp. $0 \leq t < \tau$), define $\Phi(t, \tau) = \Phi(\tau, t)^{-1}$. Then for all $t, t_0, t_1 \in \mathbb{Z}$ (resp. $t, t_0, t_1 \geq 0$), the following properties are true:

- i.) $\Phi(t, t) = I_n$
- ii.) $\Phi(t, t_1)\Phi(t_1, t_0) = \Phi(t, t_0)$
- iii.) $\Phi(t_0, t_1)^{-1} = \Phi(t_1, t_0)$
- iv.) $\Phi(t+1, t_0) = A(t)\Phi(t, t_0)$
- v.) $\Phi(t_0, t+1) = \Phi(t_0, t)A(t)^{-1}$

Proof Properties i) and iii) follow immediately from the definitions.

(ii): There are a number of cases to consider here.

ii.0) Case 0: One of the two indices are equal. If $t_1 = t_0$ or $t = t_1$, then either $\Phi(t_1, t_0) = I_n$ or $\Phi(t, t_1) = I_n$ and the result is immediate. If $t = t_0$, then $\Phi(t, t_0) = I_n$ by definition and

$$\begin{aligned} \Phi(t, t_1)\Phi(t_1, t_0) &= \Phi(t_0, t_1)\Phi(t_1, t_0) \\ &= \begin{cases} I_n I_n, & \text{if } t_0 = t_1 \\ \Phi(t_1, t_0)^{-1} \Phi(t_1, t_0), & \text{if } t_0 < t_1 \\ \Phi(t_0, t_1)\Phi(t_0, t_1)^{-1}, & \text{if } t_1 < t_0 \end{cases} \end{aligned}$$

In all cases we obtain the identity matrix on both sides of the equation.

Assume now that none of the indices are equal. This leaves six remaining subcases, corresponding to the permutations of the three unequal indices.

We will use the following fact repeatedly: If $i < j < k$, then

$$\Phi(k, i) = \Phi(k, j)\Phi(j, i) \quad (1.50)$$

which is true because, by the definition of the fundamental matrix,

$$\begin{aligned} \Phi(k, j)\Phi(j, i) &= [A(k-1)A(k-2)\cdots A(j)][A(j-1)A(j-2)\cdots A(i)] \\ &= \Phi(k, i) \end{aligned}$$

ii.1) Subcase 1: $t_0 < t_1 < t$. This follows immediately from equation (1.50).

ii.2) Subcase 2: $t_0 < t < t_1$. Then

$$\begin{aligned} \Phi(t, t_1)\Phi(t_1, t_0) &= \Phi(t_1, t)^{-1}\Phi(t_1, t_0) \\ &= \Phi(t_1, t)^{-1}\Phi(t_1, t)\Phi(t, t_0) \quad \{\text{by (1.50)}\} \\ &= \Phi(t, t_0) \end{aligned}$$

ii.3) Subcase 3: $t < t_0 < t_1$. This case is similar to the one above.

ii.4) Subcase 4: $t < t_1 < t_0$. By equation (1.50),

$$\Phi(t_0, t) = \Phi(t_0, t_1)\Phi(t_1, t)$$

which implies, by the definitions given above, that

$$\Phi(t, t_0)^{-1} = \Phi(t_1, t_0)^{-1}\Phi(t, t_1)^{-1}$$

Taking inverses on both sides of this last equation gives

$$\Phi(t, t_0) = \Phi(t, t_1)\Phi(t_1, t_0)$$

The remaining two cases, $t_1 < t_0 < t$ and $t_1 < t < t_0$, are similar to the cases proved above.

(iv): There are three cases to consider here.

iv.1) Case 1: $t = t_0$. Then

$$\begin{aligned} \Phi(t+1, t_0) &= \Phi(t_0+1, t_0) \\ &= A(t_0) \\ &= A(t_0)I_n \\ &= A(t)\Phi(t, t_0) \end{aligned}$$

iv.2) Case 2: $t < t_0$. Then

$$\begin{aligned}
 A(t)\Phi(t, t_0) &= A(t)\Phi(t_0, t)^{-1} \\
 &= A(t)[A(t_0 - 1) \cdots A(t + 1)A(t)]^{-1} \\
 &= A(t)A(t)^{-1}[A(t_0 - 1) \cdots A(t + 1)]^{-1} \\
 &= I_n \Phi(t_0, t + 1)^{-1} \\
 &= \Phi(t + 1, t_0) \quad \{t + 1 \leq t_0\}
 \end{aligned}$$

iv.3) Case 3: $t_0 < t$. This case is similar to Case 2.

(v): Again, there are three cases, two of which we will prove.

v.1) Case 1: $t_0 = t$. Then

$$\begin{aligned}
 \Phi(t_0, t + 1) &= \Phi(t_0, t_0 + 1) \\
 &= \Phi(t_0 + 1, t_0)^{-1} \\
 &= A(t_0)^{-1} \\
 &= I_n A(t_0)^{-1} \\
 &= \Phi(t_0, t)A(t)^{-1}
 \end{aligned}$$

v.2) Case 2: $t_0 < t$. Then

$$\begin{aligned}
 \Phi(t_0, t + 1) &= \Phi(t + 1, t_0)^{-1} \\
 &= [A(t)A(t - 1) \cdots A(t_0)]^{-1} \\
 &= [A(t - 1) \cdots A(t_0)]^{-1} A(t)^{-1} \\
 &= \Phi(t, t_0)^{-1} A(t)^{-1} \\
 &= \Phi(t_0, t)A(t)^{-1} \quad \{t_0 < t\}
 \end{aligned}$$

v.3) Case 3: $t < t_0$. This case is similar to Case 2. ■

The remainder of this subsection is devoted to considering “time-invariant” discrete dynamical systems. This definition is identical to the one given in the continuous case—except the time scale has changed.

1.36 Definition (Time invariant discrete linear dynamical system) Suppose that the matrices $A(t)$ and $C(t)$ of equations (1.42)–(1.43) are $n \times n$ constant matrices, and the function $f(t)$ of equation (1.42) is given by the product $Bu(t)$, where the vector function $u(t)$ is the input to the system, and B is an $n \times n$ constant matrix. Then the discrete linear dynamical system

$$x(t+1) = Ax(t) + Bu(t), \quad x(t_0) = x_0 \quad (1.51)$$

$$y(t) = Cx(t) \quad (1.52)$$

is called **time-invariant**, because of the time independence of the entries of the matrices A , B , and C . ■

The fundamental matrix for a time-invariant discrete dynamical system can be expressed in an especially simple form.

1.37 Theorem Suppose A is an $n \times n$ constant matrix and $t > \tau$. Then

$$\Phi(t, \tau) = A^{t-\tau}$$

Proof This is immediate from the definition of the fundamental matrix, and the fact that $A(j) = A$ for all $j \geq t_0$. ■

1.38 Corollary Suppose A is an $n \times n$ constant matrix, such that A is diagonalizable, and $t > \tau$. Then there exists an invertible matrix S such that

$$\Phi(t, \tau) = S\Lambda^{t-\tau}S^{-1}$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is a diagonal matrix, and $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A in some order.

Proof Since A is diagonalizable, there exists an invertible matrix S such that $S^{-1}AS = \Lambda$, where Λ is given above. Solving this equation for A and substituting into Theorem 1.37 gives

$$\begin{aligned} \Phi(t, \tau) &= A^{t-\tau} \\ &= (S\Lambda S^{-1})^{t-\tau} \\ &= S\Lambda^{t-\tau}S^{-1} \end{aligned} \quad \blacksquare$$

1.2.3 Discrete Dynamical Systems and Z-transform

The discrete equivalent of Laplace Transform is Z-transform, which is a function transformation that reduces the solution of linear difference equations with constant coefficients to the solution of algebraic equations, which can be solved using partial fractions ([SB], pp. 74–80). As in the continuous case, our main focus here will be on how Z-transform can be applied to discrete linear dynamical systems as given by equation (1.42). For reference purposes, we present the key terms and properties of Z-transform that will be useful to us. The reader is referred to [Mil] for a complete review of Z-transform and its application to the solution of linear difference equations with constant coefficients.

1.39 Definition (Z-transform) Suppose $\mathbf{x}(t)$ is a vector valued function defined for all $t \in \mathbb{N}$. Then the **Z-transform** of $\mathbf{x}$ is given by the following infinite series:

$$X(z) = \sum_{t=0}^{\infty} \frac{\mathbf{x}(t)}{z^t} \quad (1.53)$$

([SB], p. 74) ■

The *radius of convergence* of series (1.53) is the smallest real number $r \geq 0$ such that the series converges whenever $|z| > r$. The key properties of Z-transform are given in the next theorem, and we formulate as in Szidarovsky and Bahill.

1.40 Theorem (Properties of Z-transform)

i) Z-transform is a linear operator.

ii) The following are true:

ii.1) If $\mathbf{x}(t) = \mathbf{x}_1(t)a^t$ where a is a constant, then $X(z) = X_1\left(\frac{z}{a}\right)$.

ii.2) If $x(t) = x_1(t+1)$, then $X(z) = zX_1(z) - z \cdot x_1(0)$, and in general, if

$x(t) = x_1(t+n)$, where $n > 0$, then

$$X(z) = z^n X_1(z) - z^n x_1(0) - z^{n-1} x_1(1) - \dots - z x_1(n-1)$$

ii.3) If $x(t) = x_1(t-1)$ and $x_1(j) = 0$, for all $j < 0$, then $X(z) = \frac{1}{z} X_1(z)$. In

general, if $x(t) = x_1(t-n)$ for some $n > 0$, and $x_1(j) = 0$ for all $j < 0$, then

$$X(z) = \frac{1}{z^n} X_1(z)$$

■

This theorem is useful in obtaining the following:

Table 1.1 Z-transforms of Common Functions

No.	$x(t)$	$X(z)$
1.	1	$\frac{z}{z-1}$
2.	a^t	$\frac{z}{z-a}$
3.	ta^{t-1}	$\frac{z}{(z-a)^2}$
4.	t	$\frac{z}{(z-1)^2}$
5.	$\begin{cases} a^{t-1}, & \text{if } t \geq 1 \\ 0, & \text{if } t = 0 \end{cases}$	$\frac{1}{z-a}$
6.*	$\begin{cases} \binom{t-1}{k-1} a^{t-k}, & t \geq 1 \\ 0, & \text{if } t = 0 \end{cases}$	$\frac{1}{(z-a)^k}$
7.	$\begin{cases} a, & \text{if } t = 0 \\ 0, & \text{if } t \geq 1 \end{cases}$	a

$$* \quad \binom{t-1}{k-1} = \frac{(t-1)!}{(k-1)!(t-k)!}$$

We are now ready to present the main result of this subsection. As in the continuous case, this theorem gives us an alternate method to solving discrete linear dynamical systems that is based on Z-transform, and does not require the computation of the fundamental matrix.

1.41 Theorem Suppose that for all $t = 0, 1, 2, \dots$,

$$\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{f}(t), \quad \mathbf{x}_0 = \mathbf{x}(0) \quad (1.54)$$

where $\mathbf{A}$ is an $n \times n$ constant matrix and $\mathbf{f}$ is vector-valued function. Then the Z-transform of $\mathbf{x}$ is given by

$$\mathbf{X}(z) = (z\mathbf{I} - \mathbf{A})^{-1}(z\mathbf{x}_0 + \mathbf{F}(z)) \quad (1.55)$$

where $\mathbf{X}$ and $\mathbf{F}$ are the Z-transforms of $\mathbf{x}$ and $\mathbf{f}$, respectively, and z is any value not equal to an eigenvalue of $\mathbf{A}$.

Proof Consider the resulting equation of the j th entry of system (1.54):

$$x_j(t+1) = \mathbf{a}^{(j)}\mathbf{x} + f_j(t)$$

where $\mathbf{a}^{(j)}$ is the j th row vector of $\mathbf{A}$. Applying Z-transform to both sides of this equation, and using property ii.2) of Theorem 1.40 on the left side and the linearity of the Z-transform operator on the right-side of the equation, gives

$$\begin{aligned} -zx_j(0) + zX_j(z) &= \sum_{k=1}^n a_{jk}X_k(z) + F_j(z) \\ &= \mathbf{a}^{(j)}\mathbf{X}(z) + F_j(z) \end{aligned} \quad (1.56)$$

Equation (1.56) implies that

$$-z\mathbf{x}(0) + z\mathbf{X}(z) = \mathbf{A}\mathbf{X}(z) + \mathbf{F}(z)$$

Since z is not an eigenvalue of $\mathbf{A}$, $(z\mathbf{I} - \mathbf{A})$ is invertible and hence

$$\mathbf{X}(z) = (z\mathbf{I} - \mathbf{A})^{-1}(z\mathbf{x}_0 + \mathbf{F}(z)) \quad \blacksquare$$

1.2.4 Solving Discrete Linear Dynamical Systems: Examples

Analogous to the continuous case, in this subsection we present examples that demonstrate the methodology developed thus far to solve discrete linear dynamical systems. In the first of two examples we consider a discrete linear dynamical system that is not time invariant.

1.42 Example Consider the discrete linear dynamical system

$$\begin{aligned} \mathbf{x}(t+1) &= \begin{bmatrix} t & 0 \\ 0 & t \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(t), \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ y(t) &= [1 \quad 1] \mathbf{x}(t) \end{aligned}$$

where

$$\begin{aligned} A(t) &= \begin{bmatrix} t & 0 \\ 0 & t \end{bmatrix} \\ B(t) &= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ C(t) &= [1 \quad 1] \end{aligned}$$

This system is not time-invariant, because the matrix A is not a constant matrix. Since matrix $A(t)$ is a diagonal matrix, the fundamental matrix can be computed quite easily for $t > \tau \geq 0$:

$$\begin{aligned} \Phi(t, \tau) &= A(t-1)A(t-2)\cdots A(\tau) \\ &= \begin{bmatrix} (t-1)(t-2)\cdots\tau & 0 \\ 0 & (t-1)(t-2)\cdots\tau \end{bmatrix} \end{aligned}$$

So by Corollary 1.34, the solution is given by

$$\begin{aligned} \mathbf{x}(t) &= \Phi(t, 0) \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \sum_{\tau=0}^{t-1} \Phi(t, \tau+1) \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(\tau) \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \sum_{\tau=0}^{t-1} \begin{bmatrix} (t-1)(t-2)\cdots(\tau+1) & 0 \\ 0 & (t-1)(t-2)\cdots(\tau+1) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(\tau) \\ &= \left\{ \sum_{\tau=0}^{t-1} u(\tau)(t-1)(t-2)\cdots(\tau+1) \right\} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \end{aligned}$$

where we used the fact that $\Phi(t, 0)$ is the zero matrix. For $t = 1, 2$, and 3 , we have

$$\mathbf{x}(1) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(0)$$

(recall that when $\tau = t - 1$, the sum reduces to $\begin{bmatrix} 1 & 1 \end{bmatrix}^T u(\tau)$)

$$\mathbf{x}(2) = \{u(0) + u(1)\} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and

$$\mathbf{x}(3) = \{u(0) \cdot 2 \cdot 1 + u(1) \cdot 2 + u(2)\} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

The output of the system is computed to be

$$\begin{aligned} y(t) &= \begin{bmatrix} 1 & 1 \end{bmatrix} \mathbf{x}(t) \\ &= 2 \left\{ \sum_{\tau=0}^{t-1} u(\tau)(t-1)(t-2) \cdots (\tau+1) \right\} \end{aligned}$$

■

In our next example, we consider a time-invariant discrete linear dynamical system. Recall that for such systems, the fundamental matrix has a simple form, and that one of two methods can be used to obtain the solution: the first is based on the fundamental matrix (Corollary 1.34), and the second is based on Z-transform (Theorem 1.41). For purposes of comparison, the form of the matrix equation given on the right in the example below is identical to that of Example 1.25 in the special case that $u(t) \equiv 1$; however, in Example 1.25 the matrix equation expressed the derivative of $\mathbf{x}$ at time t , and hence expresses the rate of change of the state at time t —while in Example 1.43, the matrix equation gives the state at time $t+1$ —so the context is quite different.

1.43 Example For the time-invariant discrete linear dynamical system

$$\mathbf{x}(t+1) = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- i) solve the system using Corollary 1.34 for $t \geq 2$;
- ii) use the fact that $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$ is diagonalizable and Corollary 1.38 to give an alternate computation of the fundamental matrix;
- iii) solve the system using Z-transform (Theorem 1.41) for $t \geq 2$.

Solution

- i) First note that by Theorem 1.37, $\Phi(t, \tau) = A^{t-\tau}$. We also showed that

$$A^k = 3^{k-1}A$$

in Example 1.25. Hence, by Corollary 1.34, the solution for $t \geq 2$ is given by

$$\begin{aligned} \mathbf{x}(t) &= A^t \mathbf{x}_0 + \sum_{\tau=0}^{t-1} A^{t-\tau-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= 3^{t-1} \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \sum_{\tau=0}^{t-2} 3^{t-\tau-2} \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= 3^{t-1} \begin{bmatrix} 2 \\ 4 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \sum_{\tau=0}^{t-2} 3^{t-\tau-2} \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= 3^{t-1} \begin{bmatrix} 2 \\ 4 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \left(\frac{3^{t-1} - 1}{2} \right) \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{5}{2} 3^{t-1} - \frac{1}{2} \\ 5 \cdot 3^{t-1} \end{bmatrix} \end{aligned}$$

The second equation above results from the fact that when $\tau = t - 1$, the term in

the sum reduces to: $I_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. Referring back to the solution obtained in

Example 1.25, we see that the solution above is quite different from the solution obtained in that example—which is what we expect, since the dynamical system here is discrete rather than continuous.

ii) A diagonalization for matrix A is given by

$$\begin{aligned} A &= S\Lambda S^{-1} \\ &= \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \end{aligned}$$

So by Corollary 1.38,

$$\begin{aligned} \Phi(t, \tau) &= S\Lambda^{t-\tau}S^{-1} \\ &= \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}^{t-\tau} \begin{bmatrix} 1 & 1 \\ 2 & -1 \end{bmatrix} \\ &= 3^{t-\tau-1} \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \\ &= 3^{t-\tau-1} A \end{aligned}$$

which by Example 1.25 equals $A^{t-\tau}$, and $A^{t-\tau}$ is the fundamental matrix by Theorem 1.37.

iii) By Theorem 1.41, the Z-transform of the solution is given by

$$\begin{aligned} X(z) &= (zI_n - A)^{-1}(zx_0 + F(z)) \\ &= \begin{bmatrix} z-1 & -1 \\ -2 & z-2 \end{bmatrix}^{-1} \left(z \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ z \\ z-1 \end{bmatrix} \right) \\ &= \frac{1}{z(z-3)} \begin{bmatrix} z-2 & 1 \\ 2 & z-1 \end{bmatrix} \begin{bmatrix} z \\ z \\ z-1 \end{bmatrix} \\ &= \begin{bmatrix} \frac{z-1}{z-3} + \frac{1}{(z-1)(z-3)} \\ 1 + \frac{5}{z-3} \end{bmatrix} \end{aligned}$$

Compute a partial fractions decomposition of $\frac{1}{(z-1)(z-3)}$, write $\frac{z-1}{z-3}$ as a

proper fraction, and simplify the sum in the first component to obtain:

$$X(z) = \begin{bmatrix} 1 - \frac{1}{2} \left(\frac{1}{z-1} \right) + \frac{5}{2} \left(\frac{1}{z-3} \right) \\ 1 + 5 \left(\frac{1}{z-3} \right) \end{bmatrix}$$

Now, using the linearity of the Z-transform operator, Table 1.1, and the fact that $t \geq 2$, apply inverse Z-transform to both sides of this last equation to obtain $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \begin{bmatrix} \frac{5}{2} \cdot 3^{t-1} - \frac{1}{2} \\ 5 \cdot 3^{t-1} \end{bmatrix}$$

which agrees with the solution obtained in part i) above. ■

1.3 CONSIDERING HIGHER-ORDER EQUATIONS

The reader may have noticed that higher-order differential and difference equations have not been mentioned thus far. In this section, we describe a simple substitution method that can be used to convert higher-order differential (resp. difference) equations into higher-dimensional continuous (resp. discrete) first order dynamical systems as given by the definitions presented in this chapter ([SB], p. 51 (respectively, p. 70)).

1.3.1 Transforming Higher-Order Differential Equations

Suppose we are given an n th-order differential equation

$$x^{(n)} = f(t, x, \dot{x}, \ddot{x}, \dots, x^{(n-1)}) \quad (1.57)$$

with initial conditions

$$\begin{aligned} x(t_0) &= x_0 \\ \dot{x}(t_0) &= \dot{x}_0 \\ &\vdots \\ x^{(n-1)}(t_0) &= x_0^{(n-1)} \end{aligned} \quad (1.58)$$

where $\dot{x}$ and $\ddot{x}$ are the first and second derivatives, respectively, and $x^{(k)}$ denotes the k th derivative of x . Next, we introduce the variables

$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x} \\ x_3 &= \ddot{x} \\ &\vdots \\ x_n &= x^{(n-1)} \end{aligned}$$

Then $x^{(n)} = \dot{x}_n$ and we obtain the following continuous dynamical system:

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ \vdots \\ x_n \\ f(t, x_1, x_2, \dots, x_n) \end{bmatrix} = \begin{bmatrix} g_1(t, x_1, \dots, x_n) \\ g_2(t, x_1, \dots, x_n) \\ \vdots \\ g_{n-1}(t, x_1, \dots, x_n) \\ g_n(t, x_1, \dots, x_n) \end{bmatrix} = \mathbf{g}(t, \mathbf{x}(t)) \quad (1.59)$$

with initial condition

$$\mathbf{x}(t_0) = \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \\ \vdots \\ x_{n-1}(t_0) \\ x_n(t_0) \end{bmatrix} = \begin{bmatrix} x_0 \\ \dot{x}_0 \\ \vdots \\ x_0^{(n-2)} \\ x_0^{(n-1)} \end{bmatrix} = \mathbf{x}_0 \quad (1.60)$$

Note that Equations (1.59)–(1.60) define an n -dimensional continuous dynamical system with initial condition $\mathbf{x}_0$. This technique is demonstrated in the following example.

1.44 Example Rewrite the second-order differential equation

$$\ddot{x} + 3\dot{x} + 2x = 1, \quad x(0) = \dot{x}(0) = 0$$

as a continuous dynamical system written in the form given by equation (1.6), and then solve the system.

Solution. Solving for $\ddot{x}$ we have $\ddot{x} = 1 - 3\dot{x} - 2x = f(t, x, \dot{x})$. Introduce the variables

$$\begin{aligned} x_1 &= x \\ x_2 &= \dot{x} \end{aligned}$$

Then equation (1.59) takes the particular form

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ 1 - 3x_2 - 2x_1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{aligned}$$

and

$$\mathbf{x}_0 = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} x(0) \\ \dot{x}(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Note that the resulting dynamical system is linear and time invariant. Use

Theorem 1.23 of Section 1.1.4 to get

$$\begin{aligned} \mathbf{X}(s) &= (sI - A)^{-1}(\mathbf{x}_0 + \mathbf{F}(s)) \\ &= \begin{bmatrix} s & -1 \\ 2 & s+3 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1/s \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{s(s+1)(s+2)} \\ \frac{1}{(s+1)(s+2)} \end{bmatrix} \end{aligned}$$

Since we are interested in the component $x_1 = x$ of the vector $\mathbf{x}$, apply a partial fractions decomposition to the first component of this vector and then apply inverse Laplace transform. The reader can verify that

$$x_1(t) = x(t) = \frac{1}{2} - e^{-t} + \frac{1}{2}e^{-2t}$$

is the solution of the original initial-value problem. Since the coefficients of the given second-order differential equation are constant, this problem could also be solved directly with Laplace transforms, which would actually be simpler. ■

1.3.2 Transforming Higher-Order Difference Equations

The technique for transforming higher order difference equations to higher dimensional discrete dynamical systems is essentially the same as for the continuous case. Suppose we are given the n th order difference equation

$$x(t+n) = f(t, x(t), x(t+1), \dots, x(t+n-1)) \quad (1.61)$$

with initial conditions

$$x(0) = x_0; x(1) = x_1; \dots; x(n-1) = x_{n-1}$$

Note that we have taken $t_0 = 0$ to simplify notation. The initial conditions guarantee the uniqueness of the solution to equation (1.61). As in the continuous case above, we introduce the variables

$$\begin{aligned} x_1(t) &= x(t) \\ x_2(t) &= x(t+1) \\ &\vdots \\ x_n(t) &= x(t+n-1) \end{aligned}$$

and note that $x(t+n) = x_n(t+1)$, whence we obtain the following first-order discrete dynamical system

$$\mathbf{x}(t+1) = \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \\ \vdots \\ x_{n-1}(t+1) \\ x_n(t+1) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ x_3(t) \\ \vdots \\ x_n(t) \\ f(t, x_1(t), x_2(t), \dots, x_n(t)) \end{bmatrix} = \begin{bmatrix} g_1(t, x_1, \dots, x_n) \\ g_2(t, x_1, \dots, x_n) \\ \vdots \\ g_{n-1}(t, x_1, \dots, x_n) \\ g_n(t, x_1, \dots, x_n) \end{bmatrix} \quad (1.62)$$

that is, is the discrete first-order dynamical system $\mathbf{x}(t+1) = \mathbf{g}(t, \mathbf{x}(t))$. The initial condition is given by

$$\mathbf{x}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \\ \vdots \\ x_{n-1}(0) \\ x_n(0) \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-2} \\ x_{n-1} \end{bmatrix} = \mathbf{x}_0 \quad (1.63)$$

Again, note that Equations (1.62)–(1.63) define an n -dimensional first order discrete dynamical system with initial condition $\mathbf{x}_0$. The technique above is demonstrated in the following example.

1.45 Example Rewrite the second-order difference equation

$$x(t+2) + 3x(t+1) + 2x(t) = 1, \quad x(0) = x(1) = 0$$

as a discrete 2-dimensional dynamical system.

Solution. Solving for $x(t+2)$ we have

$$x(t+2) = 1 - 3x(t+1) - 2x(t) = f(t, x(t), x(t+1))$$

Introduce the variables

$$x_1(t) = x(t)$$

$$x_2(t) = x(t+1)$$

Then equation (1.62) takes the particular form

$$\begin{aligned} \mathbf{x}(t+1) &= \begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ 1 - 3x_2(t) - 2x_1(t) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{aligned}$$

and

$$\mathbf{x}_0 = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} x(0) \\ x(1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Finally, note that other than the fact that we are considering difference equations as opposed to differential equations, the matrix expression on the right of the difference equation above is identical to the right side matrix expression in the differential equation of Example 1.44. ■

1.4 GENERALIZED DEFINITION OF A DYNAMICAL SYSTEM

We are now ready to present a generalized definition of a dynamical system that will be of use later when we consider stability. The definitions that were given earlier are special cases of this more general definition—as we shall see. This section begins by presenting this general definition, and then follows the definition with a discussion of the various conditions given in the definition. Individual subsections will be devoted to showing that the definitions of a dynamical system given previously are special cases of this definition.

1.46 Definition (Deterministic system) A **deterministic system** is comprised of the following five sets: T (*time set*); X (*state set*); U (*input set*); Y (*output set*); and Ω (*admissible input functions*), together with the following two functions:

$$\phi: T \times T \times X \times \Omega \rightarrow X$$

is the *state transition function*, and

$$\eta: T \times X \rightarrow Y$$

is the *output function*. Additionally, we require:

i) Ω is a subset of the set of all functions from T to U such that Ω is closed under *splicing*, that is, for all $u_1, u_2 \in \Omega$ and all times $\tau \in T$, there exists $u \in \Omega$ such that

$$u(t) = \begin{cases} u_1(t) & \text{if } t \leq \tau \\ u_2(t) & \text{if } \tau < t \end{cases}$$

ii) ϕ satisfies the following *consistency conditions*:

$$A) \phi(t_0, t_0, x_0, u) = x_0$$

$$B) \phi(t_2, t_1, \phi(t_1, t_0, x_0, u), u) = \phi(t_2, t_0, x_0, u)$$

for all times t_0, t_1, t_2 , all states $x_0 \in X$, and all admissible input functions u .

iii) ϕ satisfies the following *causality condition*:

$$C) \phi(t_1, t_0, x_0, u) = \phi(t_1, t_0, x_0, u_1)$$

if $u(t) = u_1(t)$ for $t_0 \leq t < t_1$. ([PA], p. 194) ■

Next, we discuss conditions A–C of the definition above.

Consistency condition 1.46A states that if the dynamical system begins at the initial state x_0 at time t_0 , then, no matter what input u is applied, the state at time t_0 must be x_0 ([PA], p. 27).

Consistency condition 1.46B results from the fact that the state of the system at time t_2 cannot depend on the way it is computed. For example, say we begin at the state x_0 at time t_0 and apply the input function u . Then the state of the system at time t_1 will be $\phi(t_1, t_0, x_0, u)$. If we continue to apply u from time t_1 to t_2 , then the new state can be computed in two ways: (i) we may update the state $\phi(t_1, t_0, x_0, u)$ under the influence of u , from time t_1 to t_2 , to the new state $\phi(t_2, t_1, \phi(t_1, t_0, x_0, u), u)$; (ii) alternatively, we can begin at the state x_0 at time t_0 and apply u over the period $[t_0, t_2]$, and obtain the state $\phi(t_2, t_0, x_0, u)$. Finally, since the state at time t_2 does not depend on how it is computed, we obtain the equality given in consistency condition 1.46B. ([PA], p. 27)

Lastly, the causality condition 1.46C states that the transition from the state x_0 at time t_0 to the state $\phi(t_1, t_0, x_0, u)$ at time t_1 only depends on the values that the input function u takes on from t_0 to t_1 ; that is, if u and u_1 agree over the interval $[t_0, t_1]$, then the behavior of the state over this interval must be identical under the influence of either input function—even if the input functions agree only on this interval. Two consequences of the causality condition are the following. One, $\phi(t_1, t_0, x_0, u)$ is not affected by input to the dynamical system prior to time t_0 . Two, $\phi(t_1, t_0, x_0, u)$ is not affected by input to the dynamical system after time t_1 , so that starting in state x_0 , the output depends on input received between t_0 and t_1 . ([PA], pp. 28–30)

As a first step toward connecting the definitions we have used thus far to the generalized definition given in this section, we make the following observations. First, when the input function $\mathbf{u}$ over the interval $[t_0, t_1)$ is fixed, and t_0 is fixed, we may view ϕ as a function from $T \rightarrow X$, and in this case only, we will write:

$$\mathbf{x}(t_1) = \phi(t_1, t_0, \mathbf{x}(t_0), \mathbf{u})$$

to simplify notation. Notice that this is the same abuse of notation that we made following Definition 1.2 ([PA], p. 26). Secondly, we emphasize that on the interval $[t_0, t_1)$, only the values of the input $\mathbf{u}$ over this interval affect the state transition from $\mathbf{x}(t_0)$ to $\mathbf{x}(t_1)$. It will often be convenient to write

$$\mathbf{x}(t_1) = \phi(t_1, t_0, \mathbf{x}_0, \mathbf{u}_{[t_0, t_1)})$$

where the restriction of the function $\mathbf{u}$ to the subset $[t_0, t_1) \subseteq T$ is denoted $\mathbf{u}_{[t_0, t_1)}$ ([PA], p. 28).

In Sections 1.1–1.2, we defined continuous, discrete, and time invariant dynamical systems. These terms have analogous counterparts in the (more general) context of deterministic systems.

1.47 Definition (Continuous-time (resp. discrete-time) deterministic system) A deterministic system as described above is **continuous-time** if $T = \mathbb{R}$, the set of real numbers (resp. **discrete-time** if $T = \mathbb{Z}$, the integers). ■

We will usually abbreviate these terms as “continuous deterministic system” or “discrete deterministic system.”

In Sections 1.4.2 and 1.4.4 below, we will show that the continuous and discrete definitions of a dynamical system given in Sections 1.1 and 1.2 are particular cases of Definition 1.46.

1.4.1 Time Invariant Deterministic Systems

Recall that in updating the state of a dynamical system, the new state $\phi(t_1, t_0, \mathbf{x}_0, \mathbf{u})$ depends on the initial state $\mathbf{x}_0$, the input $\mathbf{u}$ over the interval $[t_0, t_1)$, as well as the initial time t_0 . However, the behavior of many dynamical systems is independent of the initial time used. For example, computer output depends only on the program being run and the data supplied—not on the time at which the program starts executing.

Toward stating the notion of time invariance mathematically, we begin by introducing the notion of a *translation operator* $z^{-\tau}$ which, given a function $\mathbf{u}$, $z^{-\tau}\mathbf{u}$ translates $\mathbf{u}$ by τ units according to the property:

$$z^{-\tau}\mathbf{u}(t) \equiv \mathbf{u}(t - \tau) \text{ for all } t, \tau \text{ in the domain of } \mathbf{u}$$

Notice that if D is the domain of $\mathbf{u}$, then D must have enough structure for $t - \tau \in D$ for all $t, \tau \in D$. In the context of deterministic systems, the translation operator will be applied to functions in Ω , the space of admissible input functions. Recall that elements of Ω were functions $\mathbf{u}: T \rightarrow U$. We usually will take T to be the real numbers or the integers—which certainly have enough structure; however, T was not required in Definition 1.46 to be either of these sets, so we will require here that T is closed with respect to addition. We also need the shift of each input function to be an admissible input function, i.e., that Ω is closed with respect to translation; $z^{-\tau}\mathbf{u} \in \Omega$ for each $\mathbf{u} \in \Omega$ and $\tau \in T$.

Now fix the input function $\mathbf{u}$ and initial state $\mathbf{x}_0$. At the end of the interval $[t_0, t_1)$ the state is given by the value $\phi(t_1, t_0, \mathbf{x}_0, \mathbf{u})$. If instead we begin at time $t_0 + \tau$, with the same initial state $\mathbf{x}_0$ and use the same input function, then we want the value of the state at the end of the interval $[t_0 + \tau, t_1 + \tau)$ to be equal to

$\phi(t_1, t_0, \mathbf{x}_0, \mathbf{u})$ —the value of the state at the end of the interval $[t_0, t_1)$. Precisely, we need to require that

$$\phi(t_1, t_0, \mathbf{x}_0, \mathbf{u}) = \phi(t_1 + \tau, t_0 + \tau, \mathbf{x}_0, z^{-\tau} \mathbf{u})$$

for all times $t_0, t_1 \in T$, all (positive or negative) delays $\tau \in T$, all (initial) states $\mathbf{x}_0$, and all input functions $\mathbf{u}$, where $z^{-\tau}$ is the *shift operator*: $[z^{-\tau} \mathbf{u}](t) = \mathbf{u}(t - \tau)$, for all $t \in T$. ([PA], pp. 39–40)

Note that the translation operator $z^{-\tau} \mathbf{u}$ accomplishes the task of applying the same sequence of inputs over the interval $[t_0 + \tau, t_1 + \tau)$ as $\mathbf{u}(t)$ applies over the interval $[t_0, t_1)$. As a simple example of why this works, consider Figure 1.3 below, in the case that $u: \mathfrak{R} \rightarrow \mathfrak{R}$. We see that at time $t \in [t_0 + \tau, t_1 + \tau)$ the input applied is $z^{-\tau} u(t) = u(t - \tau)$.

Figure 1.3 The translation operator $z^{-\tau}$

It also seems reasonable that a time invariant system's output should depend only on the current (fixed) state $\hat{\mathbf{x}}$ —not on the time the system is in this state.

This condition is expressed mathematically by the equation

$$\eta(t_0, \hat{\mathbf{x}}) = \eta(t_1, \hat{\mathbf{x}})$$

for all $t_0, t_1 \in T$ and all states $\hat{\mathbf{x}} \in X$.

We summarize the above considerations in the following definition.

1.48 Definition (Time invariant deterministic system) A deterministic system as described above is called **time invariant** if:

i) T is closed under addition;

ii) Ω is closed with respect to the *shift operator* $z^{-\tau}$ for every $\tau \in T$, where

$$[z^{-\tau}\mathbf{u}](t) = \mathbf{u}(t - \tau), \text{ for all } t \in T;$$

iii) the following two additional conditions are satisfied:

$$D) \phi(t_1, t_0, \mathbf{x}_0, \mathbf{u}) = \phi(t_1 + \tau, t_0 + \tau, \mathbf{x}_0, z^{-\tau}\mathbf{u})$$

for all $t_0, t_1, \tau \in T$, $\mathbf{x}_0 \in X$, and $\mathbf{u} \in \Omega$;

$$E) \eta(t_0, \hat{\mathbf{x}}) = \eta(t_1, \hat{\mathbf{x}})$$

for all $t_0, t_1 \in T$ and all states $\hat{\mathbf{x}} \in X$ ([PA], pp. 194–195). ■

Recall that in Sections 1.1.3 and 1.2.2 we defined time invariant linear dynamical systems to be ones in which the matrices were constant.

A natural question to ask at this point is, if we are given a continuous time or discrete time linear dynamical system which satisfies Definition 1.48, does that imply the matrices are constant? Conversely, if a system satisfies Definition 1.19 or 1.36, will it satisfy Definition 1.48? We verify that the answer is a qualified yes in the following proposition.

1.49 Proposition Suppose we are given a continuous time linear dynamical system defined by

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.64)$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) \text{ (output equation)} \quad (1.65)$$

or a discrete time linear system defined by

$$\mathbf{x}(t+1) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \quad (1.66)$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t) \quad (1.67)$$

Suppose further that the set Ω , of admissible input functions, includes all constant functions $\mathbf{u}(t) \equiv \mathbf{v} \in \mathfrak{R}^n$ for all $t \in T$, for each (fixed) $\mathbf{v} \in \mathfrak{R}^n$, and that $A(t)$, $B(t)$, and $C(t)$ are continuous. Then these matrices are constant if and only if conditions 1.48 D–E are satisfied.

Proof

Begin by noting that under the assumptions of the proposition, the output function η takes the particular form

$$\eta(t, \bar{\mathbf{x}}) = C(t)\bar{\mathbf{x}},$$

where $\mathbf{x}_0, \bar{\mathbf{x}} \in \mathfrak{R}^n$ are arbitrary (and fixed). Thus Condition 1.48E,

$$\eta(t_1, \bar{\mathbf{x}}) = \eta(t_2, \bar{\mathbf{x}}) \text{ for each } t_1, t_2 \in T,$$

is true if and only if $C(t_1)\bar{\mathbf{x}} = C(t_2)\bar{\mathbf{x}}$ for every $t_1, t_2 \in T$ and $\bar{\mathbf{x}} \in \mathfrak{R}^n$, if and only if $C(t_1) = C(t_2)$ for every $t_1, t_2 \in T$; because $C(t)$ is continuous, this means C is a constant matrix. Since the output equation in the discrete case is identical to the continuous case, this also proves the discrete case.

Therefore it remains to see that Condition 1.48D holds if and only if A and B are constant matrices. The continuous time case is proved first, and then the discrete time case.

First observe that under the assumptions of the proposition, and by Corollary 1.18, the state transition function ϕ takes the particular form

$$\phi(t, t_0, \mathbf{x}_0, \mathbf{u}) = \Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^t \Phi(t, w)B(w)\mathbf{u}(w)dw, \quad (1.68)$$

where $\Phi(t, w)$ is the fundamental matrix of Definition 1.11.

($\Rightarrow$) Begin by assuming that A and B are constant. Then by Theorem 1.21, $\Phi(t, w) = e^{(t-w)A}$, and it follows that

$$\phi(t, t_0, \mathbf{x}_0, \mathbf{u}) = e^{(t-t_0)A}\mathbf{x}_0 + \int_{t_0}^t e^{(t-w)A}B\mathbf{u}(w)dw \quad (1.69)$$

and

$$\begin{aligned}
\phi(t+\tau, t_0+\tau, \mathbf{x}_0, z^{-\tau}\mathbf{u}) &= \\
& e^{((t+\tau)-(t_0+\tau))A} \mathbf{x}_0 + \int_{t_0+\tau}^{t+\tau} e^{(t+\tau-w)A} B[z^{-\tau}\mathbf{u}](w) dw \\
&= e^{(t-t_0)A} \mathbf{x}_0 + \int_{t_0+\tau}^{t+\tau} e^{(t-(w-\tau))A} B\mathbf{u}(w-\tau) dw
\end{aligned}$$

Performing the change of variables, $s = w - \tau$, in the integrand of the last equation above gives

$$\begin{aligned}
\phi(t+\tau, t_0+\tau, \mathbf{x}_0, z^{-\tau}\mathbf{u}) &= e^{(t-t_0)A} \mathbf{x}_0 + \int_{t_0}^t e^{(t-s)A} B\mathbf{u}(s) ds \\
&= \phi(t, t_0, \mathbf{x}_0, \mathbf{u})
\end{aligned}$$

by equation (1.68) above .

($\Leftarrow$) Conversely, suppose Condition 1.48D is true. Then

$$\phi(t, t_0, \mathbf{x}_0, \mathbf{u}) = \phi(t+\tau, t_0+\tau, \mathbf{x}_0, z^{-\tau}\mathbf{u})$$

for all $t, t_0, \tau \in \mathbb{T}$, initial states $\mathbf{x}_0 \in X = \mathfrak{R}^n$, and $\mathbf{u} \in \Omega$. By equation (1.68) this last equation can be rewritten

$$\begin{aligned}
\Phi(t, t_0) \mathbf{x}_0 + \int_{t_0}^t \Phi(t, w) B(w) \mathbf{u}(w) dw &= \\
\Phi(t+\tau, t_0+\tau) \mathbf{x}_0 + \int_{t_0+\tau}^{t+\tau} \Phi(t+\tau, w) B(w) \mathbf{u}(w-\tau) dw & \quad (1.70)
\end{aligned}$$

To see that $B(t)$ is constant, take $\mathbf{x}_0 = \theta$ in (1.70) to get

$$\begin{aligned}
\int_{t_0}^t \Phi(t, s) B(s) \mathbf{u}(s) ds &= \int_{t_0+\tau}^{t+\tau} \Phi(t+\tau, w) B(w) \mathbf{u}(w-\tau) dw \\
&= \int_{t_0}^t \Phi(t+\tau, s+\tau) B(s+\tau) \mathbf{u}(s) ds
\end{aligned}$$

where the change of variables $s = w - \tau$ was used in obtaining the second integral. Taking $\mathbf{u}(t) \equiv \mathbf{e}_1, \dots, \mathbf{e}_n$ in this last equation gives integrals of continuous

functions over the same interval, and hence the integrands are equal for such $\mathbf{u}(t)$. Thus

$$\Phi(t, s)B(s) = \Phi(t + \tau, s + \tau)B(s + \tau)$$

Now let $s = t$, and since $\Phi(t, t)$ is the identity matrix for any t by Theorem 1.14, it follows that $B(s) = B(s + \tau)$ for every $\tau \in T$. Since $B(t)$ is continuous, this implies that $B(t)$ is constant.

Finally, to see $A(t)$ is constant, take $\mathbf{u}(t) \equiv \theta$ in (1.70) to obtain

$$\Phi(t, t_0)\mathbf{x}_0 = \Phi(t + \tau, t_0 + \tau)\mathbf{x}_0,$$

which holds for arbitrary $\mathbf{x}_0 \in \mathfrak{R}^n$, whence it follows that

$$\Phi(t, t_0) = \Phi(t + \tau, t_0 + \tau)$$

Taking partial derivatives with respect to t , it follows from Theorem 1.14 that

$$A(t)\Phi(t, t_0) = A(t + \tau)\Phi(t + \tau, t_0 + \tau)$$

Let $t = t_0$, and note that this last equation reduces to $A(t) = A(t + \tau)$. This holds for every $\tau \in T$, and $A(t)$ is continuous, so $A(t)$ is constant. This completes the continuous time case.

For the discrete case, it follows from Theorem 1.32 that

$$\phi(t, t_0, \mathbf{x}_0, \mathbf{u}) = \Phi(t, t_0)\mathbf{x}_0 + \sum_{w=t_0}^{t-1} \Phi(t, w+1)B(w)\mathbf{u}(w) \quad (1.71)$$

where $\Phi(t, \tau)$ is the fundamental matrix of Definition 1.31.

($\Rightarrow$) Assume that A and B are constant. Theorem 1.37 gives $\Phi(t, w) = A^{t-w}$, and it follows that

$$\begin{aligned} \phi(t + \tau, t_0 + \tau, \mathbf{x}_0, z^{-\tau}\mathbf{u}) &= \Phi(t + \tau, t_0 + \tau)\mathbf{x}_0 + \sum_{w=t_0+\tau}^{t+\tau-1} \Phi(t + \tau, w+1)B\mathbf{u}(w - \tau) \\ &= A^{t-t_0}\mathbf{x}_0 + \sum_{w=t_0+\tau}^{t+\tau-1} A^{t+\tau-(w+1)}B\mathbf{u}(w - \tau) \end{aligned}$$

Perform the change of variables, $s = w - \tau$, in the sum of this last equation to obtain

$$\begin{aligned}\phi(t + \tau, t_0 + \tau, \mathbf{x}_0, z^{-\tau} \mathbf{u}) &= A^{t-t_0} \mathbf{x}_0 + \sum_{s=t_0}^{t-1} A^{t-(s+1)} B \mathbf{u}(s) \\ &= \phi(t, t_0, \mathbf{x}_0, \mathbf{u})\end{aligned}$$

by equation (1.71) above.

($\Leftarrow$) If Condition 1.48D is true, then, as in the continuous time case, it follows that

$$\begin{aligned}\Phi(t, t_0) \mathbf{x}_0 + \sum_{w=t_0}^{t-1} \Phi(t, w+1) B(w) \mathbf{u}(w) &= \\ \Phi(t + \tau, t_0 + \tau) \mathbf{x}_0 + \sum_{w=t_0 + \tau}^{t + \tau - 1} \Phi(t + \tau, w+1) B(w) \mathbf{u}(w - \tau)\end{aligned}\tag{1.72}$$

To see that $B(w)$ is constant, take $\mathbf{x}_0 = \theta$ and $s = w - \tau$ in (1.72) to obtain

$$\sum_{s=t_0}^{t-1} \Phi(t, s+1) B(s) \mathbf{u}(s) = \sum_{s=t_0}^{t-1} \Phi(t + \tau, s + \tau + 1) B(s + \tau) \mathbf{u}(s)$$

As in the continuous time case, taking $\mathbf{u}(t) \equiv \mathbf{e}_1, \dots, \mathbf{e}_n$ and $t = t_0 + 1$, this last equation implies $B(t_0) = B(t_0 + \tau)$ for all $\tau \in T$; therefore $B(w)$ is constant because $B(w)$ is continuous.

To see that $A(t)$ is constant, proceed exactly as in the continuous case to obtain $\Phi(t, t_0) = \Phi(t + \tau, t_0 + \tau)$. Now use Theorem 1.35 (iv) to get

$$A(t) \Phi(t, t_0) = A(t + \tau) \Phi(t + \tau, t_0 + \tau),$$

The rest of the proof is identical to the argument in the continuous case. This completes the proof of the discrete case, and hence the proof. ■

It is often convenient to take the initial time t_0 of a time invariant dynamical system to be 0. This is justified since if $\tau = -t_0$ in condition 1.48D) of the definition above, then we obtain

$$\begin{aligned}\phi(t, t_0, \mathbf{x}_0, \mathbf{u}) &= \phi(t - t_0, t_0 - t_0, \mathbf{x}_0, z^{+t_0} \mathbf{u}) \\ &= \phi(t - t_0, 0, \mathbf{x}_0, z^{+t_0} \mathbf{u})\end{aligned}\tag{1.73}$$

which shows how to shift time from t_0 to zero, and conversely; and system output of a time invariant dynamical system is dependent only on the state. Moreover, we can use this observation to show that for a time invariant deterministic system, one can remove the mention of one time variable in the definition. We state this as another proposition.

1.50 Proposition Suppose we have a deterministic system that is time invariant. Then we can remove the reference to one time set T in Definition 1.46 if we define

$$\begin{aligned}\tilde{\phi}(t, \tilde{\mathbf{x}}, \mathbf{u}) &= \phi(t, 0, \tilde{\mathbf{x}}, \mathbf{u}) \\ \tilde{\eta}(\hat{\mathbf{x}}) &= \eta(0, \hat{\mathbf{x}})\end{aligned}$$

where ϕ is the state transition function, and η is the output function of the system. Furthermore,

i) the consistency conditions of Definition 1.46 can be written

$$\begin{aligned}A') \quad &\tilde{\phi}(0, \tilde{\mathbf{x}}, \mathbf{u}) = \tilde{\mathbf{x}} \\ B') \quad &\tilde{\phi}(t_2 - t_1, \tilde{\phi}(t_1 - t_0, \tilde{\mathbf{x}}, \mathbf{u}), z^{t_1 - t_0} \mathbf{u}) = \tilde{\phi}(t_2 - t_0, \tilde{\mathbf{x}}, \mathbf{u}) \text{ for } t_0, t_1, t_2 \in T;\end{aligned}$$

and

ii) the causality condition can be written

$$C') \quad \tilde{\phi}(t, \tilde{\mathbf{x}}, \mathbf{u}) = \tilde{\phi}(t, \tilde{\mathbf{x}}, \mathbf{u}_1) \text{ if } \mathbf{u} = \mathbf{u}_1 \text{ on } [0, t]$$

Proof

A'): By definition, $\tilde{\phi}(0, \tilde{\mathbf{x}}, \mathbf{u}) = \phi(0, 0, \tilde{\mathbf{x}}, \mathbf{u}) = \tilde{\mathbf{x}}$, where the last equality is by Condition 1.46A.

B'): We begin by proving the following claim: $z^{+t_1}[z^{-t_0} \mathbf{u}] = z^{t_1 - t_0} \mathbf{u}$ for all $t_0, t_1 \in T$.

Proof of Claim: For any $w \in T$, we have by definition of the shift operator

$$\begin{aligned}
 \left(z^{+t_1} [z^{-t_0} \mathbf{u}] \right) (w) &= [z^{-t_0} \mathbf{u}] (w + t_1) \\
 &= \mathbf{u}((w + t_1) - t_0) \\
 &= \mathbf{u}(w + (t_1 - t_0)) \\
 &= [z^{t_1 - t_0} \mathbf{u}] (w)
 \end{aligned}$$

Now,

$$\begin{aligned}
 &\check{\phi}(t_2 - t_1, \check{\phi}(t_1 - t_0, \bar{\mathbf{x}}, \mathbf{u}), z^{t_1 - t_0} \mathbf{u}) \\
 &= \phi(t_2 - t_1, 0, \phi(t_1 - t_0, 0, \bar{\mathbf{x}}, \mathbf{u}), z^{t_1 - t_0} \mathbf{u}) \text{ \{definition of } \check{\phi}\} \\
 &= \phi(t_2 - t_1, 0, \phi(t_1 - t_0, 0, \bar{\mathbf{x}}, z^{+t_0} [z^{-t_0} \mathbf{u}]), z^{+t_1} [z^{-t_0} \mathbf{u}]) \text{ \{Claim\}} \\
 &= \phi(t_2, t_1, \phi(t_1 - t_0, 0, \bar{\mathbf{x}}, z^{+t_0} [z^{-t_0} \mathbf{u}]), z^{-t_0} \mathbf{u}) \text{ \{equation (1.73)\}} \\
 &= \phi(t_2, t_1, \phi(t_1, t_0, \bar{\mathbf{x}}, z^{-t_0} \mathbf{u}), z^{-t_0} \mathbf{u}) \text{ \{equation (1.73)\}} \\
 &= \phi(t_2, t_0, \bar{\mathbf{x}}, z^{-t_0} \mathbf{u}) \text{ \{Condition 1.44B\}} \\
 &= \phi(t_2 - t_0, 0, \bar{\mathbf{x}}, z^{+t_0} [z^{-t_0} \mathbf{u}]) \text{ \{equation (1.73)\}} \\
 &= \phi(t_2 - t_0, 0, \bar{\mathbf{x}}, \mathbf{u}) \text{ \{Claim\}} \\
 &= \check{\phi}(t_2 - t_0, \bar{\mathbf{x}}, \mathbf{u}) \text{ \{definition of } \check{\phi}\}
 \end{aligned}$$

This completes the proof of B'.

C'):

$$\begin{aligned}
 \check{\phi}(t, \bar{\mathbf{x}}, \mathbf{u}) &= \phi(t, 0, \bar{\mathbf{x}}, \mathbf{u}) \\
 &= \phi(t, 0, \bar{\mathbf{x}}, \mathbf{u}_1) \text{ \{since } \mathbf{u} = \mathbf{u}_1 \text{ on } [0, t)\} \\
 &= \check{\phi}(t, \bar{\mathbf{x}}, \mathbf{u}_1) \text{ \{by definition of } \check{\phi}\}
 \end{aligned}$$

The second equality above is by Condition 1.46C. ■

1.4.2 Connecting Notions of Continuous-Time Dynamical Systems

Our purpose in this section is to show that Definitions 1.1–1.2 are a particular case of Definition 1.46. Recall that Definition 1.1 says a continuous dynamical system is described by a differential equation of the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad (1.74)$$

where $\mathbf{f}: T \times X \times U \rightarrow \mathbb{R}^n$, $T \subseteq \mathbb{R}$, and $X \subseteq \mathbb{R}^n$. We saw that if $\mathbf{f}$ is sufficiently well behaved and we fix $\mathbf{u}(t)$, an initial time t_0 , and an initial state $\mathbf{x}_0$, then there will be a unique solution to (1.74) such that $\mathbf{x}(t_0) = \mathbf{x}_0$. We can then define

$$\phi: T \times T \times X \times U \rightarrow X$$

by $\phi(t, t_0, \mathbf{x}_0, \mathbf{u}) = \mathbf{x}(t)$.

This function ϕ will satisfy Definition 1.46 if we also choose the set Ω of input functions so that the splicing condition 1.46(i) holds.

To see that Conditions 1.46(ii) and 1.46(iii) hold, choose any t_0 , $\mathbf{x}_0$, and $\mathbf{u}$, let $\mathbf{x}(t)$ be the unique solution to (1.74) so that $\mathbf{x}(t_0) = \mathbf{x}_0$, and calculate:

$$(ii-A) \quad \phi(t_0, t_0, \mathbf{x}_0, \mathbf{u}) = \mathbf{x}_0 = \mathbf{x}(t_0)$$

(ii-B) First note that $\phi(t_2, t_1, \phi(t_1, t_0, \mathbf{x}_0, \mathbf{u}), \mathbf{u}) = \phi(t_2, t_1, \mathbf{x}(t_1), \mathbf{u})$; let $\mathbf{y}(t) = \phi(t, t_1, \mathbf{x}(t_1), \mathbf{u})$ so $\mathbf{y}(t)$ is the unique solution to (1.74) such that $\mathbf{y}(t_1) = \mathbf{x}(t_1)$. But only one solution has the value $\mathbf{x}(t_1)$ at t_1 and that is $\mathbf{x}(t)$, i.e., $\mathbf{y}(t)$ must be $\mathbf{x}(t)$. In particular, $\mathbf{y}(t_2) = \phi(t_2, t_1, \mathbf{x}(t_1), \mathbf{u}) = \mathbf{x}(t_2)$ and $\mathbf{x}(t_2) = \phi(t_2, t_1, \mathbf{x}_0, \mathbf{u})$. Hence we have

$$\phi(t_2, t_1, \phi(t_1, t_0, \mathbf{x}_0, \mathbf{u}), \mathbf{u}) = \phi(t_2, t_0, \mathbf{x}_0, \mathbf{u})$$

as desired. Condition (iii) is also valid because if $\mathbf{u} = \mathbf{u}_1$ on $[t_0, t_1]$, then integration to solve equation (1.74) on this interval will give the same solution $\mathbf{x}(t)$ whether one uses $\mathbf{u}$ or $\mathbf{u}_1$, that is for any interval $[t_0, t_1]$ and initial state $\mathbf{x}_0$,

$$(iii-C) \quad \phi(t_1, t_0, \mathbf{x}_0, \mathbf{u}) = \phi(t_1, t_0, \mathbf{x}_0, \mathbf{u}_1)$$

The output function $g: T \times X \rightarrow \mathbb{R}^p$ given in Definition 1.1 qualifies as an output function according to Definition 1.46 when we take $Y = \mathbb{R}^p$.

1.4.3 More General Examples of Deterministic Systems Described by Differential Equations

Given a deterministic system defined by Definition 1.46, what kind of additional assumptions are necessary to describe the state in terms of a differential equation like equation (1.1) of Definition 1.1? As in Section 1.4, it will be convenient to define $x(t) = \phi(t, t_0, x_0, u)$ when t_0 , u , and x_0 have been fixed.

Not every continuous dynamical system will arise as a differential equation like (1.1), as shown in Example 1.5. For another example, consider a continuous dynamical system that has a “reset to zero” input function in U —that is, when this input is applied at any time t , the state is set to be the initial state. Then the splicing condition could yield highly noncontinuous input functions, and such a system might not be describable by a differential equation ([PA], p. 66).

Begin by assuming that a continuous-time dynamical system is described by Definition 1.46. A natural question at this point is what kind of structure must be imposed on the sets U , X , and Y (output set) in order for the system’s behavior to be described in terms of a differential equation as given in equation (1.1) of Definition 1.1? Since computation of the derivative requires that we be able to form differences and take limits, these sets must have enough structure to perform basic arithmetic operations and take limits. Typically, U , X , and Y are taken to be $\mathbb{R}^n$, C^k , or more generally, **Banach spaces** (a vector space which has a norm and is complete with respect to the metric induced by the norm). In the special case $X = \mathbb{R}^n$, and T is an interval $[a, b] \subseteq \mathbb{R}$, the mapping $x: T \rightarrow X$ may be

visualized as a curve in n – space traced out by the state vector $\mathbf{x}(t)$, as t increases from a to b . ([PA], pp. 65, 70)

Consider the derivative of the state at time t , which is given by

$$\dot{\mathbf{x}}(t) = \lim_{\Delta t \rightarrow 0} \frac{\mathbf{x}(t + \Delta t) - \mathbf{x}(t)}{\Delta t}$$

Assuming that the input is $\mathbf{u}$, and using function ϕ , this equation may be rewritten as follows:

$$\dot{\mathbf{x}}(t) = \lim_{\Delta t \rightarrow 0} \frac{\phi(t + \Delta t, t, \mathbf{x}(t), \mathbf{u}) - \phi(t, t, \mathbf{x}(t), \mathbf{u})}{\Delta t}$$

If this limit depends only on the value of $\mathbf{u}$ at t , then we have

$$\dot{\mathbf{x}}(t) = \left. \frac{\partial}{\partial t_1} \phi(t_1, t, \mathbf{x}(t), \mathbf{u}(t)) \right|_{t_1=t}$$

so we can define

$$\mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) = \left. \frac{\partial}{\partial t_1} \phi(t_1, t, \mathbf{x}(t), \mathbf{u}(t)) \right|_{t_1=t}$$

and then the system is defined by the differential equation

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

([PA], pp. 66–67)

However, $\dot{\mathbf{x}}(t)$ could fail to be this simple, as demonstrated in the following example.

1.51 Example Consider a capacitor in an electrical circuit, where the voltage across the capacitor is input to the system and the current running through the capacitor is the state of the system. It can be shown that the differential equation for this system is given by

$$\dot{\mathbf{x}}(t) = k\ddot{u}(t)$$

for some constant k . In this case, the derivative of the state depends on the second derivative of u at t , and not just $u(t)$. ([PA], p. 68) ■

Observe that the requirement that the system be described by a differential equation does *not* force $\mathbf{f}$ to be differentiable. Piecewise continuous input functions or piecewise continuous functions are acceptable, as long as the system differential equation results in an integrable differential equation after the input function is substituted into the system differential equation ([PA], p. 67). This situation is demonstrated in the following example.

1.52 Example. Consider the one-dimensional scalar differential equation,

$$\dot{x}(t) = u(t)$$

where $u(t)$ is taken to be the **Heaviside step function**,

$$u(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1 & \text{if } t > 0 \end{cases}$$

so that $u(t)$ is piecewise continuous but not differentiable. However we can integrate this differential equation to get the state trajectory

$$x(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ t & \text{if } t > 0 \end{cases}$$

where the initial condition is $x(t_0) = 0$ for some $t_0 \leq 0$. ■

1.4.4 Connecting Notions of Discrete-Time Dynamical Systems

In this subsection we show that discrete dynamical system as characterized by equation (1.41) and Definition 1.27 is a special case of Definition 1.46. This case is much simpler because sums and limits in the state space are not needed.

Consider a discrete dynamical system with fixed input, as defined in Definition 1.27:

$$\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t)), \quad (1.75)$$

where $\mathbf{x}: \mathbb{N} \rightarrow \mathbb{R}^n$, $\mathbf{f}: \mathbb{N} \times X \rightarrow \mathbb{R}^n$, and $X \subseteq \mathbb{R}^n$. When $\mathbf{x}_0$ is fixed, the solution, or state transition function $\tilde{\phi}: \mathbb{N} \times X \rightarrow \mathbb{R}^n$ is defined by $\tilde{\phi}(0; \mathbf{x}_0) = \mathbf{x}_0$ and $\tilde{\phi}(t+1, \mathbf{x}_0) = \mathbf{f}(t, \tilde{\phi}(t, \mathbf{x}_0))$. As shown in Example 1.29, it is possible to write equations $\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t))$ for which no solution exists, but we assume here that one does exist. In Definition 1.46, T was an arbitrary set, so now take $T = \mathbb{N}$.

If a discrete-time system allows input functions, then it would be described by $\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$. If we assume that any state $\hat{\mathbf{x}}$ can be reached at any time n , i.e., there is $\mathbf{x}_0$ so that defining $\mathbf{x}(0) = \mathbf{x}_0$ and $\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$ yields $\mathbf{x}(n) = \hat{\mathbf{x}}$, then we can define a transition function thus: if $\mathbf{u}$ and n are fixed and $\mathbf{x}$ is a particular solution such that $\mathbf{x}(n) = \hat{\mathbf{x}}$, then $\phi(n+k, n, \hat{\mathbf{x}}, \mathbf{u}) = \mathbf{x}(n+k)$. That is, for any solution $\mathbf{x}$, we have

$$\mathbf{x}(n+k) = \phi(n+k, n, \mathbf{x}(n), \mathbf{u}) \quad (1.76)$$

We can use equation (1.76) to see that the causality conditions of Definition 1.46 are satisfied for function ϕ :

- (i-A) $\phi(n, n, \hat{\mathbf{x}}, \mathbf{u}) = \mathbf{x}(n)$ by (1.76) and $\mathbf{x}(n) = \hat{\mathbf{x}}$;
- (i-B) $\phi(n+k, n, \phi(n, 0, \mathbf{x}_0, \mathbf{u}), \mathbf{u}) = \phi(n+k, n, \mathbf{x}(n), \mathbf{u})$ by (1.76) and the latter equals $\mathbf{x}(n+k)$; in addition, $\mathbf{x}(n+k) = \phi(n+k, 0, \mathbf{x}_0, \mathbf{u})$ by (1.76);
- (ii-C) if $\mathbf{u} = \mathbf{u}_1$ on $\{0, 1, \dots, n-1\}$, then $\phi(n, 0, \mathbf{x}_0, \mathbf{u}) = \phi(n, 0, \mathbf{x}_0, \mathbf{u}_1)$ is clear.

CHAPTER 2

STABILITY OF DYNAMICAL SYSTEMS

2.0 ABOUT THIS CHAPTER

In this chapter the question of the stability of a dynamical system is considered. This stability will be defined with respect to an equilibrium state—which can be thought of intuitively as a place of rest or no movement for the dynamical system. Loosely speaking, a dynamical system is stable if as long as the initial state is selected close enough to the equilibrium state, the state itself remains near the equilibrium state—perhaps even approaching equilibrium as time increases.

In Section 2.1 a definition of an equilibrium state is given, and Liapunov's "first method" of linear approximation of a dynamical system near an equilibrium state is discussed. In Section 2.2 all of the stability types discussed in this chapter are defined, and Liapunov's "second" or "indirect" method of stability analysis is described. The stability of time invariant linear dynamical

systems is discussed in Section 2.3, and BIBO stability is considered in Section 2.4. Finally, in Section 2.5 we generalize the notion of an equilibrium state and stability to the more general context of deterministic systems—which were first defined and discussed in Section 1.4.

The origins of stability theory for dynamical systems can be traced back to one man: Aleksandr Mikhaylovich Lyapunov (1857–1918). The spelling “Liapunov” occurs frequently in the literature, and is used throughout this chapter. Moreover, recent works in the theory of dynamical systems still reference theorems Liapunov proved in his doctoral dissertation, which is dated “5 April 1892” [Lya]. Little biographical information about his life appears in most books that discuss his work, but there is a biography of him written by V. I. Smirnov in the collected works of Liapunov ([Lya], v. 1, pp. 3–15), and we present a few highlights from that.

Born on May 25, 1857 in Yaroslavl, which is just north of Moscow, he was the eldest of the three sons of Mikhail V. Liapunov. His father graduated from Kazan University in 1839, became an observing astronomer at Kazan University in 1840, and became director of the Demidov Lyceum in Yaroslavl in 1856. His father married Sophiya Aleksandrovna Shipilova in 1852, and the couple had seven children—four of whom died while infants. Liapunov’s second youngest brother, Sergey Mikhaylovich (1859–1924), was a well-known composer, and his youngest brother, Boris Mikhaylovich (1864–1942), specialized in Slavic philology. ([Lya], v.1, p. 3)

In 1876 Liapunov entered St. Petersburg University to study natural sciences, but transferred a month later to study mathematics. At the university, Liapunov was especially interested in lectures given by P. L. Chebyshev. However, it was

under the guidance of D. K. Bobylev that Liapunov began to work independently, winning a gold medal for a paper in hydrostatics submitted to the math department, and later publishing two papers on hydrostatics in the "Journal of the Physical Chemistry Society." ([Lya], v.1, p.5)

Liapunov passed his master's examinations in 1882, and defended his master's dissertation, "On the Stability of Ellipsoidal Forms of Equilibrium in Revolving Liquid," in January 1885, at St. Petersburg University. In the spring of 1885 (?) he accepted an offer to occupy the chair of mechanics at Khar'kov University.

The first part of his time at Khar'kov University was dedicated to composing notes for the courses in mechanics that he taught. However, in 1888 papers by Liapunov on the stability of motion for mechanical systems began to appear in the press, and in September 1892 he defended his doctoral dissertation, "The General Problem of Stability of Motion," at Moscow University. This dissertation, as well as his master's, was translated into French and printed in "Annales de l'Université de Toulouse." ([Lya], v.1, pp. 6–9) Moreover, Liapunov developed the theory that comprises the "first" and "second" methods that we consider later in this chapter. In his own words:

The problem which I set myself in starting the present investigation may be formulated thus: to indicate those cases in which a first approximation indeed resolves the question of stability and to give some means enabling it to be resolved at least in certain of the cases where it is impossible to judge stability by a first approximation. ([Lya], v.2, p. 7)

On November 6, 1901 Liapunov accepted an offer to occupy the chair of applied mathematics at St. Petersburg, which had been left vacant since the death of P. L. Chebyshev in 1894. In the spring of 1902 he returned to St. Petersburg,

and the rest of his life was not dedicated to teaching, but rather to scientific research. Liapunov authored a number of papers that continued his research in stability theory up until his death on November 3, 1918. ([Lya], v.1, pp. 10–14)

2.1 EQUILIBRIUM AND LINEARIZATION

We will define stability, essentially as Liapunov did, in reference to an equilibrium state. In this first section, we will define an equilibrium state, and also consider how a nonlinear dynamical system can be approximated by a linear dynamical system around an equilibrium state. In Chapter 1 we saw that linear dynamical systems can be solved quite easily, so linear approximation is a powerful tool to help analyze nonlinear systems. In many cases, linear approximations reveal stability properties of the nonlinear system ([Lue], p. 324). This method of linearization is often referred to as “Liapunov’s first method” or “Liapunov’s indirect method” ([Lue], p. 325).

We begin by presenting our working definition of equilibrium. In Section 2.5 we will consider a generalization of this definition, which can be used in the more general context of deterministic systems as given by Definition 1.46.

2.1 Definition (Equilibrium state). Suppose we are given a dynamical system according to Definitions 1.1–1.2 in the continuous case, and Definitions 1.26–1.27 in the discrete case. Let $\mathbf{u}$ be the input function and take $T = [t_0, \infty)$ in the continuous case, $T = \mathbf{N}$ in the discrete case. Then a vector $\mathbf{x}_e \in X$ is an **equilibrium state** for the continuous dynamical system described by the differential equation $\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$, if

$$\mathbf{f}(t, \mathbf{x}_e, \mathbf{u}(t)) = \mathbf{0} \text{ for all } t \geq t_0, \quad (2.1)$$

and for the discrete dynamical system described by the difference equation $\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$, if

$$\mathbf{f}(t, \mathbf{x}_e, \mathbf{u}(t)) = \mathbf{x}_e \text{ for all } t = 0, 1, 2, \dots \quad (2.2)$$

([SB], pp. 87–88). ■

Note that this definition agrees with our intuitive sense that a dynamical system in equilibrium is one that is not changing with respect to time, or motionless. Equation (2.1) just expresses the fact that the derivative of a constant function is zero, and (2.2) expresses no movement in the context of discrete systems.

2.1.1 Linearization of Nonlinear Dynamical Systems with Known Equilibrium

As mentioned in the beginning of this section, we can check stability properties of a nonlinear dynamical system by linearizing the system near an equilibrium state. This technique will be demonstrated first for a one dimensional state space $X \subseteq \mathbb{R}$ and then for the more general case of $X \subseteq \mathbb{R}^n$.

Suppose for convenience that the input is zero and we have a continuous one-dimensional dynamical system

$$\dot{x}(t) = f(x(t)) \quad (2.3)$$

where $f: X \rightarrow \mathbb{R}$, $x: T \rightarrow \mathbb{R}$, and $X \subseteq \mathbb{R}$. Further, suppose f is a nonlinear differentiable function of x , and we wish to obtain a linear approximation to f near the equilibrium point x_e .

First, if f' is continuous at x then it's true that for small values of h we may approximate values $f(x+h)$ according to the linear approximation

$$f(x+h) \approx f(x) + f'(x)h \quad (2.4)$$

which follows immediately from the fact that $f'(x) = \frac{f(x+h) - f(x)}{h}$ for small

values of h . If $x = x_e$, $f(x_e) = 0$ by (2.1) so we obtain

$$f(x_e + h) \approx f'(x_e)h \quad (2.5)$$

Now let $h = x_\delta = x - x_e$ be the distance of the state from the equilibrium x_e . Then

$\dot{x}_\delta = \dot{x}$ and by (2.3) and (2.5) we obtain

$$\dot{x}_\delta = \dot{x} = f(x) = f(x_e + x_\delta) \approx f'(x_e)x_\delta,$$

where the second equality is by (2.3) and the last approximation is by (2.5). That is,

$$\dot{x}_\delta \approx f'(x_e)x_\delta, \quad (2.6)$$

where the initial state of this system is $x_\delta(t_0) = x(t_0) - x_e$. Using equality in (2.6) gives a time invariant linear (one-dimensional) dynamical system, which we developed methods to solve in the last chapter. In this case, the matrix A equals $f'(x_e)$, and by Corollary 1.13 and Theorem 1.21 we have

$$x_\delta(t) \approx e^{(t-t_0)f'(x_e)}x_\delta(t_0) \quad (2.7)$$

Solution (2.7) can now be used to approximate the original state $x(t)$ according to the equation $x(t) = x_e + x_\delta(t)$, for $x(t)$ sufficiently near x_e .

Next, we consider a discrete dynamical system given by

$$x(t+1) = f(x(t)) \quad (2.8)$$

where the only difference here from the case above is that $T = N$. Since f is assumed to be a differentiable function of x such that f' is continuous at x , we can use the same derivation to obtain the linear approximation given in equation (2.4): $f(x+h) \approx f(x) + f'(x)h$. Let $x(k) = x_\delta(k) + x_e$, and note that by (2.8) we have

$$x(k+1) = f(x(k)) = f(x_e + x_\delta(k)) \approx f(x_e) + f'(x_e)x_\delta(k) \quad (2.9)$$

Also, $x(k+1) = x_e + x_\delta(k+1)$ and by definition of an equilibrium state for a discrete system, $f(x_e) = x_e$, so (2.9) can be rewritten

$$x_\delta(k+1) + x_e \approx x_e + f'(x_e)x_\delta(k)$$

Hence

$$x_\delta(k+1) \approx f'(x_e)x_\delta(k) \quad (2.10)$$

Using equality in (2.10) gives a time invariant one-dimensional linear discrete dynamical system, where matrix $A = f'(x_e)$ as it did in the continuous case

above. By Corollary 1.33 and Theorem 1.37, we obtain the approximation for x_δ given by

$$x_\delta(k) = [f''(x_e)]^k x_\delta(0), \quad x_\delta(0) = x(0) - x_e, \quad (2.11)$$

for $k = 0, 1, 2, \dots$, where, as above, we have assumed input is zero here. As in the continuous case, solution (2.11) can be used to approximate the state $x(k)$ according to the equation $x(k) = x_\delta(k) + x_e$ and for initial states $x(0) = x_0$ close to the equilibrium point.

Now we will generalize the approximations above to the n -dimensional case. The following theorem is based on [BL], pp. 113–121, and is the basis for the approximations given below.

2.2 Theorem Suppose X is an open subset of $\mathfrak{R}^n$ and $f: X \subseteq \mathfrak{R}^n \rightarrow \mathfrak{R}$ has partial derivatives throughout a neighborhood $N(x, \varepsilon)$ of $x \in X$. Let

$D_x = \{v \in \mathfrak{R}^n: x + v \in X\}$, and suppose further that

$$\frac{\partial f}{\partial x_j}: N(x, \varepsilon) \subseteq \mathfrak{R}^n \rightarrow \mathfrak{R}$$

is continuous at x for $j = 1, 2, \dots, n$. Then there exists a function $\eta: D_x \rightarrow \mathfrak{R}$ such that for all $v \in D_x$

$$f(x + v) - f(x) = v_1 \frac{\partial f}{\partial x_1}(x) + \dots + v_n \frac{\partial f}{\partial x_n}(x) + \|v\| \eta(v)$$

where $\lim_{v \rightarrow 0} \eta(v) = 0$, $x = [x_1 \ \dots \ x_n]^T$, and $v = [v_1 \ \dots \ v_n]^T$. ■

If $\|v\|$ is small, so $x + v$ is near x , we can drop the term $\|v\| \eta(v)$ and obtain the approximation

$$f(x + v) \approx f(x) + v_1 \frac{\partial f}{\partial x_1}(x) + \dots + v_n \frac{\partial f}{\partial x_n}(x),$$

which will be used repeatedly below.

Suppose we are given the nonlinear functions $f(t, x(t), u(t))$ and $g(t, x(t))$, where $x(t)$, $u(t)$, $f(t, x(t), u(t))$, and $g(t, x(t))$ are all in $\mathfrak{R}^n$, and suppose

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \text{ (respectively, } \mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) \text{)}$$

in the continuous (resp. discrete) case. Let

$$\mathbf{y}(t) = \mathbf{g}(t, \mathbf{x}(t))$$

be the output function in both cases. Assume further that the first order partial derivatives of $\mathbf{f}$ and $\mathbf{g}$ exist in a neighborhood $N(\mathbf{x}_e, \varepsilon)$ of the equilibrium state $\mathbf{x}_e$ and are continuous at $\mathbf{x}_e$. Fix an input function $\bar{\mathbf{u}}(t)$ from the set of admissible input functions, and let $\mathbf{x}_e = (x_i^e)$, where each $x_i^e \in \mathcal{R}$, be an equilibrium state of the system associated with this particular choice of input.

Suppose first that the input is zero. Then we obtain the simpler form

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) = (f_i(\mathbf{x}(t))) \text{ (resp. } \mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t)) \text{ in the discrete case)}$$

By Theorem 2.2, each function f_i can be approximated by the relation

$$\begin{aligned} & f_i(x_1 + x_1^e, x_2 + x_2^e, \dots, x_n + x_n^e) \\ &= f_i(x_1^e, x_2^e, \dots, x_n^e) + x_1 \frac{\partial}{\partial x_1} f_i(x_1^e, \dots, x_n^e) \\ & \quad + x_2 \frac{\partial}{\partial x_2} f_i(x_1^e, \dots, x_n^e) + \dots + x_n \frac{\partial}{\partial x_n} f_i(x_1^e, \dots, x_n^e) \end{aligned}$$

In this special case of zero input, the linearized version of $\mathbf{f}$ can be expressed in vector notation as

$$\mathbf{f}(\mathbf{x} + \mathbf{x}_e) \approx \mathbf{f}(\mathbf{x}_e) + \mathbf{J}(\mathbf{x}_e) \cdot \mathbf{x} \quad (2.12)$$

where

$$\mathbf{J}(\mathbf{x}_e) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}_{\mathbf{x}=\mathbf{x}_e}$$

([Lue], p. 326).

2.3 Definition. The matrix of partial derivatives above is called the **Jacobian matrix** of the function f . ■

Continuing to assume the input is zero, we can approximate much as we did in the one dimensional case. Consider first the continuous system $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$. Let $\mathbf{x}_\delta(t) = \mathbf{x}(t) - \mathbf{x}_e$ be the deviation of the state from the equilibrium state. Then $\mathbf{x}(t) = \mathbf{x}_e + \mathbf{x}_\delta(t)$ and by (2.12)

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) = \mathbf{f}(\mathbf{x}_e + \mathbf{x}_\delta(t)) \approx \mathbf{f}(\mathbf{x}_e) + \mathbf{J}(\mathbf{x}_e) \cdot \mathbf{x}_\delta(t),$$

where the approximation is valid as long as $\|\mathbf{x}_\delta(t)\|$ is small, i.e., $\mathbf{x}(t)$ is near $\mathbf{x}_e$.

Since $\dot{\mathbf{x}} = \dot{\mathbf{x}}_\delta$ and $\mathbf{f}(\mathbf{x}_e) = \mathbf{0}$, we have

$$\dot{\mathbf{x}}_\delta(t) \approx \mathbf{J}(\mathbf{x}_e) \cdot \mathbf{x}_\delta(t), \text{ with initial state } \mathbf{x}_\delta(t_0) = \mathbf{x}(t_0) - \mathbf{x}_e \quad (2.13)$$

In the discrete case, where $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$, again let $\mathbf{x}(t) = \mathbf{x}_e + \mathbf{x}_\delta(t)$, substitute this into the dynamical system, and use equation (2.12) to get

$$\mathbf{x}_e + \mathbf{x}_\delta(t+1) \approx \mathbf{f}(\mathbf{x}_e) + \mathbf{J}(\mathbf{x}_e) \cdot \mathbf{x}_\delta(t)$$

Since $\mathbf{f}(\mathbf{x}_e) = \mathbf{x}_e$, this gives

$$\mathbf{x}_\delta(t+1) \approx \mathbf{J}(\mathbf{x}_e) \cdot \mathbf{x}_\delta(t), \text{ with initial state } \mathbf{x}_\delta(0) = \mathbf{x}(0) - \mathbf{x}_e \quad (2.14)$$

Notice that equations (2.13)–(2.14) give time invariant linear dynamical systems, and hence can be solved using methods given in Chapter 1. Assuming that the initial state $\mathbf{x}_0$ is fixed, as usual, let the unique solution to these equations also be $\mathbf{x}_\delta(t)$. Then the state of the original system can be approximated by $\mathbf{x}(t) = \mathbf{x}_e + \mathbf{x}_\delta(t)$, for small t , when the initial state is sufficiently close to $\mathbf{x}_e$.

We have assumed that input to the system is zero in the discussion above. The case of nonzero input is developed similarly, and we summarize the results ([SB], pp. 89–90).

To begin, define two matrices that are a generalization of the Jacobian matrix defined in Definition 2.3, for a nonlinear function $\mathbf{f}$ with continuous partial derivatives in a neighborhood $N(\mathbf{x}_e, \varepsilon)$ of $\mathbf{x}_e$ and $N(\bar{\mathbf{u}}(t), \mu)$ of $\bar{\mathbf{u}}(t)$.

2.4 Definition. Let $\mathbf{x}(t) = (x_i(t))$ and $\mathbf{f}(t) = (f_i(t))$ for $i = 1, 2, \dots, n$, and let $\mathbf{u}(t) = (u_j(t))$ for $j = 1, 2, \dots, m$. Define the matrices $J_x(t, \mathbf{x}, \mathbf{u})$ and $J_u(t, \mathbf{x}, \mathbf{u})$, the **Jacobian of $\mathbf{f}$ with respect to $\mathbf{x}$** and the **Jacobian of $\mathbf{f}$ with respect to $\mathbf{u}$** , respectively, by the following equations:

$$J_x(t, \mathbf{x}, \mathbf{u}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(t, \mathbf{x}, \mathbf{u}) & \frac{\partial f_1}{\partial x_2}(t, \mathbf{x}, \mathbf{u}) & \cdots & \frac{\partial f_1}{\partial x_n}(t, \mathbf{x}, \mathbf{u}) \\ \frac{\partial f_2}{\partial x_1}(t, \mathbf{x}, \mathbf{u}) & \frac{\partial f_2}{\partial x_2}(t, \mathbf{x}, \mathbf{u}) & \cdots & \frac{\partial f_2}{\partial x_n}(t, \mathbf{x}, \mathbf{u}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(t, \mathbf{x}, \mathbf{u}) & \frac{\partial f_n}{\partial x_2}(t, \mathbf{x}, \mathbf{u}) & \cdots & \frac{\partial f_n}{\partial x_n}(t, \mathbf{x}, \mathbf{u}) \end{bmatrix}$$

and

$$J_u(t, \mathbf{x}, \mathbf{u}) = \begin{bmatrix} \frac{\partial f_1}{\partial u_1}(t, \mathbf{x}, \mathbf{u}) & \frac{\partial f_1}{\partial u_2}(t, \mathbf{x}, \mathbf{u}) & \cdots & \frac{\partial f_1}{\partial u_m}(t, \mathbf{x}, \mathbf{u}) \\ \frac{\partial f_2}{\partial u_1}(t, \mathbf{x}, \mathbf{u}) & \frac{\partial f_2}{\partial u_2}(t, \mathbf{x}, \mathbf{u}) & \cdots & \frac{\partial f_2}{\partial u_m}(t, \mathbf{x}, \mathbf{u}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial u_1}(t, \mathbf{x}, \mathbf{u}) & \frac{\partial f_n}{\partial u_2}(t, \mathbf{x}, \mathbf{u}) & \cdots & \frac{\partial f_n}{\partial u_m}(t, \mathbf{x}, \mathbf{u}) \end{bmatrix}$$

Notice that J_u need not be square. ■

As before, choose a particular input function $\bar{\mathbf{u}}(t)$ and let $\mathbf{x}_e$ be an equilibrium state of a continuous or discrete nonlinear dynamical system, computed using this particular input function. Then by Theorem 2.2 applied to each component function f_i of $\mathbf{f}$, we have

$$\mathbf{f}(t, \mathbf{x}, \mathbf{u}) \approx \mathbf{f}(t, \mathbf{x}_e, \bar{\mathbf{u}}) + J_x(t, \mathbf{x}_e, \bar{\mathbf{u}})(\mathbf{x} - \mathbf{x}_e) + J_u(t, \mathbf{x}_e, \bar{\mathbf{u}})(\mathbf{u} - \bar{\mathbf{u}}), \quad (2.15)$$

which is a good approximation if $\mathbf{x}(t)$ and $\mathbf{u}(t)$ are sufficiently near $\mathbf{x}_e$ and $\bar{\mathbf{u}}(t)$.

In the continuous case, where the state is given by the differential equation $\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \bar{\mathbf{u}}(t))$, the linear approximation gives

$$\dot{\mathbf{x}}_\delta(t) = \mathbf{A}(t)\mathbf{x}_\delta(t) + \mathbf{B}(t)\mathbf{u}_\delta(t), \quad (2.16)$$

and in the discrete case, with $\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \bar{\mathbf{u}}(t))$, one obtains

$$\mathbf{x}_\delta(t+1) = \mathbf{A}(t)\mathbf{x}_\delta(t) + \mathbf{B}(t)\mathbf{u}_\delta(t), \quad (2.17)$$

where

- 1) $\mathbf{A}(t) = \mathbf{J}_x(t, \mathbf{x}_e, \bar{\mathbf{u}}(t));$
- 2) $\mathbf{B}(t) = \mathbf{J}_u(t, \mathbf{x}_e, \bar{\mathbf{u}}(t));$
- 3) $\mathbf{x}_\delta(t) = \mathbf{x}(t) - \mathbf{x}_e$ and $\mathbf{u}_\delta(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t)$

Since systems (2.16) and (2.17) are linear, we can solve them easily using methods described in the first chapter. Abusing notation and calling the unique solution of (2.16) or (2.17) $\mathbf{x}_\delta$, we can approximate the state $\mathbf{x}(t)$ of the nonlinear system using 3): $\mathbf{x}(t) = \mathbf{x}_e + \mathbf{x}_\delta(t)$. Again, we emphasize that the initial state of the nonlinear system, $\mathbf{x}(t_0) = \mathbf{x}_0$, must be chosen near the equilibrium and t must be near t_0 for the approximation to be accurate.

Finally, we consider a linear approximation of a nonlinear output relation,

$$\mathbf{y}(t) = \mathbf{g}(t, \mathbf{x}(t))$$

where $\mathbf{g}$ is a nonlinear function with continuous partial derivatives in a neighborhood $N(\mathbf{x}_e, \varepsilon)$ of $\mathbf{x}_e$. Suppose $\mathbf{x}_e$ is an equilibrium of the dynamical system. Let $\hat{\mathbf{y}}(t) = \mathbf{g}(t, \mathbf{x}_e)$. As in (2.12), linearize $\mathbf{g}$ around $\mathbf{x}_e$ to obtain the approximation:

$$\begin{aligned} \mathbf{y}(t) &\approx \mathbf{g}(t, \mathbf{x}_e) + \mathbf{J}_g(t, \mathbf{x}_e)(\mathbf{x}(t) - \mathbf{x}_e) \\ &= \hat{\mathbf{y}}(t) + \mathbf{J}_g(t, \mathbf{x}_e)(\mathbf{x}(t) - \mathbf{x}_e) \end{aligned}$$

where $\mathbf{J}_g$ is the Jacobian of $\mathbf{g}$ with respect to $\mathbf{x}$. Let $\mathbf{y}_\delta(t) = \mathbf{y}(t) - \hat{\mathbf{y}}(t)$ and $\mathbf{C}(t) = \mathbf{J}_g(t, \mathbf{x}_e)$, so that the linearized output relation can be written

$$\mathbf{y}_\delta(t) = \mathbf{C}(t)\mathbf{x}_\delta(t) \quad (2.18)$$

where $\mathbf{x}_\delta$ is the same as above.

Remark

It is worth noting that if any component f_i of the function $\mathbf{f}$ is linear, then the linear approximation of f_i will be f_i itself. That is, the linearization process need only be applied to nonlinear component functions. This is easy to see in the case of zero input. Suppose $f_i(x_1, \dots, x_n) = a_1x_1 + \dots + a_nx_n$, where the a_i are constants. Then the i^{th} row of the Jacobian of $\mathbf{f}$ with respect to $\mathbf{x}$, $\mathbf{j}^i$, is then

$$\mathbf{j}^i = [a_1 \quad \dots \quad a_n],$$

and the linear approximation of f_i is the i^{th} element of $\mathbf{J}\mathbf{x}$, which is given by

$$\begin{aligned} (\mathbf{J}\mathbf{x})_i &= [a_1 \quad \dots \quad a_n] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= a_1x_1 + \dots + a_nx_n \\ &= f_i(x_1, \dots, x_n) \end{aligned}$$

For the case of nonzero input, where f_i is a linear function of $\mathbf{x}$ and input $\mathbf{u}$, the same argument can be applied to each term in the sum

$$f_i(x_1, \dots, x_n, u_1, \dots, u_m) = (a_1x_1 + \dots + a_nx_n) + (b_1u_1 + \dots + b_mu_m)$$

whence it follows that the i^{th} component in the linear approximation of f_i is

$$[\mathbf{J}_x(t, \mathbf{x}_e, \bar{\mathbf{u}})\mathbf{x} + \mathbf{J}_u(t, \mathbf{x}_e, \bar{\mathbf{u}})\mathbf{u}]_i = f_i(x_1, \dots, x_n, u_1, \dots, u_m)$$

2.1.2 Linearization with Unknown Equilibrium State

We begin this section by showing that any dynamical system can be transformed so that the zero vector is an equilibrium state of the transformed system. This is useful when the system does not have an equilibrium state, or when we do not know what the equilibrium state is. Then we outline a general procedure of linearization that can be used when the equilibrium state of a given dynamical system is unknown or does not exist.

2.5 Proposition. Suppose we are given either (i) a continuous dynamical system

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

as given in Definitions 1.1–1.2, or (ii) a discrete dynamical system

$$\mathbf{x}(t+1) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))$$

as given in Definitions 1.26–1.27. Suppose we proceed thus:

A. Select an arbitrary initial input function $\tilde{\mathbf{u}}(t)$ and solve the differential or the difference equation above to obtain the initial solution $\mathbf{x}_1(t)$, so that

$$\dot{\mathbf{x}}_1(t) = \mathbf{f}(t, \mathbf{x}_1(t), \tilde{\mathbf{u}}(t))$$

in the continuous case, or

$$\mathbf{x}_1(t+1) = \mathbf{f}(t, \mathbf{x}_1(t), \tilde{\mathbf{u}}(t))$$

in the discrete case.

B. Introduce the new variable $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{x}_1(t)$, and define a *new* function:

$$(\text{Newf})(t, \mathbf{z}(t), \tilde{\mathbf{u}}(t)) = \mathbf{f}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(t)) - \mathbf{f}(t, \mathbf{x}_1(t), \tilde{\mathbf{u}}(t))$$

Then $\mathbf{z}(t) \equiv \theta$ is an equilibrium for $\dot{\mathbf{z}}(t) = (\text{Newf})(t, \mathbf{z}(t), \tilde{\mathbf{u}}(t))$ and for

$\mathbf{z}(t+1) = (\text{Newf})(t, \mathbf{z}(t), \tilde{\mathbf{u}}(t))$. That is,

$$(\text{Newf})(t, \theta, \tilde{\mathbf{u}}(t)) = \theta$$

for all $t \geq t_0$ in the continuous case, and for all $t = 0, 1, 2, \dots$ in the discrete case.

Proof. To say $\mathbf{z}(t) \equiv \theta$ is to say $\mathbf{x}(t) - \mathbf{x}_1(t) = \theta$ for all $t \geq t_0$ in the continuous case

and all $t = 0, 1, 2, \dots$ in the discrete case, i.e., $\mathbf{x}$ and $\mathbf{x}_1$ are identical functions. A

simple computation then gives

$$\begin{aligned} (\text{Newf})(t, \theta, \tilde{\mathbf{u}}(t)) &= (\text{Newf})(t, \mathbf{x}(t) - \mathbf{x}_1(t), \tilde{\mathbf{u}}(t)) \\ &= \mathbf{f}(t, \mathbf{x}(t), \tilde{\mathbf{u}}(t)) - \mathbf{f}(t, \mathbf{x}_1(t), \tilde{\mathbf{u}}(t)) \\ &= \mathbf{f}(t, \mathbf{x}_1(t), \tilde{\mathbf{u}}(t)) - \mathbf{f}(t, \mathbf{x}_1(t), \tilde{\mathbf{u}}(t)) \\ &= \theta \end{aligned}$$

■

Since the transformed system introduced in the proof above has an equilibrium state, it can be linearized using the techniques of Section 2.1.1. This suggests a general procedure of linearization in the case that a given dynamical system does not have an equilibrium state, or the equilibrium state is unknown. For convenience we consider only the continuous case here. The discrete case is similar.

- *Step 1:* By Proposition 2.5, $\mathbf{z}_e = \theta$ is an equilibrium state for the new dynamical system

$$\begin{aligned}\dot{\mathbf{z}}(t) &= (\text{Newf})(t, \mathbf{z}(t), \bar{\mathbf{u}}(t)) \\ &= \mathbf{f}(t, \mathbf{x}(t), \bar{\mathbf{u}}(t)) - \mathbf{f}(t, \mathbf{x}_1(t), \bar{\mathbf{u}}(t))\end{aligned}\tag{2.19}$$

- *Step 2:* Linearize the new system in Step 1 about the zero equilibrium state, as in Section 2.1.1. Recall that by equation (2.16) the linearized system has the form:

$$\dot{\mathbf{z}}_\delta(t) \approx \mathbf{A}(t)\mathbf{z}_\delta(t) + \mathbf{B}(t)\mathbf{u}_\delta(t),\tag{2.20}$$

where $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are the Jacobian matrices with respect to $\mathbf{z}$ and $\mathbf{u}$, $\mathbf{z}_\delta(t) = \mathbf{z}(t) - \mathbf{z}_e = \mathbf{z}(t) - \theta = \mathbf{z}(t)$, and $\mathbf{u}_\delta(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t)$. Abuse notation and call the unique solution of (2.20) $\mathbf{z}(t)$.

- *Step 3:* The state $\mathbf{x}(t)$ of the original dynamical system can now be approximated by $\mathbf{x}(t) \approx \mathbf{x}_1(t) + \mathbf{z}(t)$. ([SB], p. 90)

2.1.3 Example: The Satellite Problem

In this subsection we present an example that demonstrates Proposition 2.5, the method of linearization in the case that a dynamical system does not have an equilibrium state. This example comes from Szidarovsky and Bahill, but our presentation is more complete ([SB], pp. 85–93).

2.6 Example In this example we consider the so called *satellite problem*, which concerns a point mass m in the presence of a force field such as gravity. The name for the problem comes from the fact that this problem is concerned with the dynamics of an orbiting satellite where there is no friction. The setup for the problem is illustrated in Figure 2.1.

Figure 2.1 Setup for the satellite problem

As in [SB] we will assume that the force field is given by the quantity $\frac{-k}{r^2}$, where k is a constant. Since we are interested in the mathematics of this dynamical system, we will omit the energy argument and use of Lagrange's equations given in [SB] to derive the equations below. Instead we just assume that the quantities in the figure are related by the equations

$$\ddot{r} = r\dot{\alpha}^2 - \frac{k}{r^2} + u_1 \quad (2.21)$$

$$\ddot{\alpha} = -\frac{2\dot{\alpha}\dot{r}}{r} + \frac{1}{r}u_2 \quad (2.22)$$

where r , α , u_1 , and u_2 are functions of t . Introduce the notation

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} r \\ \dot{r} \\ \alpha \\ \dot{\alpha} \end{bmatrix}$$

and observe that

$$\begin{aligned} \dot{x}_1 &= \dot{r} = x_2 \\ \dot{x}_2 &= x_1(x_4)^2 - \frac{k}{(x_1)^2} + u_1 \text{ (by equation (2.21))} \\ \dot{x}_3 &= \dot{\alpha} = x_4 \\ \dot{x}_4 &= -\frac{2x_4x_2}{x_1} + \frac{1}{x_1}u_2 \text{ (by equation (2.22))} \end{aligned} \quad (2.23)$$

These four equations can now be expressed in vector notation as follows:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad (2.24)$$

where $\mathbf{f}$ is defined by (2.23) and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$. Notice that we have used the

techniques outlined in Section 1.3 to reduce the second-order differential equations to higher-dimensional first-order differential equations. Moreover, the formulation in (2.24) agrees with the general form of a continuous dynamical system, as given by Definitions 1.1–1.2. However, notice that the second and fourth component functions of $\mathbf{f}$ in (2.23) are nonlinear. We will proceed as follows:

- i) Show that system (2.24) has no equilibrium state if input is zero.
- ii) Use the linearization procedure outlined in Section 2.1.2 to approximate the solution, $\mathbf{x}(t)$, of equation (2.24), when input is zero.

i): Assume $u_1(t) = u_2(t) = 0$ for all $t \geq 0$. By definition of an equilibrium state, we seek a fixed state $\mathbf{x}_e$ so that, with zero input, we have

$$\mathbf{f}(t, \mathbf{x}_e, \theta) = \theta,$$

which, by equations (2.23), can be written in component form as

$$\begin{aligned} x_2 &= 0 \\ x_1(x_4)^2 - \frac{k}{(x_1)^2} &= 0 \\ x_4 &= 0 \\ -\frac{2x_4x_2}{x_1} &= 0 \end{aligned}$$

Notice that if $k \neq 0$ (i.e., the force acting on the point mass m is non zero), then, since $x_1 = r > 0$, the second and third equations directly above contradict each other. Hence, this dynamical system does not have an equilibrium state if the input is zero. (However, it should be pointed out that if input is non zero, the dynamical system may have an equilibrium state. For example, the input

$$\bar{\mathbf{u}}(t) = [k \ 0]^T$$

results in the equilibrium states $\mathbf{x}_e = [1 \ 0 \ c \ 0]^T, c \in \mathbb{R}$.)

ii): To use Proposition 2.5, we must first find a solution $\mathbf{x}_1(t)$ to equation (2.24) for the case of zero input: $\bar{\mathbf{u}}(t) \equiv [0 \ 0]^T = \theta_2$. It will simplify notation to let

$k = \omega^2$. We claim the function

$$\mathbf{x}_1(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \omega t \\ \omega \end{bmatrix}$$

is a solution of equation (2.24), that is, $\dot{\mathbf{x}}_1(t) = \mathbf{f}(t, \mathbf{x}_1(t), \theta_2)$. To see this, note that, equivalently, equations (2.23) hold for this function:

$$\begin{aligned}
\dot{x}_1 &= \frac{d1}{dt} = 0 = x_2 \\
\dot{x}_2 &= \frac{d0}{dt} = 0 = x_1(x_4)^2 - \frac{k}{(x_1)^2} = \omega^2 - \omega^2 \\
\dot{x}_3(t) &= \frac{d(\omega t)}{dt} = \omega = x_4(t) \\
\dot{x}_4(t) &= \frac{d\omega}{dt} = 0 = -\frac{2x_4(t)x_2(t)}{x_1(t)} + \frac{1}{x_1(t)}u_2(t),
\end{aligned}$$

The last equation is true because $x_2(t) = u_2(t) = 0$ for all $t \geq 0$.

As in Proposition 2.5B, introduce the new variable $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{x}_1(t)$, that is,

$$\begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \\ z_4(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \\ \omega t \\ w \end{bmatrix} \quad (2.25)$$

Before proceeding further, we comment on what is meant by the equivalent component form given above. The functions $x_i(t)$ are the components of the solution to the *original* system given by the differential equation (2.24). The functions $z_i(t)$ of $\mathbf{z}(t)$ are the components of the *new* dynamical system

$$\begin{aligned}
\dot{\mathbf{z}}(t) &= (\text{Newf})(t, \mathbf{z}(t), \theta_2) \\
&= \mathbf{f}(t, \mathbf{x}(t), \theta_2) - \mathbf{f}(t, \mathbf{x}_1(t), \theta_2)
\end{aligned} \quad (2.26)$$

By Proposition 2.5, system (2.26) has an equilibrium state at zero. We can linearize (2.26) and obtain a linear approximation to each of the functions $z_i(t)$ of $\mathbf{z}(t)$. Then, by Step 3 at the end of Section 2.1.2, we can approximate the original state, $\mathbf{x}(t)$, using the formula $\mathbf{x}(t) = \mathbf{x}_1(t) + \mathbf{z}(t)$.

The first step in the linearization of (2.26) is to write this equation in component form:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} x_2 \\ x_1(x_4)^2 - \frac{\omega^2}{(x_1)^2} + u_1 \\ x_4 \\ -\frac{2x_4x_2}{x_1} + \frac{1}{x_1}u_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \omega \\ 0 \end{bmatrix} \quad (2.27)$$

Notice in the first term on the right of (2.27) that we have included the input function variables, even though they will be taken to be zero in the new system (see equation (2.26)). This is because the linearization procedure requires that we compute partial derivatives of this function with respect to u_1, u_2 , and *then* substitute zero for each value. The second term of (2.27) was computed from the fact that $x_1(t)$ is a solution of equation (2.26), that is, $f(t, x_1(t), \theta_2) = \dot{x}_1(t)$.

By equation (2.25) we can write each $x_i(t)$ as a function of z_i, t :

$$x_1 = z_1 + 1$$

$$x_2 = z_2$$

$$x_3 = z_3 + \omega t$$

$$x_4 = z_4 + \omega$$

Substituting these values into (2.27) results in the equivalent form:

$$\begin{bmatrix} \dot{z}_1 \\ \dot{z}_2 \\ \dot{z}_3 \\ \dot{z}_4 \end{bmatrix} = \begin{bmatrix} z_2 \\ (z_1 + 1)(z_4 + \omega)^2 - \frac{\omega^2}{(z_1 + 1)^2} + u_1 \\ z_4 + \omega \\ -\frac{2(z_4 + \omega)z_2}{z_1 + 1} + \frac{1}{z_1 + 1}u_2 \end{bmatrix} \quad (2.28)$$

Recall that at the end of Section 2.1.1, we observed that the linearization procedure need only be applied to component functions that are nonlinear. Thus, we need only linearize the second and fourth component function on the right of equation (2.28). In what follows, we compute the partial derivatives of the components of Newf to obtain the first equality, and then substitute the zero

vector in for $\mathbf{z}$ (recall that this was the equilibrium state) and the zero input to get the second equality. To simplify notation, we will write F_i instead of $(\text{Newf})_i$ below.

$$\begin{aligned}\frac{\partial F_2}{\partial z_1} &= (z_4 + \omega)^2 + \frac{2\omega^2}{(z_1 + 1)^3} = \omega^2 + 2\omega^2 = 3\omega^2 \\ \frac{\partial F_2}{\partial z_2} &= \frac{\partial f_2}{\partial z_3} = 0 \\ \frac{\partial F_2}{\partial z_4} &= 2(z_1 + 1)(z_4 + \omega) = 2\omega \\ \frac{\partial F_2}{\partial u_1} &= 1 \\ \frac{\partial F_2}{\partial u_2} &= 0\end{aligned}$$

and for $F_4 = (\text{Newf})_4$:

$$\begin{aligned}\frac{\partial F_4}{\partial z_1} &= \frac{2(z_4 + \omega)z_2}{(z_1 + 1)^2} - \frac{1}{(z_1 + 1)^2} u_2 = 0 \\ \frac{\partial F_4}{\partial z_2} &= \frac{-2(z_4 + \omega)}{z_1 + 1} = -2\omega \\ \frac{\partial F_4}{\partial z_3} &= 0 \\ \frac{\partial F_4}{\partial z_4} &= -\frac{2z_2}{z_1 + 1} = 0 \\ \frac{\partial F_4}{\partial u_1} &= 0 \\ \frac{\partial F_4}{\partial u_2} &= \frac{1}{z_1 + 1} = 1\end{aligned}$$

In the notation below, we will use the notation θ_4 to represent the (zero) equilibrium state of Newf under zero input $\tilde{\mathbf{u}}(t) \equiv \theta_2$. Hence, equation (2.16) takes the particular form

$$\begin{aligned}
\dot{\mathbf{z}}(t) &= \mathbf{J}_{x_\delta}(t, \theta_4, \theta_2) \mathbf{z}(t) + \mathbf{J}_u(t, \theta_4, \theta_2) \mathbf{u}_\delta(t) \\
&= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega & 0 & 0 \end{bmatrix} \begin{bmatrix} z_1(t) \\ z_2(t) \\ z_3(t) \\ z_4(t) \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} \\
&= \mathbf{A}\mathbf{z}(t) + \mathbf{B}\mathbf{u}(t)
\end{aligned} \tag{2.29}$$

where we have substituted $\mathbf{z}(t)$ for $\mathbf{z}_\delta(t)$ in (2.16), since

$$\mathbf{z}_\delta(t) = \mathbf{z}(t) - \mathbf{z}_e = \mathbf{z}(t) - \boldsymbol{\theta} = \mathbf{z}(t)$$

In equation (2.29) we have used the fact that since input is zero,

$$\mathbf{u}_\delta(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t) = \mathbf{u}(t) - \boldsymbol{\theta}_2 = \mathbf{u}(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}$$

Note also that the linearized system (2.29) is time invariant. By Corollary 1.18, the solution of system (2.29) is given by

$$\mathbf{z}(t) = \Phi(t, t_0) \mathbf{z}_0 + \int_{t_0}^t \Phi(t, \tau) \mathbf{B} \mathbf{u}(\tau) d\tau, \tag{2.30}$$

where $\mathbf{z}_0 = \mathbf{z}(t_0) = \mathbf{x}(t_0) - \mathbf{x}_1(t_0)$. The state of the satellite problem, $\mathbf{x}(t)$, can now be approximated using $\mathbf{x}(t) \approx \mathbf{x}_1(t) + \mathbf{z}(t)$.

Finally we note that by Theorem 1.21, because $\mathbf{A}$ is constant, the fundamental matrix of system (2.30) can be written $\Phi(t, t_0) = e^{(t-t_0)\mathbf{A}}$. Using the Jordan form of $\mathbf{A}$, this matrix exponential was computed to be

$$\Phi(t, t_0) = \begin{bmatrix} \frac{4\omega^2 - 3 \cos \omega(t-t_0)}{4\omega^2 - 3} & \frac{1}{\omega} \sin \omega(t-t_0) & 0 & \frac{2\omega(1 - \cos \omega(t-t_0))}{4\omega^2 - 3} \\ \frac{3\omega \sin \omega(t-t_0)}{4\omega^2 - 3} & \cos \omega(t-t_0) & 0 & \frac{2\omega^2 \sin \omega(t-t_0)}{4\omega^2 - 3} \\ \frac{6(\sin \omega(t-t_0) - \omega(t-t_0))}{4\omega^2 - 3} & \frac{2(\cos \omega(t-t_0) - 1)}{\omega} & 1 & \frac{-3(t-t_0) + 4\omega \sin \omega(t-t_0)}{4\omega^2 - 3} \\ \frac{-6\omega(1 - \cos \omega(t-t_0))}{4\omega^2 - 3} & -2 \sin \omega(t-t_0) & 0 & 1 + \frac{4\omega^2(\cos \omega(t-t_0) - 1)}{4\omega^2 - 3} \end{bmatrix}$$

It is worth noting that the matrix given in Szidarovsky and Bahill is incorrect if $\omega \neq 1$ (see [SB], p. 92). ■

2.1.4 Asymptotic Stability Via Linearization

In this last subsection of Section 2.1, we consider one more use of the linearization procedure described in this section. In Section 2.2 a formal definition of asymptotic stability will be given, but for now, an equilibrium state $\mathbf{x}_e$ is *asymptotically stable* if whenever we choose $\mathbf{x}(t_0)$ close enough to $\mathbf{x}_e$, and $\mathbf{x}(t)$ is the corresponding solution, then $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$.

Linearization can be used in investigating asymptotic stability as follows. Consider the state transition equation of a continuous (resp. discrete) nonlinear dynamical system with zero input, $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$ (resp. $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$). Let $\mathbf{J}(\mathbf{x})$ denote the Jacobian matrix of $\mathbf{f}$ as given by Definition 2.3, and let $\mathbf{x}_e$ be an equilibrium state for this system. By Proposition 2.5 such an equilibrium state may be assumed to exist.

We showed in Section 2.1.1 that the nonlinear dynamical system can be linearized around the equilibrium state to obtain the time invariant continuous (resp. discrete) linear system:

$$\dot{\mathbf{x}}_\delta(t) = \mathbf{J}(\mathbf{x}_e)\mathbf{x}_\delta(t) \text{ (resp. } \mathbf{x}_\delta(t+1) = \mathbf{J}(\mathbf{x}_e)\mathbf{x}_\delta(t)),$$

where $\mathbf{x}_\delta(t) = \mathbf{x}(t) - \mathbf{x}_e$, and $\mathbf{x}(t)$ is the state of the original nonlinear system. Note that the zero vector is an equilibrium state for the linear system in both cases.

The asymptotic stability of the zero vector in the linearized system implies the asymptotic stability of the equilibrium state $\mathbf{x}_e$ in the original nonlinear system (see discussion in [SB], p. 172). Therefore, the question of whether the equilibrium state in a nonlinear system is asymptotically stable can be reduced to the question of whether the zero vector in a time invariant linear dynamical system is asymptotically stable.

In the next section, Liapunov's "indirect method" or "second" method is presented that gives conditions that imply asymptotic stability without linearizing the nonlinear dynamical system. This method has distinct advantages over linearization, as we shall see.

2.2 TYPES OF STABILITY AND LIAPUNOV'S SECOND METHOD

In the previous section, we showed that it is possible to linearize a nonlinear dynamical system near an equilibrium state, but we have not yet discussed system behavior as time increases. We know that if the initial state, $\mathbf{x}_0$, is chosen as an equilibrium, the state will remain at the equilibrium state, $\mathbf{x}_e$, for all future time. However, if the initial state is chosen close to an equilibrium state, so $\|\mathbf{x}_0 - \mathbf{x}_e\|$ is small, the state might remain close to the equilibrium state for all future time, possibly even converging to $\mathbf{x}_e$, or the state may move off so that $\|\mathbf{x}(t) - \mathbf{x}_e\| \rightarrow \infty$ as $t \rightarrow \infty$.

In this section we first give a number of definitions to describe this behavior. Then we discuss Liapunov's "Second Method," which is a way to analyze nonlinear dynamical systems that can be simple to use and does not require linearization nor knowledge of the actual solution to the dynamical system. This alternative to linearization is attractive because it allows the analysis to extend beyond a small neighborhood of the equilibrium—whereas linearization is valid only in a small neighborhood of the equilibrium ([Lue], p. 332).

In practice, Liapunov's Second Method is usually applied in the analysis of nonlinear dynamical systems ([SB], p. 176). It can also be applied to linear dynamical systems, but we will see in Section 2.3 that the stability of time invariant linear dynamical systems can be analyzed by considering the eigenvalues of the fundamental matrix for the system and these can often be computed. Time variant linear dynamical systems will be considered in Section 2.4 when we consider BIBO stability.

We will discuss Liapunov's second method only for continuous dynamical systems given by

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t)) \quad (2.31)$$

and similarly for discrete dynamical systems given by

$$\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t)), \quad (2.32)$$

where $\mathbf{f}: X \rightarrow \mathfrak{R}^n$ is a continuous function and $X \subseteq \mathfrak{R}^n$ is the state space.

2.2.1 Types of Stability

In this first subsection, we first define four types of stability, and then discuss each type intuitively and from a geometric perspective. These four types are all types of *internal* stability—that is, they describe the behavior of the state $\mathbf{x}(t)$ as $t \rightarrow \infty$. *External* stability is the phrase used to refer to the input-output behavior only. In Section 2.4 we discuss one type of external stability, BIBO (Bounded Input-Bounded Output). The stability types defined here will be sufficient for our work in this chapter, but are not exhaustive. Kalman and Bertram [KB] define eight types of stability, and numerous other variations of the basic types presented here appear in the literature. Our definitions below are a synthesis of those given by Szidarovsky and Bahill, and by Glendinning ([SB], pp. 167–168 and [Gle], pp. 27–29).

2.7 Definition (Stability types) Suppose $\mathbf{x}_e$ is an equilibrium state for a continuous dynamical system (resp. discrete dynamical system) given by equation (2.31) (resp. (2.32)). If $\mathbf{x}_0 \in X$, let $\mathbf{x}(t)$ denote the solution such that $\mathbf{x}(t_0) = \mathbf{x}_0$ (resp. $\mathbf{x}(0) = \mathbf{x}_0$). Then $\mathbf{x}_e$ is

A. stable (“start near stay near”) if for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\|\mathbf{x}_e - \mathbf{x}_0\| < \delta$, then $\|\mathbf{x}_e - \mathbf{x}(t)\| < \varepsilon$ for all $t \geq t_0$ (resp. $t \geq 0$);

B. quasi-asymptotically stable (“tends to eventually”) if there exists $\delta > 0$ such that if $\|\mathbf{x}_e - \mathbf{x}_0\| < \delta$, then $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$;

C. asymptotically stable (“tends to directly”) if $\mathbf{x}_e$ is both stable and quasi-asymptotically stable;

D. globally asymptotically stable if $\mathbf{x}_e$ is stable and for any $\mathbf{x}_0 \in X$, $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$;

E. unstable if it is not stable. ■

We will often refer to the trajectory of the state, in discussing stability.

2.8 Definition (State trajectory) Let $\mathbf{x}(t)$ be the state of a continuous (resp. discrete) dynamical system with given initial time t_0 (resp. 0). Then the **state trajectory** is the set $\{\mathbf{x}(t): t \geq t_0\}$ (resp. $\{\mathbf{x}(t): t \geq 0\}$). ■

The following figure illustrates three of the definitions. In each case of

Figure 2.2 Forms of stability

stability above, the ball (i.e., the state) is at rest, i.e., the system is in equilibrium. Now consider a slight initial displacement of the initial state away from the equilibrium state. In the first sketch the equilibrium is unstable, for clearly the ball will roll down either side of the half circle, getting farther away from the equilibrium as time increases. In the second sketch the equilibrium is asymptotically stable, for a slight initial displacement of the initial state from the equilibrium state will eventually be “overcome” as the state returns to equilibrium at the bottom of the bowl. In the case of global asymptotic stability, the amount of initial displacement is arbitrary. The third sketch shows a stable equilibrium, for a slight displacement of the initial state away from the equilibrium does not return the state to equilibrium as time increases, but neither does the state continue to get farther away. ([PA], p. 574)

We will give another example of a stable but not asymptotically stable equilibrium in Example 2.17 (see Section 2.2.4). There we will see that the entire trajectory oscillates around a circle of radius k centered at the equilibrium state, as $t \rightarrow \infty$.

Another sketch that helps to visualize the terms geometrically is given in Figure 2.3. Definition 2.7A says that the equilibrium is stable (i.e., “start near stay near”) if with x_0 near x_e the state $x(t)$ remains near x_e , but may not converge to x_e . In the case of quasi-asymptotic stability (i.e., “tends to eventually”) the state $x(t)$ “eventually” converges to x_e as $t \rightarrow \infty$, but may oscillate wildly before converging. If x_e is asymptotically stable (i.e., “tends to directly”), then with x_0 near x_e the state $x(t)$ both remains near x_e and tends to x_e , as $t \rightarrow \infty$. If the equilibrium is globally asymptotic, then every state $x(t)$ converges to x_e , no matter what x_0 is. ([Gle], pp. 27–29)

Figure 2.3 Geometric visualization of Definition 2.7

2.2.2 Definition of a Liapunov Function

A Liapunov function can be thought of as a generalization of the energy function in mechanical systems ([SB], p. 169). Since the energy of a mechanical system in the presence of friction is constantly decreasing until it reaches its lowest point, there must be an equilibrium, and the system must converge to that as $t \rightarrow \infty$. As we shall see, the existence of a Liapunov function for a given dynamical system with equilibrium state $\mathbf{x}_e$ will imply that $\mathbf{x}_e$ is stable. This method works directly with the nonlinear dynamical system—rather than using a linear approximation to the system as derived in Section 2.1.

2.9 Definition (Liapunov function for $\mathbf{x}_e$) Fix the state transition function of a continuous (resp. discrete) dynamical system of the form of equation (2.31) (resp. (2.32)) with equilibrium state $\mathbf{x}_e$. A Liapunov function for $\mathbf{x}_e$ is a pair (V, S) , where $S \subseteq X$ (state space), $\mathbf{x}_e \in S$, and $V: S \rightarrow \mathfrak{R}$ satisfies the following:

- A. V is continuous;
- B. V has a unique global minimum at $\mathbf{x}_e$ with respect to all other elements of S ;
- C. for any state trajectory $\mathbf{x}(t)$ contained in S , $V(\mathbf{x}(t))$ is nonincreasing in t . ■

Usually S is taken to be an open sphere containing the state trajectory, or the closure of such an open sphere.

Conditions A–C of Definition 2.9 have nice geometrical interpretations when $X = \mathfrak{R}$ or $\mathfrak{R}^2$. Condition 2.9A says that the graph of V has no breaks, and 2.9B says that the graph of V has a unique minimum point $\mathbf{x}_e$ over its entire domain S .

Observe that 2.9A–B are system independent; it is Condition 2.9C that depends on the particular dynamical system given. Let $k_0 < k_1 < k_2 < \dots$ be constants and define $C_i = \{\mathbf{x} \in S : V(\mathbf{x}) = k_i\}$. Suppose $\mathbf{x}(t)$ is the state of a continuous (resp. discrete) dynamical system and suppose $C_0 = \{\mathbf{x}_e\}$. Then

Condition 2.9C says that values $\mathbf{x}(t)$, for $t \geq t_0$ (resp. $t \geq 0$), pass from contour C_i to C_j with $i > j$. In two dimensions, this means that starting with $\mathbf{x}_0$, subsequent values $\mathbf{x}(t)$ for $t \geq t_0$ (resp. $t \geq 0$) must cross contours in the direction toward the center $C_0 = \{\mathbf{x}_e\}$, as seen in Figure 2.4, and never outward. ([Lue], p. 333)

Figure 2.4 Contours C_i in two dimensions

In practice, it can be hard to find a Liapunov function for a dynamical system. There are clearly many functions that satisfy 2.9A–B, but one must usually work hard at finding one that also satisfies 2.9C. In cases where the entire state trajectory lies at a finite distance from the equilibrium state, the reader is referred to [Rod] for computer methods for testing whether a given function qualifies as a Liapunov function.

2.2.3 Liapunov Theorems for Discrete Dynamical Systems

Recall that by the assumptions made at the beginning of Section 2.2, our discrete dynamical systems are given by $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$, where $\mathbf{f}: X \rightarrow \mathbb{R}^n$, $\mathbf{f}$ is continuous, and $t = 0, 1, 2, \dots$.

Given a Liapunov function (V, S) for $\mathbf{x}_e$, define ΔV as follows:

$$\begin{aligned}\Delta V(\mathbf{x}(t)) &= V(\mathbf{x}(t+1)) - V(\mathbf{x}(t)) \\ &= V(\mathbf{f}(\mathbf{x}(t))) - V(\mathbf{x}(t))\end{aligned}\tag{2.33}$$

We will often suppress t and just write $\mathbf{x}$ and $\Delta V(\mathbf{x}) = V(\mathbf{f}(\mathbf{x})) - V(\mathbf{x})$. Then

Condition 2.9C that $V(\mathbf{x}(t))$ is nonincreasing in t , can be abbreviated as follows:

$$\Delta V(\mathbf{x}) \equiv V(\mathbf{f}(\mathbf{x})) - V(\mathbf{x}) \leq 0 \text{ for all } \mathbf{x} \in S \tag{2.34}$$

In situations where Liapunov's indirect method is applied, we often are unable to compute the state, especially when the dynamical system under consideration is nonlinear. However sometimes we can show that every possible solution $\mathbf{x}(t)$ of the dynamical system lies in a region S in the state space, and that (2.34) holds in S , and then we can conclude that $V(\mathbf{x}(t))$ is nonincreasing in t —without ever solving the dynamical system.

The following is a restatement of Definition 2.9 for the case of a discrete dynamical system.

2.10 Definition Let $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$ be a discrete dynamical system with equilibrium state $\mathbf{x}_e$. A **Liapunov function for $\mathbf{x}_e$** is a pair (V, S) ($S \subseteq X$ and $V: S \rightarrow \mathbb{R}$) satisfying the following conditions:

- A. V is continuous;
- B. V has a unique global minimum at $\mathbf{x}_e$ with respect to all other elements of S ;
- C. for all $\mathbf{x} \in S$, $\Delta V(\mathbf{x}) \equiv V(\mathbf{f}(\mathbf{x})) - V(\mathbf{x}) \leq 0$. ([Lue], p. 335) ■

Note that if the entire state trajectory of $\mathbf{x}(t)$ lies in S , then 2.10C implies 2.9C, that is, $V(\mathbf{x}(t))$ is nonincreasing in t .

Our first major stability theorem, and subsequent proof, follows [Lue] and [SB].

2.11 Theorem (Liapunov Theorem: Discrete Case) Let $\mathbf{x}(t+1) = \mathbf{f}(\mathbf{x}(t))$ be a discrete dynamical system, where $\mathbf{f}$ is continuous, and $\mathbf{x}_e$ is an equilibrium state for the system. Suppose there exists a Liapunov function for $\mathbf{x}_e$, $(V, S(\mathbf{x}_e, \varepsilon_0))$, where

$$S(\mathbf{x}_e, \varepsilon_0) = \{\mathbf{x} \in \mathcal{R}^n : \|\mathbf{x} - \mathbf{x}_e\| \leq \varepsilon_0\} \quad (2.35)$$

is a spherical region in the state space X , centered at $\mathbf{x}_e$, and $\varepsilon_0 > 0$ is given.

Then

A. $\mathbf{x}_e$ is stable; and

B. if, in addition, $t_2 > t_1$ implies $V(\mathbf{x}(t_2)) < V(\mathbf{x}(t_1))$, unless $\mathbf{x}(t_1) = \mathbf{x}_e$, for all solutions $\mathbf{x}(t)$ with trajectories in $S(\mathbf{x}_e, \varepsilon_0)$, then $\mathbf{x}_e$ is asymptotically stable.

Proof

Figure 2.5 Regions in the proof of Theorem 2.11A

A: We must show that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathbf{x}(t)$ is a solution with $\|\mathbf{x}_e - \mathbf{x}(0)\| < \delta$, then $\|\mathbf{x}_e - \mathbf{x}(t)\| < \varepsilon$, for all $t \geq 0$. Let $0 < \varepsilon < \varepsilon_0$, and choose ε_1 such that $0 < \varepsilon_1 \leq \varepsilon$ and

$$\|\mathbf{x} - \mathbf{x}_e\| < \varepsilon_1 \text{ implies } \|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_e)\| = \|\mathbf{f}(\mathbf{x}) - \mathbf{x}_e\| < \varepsilon,$$

which is possible since $\mathbf{f}$ is continuous.

Let $m = \min\{V(\mathbf{x}) : \varepsilon_1 \leq \|\mathbf{x} - \mathbf{x}_e\| \leq \varepsilon_0\}$, which exists since V is continuous and the closed annulus is compact. Note further that $\mathbf{x}_e$ is not in the annulus, and since V has its unique minimum at $\mathbf{x}_e$, we must have $V(\mathbf{x}_e) < m$. The continuity of function V implies that there exists $\delta > 0$ such that if $\|\mathbf{x} - \mathbf{x}_e\| < \delta$, then $V(\mathbf{x}) < m$.

Now let $\mathbf{x}(t)$ be any solution for which $\|\mathbf{x}(0) - \mathbf{x}_e\| < \delta$; then $V(\mathbf{x}(0)) < m$. We claim that for all $t \geq 0$, $\|\mathbf{x}(t) - \mathbf{x}_e\| < \varepsilon_1$. Since $\varepsilon_1 < \varepsilon$, proving this will complete the proof. For a contradiction let $k+1$ be least such that $\|\mathbf{x}(k+1) - \mathbf{x}_e\| \geq \varepsilon_1$. Then $\|\mathbf{x}(k) - \mathbf{x}_e\| < \varepsilon_1$; however, $\mathbf{x}(k+1) = \mathbf{f}(\mathbf{x}(k))$, and we chose ε_1 above so that $\|\mathbf{x}(k) - \mathbf{x}_e\| < \varepsilon_1$ implies $\|\mathbf{x}(k+1) - \mathbf{x}_e\| = \|\mathbf{f}(\mathbf{x}(k)) - \mathbf{x}_e\| < \varepsilon$. Hence $\mathbf{x}(k+1)$ is in the annulus, and by definition of m , $V(\mathbf{x}(k+1)) \geq m$, contradicting the fact that $V(\mathbf{x}(k+1)) \leq V(\mathbf{x}(0)) < m$.

B: Let $\mathbf{x}(t)$ be a solution such that $\|\mathbf{x}(0) - \mathbf{x}_e\| < \delta$, where δ is defined as in the proof of part A above. We claim that $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$. By the last paragraph of the proof above, we have that $\|\mathbf{x}(t) - \mathbf{x}_e\| < \varepsilon_1$ for all $t \geq 0$. Hence, $\{\mathbf{x}(t) : t \geq 0\}$ is bounded. Let $\mathbf{x}(t_k)$ be a convergent subsequence such that $\mathbf{x}(t_k) \rightarrow \mathbf{x}^*$. We need to prove $\mathbf{x}^* = \mathbf{x}_e$.

The sequence $\mathbf{x}(t_k + 1)$ ($k = 0, 1, 2, \dots$) is also bounded, whence it must have a convergent subsequence $\mathbf{x}(t_{k_i} + 1) \rightarrow \mathbf{x}^{**}$. Since $t_{k_i} < t_{k_i} + 1 \leq t_{k_{i+1}}$, it follows by hypothesis that $V(\mathbf{x}(t_{k_{i+1}})) \leq V(\mathbf{x}(t_{k_i} + 1)) \leq V(\mathbf{x}(t_{k_i}))$. Let $i \rightarrow \infty$, and note that by

the continuity of the function V , we have $V(\mathbf{x}^*) \leq V(\mathbf{x}^{**}) \leq V(\mathbf{x}^*)$, which implies that

$$V(\mathbf{x}^*) = V(\mathbf{x}^{**}) \quad (2.36)$$

Since f is continuous, we have that

$$\begin{aligned} \mathbf{x}^{**} &= \lim_{i \rightarrow \infty} \mathbf{x}(t_{k_i} + 1) \\ &= \lim_{i \rightarrow \infty} f(\mathbf{x}(t_{k_i})) \\ &= f\left(\lim_{i \rightarrow \infty} \mathbf{x}(t_{k_i})\right) \\ &= f(\mathbf{x}^*) \end{aligned}$$

Finally, let $\mathbf{x}_1(t)$ be the unique solution to the discrete dynamical system $\mathbf{x}(t+1) = f(\mathbf{x}(t))$, $\mathbf{x}(0) = \mathbf{x}^*$. Then $\mathbf{x}_1(1) = f(\mathbf{x}_1(0)) = f(\mathbf{x}^*) = \mathbf{x}^{**}$. If $\mathbf{x}^* \neq \mathbf{x}_e$ then $V(\mathbf{x}^{**}) < V(\mathbf{x}^*)$ since V is strictly decreasing in t . But this contradicts equation (2.36), and completes the proof. ■

2.12 Corollary Suppose the function V of Theorem 2.11 is defined on the entire state space X , $V(\mathbf{x}(t))$ is strictly decreasing in t unless $\mathbf{x}(t) = \mathbf{x}_e$, and $V(\mathbf{x}) \rightarrow \infty$ as any component of $\mathbf{x} \in X$ gets arbitrarily large in magnitude. Then $\mathbf{x}_e$ is a globally asymptotic equilibrium state. ([SB], p. 170)

Proof Let $\mathbf{x}_0 \in X$ be arbitrary, and let $\mathbf{x}(t)$ be the solution with this initial state.

Our assumption about the monotonicity of V implies that

$$V(\mathbf{x}(t)) \leq V(\mathbf{x}_0) \text{ for all } t \geq 0 \quad (2.37)$$

Now $V(\mathbf{x}) \rightarrow \infty$ as any component of $\mathbf{x} \in X$ gets arbitrarily large, by hypothesis.

This fact and (2.37) imply that the set $\{\mathbf{x}(t): t \geq 0\}$ is bounded. Now employ the same reasoning that was used in part B of the proof above to show that $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$. Since $\mathbf{x}_0$ is arbitrary, we have that $\mathbf{x}_e$ is globally asymptotic. ■

We close this section with an example.

2.13 Example Consider the discrete dynamical system,

$$\mathbf{x}(t+1) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} \frac{x_2(t)}{1+x_2(t)^2} \\ \frac{x_1(t)}{1+x_2(t)^2} \end{bmatrix} = \mathbf{f}(\mathbf{x}(t)) \quad (2.38)$$

Notice that $\mathbf{x}_e = \theta = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is an equilibrium state for this dynamical system. Define

function V on the entire state space, $\mathcal{R}^2$, by

$$V\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}\right) = \alpha^2 + \beta^2, \quad (2.39)$$

and note that V has a unique minimum at the equilibrium state $\mathbf{x}_e = \theta$. To see that

$(V, \mathcal{R}^2)$ is a Liapunov function for $\mathbf{x}_e$, it suffices to see that for any solution to

(2.38), $V(\mathbf{x}(t+1)) \leq V(\mathbf{x}(t))$ for all $t \geq 0$. Note that by (2.38) and (2.39)

$$\begin{aligned} V(\mathbf{x}(t+1)) &= \frac{x_2(t)^2 + x_1(t)^2}{[1+x_2(t)^2]^2} \\ &= \frac{V(\mathbf{x}(t))}{[1+x_2(t)^2]^2} \\ &\leq V(\mathbf{x}(t)) \end{aligned}$$

Thus $(V, \mathcal{R}^2)$ is a Liapunov function for $\mathbf{x}_e$, so by Theorem 2.11, $\mathbf{x}_e$ is a stable

equilibrium state for this discrete dynamical system. Since V is not strictly

decreasing in t , we cannot infer asymptotic stability. ([Lue], pp. 333–334) ■

2.2.4 Liapunov Theorems for Continuous Dynamical Systems

Recall that by the assumptions made at the beginning of Section 2.2, our continuous dynamical systems are given by $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$, where $\mathbf{f}: X \rightarrow \mathbb{R}^n$, $\mathbf{f}$ is continuous, $X \subseteq \mathbb{R}^n$, and $t \in [t_0, \infty)$. Let $\mathbf{x}_e$ be an equilibrium state for the system.

Suppose $\mathbf{x}(t)$ is a solution of the state transition equation above. Let $S \subseteq X$ be a spherical region centered at $\mathbf{x}_e$ that contains the entire state trajectory, and suppose there is a Liapunov function (V, S) for $\mathbf{x}_e$. So V is continuous in S , and if, in addition, V is differentiable in S , then the condition that $V(\mathbf{x}(t))$ is nonincreasing in t can be written

$$\dot{V}(\mathbf{x}(t)) \leq 0 \text{ for all } t \geq t_0$$

For the case $n = 1$, so that $x(t)$ is a one-dimensional vector function of t , the derivative above can be expressed using the chain rule:

$$\dot{V}(x(t)) = \frac{dV}{dx} \dot{x}(t)$$

In the case that $n > 1$, where $\mathbf{x}(t) = [x_1(t) \cdots x_n(t)]^T$ and $\mathbf{f}(\mathbf{x}(t)) = [f_1(\mathbf{x}(t)), \dots, f_n(\mathbf{x}(t))]^T$, the chain rule for differentiation gives

$$\dot{V}(\mathbf{x}(t)) = \frac{\partial V}{\partial x_1} \dot{x}_1(t) + \frac{\partial V}{\partial x_2} \dot{x}_2(t) + \cdots + \frac{\partial V}{\partial x_n} \dot{x}_n(t) \quad (2.40)$$

[BL]. Since $\dot{x}_i(t) = f_i(\mathbf{x}(t))$, equation (2.40) can be rewritten

$$\dot{V}(\mathbf{x}(t)) = \frac{\partial V}{\partial x_1} f_1(\mathbf{x}(t)) + \frac{\partial V}{\partial x_2} f_2(\mathbf{x}(t)) + \cdots + \frac{\partial V}{\partial x_n} f_n(\mathbf{x}(t)) \quad (2.41)$$

Define the *gradient row vector* $\nabla V(\mathbf{x}) = \left[\frac{\partial V(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial V(\mathbf{x})}{\partial x_n} \right]$, and note that (2.40) can

be written as follows:

$$\dot{V}(\mathbf{x}(t)) = \nabla V(\mathbf{x}(t)) \cdot \mathbf{f}(\mathbf{x}(t)) \quad (2.42)$$

We will often suppress t and write $\mathbf{x}$ and $\dot{V}(\mathbf{x}) = \nabla V(\mathbf{x})\mathbf{f}(\mathbf{x})$. Then Condition 2.9C, that $V(\mathbf{x}(t))$ is nonincreasing in t , can be abbreviated as follows:

$$\dot{V}(\mathbf{x}) = \nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) \leq 0 \text{ for all } \mathbf{x} \in S,$$

and we use this abbreviation below. ([Lue], pp. 336–337)

2.14 Definition A **Liapunov function** for $\mathbf{x}_e$, where $\mathbf{x}_e$ is an equilibrium state of a continuous dynamical system as described above, is a pair (V, S) ($S \subseteq X$ and $V: S \rightarrow \mathfrak{R}$) satisfying the following conditions:

- A. V is continuous and differentiable over S (often V is required to have continuous first partial derivatives, as in [Lue]);
- B. V has a unique global minimum at $\mathbf{x}_e$ with respect to all other elements of S ;
- C. for all $\mathbf{x} \in S$, $\dot{V}(\mathbf{x}) \equiv \nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) \leq 0$. ([Lue], p. 337) ■

Note that if an entire state trajectory of $\mathbf{x}(t)$ is in S , then 2.14C implies 2.9C, that is, $V(\mathbf{x}(t))$ is nonincreasing in t .

Our first Liapunov theorem parallels Theorem 2.11, but the proof follows that of Sanchez ([San], pp. 105–108).

2.15 Theorem (Liapunov Theorem: Continuous Case) Let $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$ be a continuous dynamical system, where $\mathbf{f}$ is continuous, and $\mathbf{x}_e$ is an equilibrium state for the system. Suppose there exists a Liapunov function for $\mathbf{x}_e$,

$(V, S(\mathbf{x}_e, \epsilon_0))$, where

$$S(\mathbf{x}_e, \epsilon_0) = \left\{ \mathbf{x} \in \mathfrak{R}^n : \|\mathbf{x} - \mathbf{x}_e\| \leq \epsilon_0 \right\} \quad (2.43)$$

is a spherical region in the state space X , centered at $\mathbf{x}_e$, and $\epsilon_0 > 0$ is given.

Then

- A. $\mathbf{x}_e$ is stable; and
- B. if it is also true that for every $\epsilon > 0$ there exists $m_\epsilon < 0$ such that

$$\text{if } 0 < \|\mathbf{x} - \mathbf{x}_e\| \leq \epsilon, \text{ then } \dot{V}(\mathbf{x}) \equiv \nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) \leq m_\epsilon < 0, \quad (2.44)$$

then $\mathbf{x}_e$ is asymptotically stable. ([San], pp. 106–107)

Notice that if V has continuous first partial derivatives on a neighborhood of $\mathbf{x}_e$, and $\dot{V}(\mathbf{x}) < 0$ over the deleted neighborhood, then Condition 2.15B is satisfied.

Proof

Figure 2.6 Regions in the proof of Theorem 2.15A

A: We must show that for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $\mathbf{x}(t)$ is a solution with $\|\mathbf{x}_e - \mathbf{x}(t_0)\| < \delta$, then $\|\mathbf{x}_e - \mathbf{x}(t)\| < \varepsilon$, for all $t \geq t_0$. Let $0 < \varepsilon < \varepsilon_0$ and define $m = \min\{V(\mathbf{x}) : \varepsilon \leq \|\mathbf{x}_e - \mathbf{x}\| \leq \varepsilon_0\}$, which exists since V is continuous and the closed annulus is compact. Note that $\mathbf{x}_e$ is not in the annulus, and since V has a unique minimum at $\mathbf{x}_e$, $V(\mathbf{x}_e) < m$. The continuity of function V implies that there exists $\delta > 0$ such that if $\|\mathbf{x} - \mathbf{x}_e\| < \delta$ then $V(\mathbf{x}) < m$.

Let $\mathbf{x}(t)$ be any solution for which $\|\mathbf{x}(t_0) - \mathbf{x}_e\| < \delta$; then $V(\mathbf{x}(t_0)) < m$. Since $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$ for all $t \geq t_0$, and $\mathbf{f}$ is continuous, it follows that $\mathbf{x}(t)$ is continuous on $[t_0, \infty)$, and since $[t_0, \infty)$ is connected, $\{\mathbf{x}(t) : t \geq t_0\}$ is connected. We claim that $\|\mathbf{x}(t) - \mathbf{x}_e\| < \varepsilon$ for all $t \geq t_0$. For a contradiction, suppose there exists a value $\mathbf{x}(t)$ such that $\|\mathbf{x}(t) - \mathbf{x}_e\| \geq \varepsilon$. Then since $\{\mathbf{x}(t) : t \geq t_0\}$ is connected, there exists an $\mathbf{x}(t_1)$ on the circle $\|\mathbf{x} - \mathbf{x}_e\| = \varepsilon$; but then $V(\mathbf{x}(t_1)) \geq m$, contradicting the fact that since V is nonincreasing in t , $V(\mathbf{x}(t_1)) \leq V(\mathbf{x}(t_0)) < m$. Therefore $\|\mathbf{x}(t) - \mathbf{x}_e\| < \varepsilon$ for all $t \geq t_0$, whence $\mathbf{x}_e$ is stable.

B: Suppose the hypothesis of part B is true. Let $\varepsilon > 0$, and let $\mathbf{x}(t)$ be any solution such that $\|\mathbf{x}(t_0) - \mathbf{x}_e\| < \delta$, where δ is defined as in the proof of part A. We proved that $\|\mathbf{x}(t) - \mathbf{x}_e\| < \varepsilon$ for all $t \geq t_0$. We wish to prove that $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$. Suppose first that there exists $\gamma > 0$ and $t_1 \geq t_0$ such that $\|\mathbf{x}(t) - \mathbf{x}_e\| \geq \gamma$ for all $t \geq t_1$. By (2.44) we can find m_γ such that $\dot{V}(\mathbf{x}(t)) \leq m_\gamma < 0$. Hence

$$\begin{aligned} V(\mathbf{x}(t)) &= V(\mathbf{x}(t_0)) + \int_{t_0}^t \frac{d}{ds} V(\mathbf{x}(s)) ds \\ &\leq V(\mathbf{x}(t_0)) + \int_{t_0}^t m_\gamma ds \\ &\leq V(\mathbf{x}(t_0)) + m_\gamma(t - t_0) \end{aligned}$$

But the right side of this last inequality approaches $-\infty$ as $t \rightarrow \infty$, whereas $V(\mathbf{x}(t)) \geq V(\mathbf{x}_e)$ by definition of V . This is a contradiction. Therefore any neighborhood of $\mathbf{x}_e$ contains some $\mathbf{x}(t)$.

Finally, to see $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$, begin by noting that since $V(\mathbf{x}(t)) \leq V(\mathbf{x}(t_\gamma))$ for all $t \geq t_\gamma$, it follows that $\mathbf{x}(t) \in V^{-1}([V(\mathbf{x}_e), V(\mathbf{x}(t_\gamma))])$. Let $m_e = V(\mathbf{x}_e)$ be the unique minimum value achieved by V , which occurs only at $\mathbf{x}_e$. We claim that there exists $s > m_e$ such that $V^{-1}([m_e, s]) \subseteq S(\mathbf{x}_e, \varepsilon) = \{\mathbf{x} \in X : \|\mathbf{x} - \mathbf{x}_e\| \leq \varepsilon\}$. If this were not the case, then there exists $\mathbf{y}_s \in S(\mathbf{x}_e, \varepsilon_0)$ outside $S(\mathbf{x}_e, \varepsilon)$ with $V(\mathbf{y}_s) < s$, for values s arbitrarily close to m_e . Since $S(\mathbf{x}_e, \varepsilon_0)$ is compact, $\{\mathbf{y}_s\}$ has a cluster point $\bar{\mathbf{y}} \in S(\mathbf{x}_e, \varepsilon_0)$, and $\|\bar{\mathbf{y}} - \mathbf{x}_e\| \geq \varepsilon$, so $\bar{\mathbf{y}} \neq \mathbf{x}_e$. But

$$V(\bar{\mathbf{y}}) = \lim_{s \rightarrow m_e} V(\mathbf{y}_s) = m_e,$$

which contradicts the fact that V achieves its minimum only at $\mathbf{x}_e$. Therefore s must exist. Because $V^{-1}([m_e, s])$ is a neighborhood of $\mathbf{x}_e$, there exists $\mathbf{x}(t_\varepsilon)$ such that $\|\mathbf{x}(t_\varepsilon) - \mathbf{x}_e\| < \varepsilon$ and $V(\mathbf{x}(t_\varepsilon)) < s$. Since V is strictly decreasing in $\mathbf{x}(t)$ by hypothesis, $V(\mathbf{x}(t)) < s$ for all $t > t_\varepsilon$. Thus, $\mathbf{x}(t) \in V^{-1}([m_e, s]) \subseteq S(\mathbf{x}_e, \varepsilon)$, i.e., a “tail”

of the set $\{x(t): t \geq t_0\}$ is in $S(x_e, \varepsilon)$. This is true for ε arbitrarily small, hence $x(t) \rightarrow x_e$ as $t \rightarrow \infty$, as desired. ■

As in the discrete case, the usefulness of this theorem is that if we can find a Liapunov function on S , and we know that every solution $x(t)$ has its trajectory contained in S , then we can conclude that x_e is stable—without ever computing the solution $x(t)$ of the state transition equation.

The following corollary is analogous to Corollary 2.12, which was proved in the last subsection.

2.16 Corollary Suppose X is an open sphere, the function V of Theorem 2.15 is defined on the entire state space X , $\dot{V}(x) = \nabla V(x)f(x) < 0$ for all $x \neq x_e$, and $V(x) \rightarrow \infty$ as any component of $x \in X$ gets arbitrarily large in magnitude. Then x_e is a globally asymptotic equilibrium state. ([SB], p. 170 and [Lue], p. 339)

Proof

We must show that for any $x_0 \in X$ and for any solution $x(t)$ for which $x(t_0) = x_0$, that $x(t) \rightarrow x_e$ as $t \rightarrow \infty$. Let $x_0 \in X$ be an arbitrary initial state of a solution $x(t)$. Our assumption about the monotonicity of V and the fact that $V(x) \rightarrow \infty$ as any component of $x \in X$ gets arbitrarily large implies that the set $\{x(t): t \geq t_0\}$ is bounded. Hence we can find an $\varepsilon_0 > 0$ such that the entire state trajectory of $x(t)$ is in $S(x_e, \varepsilon_0)$. Now restrict V to $S(x_e, \varepsilon_0)$ and apply Theorem 2.15 to conclude that $x(t) \rightarrow x_e$ as $t \rightarrow \infty$. Since $x_0 \in X$ was arbitrary, we have that x_e is globally asymptotic. ■

We close this subsection with an example that will also be used in the next subsection.

2.17 Example Consider the continuous time dynamical system

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ -x_1(t) \end{bmatrix} = \mathbf{f}(\mathbf{x}(t))$$

Since $\mathbf{f}(\mathbf{0}) = \mathbf{0}$, it follows that the zero vector is an equilibrium state for this dynamical system. Define the function $V: \mathbb{R}^2 \rightarrow \mathbb{R}$ by

$$V(x_1, x_2) = x_1^2 + x_2^2$$

This function is continuous, has continuous first partial derivatives, and has a unique minimum at $\mathbf{x}_e = \mathbf{0}$, that is, Conditions 2.14A–B of Definition 2.14 are satisfied. To see that 2.14C is satisfied, note that, at any $\mathbf{x}$,

$$\begin{aligned} \dot{V}(\mathbf{x}) &= \nabla V(\mathbf{x})\mathbf{f}(\mathbf{x}) \\ &= 2x_1f_1(\mathbf{x}) + 2x_2f_2(\mathbf{x}) \\ &= 2x_1x_2 + 2x_2(-x_1) \\ &= 0 \end{aligned}$$

Thus $(V, \mathbb{R}^2)$ is a Liapunov function for $\mathbf{x}_e = \mathbf{0}$, which is therefore a stable equilibrium state for this dynamical system by Theorem 2.15A; however, we cannot infer asymptotic stability for this equilibrium state. ([Lue], p. 338)

In fact, $\mathbf{x}_e = \mathbf{0}$ is not asymptotically stable: the general solution of this dynamical system is $\mathbf{x}(t) = [k \sin t \quad k \cos t]^T$, and for any $k \neq 0$, the trajectory with $\mathbf{x}(0) = [0 \quad k]^T$ is a circle of radius k , hence $\mathbf{x}(t) \not\rightarrow \mathbf{0}$ as $t \rightarrow \infty$. ■

Before leaving this subsection, note that the function V in the preceding example has the property that for any state $\mathbf{x}(t)$ of the dynamical system, $\dot{V}(\mathbf{x}(t)) = 0$ for all $t \geq t_0$, so that $V(\mathbf{x}(t))$ never increases nor decreases over the entire state trajectory. That is, the entire trajectory lies along a contour of the function V . This special case is considered in the next subsection.

2.2.5 A Special Case: “Constants of Motion”

At the end of the last subsection, in Example 2.17, it turned out that the state trajectory lies along a contour of the Liapunov function V . This is a very special situation but when it occurs, and we know the general geometric form of the contour, then we know a lot about the path of the state as time increases. Moreover, the particular contour is determined once the initial state is chosen. Luenberger gives the following definition for the special case that V is constant over an entire trajectory.

2.18 Definition (Constant of motion) Suppose (V, S) is a Liapunov function for the state of a continuous (resp. discrete) dynamical system. Let $\mathbf{x}(t)$ be a solution such that $V(\mathbf{x}(t)) = k$ for all $t \geq t_0$ (resp. $t \geq 0$). Then V is said to be a **constant of motion**. ([Lue], p. 338) ■

Recall the continuous dynamical system and Liapunov function of Example 2.17:

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ -x_1(t) \end{bmatrix}, \quad V(x_1, x_2) = x_1^2 + x_2^2$$

Assume that the initial condition is given by $\mathbf{x}_0 = \begin{bmatrix} k \\ 0 \end{bmatrix}$. In Example 2.17 we observed that the solution is of the form $\mathbf{x}(t) = \begin{bmatrix} k \sin t \\ k \cos t \end{bmatrix}$, and the trajectory does lie along a contour of V , $V(\mathbf{x}) = k^2$.

For a less trivial example, we consider a dynamical system that models the behavior of a disease as it spreads throughout a population. Our formulation follows that given in Luenberger ([Lue], pp. 376–378).

2.19 Example Consider a population of n individuals, and a disease that is spreading through contact of the members of the population. At any time t let

$x_1(t)$ = number of susceptible non infected individuals

$x_2(t)$ = number of infected and circulating individuals

$x_3(t)$ = number of individuals removed by death or isolation or immunity

Clearly $x_1(t) + x_2(t) + x_3(t) = n$ for all t . Additionally, it is assumed that the contact rate between susceptible individuals and those infected is proportional to the product $x_1(t)x_2(t)$, so that the generation rate of new infected members is $\beta x_1(t)x_2(t)$, where β is the *infection rate constant*. Further, the number of infected members who are removed or become immune is assumed to be proportional to the total number of infected individuals $x_2(t)$. Hence,

$$\dot{x}_2(t) = \beta x_1(t)x_2(t) - \gamma x_2(t),$$

where $\beta, \gamma > 0$, and γ is the *infective depletion constant*. It follows that the other differential equations for this model are

$$\dot{x}_1(t) = -\beta x_1(t)x_2(t)$$

$$\dot{x}_3(t) = \gamma x_2(t)$$

Suppressing t , this dynamical system can be written in vector form as follows:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -\beta x_1 x_2 \\ \beta x_1 x_2 - \gamma x_2 \\ \gamma x_2 \end{bmatrix}$$

Observe that

$$\frac{d}{dt}(x_1 + x_2 + x_3) = -\beta x_1 x_2 + (\beta x_1 x_2 - \gamma x_2) + \gamma x_2 = 0,$$

so that the sum of the number of susceptible individuals, infected and circulating individuals, and those removed by death, isolation, or immunity is constant at any time t .

Observe that the only equilibrium states for this system occur when $x_2(t) \equiv 0$, that is, when there are no infected and circulating individuals in the population. Therefore in this model an equilibrium state $\mathbf{x}_e$ must be of the form

$$\mathbf{x}_e = \begin{bmatrix} x_1^e & 0 & x_3^e \end{bmatrix}^T$$

Since x_3 does not appear in the differential equations for x_1 and x_2 , we can analyze the simpler system formed by these two equations. Notice that

$$\frac{\dot{x}_1}{\dot{x}_2} = \frac{-\beta x_1 x_2}{\beta x_1 x_2 - \gamma x_2} = \frac{-\beta x_1}{\beta x_1 - \gamma}$$

Rearranging this equation gives

$$\dot{x}_1 - \frac{\sigma \dot{x}_1}{x_1} + \dot{x}_2 = 0, \quad \sigma = \frac{\gamma}{\beta}$$

Integrate this last equation with respect to t to get

$$x_1(t) - \sigma \ln x_1(t) + x_2(t) = C$$

Define $V(x_1, x_2)$ by

$$V(x_1, x_2) = x_1 - \sigma \ln x_1 + x_2$$

Clearly this function is constant on any trajectory. It is also a Liapunov function for the equilibrium state

$$\mathbf{x}_e = \begin{bmatrix} \sigma & 0 \end{bmatrix}^T$$

since clearly V is continuous over its domain, and $\dot{V}(x_1(t), x_2(t)) = 0$ for all t by construction, so it is nonincreasing in t . To see V has its unique minimum at $\mathbf{x}_e$, first observe that with $x_2 = 0$ —which is required of all equilibrium states, V simplifies to a function of one variable:

$$V(x_1, 0) = x_1 - \sigma \ln x_1$$

Differentiating this function with respect to x_1 and setting the resulting function equal to zero gives $1 - \frac{\sigma}{x_1} = 0$, i.e., $x_1 = \sigma = \frac{\gamma}{\beta}$. The second derivative with respect to x_1 is $\frac{\sigma}{x_1^2}$, which is positive at $x_1 = \sigma$, so $V(x_1, 0)$ has its unique minimum at $x_1 = \sigma$ for all $x_1 > 0$. Since

$$V(\sigma, 0) \leq V(x_1, 0) < V(x_1, x_2)$$

for any $x_2 > 0$, V does have a unique global minimum with respect to

$$X = \left\{ \begin{bmatrix} x_1 & x_2 \end{bmatrix}^T \in \mathbb{R}^2 \mid x_1 > 0, x_2 \geq 0 \right\}$$

at $\mathbf{x}_e = [\sigma \ 0]^T$. Therefore, (V, X) is a Liapunov function for $\mathbf{x}_e$.

Fix an initial state $\mathbf{x}_0 = [x_{01} \ x_{02}]^T$. Then the solution originating at $\mathbf{x}_0$ satisfies $V(x_1(t), x_2(t)) = V(x_{01}, x_{02})$ for all $t \geq 0$, which can be rewritten as follows (and suppressing t):

$$x_1 - \sigma \ln x_1 + x_2 = x_{01} - \sigma \ln x_{01} + x_{02}$$

Solving for x_2 makes this easy to plot:

$$x_2 = x_{02} + x_{01} - x_1 + \sigma \ln\left(\frac{x_1}{x_{01}}\right)$$

Some of the contours of function V , that solutions $\mathbf{x}(t) = [x_1(t) \ x_2(t)]^T$ must lie along, are plotted below in Figure 2.7. The arrows indicate the direction of movement of the trajectory $\mathbf{x}(t)$ as time increases, as dictated by the governing differential equations of the dynamical system.

A number of interesting conclusions can be drawn from these contours about the trajectories $\mathbf{x}(t)$ that describe the epidemic.

First, the maximum value of $x_2(t)$, the number of infected and circulating individuals, occurs when $x_1(t) = \sigma$, where $x_1(t)$ is the number of susceptible individuals. Following the direction of movement along a particular contour of function V , we see that if the initial number of susceptible individuals, x_{01} , is greater than σ , then the number of infected individuals, $x_2(t)$, increases over time—that is, an epidemic occurs. However, if $x_{01} < \sigma$, which could happen for example if enough individuals were immune in the beginning, then $x_2(t)$ decreases monotonically to zero and an epidemic does not occur. Thus σ acts as a threshold value for an epidemic. ([Lue], pp. 377–378)

Second, since $x_2(t) \rightarrow -\infty$ as $x_1(t) \rightarrow 0$, it follows that $x_2(t_1) = 0$ for some positive t_1 , but $x_1(t_1) > 0$. Thus an occurrence of the disease terminates because of a shortage of infected individuals, not a lack of susceptible individuals; moreover, some susceptible individuals will escape the disease altogether, since $x_1(t) > 0$ when $x_2(t)$ vanishes. ([Lue], p. 378) ■

Figure 2.7 Contours of function V in Example 2.19

2.3 STABILITY ANALYSIS: TIME INVARIANT LINEAR SYSTEMS

In analyzing whether an equilibrium state $\mathbf{x}_e$ is stable, we are concerned with the difference $\mathbf{x}(t) - \mathbf{x}_e$ and, in particular, whether $\|\mathbf{x}(t) - \mathbf{x}_e\|$ is bounded, tends to zero, or increases without bound, as $t \rightarrow \infty$. In the case of time invariant linear dynamical systems, this difference can be expressed in a simple closed form.

2.20 Theorem Suppose we are given a time invariant continuous linear dynamical system, $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$, or a time invariant discrete linear system, $\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$. Let $\mathbf{x}(t)$ be a solution with initial state $\mathbf{x}_0 = \mathbf{x}(t_0)$ (resp. $\mathbf{x}_0 = \mathbf{x}(0)$), and let $\mathbf{x}_e$ be an equilibrium. Then

$$\mathbf{x}(t) - \mathbf{x}_e = e^{(t-t_0)\mathbf{A}}(\mathbf{x}_0 - \mathbf{x}_e) \quad (2.45)$$

in the continuous case, and

$$\mathbf{x}(t) - \mathbf{x}_e = \mathbf{A}^t(\mathbf{x}_0 - \mathbf{x}_e) \quad (2.46)$$

in the discrete case.

Proof

We prove the continuous case first. By definition of an equilibrium state,

$$\mathbf{A}\mathbf{x}_e + \mathbf{B}\mathbf{u}(t) = \mathbf{0},$$

so we have

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \\ &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) - (\mathbf{A}\mathbf{x}_e + \mathbf{B}\mathbf{u}(t)) \\ &= \mathbf{A}(\mathbf{x}(t) - \mathbf{x}_e) \end{aligned}$$

Letting $\mathbf{x}_\delta(t) = \mathbf{x}(t) - \mathbf{x}_e$, we have

$$\dot{\mathbf{x}}_\delta(t) = \mathbf{A}\mathbf{x}_\delta(t)$$

By Corollary 1.13 and Theorem 1.21, the solution of this equation is given as

$$\mathbf{x}_\delta(t) = e^{(t-t_0)\mathbf{A}}\mathbf{x}_\delta(t_0) \quad (2.47)$$

Notice that $\mathbf{x}_\delta(t_0) = \mathbf{x}(t_0) - \mathbf{x}_e = \mathbf{x}_0 - \mathbf{x}_e$. Substitute this information into (2.47) to obtain equation (2.45).

Next, we turn to the discrete case. In this case the definition of an equilibrium state gives

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

which implies that

$$\begin{aligned}\mathbf{x}(t+1) - \mathbf{x}_e &= (A\mathbf{x}(t) + B\mathbf{u}(t)) - (A\mathbf{x}_e + B\mathbf{u}(t)) \\ &= A(\mathbf{x}(t) - \mathbf{x}_e)\end{aligned}$$

Define $\mathbf{x}_\delta(t)$ as above and note that this last equation implies that

$$\mathbf{x}_\delta(t+1) = A\mathbf{x}_\delta(t)$$

which, by Corollary 1.34 and Theorem 1.37, has the solution

$$\mathbf{x}_\delta(t) = A^t \mathbf{x}_\delta(0) \tag{2.48}$$

Substituting $\mathbf{x}_\delta(0) = \mathbf{x}(0) - \mathbf{x}_e = \mathbf{x}_0 - \mathbf{x}_e$ into (2.48), we obtain equation (2.46). ■

Using the closed forms (2.45)–(2.46) we can see that the stability of an equilibrium state for a time invariant continuous (resp. discrete) linear dynamical system depends entirely on properties of the fundamental matrix.

2.21 Corollary The equilibrium state of the time invariant continuous (resp. discrete) linear dynamical system in Theorem 2.20 is

- A. stable if and only if $e^{(t-t_0)A}$ (resp. A^t) is bounded for $t \geq t_0$ (resp. $t \geq 0$);
- B. asymptotically stable if and only if $e^{(t-t_0)A}$ (resp. A^t) tends to the zero matrix as $t \rightarrow \infty$. ■

This theorem is an immediate consequence of the fact that the difference $\mathbf{x}_0 - \mathbf{x}_e$ on the right side of equation (2.45) (resp. (2.46)) is constant.

Corollary 2.21 says that the entire question of stability for time invariant linear dynamical systems can be reduced to an examination of the fundamental matrix for the system. Moreover, we have the following result.

2.22 Corollary Suppose the state space, X , is taken to be $\mathfrak{R}^n$. Then equilibrium state of the time invariant continuous or discrete linear dynamical system in Theorem 2.20 is asymptotically stable if and only if it is globally asymptotically stable. Therefore if there is an asymptotically stable equilibrium point, it is the only equilibrium point of the system.

Proof

It is clear that global asymptotic stability of $\mathbf{x}_e$ implies asymptotic stability of $\mathbf{x}_e$. Note that with $X = \mathfrak{R}^n$, to see that asymptotic stability of $\mathbf{x}_e$ implies global asymptotic stability of $\mathbf{x}_e$, it suffices to see that for every $\varepsilon > 0$, if $\|\mathbf{x}_e - \mathbf{x}_0\| < \varepsilon$, then $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ as $t \rightarrow \infty$. But this is immediate from Corollary 2.21, since if $\mathbf{x}_e$ is asymptotically stable, the fundamental matrix tends to zero, whence $\mathbf{x}(t) \rightarrow \mathbf{x}_e$ by equation (2.45) (resp. (2.46)), and the distance of the initial state from the equilibrium state does not affect this convergence. Clearly an asymptotically stable equilibrium state for the system, if it exists, is the only one possible. ■

2.3.1 A Stability Representation of the Fundamental Matrix

By Corollary 2.21, the question of stability for time invariant linear systems can be reduced to whether the fundamental matrix of the system is bounded, or bounded and tends to zero as $t \rightarrow \infty$. In this section, we will draw on the theory of matrix functions to develop representations of the fundamental matrix for continuous and discrete dynamical systems that can be used to reduce the question of the stability of the equilibrium state to conditions involving the eigenvalues of the matrix A in Theorem 2.20. In Section 2.3.2 we will use these representations to develop such eigenvalue criteria.

It is well-known in matrix theory that any $n \times n$ complex matrix A can be transformed via a similarity transformation to its Jordan form,

$$A = SJS^{-1},$$

where the Jordan matrix J is given by

$$J = \begin{bmatrix} J_{n_1}(\lambda_1) & & 0 \\ & J_{n_2}(\lambda_2) & \\ & & \ddots \\ 0 & & & J_{n_s}(\lambda_s) \end{bmatrix} \quad (2.49)$$

Here $n_1 + n_2 + \dots + n_s = n$, the eigenvalues λ_i , $i = 1, \dots, s$ are not necessarily distinct, and

$$J_k(\lambda) = \begin{bmatrix} \lambda & 1 & & 0 \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix}_{k \times k}$$

([HJ85], p. 126). The following theorem is the basis of the entire derivation that follows.

2.23 Theorem Suppose $f(t)$ is a scalar-valued analytic function with a power series representation $f(t) = a_0 + a_1 t + a_2 t^2 + \dots$ that has a radius of convergence

$R > 0$. Suppose further that A is an $n \times n$ complex matrix and $|\lambda| < R$ for every eigenvalue λ of A . Then the matrix power series

$$f(A) \equiv a_0 I + a_1 A + a_2 A^2 + \dots$$

converges with respect to any matrix norm and its sum is called the *primary matrix function* $f(A)$. Also, the following is true:

$$\begin{aligned} f(A) &= S f(J) S^{-1} \\ &= S \begin{bmatrix} f(J_{n_1}(\lambda_1)) & & 0 \\ & \ddots & \\ 0 & & f(J_{n_s}(\lambda_s)) \end{bmatrix} S^{-1} \end{aligned}$$

and

$$f(J_k(\lambda)) = \begin{bmatrix} f(\lambda) & f'(\lambda) & \frac{1}{2}f''(\lambda) & \dots & \dots & \frac{1}{(k-1)!}f^{(k-1)}(\lambda) \\ 0 & f(\lambda) & f'(\lambda) & \ddots & & \vdots \\ 0 & 0 & f(\lambda) & \ddots & \ddots & \vdots \\ \vdots & \vdots & 0 & \ddots & \ddots & \frac{1}{2}f''(\lambda) \\ & & & \ddots & \ddots & f'(\lambda) \\ 0 & 0 & \dots & & 0 & f(\lambda) \end{bmatrix}$$

([HJ91], p. 412) ■

The following two corollaries to this theorem will give the desired representations of the fundamental matrix that will be used in establishing the stability criteria given in the next subsection.

2.24 Corollary Suppose A is an $n \times n$ complex matrix with Jordan form given by equation (2.49). Let $\lambda_1, \dots, \lambda_r$ be the distinct eigenvalues of A , and let m_i denote the algebraic multiplicity of λ_i . Then there are $n \times n$ constant matrices $B_{i,l}$ and $C_{i,l}$ so that

$$\text{I.} \quad A^t = \sum_{i=1}^r \lambda_i^t \sum_{l=0}^{m_i-1} t^l C_{i,l} \quad (2.50)$$

$$\text{II.} \quad e^{tA} = \sum_{i=1}^r e^{\lambda_i t} \sum_{l=0}^{m_i-1} t^l B_{i,l} \quad (2.51)$$

where in II, t is any real number, and in I, t is any real number greater than $\max\{m_i: i = 1, \dots, r\}$.

Proof

I: Take $f(z) = z^t$ in Theorem 2.23 to obtain

$$A^t = S \begin{bmatrix} [J_{n_1}(\lambda_1)]^t & & 0 \\ & \ddots & \\ 0 & & [J_{n_s}(\lambda_s)]^t \end{bmatrix} S^{-1}$$

Notice that if $\lambda = 0$, then

$$(J_k(0))^t = \begin{bmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix}^t = O,$$

since $t \geq k$ by hypothesis. If $\lambda \neq 0$, by (2.49)

$$\begin{aligned} [J_k(\lambda)]^t &= \begin{bmatrix} \lambda^t & t\lambda^{t-1} & \frac{1}{2}t(t-1)\lambda^{t-2} & \cdots & \frac{1}{(k-1)!}t(t-1)\cdots(t-k+2)\lambda^{t-k+1} \\ & \ddots & \ddots & \ddots & \vdots \\ & 0 & & \frac{1}{2}t(t-1)\lambda^{t-2} & \\ & & & t\lambda^{t-1} & \\ & & & \lambda^t & \end{bmatrix} \\ &= \lambda^t \begin{bmatrix} t^0 & t^1\lambda^{-1} & \frac{1}{2}t^2(1-\frac{1}{t})\lambda^{-2} & \cdots & \frac{1}{(k-1)!}t^{k-1}(1+q(t))\lambda^{-k+1} \\ & \ddots & \ddots & \ddots & \vdots \\ & & 0 & \ddots & \frac{1}{2}t^2(1-\frac{1}{t})\lambda^{-2} \\ & & & t^1\lambda^{-1} & \\ & & & t^0 & \end{bmatrix} \\ &= \lambda^t \left\{ t^0 I + t^1 \begin{bmatrix} 0 & \lambda^{-1} & 0 \\ & \ddots & \ddots & \lambda^{-1} \\ 0 & & 0 \end{bmatrix} + \cdots + t^{k-1} \begin{bmatrix} 0 & \cdots & 0 & \frac{1}{(k-1)!}(1+q(t))\lambda^{-k+1} \\ & \ddots & & 0 \\ & & \ddots & \vdots \\ 0 & \cdots & & 0 \end{bmatrix} \right\} \\ &= \lambda^t \sum_{i=0}^{k-1} t^i M_{k,i} \end{aligned}$$

$$= \lambda^t \sum_{i=0}^{m-1} t^i M_{k,i} \quad (2.52)$$

where m is the multiplicity of λ , $M_{k,i}$ is a $k \times k$ constant matrix for $i = 0, \dots, m-1$, with $M_{k,i} = O$ if $k < m$ and $k \leq i \leq m-1$, and $q(t)$ is a rational polynomial function.

Finally, it follows from (2.52) that

$$\begin{aligned} A^t &= S \begin{bmatrix} [J_{n_1}(\lambda_1)]^t & & 0 \\ & \ddots & \\ 0 & & [J_{n_s}(\lambda_s)]^t \end{bmatrix} S^{-1} \\ &= S \sum_{j=1}^s \begin{bmatrix} 0 & & & 0 \\ & \ddots & & \\ & & 0 & \\ & & [J_{n_j}(\lambda_j)]^t & \\ & 0 & & 0 \\ & 0 & & \ddots & \\ & & & & 0 \end{bmatrix} S^{-1} \\ &= \sum_{j=1}^s \lambda_j^t \sum_{l=0}^{m_j-1} t^l S \begin{bmatrix} 0 & & & 0 \\ & \ddots & & \\ & & 0 & \\ & & [M_{n_j,l}] & \\ & 0 & & 0 \\ & 0 & & \ddots & \\ & & & & 0 \end{bmatrix} S^{-1} \quad \text{by (2.52)} \\ &= \sum_{j=1}^r \lambda_j^t \sum_{l=0}^{m_j-1} t^l C_{j,l} \end{aligned} \quad (2.53)$$

where, in equation (2.53), each λ_j is a distinct eigenvalue of A , and in forming the $n \times n$ matrices $C_{j,l}$, we have combined the matrices $M_{n_j,l}$ associated with each distinct eigenvalue into a single block diagonal matrix. Finally, notice that the form given by (2.53) agrees with the representation (2.50) stated in part I.

II. Let $f(z) = e^{zt} = \sum_{i=0}^{\infty} \frac{(zt)^i}{i!}$ so that $f^{(k)}(z) = t^k e^{zt}$. In this case, when f is

applied to the Jordan block $J_k(\lambda)$, we have from Theorem 2.23 that

$$\begin{aligned}
 f(J_k(\lambda)) &= e^{t[J_k(\lambda)]} \\
 &= \begin{bmatrix} e^{\lambda t} & te^{\lambda t} & \frac{1}{2}t^2e^{\lambda t} & \dots & \frac{1}{(k-1)!}t^{k-1}e^{\lambda t} \\ 0 & e^{\lambda t} & te^{\lambda t} & \ddots & \vdots \\ \vdots & \ddots & & \ddots & \frac{1}{2}t^2e^{\lambda t} \\ & & & \ddots & te^{\lambda t} \\ 0 & \dots & 0 & & e^{\lambda t} \end{bmatrix} \\
 &= e^{\lambda t} \left\{ t^0 I + t \begin{bmatrix} 0 & 1 & 0 \\ & 0 & \ddots \\ 0 & & 0 \end{bmatrix} + \dots + t^{k-1} \begin{bmatrix} 0 & \dots & 0 & \frac{1}{(k-1)!} \\ & 0 & \ddots & \vdots \\ & & 0 & 0 \end{bmatrix} \right\} \\
 &= e^{\lambda t} \sum_{l=0}^{k-1} t^l \tilde{M}_{k,l} \\
 &= e^{\lambda t} \sum_{l=0}^{m-1} t^l \tilde{M}_{k,l}
 \end{aligned} \tag{2.54}$$

where m is the multiplicity of λ , $\tilde{M}_{k,l}$ is a $k \times k$ constant matrix for $l = 0, 1, \dots, m-1$, and $\tilde{M}_{k,l} = O$ if $k \leq l \leq m-1$. This is true even if $\lambda = 0$. The rest of the proof is a repetition of the steps in the proof of part I. Substitute the form of a Jordan block given by (2.54) into the function applied to the Jordan form of A as given by Theorem 2.23 to get

$$\begin{aligned}
 f(A) &= e^{tA} \\
 &= \sum_{j=1}^r e^{\lambda_j t} \sum_{l=0}^{m_j-1} t^l B_{j,l}
 \end{aligned}$$

where again, blocks $\tilde{M}_{n_j,l}$ associated with each distinct eigenvalue are combined into a single block diagonal matrix in forming $B_{j,l}$. ■

2.3.2 Stability of Time Invariant Linear Dynamical Systems

In the first part of this subsection, we consider stability criteria based on the eigenvalues of the matrix A of a time invariant linear dynamical system with transition equation given by

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \text{ (resp. } \mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)) \quad (2.55)$$

Later, for continuous systems, we consider alternate tests for stability that are not based directly on knowledge of the eigenvalues of the matrix A .

The proof of the following theorem uses the representations that we developed in the previous subsection.

2.25 Theorem Suppose we are given a continuous (resp. discrete) time invariant linear dynamical system with state transition equation given by (2.55) and suppose that $\mathbf{x}_e$ is an equilibrium state for the system. Then

- I. $\mathbf{x}_e$ is stable if for all the eigenvalues of A , $\text{Re}(\lambda) \leq 0$ (i.e., the real part of the eigenvalue) (resp. $|\lambda| \leq 1$), and all eigenvalues with the property $\text{Re}(\lambda) = 0$ (resp. $|\lambda| = 1$) have single multiplicity;
- II. $\mathbf{x}_e$ is asymptotically stable if and only if for every eigenvalue of A , $\text{Re}(\lambda) < 0$ (resp. $|\lambda| < 1$); and
- III. $\mathbf{x}_e$ is not stable if for any eigenvalue λ of A , $\text{Re}(\lambda) > 0$ (resp. $|\lambda| > 1$).

([SB], pp. 174–175)

Proof

By Theorem 2.20 and Corollary 2.21 it suffices to see that the fundamental matrix is bounded for part I, bounded and tends to zero for part II, and not bounded for part III. As in Section 2.3.1, we can factor A as follows: $A = \mathbf{S}\mathbf{J}\mathbf{S}^{-1}$, where the Jordan matrix $\mathbf{J}$ is given by equation (2.49). Furthermore, by the proof of Corollary 2.24, the fundamental matrix is given by $e^{t\mathbf{A}} = \mathbf{S}e^{t\mathbf{J}}\mathbf{S}^{-1}$ in the

continuous case, and $A^t = SJ^tS^{-1}$ in the discrete case. So it suffices to consider an arbitrary block $J_k(\lambda)$ of J , and we consider the continuous and discrete cases separately.

For the continuous case, we have by equation (2.54) that

$$e^{tJ_k(\lambda)} = e^{\lambda t} \sum_{l=0}^{m-1} t^l \tilde{M}_{k,l}$$

where each $\tilde{M}_{k,l}$ is constant and $\tilde{M}_{k,0} = I$. Clearly, if $\text{Re}(\lambda) > 0$, then the diagonal entries $e^{\lambda t}$ become arbitrarily large as $t \rightarrow \infty$, whence e^{tJ} becomes arbitrarily large so that x_e is not stable. This proves part III.

Also, if $\text{Re}(\lambda) < 0$, so $\lambda = a + ib$, $a < 0$, then

$$|t^l e^{\lambda t}| = |t^l e^{at} e^{ibt}| \leq t^l e^{at} \quad (t > 0, l \geq 0)$$

and $t^l e^{at} \rightarrow 0$ as $t \rightarrow \infty$ by L'Hospital's rule from ordinary differential calculus, whence $e^{tJ_k(\lambda)} \rightarrow 0$ as $t \rightarrow \infty$. Therefore if every eigenvalue of A has negative real part, then $e^{tJ} \rightarrow 0$ as $t \rightarrow \infty$, so that the fundamental matrix tends to the zero matrix.

Finally, suppose $\lambda = ib$, $b \neq 0$, then $|t^l e^{\lambda t}| = |t^l e^{ibt}| = |t^l|$ ($t > 0$). If $m \geq 2$, then the term with $l = 1$ appears in equation (2.54):

$$e^{ibt} t \tilde{M}_{k,1} = e^{ibt} t \begin{bmatrix} 0 & 1 & & 0 \\ & 0 & \ddots & \\ & & \ddots & 1 \\ 0 & & & 0 \end{bmatrix}$$

and this becomes arbitrarily large as $t \rightarrow \infty$, hence e^{tJ} diverges as $t \rightarrow \infty$.

However, if $m = 1$, the only term in (2.54) is

$$e^{ibt} \tilde{M}_{k,0} = e^{ibt} I$$

and this is bounded for all t because $|e^{ibt}| = 1$.

Thus, any block associated with a purely imaginary eigenvalue will be bounded if and only if the Jordan block $J_k(\lambda)$ has size equal to one, and will become arbitrarily large otherwise, as $t \rightarrow \infty$.

To summarize, we have shown that $e^{tA} \rightarrow 0$ as $t \rightarrow \infty$ if and only if every eigenvalue of A has negative real part, which proves part II, and e^{tA} will be bounded if the hypotheses of part I hold.

For the discrete case, we have from equation (2.52) that

$$[J_k(\lambda)]^t = \lambda^t \sum_{j=0}^{m-1} t^j M_{k,j}$$

The argument from here is a mild modification of the argument above. If $|\lambda| > 1$ then $|\lambda^t t^j| = |\lambda|^t t^j \rightarrow \infty$ for all $j \geq 0$, as $t \rightarrow \infty$, so that $[J_k(\lambda)]^t$ becomes arbitrarily large as $t \rightarrow \infty$ (recall $M_{k,0} = I$, so the sum has at least one nonzero matrix); hence, A^t will diverge as $t \rightarrow \infty$. If $|\lambda| < 1$, $|\lambda^t t^j| \rightarrow 0$ as $t \rightarrow \infty$, so $A^t \rightarrow 0$. If $|\lambda| = 1$ and $\text{Re}(\lambda) = 0$, then $\lambda = \pm ib$; for a 1×1 block, $[\pm ib]^t$ is bounded, but if $k \geq 2$, $[J_k(\pm ib)]^t$ diverges. ■

Notice that in part II of Theorem 2.25 we have necessary and sufficient conditions for asymptotic stability, but not in part I. The following example acts as a counterexample to converse of 2.25I.

2.26 Examples

Consider first the time invariant continuous linear dynamical system

$$\dot{x}(t) = 0x(t),$$

where 0 is the zero matrix of size $n \times n$, $n \geq 2$. Notice that every constant function $x(t) \equiv \bar{x} \in \mathbb{R}^n$ is both a solution and equilibrium state for the system. Since

$$\Phi(t, t_0) = e^{(t-t_0)0} = I$$

is constant, and hence bounded, for all $t \geq t_0$, we see that the equilibrium states are stable by Corollary 2.21. However, O has a single eigenvalue $\lambda = 0$ of multiplicity n . So this system has a stable equilibrium even though an eigenvalue with $\operatorname{Re}(\lambda) = 0$ is multiple. That is, the converse of 2.25I is false for continuous systems.

Now consider the discrete system given by

$$\mathbf{x}(t+1) = I\mathbf{x}(t),$$

where I is the $n \times n$ identity matrix and $n \geq 2$. Again, constant functions $\mathbf{x}(t) \equiv \bar{\mathbf{x}} \in \mathbb{R}^n$ are solutions and also equilibrium states for this system. In this case, $\lambda = 1$ is the only eigenvalue of this matrix, which occurs with multiplicity n , and

$$\Phi(t, 0) = I^t = I$$

which is constant—and hence bounded for all $t \geq 0$, whence by Corollary 2.21 the equilibrium states are stable. So the converse of 2.25I is false for discrete systems also. ([SB], p. 175) ■

The following corollary is an immediate consequence of our analysis of the Jordan blocks $J_k(\lambda)$ in the proof of Theorem 2.25.

2.27 Corollary The equilibrium state $\mathbf{x}_e$ of the continuous (resp. discrete) dynamical system of the hypotheses of Theorem 2.25 is stable if and only if for all eigenvalues λ of A , $\operatorname{Re}(\lambda) \leq 0$ (resp. $|\lambda| \leq 1$), and also if λ is a repeated eigenvalue of A with real part equal to zero (resp. absolute value equal to one), then the size of each Jordan block $J_k(\lambda)$ in the Jordan canonical form of A is 1×1 . ■

Remark

Since A is diagonalizable if and only if each Jordan block in the Jordan canonical form of A has size equal to one, the following result is immediate from our work above.

2.28 Corollary Suppose the matrix A of a continuous (resp. discrete) time invariant linear dynamical system, with state transition equation (2.55), is diagonalizable, and $\mathbf{x}_e$ is an equilibrium state for the system. Then $\mathbf{x}_e$ is

- I. stable if and only if $\operatorname{Re}(\lambda) \leq 0$ (resp. $|\lambda| \leq 1$), for each eigenvalue λ of A ;
- II. asymptotically stable if and only if $\operatorname{Re}(\lambda) < 0$ (resp. $|\lambda| < 1$), for each eigenvalue λ of A . ■

Referring back to Example 2.26, note that the matrices O and I used there are diagonalizable, so 2.28I implies stability of the equilibrium states for these systems.

We close this subsection with a number of results that can be used in the stability analysis of continuous time invariant linear dynamical systems. In particular, alternatives to computing the eigenvalues of the matrix A are given.

The following theorem and corollary from the classic theory of polynomials can be used as an initial test for the stability of continuous time invariant linear dynamical systems.

2.29 Theorem Let $p_A(t) = \det(tI - A) = t^n + a_{n-1}t^{n-1} + \cdots + a_1t + a_0$ be the characteristic polynomial of the *real* matrix A . Suppose further that the real part of every eigenvalue of A is negative. Then $a_i > 0$ for every $i = 0, 1, \dots, n-1$. ([SB], p. 177)

Proof

If λ_k is any real eigenvalue, then $\lambda_k < 0$ implies that the linear factor $t - \lambda_k$ of $p_A(t)$ has positive coefficients.

Now assume that $\alpha_j + i\beta_j$ is a complex eigenvalue, with $\alpha_j < 0$ and $\beta_j \neq 0$. Since A is real, the characteristic polynomial has real coefficients, whence these

eigenvalues must occur in conjugate pairs. Hence, $\alpha_j - i\beta_j$ is also an eigenvalue of the same multiplicity as $\alpha_j + i\beta_j$. Thus,

$$(t - \alpha_j - i\beta_j)(t - \alpha_j + i\beta_j) = t^2 - 2\alpha_j t + \alpha_j^2 + \beta_j^2$$

is a factor of the characteristic polynomial, and has real positive coefficients, since $\alpha_j < 0$.

Since $p_A(t)$ can be factored into linear and quadratic factors as described above, it is not difficult to see (i.e., by induction on the degree of the characteristic polynomial) that every coefficient in $p_A(t)$ appears and is positive, as desired. ■

The contrapositive of Theorem 2.29, Theorem 2.25II, and Corollary 2.21 imply the following corollary.

2.30 Corollary Suppose $\mathbf{x}_e$ is an equilibrium state for the continuous time invariant linear dynamical system $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$. Suppose further that some coefficient of the characteristic polynomial of $\mathbf{A}$ is negative or zero. Then $\mathbf{x}_e$ is not asymptotically stable. ([SB], p. 177) ■

The next result gives a characterization of asymptotic stability that is not dependent on knowledge of the eigenvalues of the matrix $\mathbf{A}$.

2.31 Theorem Suppose $\mathbf{x}_e$ is an equilibrium state of the continuous time invariant dynamical system $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$. Then $\mathbf{x}_e$ is asymptotically stable if and only if the equation

$$\mathbf{A}^T \mathbf{Q} + \mathbf{Q} \mathbf{A} = -\mathbf{M} \tag{2.56}$$

has a positive definite solution $\mathbf{Q}$ for some positive definite matrix $\mathbf{M}$. ([SB], p. 176 and [HJ91], pp. 89–99)

Proof

($\Leftarrow$) Assume that (2.56) has a positive definite solution Q for some arbitrary positive definite matrix M . Recall that the asymptotic stability of the zero equilibrium state for the homogeneous system

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

implies the existence and asymptotic stability of a (real) equilibrium $\mathbf{x}_e$ for the system $\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t)$, by the proof of Theorem 2.20, and by equation (2.45).

Define $V(\mathbf{x}) = \mathbf{x}^T Q \mathbf{x}$. We claim that the pair $(V, \mathcal{R}^n)$ is a Liapunov function for the zero state in the homogeneous system. Since Q is positive definite, $V(\mathbf{x}) > 0$ if $\mathbf{x} \neq \mathbf{0}$, and has a unique minimum of zero when $\mathbf{x} = \mathbf{0}$. Thus it suffices to show that $\dot{V}(\mathbf{x}) < 0$ for every $\mathbf{0} \neq \mathbf{x} \in \mathcal{R}^n$, where $\dot{V}(\mathbf{x})$ denotes $\frac{d}{dt} V(\mathbf{x}(t))$. To see this, note that

$$\begin{aligned} \dot{V}(\mathbf{x}) &= \dot{\mathbf{x}}^T Q \mathbf{x} + \mathbf{x}^T Q \dot{\mathbf{x}} \\ &= (A\mathbf{x})^T Q \mathbf{x} + \mathbf{x}^T Q (A\mathbf{x}) \\ &= \mathbf{x}^T A^T Q \mathbf{x} + \mathbf{x}^T Q A \mathbf{x} \\ &= \mathbf{x}^T (A^T Q + Q A) \mathbf{x} \\ &= -(\mathbf{x}^T M \mathbf{x}) < 0 \end{aligned}$$

where the last equality follows from the fact that M is positive definite. Thus V is strictly decreasing in $\mathbf{x}$ —which implies the asymptotic stability of the zero vector for the homogeneous system by Theorem 2.15 and the fact that V has continuous first partial derivatives.

($\Rightarrow$) Suppose the equilibrium state $\mathbf{x}_e$ is asymptotically stable. From Theorem 2.25II we know that every eigenvalue of matrix A has a negative real part. Let M be any positive definite matrix.

Define

$$Q = \int_0^{\infty} e^{tA^T} M e^{tA} dt,$$

which is well defined because every eigenvalue of A has negative real part. This

is easy to see if A is diagonalizable. In this case, let P be such that

$P^{-1}AP = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$. Now the integral above converges if and only if

$$\begin{aligned} & P^T \int_0^{\infty} e^{tA^T} (P^T)^{-1} P^T M P P^{-1} e^{tA} dt P \\ &= \int_0^{\infty} [P^T e^{tA^T} (P^T)^{-1}] [P^T M P] [P^{-1} e^{tA} P] dt \\ &= \int_0^{\infty} e^{(P^{-1}tA)^T} \overline{M} e^{P^{-1}tA} dt \\ &= \int_0^{\infty} e^{(t\Lambda)^T} \overline{M} e^{t\Lambda} dt \end{aligned}$$

converges. This last integral is given in component form as $\int_0^{\infty} \left[e^{(\lambda_i + \lambda_j)t} \overline{m}_{ij} \right] dt$. If

$\text{Re}(\lambda_i) < 0$ for all i , then clearly each such integral converges. Conversely, if they all converge, then in particular the i, i term converges; since $\overline{M}$ is positive definite because M is, $\overline{m}_{ii} > 0$, so the term is not zero and therefore the convergence implies that $\text{Re}(\lambda_i) < 0$.

To see that Q is positive definite, let $\theta \neq \mathbf{v} \in \mathbb{R}^n$ so that

$$\mathbf{v}^T Q \mathbf{v} = \int_0^{\infty} \mathbf{v}^T e^{tA^T} M e^{tA} \mathbf{v} dt. \text{ Take } t_0 = 0 \text{ in the homogeneous dynamical system}$$

$\dot{\mathbf{x}}(t) = A\mathbf{x}(t)$. By Theorem 1.20, $\Phi(t, 0) = e^{tA}$, which by Theorem 1.12 is invertible.

Hence, $e^{tA} \mathbf{v} \neq \theta$, and it now follows that for any $\mathbf{v} \neq \theta$

$$\mathbf{v}^T e^{tA^T} M e^{tA} \mathbf{v} = (e^{tA} \mathbf{v})^T M (e^{tA} \mathbf{v}) > 0,$$

because M is positive definite. Thus,

$$\mathbf{v}^T \mathbf{Q} \mathbf{v} = \int_0^{\infty} \mathbf{v}^T e^{t\mathbf{A}^T} \mathbf{M} e^{t\mathbf{A}} \mathbf{v} dt > 0,$$

so $\mathbf{Q}$ is positive definite.

To see $\mathbf{Q}$ satisfies equation (2.56) note that

$$\begin{aligned} \mathbf{A}^T \mathbf{Q} + \mathbf{Q} \mathbf{A} &= \int_0^{\infty} \mathbf{A}^T e^{t\mathbf{A}^T} \mathbf{M} e^{t\mathbf{A}} dt + \int_0^{\infty} e^{t\mathbf{A}^T} \mathbf{M} e^{t\mathbf{A}} \mathbf{A} dt \\ &= \int_0^{\infty} \frac{d}{dt} \left(e^{t\mathbf{A}^T} \mathbf{M} e^{t\mathbf{A}} \right) dt \\ &= \left[e^{t\mathbf{A}^T} \mathbf{M} e^{t\mathbf{A}} \right]_0^{\infty} \\ &= \mathbf{O} - \mathbf{M} \\ &= -\mathbf{M} \end{aligned}$$

The integral exists since every eigenvalue of $\mathbf{A}$ (and $\mathbf{A}^T$) has negative real part; hence both matrices $e^{t\mathbf{A}}, e^{t\mathbf{A}^T}$ tend to the zero matrix as $t \rightarrow \infty$ [HJ91]. This completes the proof. ■

The following lemma will be used in a theorem below. However, it also can be used as another test of the asymptotic stability of equilibrium states for continuous linear dynamical systems with real coefficient matrix $\mathbf{A}$.

2.32 Lemma Suppose $\mathbf{A}$ is a real square matrix, and the matrix $\mathbf{A} + \mathbf{A}^T$ is negative definite. Then every eigenvalue of $\mathbf{A}$ has negative real part, whence the equilibrium state of any continuous dynamical system $\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t)$ is asymptotically stable.

Proof

Define the conjugate transpose of a vector $\mathbf{u}$ to be $\mathbf{u}^* = \bar{\mathbf{u}}^T$. Suppose λ is an arbitrary eigenvalue of $\mathbf{A}$, so that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, $\mathbf{v} \neq \mathbf{0}$. Notice that since $\mathbf{A} + \mathbf{A}^T$ is negative definite and $\mathbf{v} \neq \mathbf{0}$,

$$\begin{aligned}
0 &> \mathbf{v}^* (\mathbf{A} + \mathbf{A}^T) \mathbf{v} \\
&= \mathbf{v}^* \mathbf{A} \mathbf{v} + \mathbf{v}^* \mathbf{A}^T \mathbf{v} \\
&= \mathbf{v}^* \lambda \mathbf{v} + (\mathbf{v}^* \mathbf{A} \mathbf{v})^* \\
&= (\lambda + \bar{\lambda}) \mathbf{v}^* \mathbf{v} \\
&= 2\text{Re}(\lambda) \mathbf{v}^* \mathbf{v}
\end{aligned}$$

Since $\mathbf{v}^* \mathbf{v} = \|\mathbf{v}\|_2^2 > 0$, it follows that $\text{Re}(\lambda) < 0$, as desired. ■

In economics, the following continuous time invariant linear dynamical system often arises. The state transition equation is given by

$$\dot{\mathbf{x}}(t) = \mathbf{K} \mathbf{A} \mathbf{x}(t) + \mathbf{B} \mathbf{u}(t), \quad (2.57)$$

where the matrices are real, and $\mathbf{K}$ is positive definite. Typically $\mathbf{K}$ is diagonal with positive diagonal elements. ([SB], p. 179) The following theorem can be used in establishing the asymptotic stability of such systems.

2.33 Theorem Suppose $\mathbf{K}$ is positive definite, and $\mathbf{A} + \mathbf{A}^T$ is negative definite. Then all the eigenvalues of $\mathbf{K} \mathbf{A}$ have negative real parts (whence any equilibrium state for (2.57) is asymptotically stable by Corollary 2.21 and Theorem 2.25II). ([SB], p. 179)

Proof

By Lemma 2.32, $\mathbf{A}$ is invertible, and since $\mathbf{K}$ is positive definite, $\mathbf{K}$ is invertible whence $\mathbf{K} \mathbf{A}$ is invertible. Let λ be an arbitrary eigenvalue of $\mathbf{K} \mathbf{A}$, which is nonzero since $\mathbf{K} \mathbf{A}$ is invertible, and let $\mathbf{v} \neq \mathbf{0}$ be an associated eigenvector. Then $\mathbf{K} \mathbf{A} \mathbf{v} = \lambda \mathbf{v}$ which implies that $\mathbf{v} = \mathbf{K} \left(\frac{1}{\lambda} \mathbf{A} \mathbf{v} \right) = \mathbf{K} \mathbf{u}$. Since $\mathbf{K}$ is nonsingular and $\mathbf{v} \neq \mathbf{0}$, it follows that $\mathbf{u} \neq \mathbf{0}$. Substituting for $\mathbf{v}$ gives

$$\mathbf{K} \mathbf{A} \mathbf{K} \mathbf{u} = \lambda \mathbf{K} \mathbf{u}$$

Now use the negative definiteness of $\mathbf{A} + \mathbf{A}^T$ and calculate:

$$\begin{aligned}
0 &> (\mathbf{K}\mathbf{u})^* (\mathbf{A} + \mathbf{A}^T) \mathbf{K}\mathbf{u} = \mathbf{u}^* \mathbf{K}^* (\mathbf{A} + \mathbf{A}^T) \mathbf{K}\mathbf{u} \\
&= \mathbf{u}^* \mathbf{K} \mathbf{A} \mathbf{K}\mathbf{u} + \mathbf{u}^* \mathbf{K} \mathbf{A}^T \mathbf{K}\mathbf{u} \quad \{\mathbf{K} = \mathbf{K}^*\} \\
&= \lambda \mathbf{u}^* \mathbf{K}\mathbf{u} + (\mathbf{u}^* \mathbf{K} \mathbf{A} \mathbf{K}\mathbf{u})^* \quad \{\text{because } \mathbf{K} \mathbf{A} \mathbf{K}\mathbf{u} = \lambda \mathbf{K}\mathbf{u}\} \\
&= \lambda \mathbf{u}^* \mathbf{K}\mathbf{u} + \bar{\lambda} \mathbf{u}^* \mathbf{K}\mathbf{u} \\
&= 2\text{Re}(\lambda) \mathbf{u}^* \mathbf{K}\mathbf{u}
\end{aligned}$$

But $\mathbf{K}$ is positive definite and $\mathbf{u} \neq \mathbf{0}$ implies that $\mathbf{u}^* \mathbf{K}\mathbf{u} > 0$. Hence, from the inequality above, and this last observation, it follows that $\text{Re}(\lambda) < 0$. Since λ was arbitrary, it follows that every eigenvalue of $\mathbf{K}\mathbf{A}$ has negative real part, which is what we wanted to show. ■

2.4 BIBO STABILITY

Until now we have studied stability only with respect to the behavior of states $\mathbf{x}(t)$. This is referred to as *internal* stability. Now we will consider *external* stability, that is, stability related to the input-output behavior of the dynamical system, and not concerned directly with the stability of the state. Our formulation is essentially that given in Szidarovsky and Bahill ([SB], pp. 184–187).

In considering external stability, we will be interested in BIBO stability, which we define next.

2.34 Definition (BIBO stability) Suppose the initial state, $\mathbf{x}_0$, of a dynamical system is zero. Then the system is called **BIBO stable** if bounded input evokes a bounded output ([SB], p. 184). ■

We consider only continuous time (resp. discrete time) linear dynamical systems of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \text{ (resp. } \mathbf{x}(t+1) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t)); \quad (2.58)$$

$$\mathbf{y}(t) = \mathbf{C}(t)\mathbf{x}(t), \quad (2.59)$$

where we also assume $\mathbf{A}(t)$, $\mathbf{B}(t)$, and $\mathbf{C}(t)$ are continuous functions of t , and the input functions $\mathbf{u}(t)$ are piecewise continuous. Equation (2.59) is the output equation for the dynamical system in both cases. First we study the time variant case, and in Section 2.4.3 we will consider time invariant systems as a special case.

Recall that by Corollaries 1.18 (Section 1.1.3) and 1.33 (Section 1.2.2), input and output of a linear dynamical system of the form (2.58)–(2.59) were related by the equations

$$\mathbf{y}(t) = \mathbf{C}(t)\Phi(t, t_0)\mathbf{x}_0 + \int_{t_0}^t \mathbf{C}(t)\Phi(t, \tau)\mathbf{B}(\tau)\mathbf{u}(\tau)d\tau \quad (2.60)$$

in the continuous time case, and

$$\mathbf{y}(t) = \mathbf{C}(t)\Phi(t,0)\mathbf{x}_0 + \sum_{\tau=0}^{t-1} \mathbf{C}(t)\Phi(t,\tau+1)\mathbf{B}(\tau)\mathbf{u}(\tau) \quad (2.61)$$

in the discrete time case, where $\Phi(t,\tau)$ is the fundamental matrix defined in Definition 1.11 for the continuous case and in Definition 1.31 for the discrete case. In the definition of BIBO stability, we assumed that the initial state was zero, so equations (2.60)–(2.61) can be rewritten

$$\mathbf{y}(t) = \int_{t_0}^t \mathbf{T}(t,\tau)\mathbf{u}(\tau)d\tau, \quad (2.62)$$

where $\mathbf{T}(t,\tau) = \mathbf{C}(t)\Phi(t,\tau)\mathbf{B}(\tau)$, and

$$\mathbf{y}(t) = \sum_{\tau=0}^{t-1} \mathbf{T}(t,\tau)\mathbf{u}(\tau), \quad (2.63)$$

where $\mathbf{T}(t,\tau) = \mathbf{C}(t)\Phi(t,\tau+1)\mathbf{B}(\tau)$, respectively.

2.4.1 BIBO Stability for Continuous Time Linear Dynamical Systems

In this first subsection we give necessary and sufficient conditions for BIBO stability for the case of time variant linear dynamical systems. Then a corollary is presented that simplifies the conditions of the theorem.

2.35 Theorem Let $\mathbf{T}(t,\tau) = [t_{ij}(t,\tau)]$ be defined as in (2.58) and (2.59) for the continuous time case and fix t_0 . Then the time variant dynamical system given by equation (2.58), with output equation given by (2.62), is BIBO stable if and only if the integral

$$\int_{t_0}^t |t_{ij}(t,\tau)|d\tau \quad (2.64)$$

is bounded for all $t \geq t_0$, i , and j ([SB], p. 185).

Proof

($\Rightarrow$) Assume BIBO stability. Then there are constants K_{ij} such that for all $t \geq t_0$, i , and j

$$\int_{t_0}^t |t_{ij}(t, \tau)| d\tau \leq K_{ij},$$

with $\mathbf{u}(t) = [u_i(t)]$ and $\mathbf{y}(t) = [y_i(t)]$. There are also constants U_j such that

$$|u_j(t)| \leq U_j \text{ for all } t \geq t_0 \text{ and } j.$$

Then

$$\begin{aligned} y_i(t) &= \left| \int_{t_0}^t \sum_j t_{ij}(t, \tau) u_j(\tau) d\tau \right| \\ &\leq \int_{t_0}^t \sum_j |t_{ij}(t, \tau)| |u_j(\tau)| d\tau \\ &\leq \int_{t_0}^t \sum_j |t_{ij}(t, \tau)| U_j d\tau \\ &= \sum_j U_j \int_{t_0}^t |t_{ij}(t, \tau)| d\tau \\ &\leq \sum_j U_j K_{ij} \end{aligned}$$

Hence the output is bounded.

($\Leftarrow$) We use the contrapositive form. Assume that the integral in (2.64) is not bounded for some pair (i_0, j_0) . Then for any $N > 0$, there exists $t_N > t_0$ such that

$$\int_{t_0}^{t_N} |t_{i_0, j_0}(t, \tau)| d\tau > N$$

Next, we define an input function that will guarantee that the output is not bounded. Define

$$u_{j_0}(\tau) = \begin{cases} 1, & \text{if } t_{i_0, j_0}(t_N, \tau) \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

which is piecewise continuous because t_{i_0, j_0} is continuous. For $j \neq j_0$, define

$u_j(\tau) \equiv 0$, that is,

$$\mathbf{u}(\tau) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ u_{j_0}(\tau) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow j_0$$

To see that output is not bounded under the assumptions above, note that

$$\begin{aligned} y_{i_0}(t_N) &= \int_{t_0}^{t_N} \sum_j t_{i_0, j}(t_N, \tau) u_j(\tau) d\tau \\ &= \int_{t_0}^{t_N} t_{i_0, j_0}(t_N, \tau) u_{j_0}(\tau) d\tau \\ &= \int_{t_0}^{t_N} |t_{i_0, j_0}(t_N, \tau)| d\tau \\ &> N \end{aligned}$$

Hence, since $N > 0$ is arbitrary, output is not bounded, and the proof is complete. ■

Notice that to use this theorem we must first compute n^2 integrals. The following corollary can be used to establish BIBO stability by computing a single integral.

2.36 Corollary The dynamical system of the hypothesis of Theorem 2.35 is BIBO stable if and only if the (single) integral

$$I(t) = \int_{t_0}^t \sum_i \sum_j |t_{ij}(t, \tau)| d\tau \quad (2.65)$$

is bounded for $t \geq t_0$ ([SB], p. 186).

Proof

This is an immediate consequence of Theorem 2.35 and the fact that, since the integral is a linear operator and the terms are nonnegative, the integral (2.65) is bounded if and only if the integrals (2.64) are bounded. ■

2.4.2 BIBO Stability for Discrete-Time Linear Dynamical Systems

We now turn to the discrete case. The proofs here are similar to the continuous case, and for simplicity we will use the notation given in the previous section when appropriate.

2.37 Theorem Let $T(t, \tau) = [t_{ij}(t, \tau)]$ be defined as in (2.58) and (2.59) for the discrete time case. Then the system with output equation given by (2.63) is BIBO stable if and only if the sum

$$\sum_{\tau=0}^{t-1} |t_{ij}(t, \tau)| \quad (2.66)$$

is bounded for all $t \geq 1$, i , and j ([SB], p. 186).

Proof

($\Leftarrow$) We use the notation of the proof of Theorem 2.35 and modify as necessary. As in the proof of Theorem 2.35, assume that for all $t \geq 1$, i , and j ,

$$\sum_{\tau=0}^{t-1} |t_{ij}(t, \tau)| \leq K_{ij} \quad (2.67)$$

and input is bounded:

$$|u_j(t)| \leq U_j \text{ for all } t \geq 0, j$$

Then

$$\begin{aligned} |y_i(t)| &= \left| \sum_{\tau=0}^{t-1} \sum_j t_{ij}(t, \tau) u_j(\tau) \right| \\ &\leq \sum_{\tau=0}^{t-1} \sum_j |t_{ij}(t, \tau)| |u_j(\tau)| \\ &\leq \sum_{\tau=0}^{t-1} \sum_j K_{ij} U_j \end{aligned}$$

so that output is bounded.

($\Rightarrow$) Again, we prove the contrapositive form. Assume that not all of the sums (2.66) are bounded, so that there exists a pair (i_0, j_0) such that for all $N > 0$,

$$\sum_{\tau=0}^{t_N-1} |t_{i_0, j_0}(t, \tau)| > N$$

for some $t_N > 0$. Define the input function $u(t)$ exactly as we did in the proof of Theorem 2.35. Then

$$\begin{aligned} y_{i_0}(t_N) &= \sum_{\tau=0}^{t_N-1} \sum_j t_{i_0, j}(t_N, \tau) u_j(\tau) \\ &= \sum_{\tau=0}^{t_N-1} t_{i_0, j_0}(t_N, \tau) u_{j_0}(\tau) \\ &= \sum_{\tau=0}^{t_N-1} |t_{i_0, j_0}(t_N, \tau)| \\ &> N \end{aligned}$$

Therefore, the output is not bounded, which completes the proof. ■

Again, we can avoid computing the n^2 sums in the hypothesis of Theorem 2.37 by computing the single sum given in the next corollary.

2.38 Corollary The discrete dynamical system given in the hypothesis of Theorem 2.37 is BIBO stable if and only if the (single) sum

$$I(t) = \sum_{\tau=0}^{t-1} \left(\sum_i \sum_j |t_{ij}(t, \tau)| \right) \quad (2.68)$$

is bounded. ([SB], p. 186) ■

2.4.3 BIBO Stability for Time Invariant Dynamical Systems

Finally, we consider the special case that the matrices $A(t)$, $B(t)$, and $C(t)$ described at the beginning of Section 2.4 are constant.

Theorem 2.39 The time invariant continuous (resp. discrete) dynamical system given by equations (2.61)–(2.62), with constant matrices $A(t) \equiv A$, $B(t) \equiv B$, and $C(t) \equiv C$, is BIBO stable, if for every eigenvalue λ of A , $\text{Re}(\lambda) < 0$ (resp. $|\lambda| < 1$) ([SB], p. 187).

Proof

By Theorem 2.35 (resp. 2.37) it suffices to show that the integrals (resp. sums) (2.64) (resp. (2.66)) are bounded for all $t \geq t_0$ (resp. $t \geq 1$), i , and j . However, in view of the fact that the matrices $B(t)$ and $C(t)$ in the definition of $T(t, \tau)$ are constant, it suffices to show that the matrix $\Phi(t, \tau)$ is bounded for all $t, \tau \geq t_0$ (resp. $t, \tau \geq 0$). It follows from the proof of Theorem 2.25 that this occurs if $\text{Re}(\lambda) < 0$ (resp. $|\lambda| < 1$) for each eigenvalue λ of the matrix A . ■

Remark

The conditions on the eigenvalues of the matrix A of Theorem 2.39 can be weakened to $\operatorname{Re}(\lambda) \leq 0$ (resp. $|\lambda| \leq 1$) if for each eigenvalue with real part equal to zero (resp. absolute value equal to one) we know, in addition, that either (i) the multiplicity of the eigenvalue is one, or (ii) if the multiplicity of the eigenvalue is greater than one, then every Jordan block in the Jordan Canonical form of A associated with this eigenvalue has size equal to one.

2.5 GENERALIZING THE CONCEPT OF STABILITY

In this last section of Chapter 2, we consider stability in the more general context of deterministic systems, as given by Definition 1.46 in Section 1.4. Recall that the state transition function was a function $\phi: T \times T \times X \times \Omega \rightarrow X$, where X is the state space, T is the time set, and Ω is the set of admissible input functions. Our definition of an equilibrium state for deterministic systems is due to Padulo and Arbib. In contrast to Definition 2.1 in Section 2.1, we assume here that input is zero, that is, $u(t) \equiv \theta$.

2.40 Definition (Equilibrium state for a deterministic system) Suppose we are given an arbitrary deterministic system according to Definition 1.46 with state transition function $\phi(t, t_0, x_0, u)$, and suppose the input to the system is zero, i.e., $u(t) \equiv \theta$. Then x_e is an **equilibrium state** if and only if

$$\phi(t, t_0, x_e, \theta) = x_e \text{ for all } t \geq t_0 \quad (2.69)$$

([PA], p. 572) ■

This definition does not assume that the dynamical system is given by a differential equation or a difference equation, as was assumed in Definition 2.1, for the continuous and discrete cases, respectively. However, Definition 2.40 does express that the deterministic system is not changing with respect to time at equilibrium.

Recall that in Definition 2.1, the equilibrium state was defined in terms of conditions on the differential equation and difference equation, in the continuous and discrete cases, respectively. Then we argued that these conditions implied the dynamical system was at rest at the equilibrium state for all future time. The question here is, if the deterministic system in Definition 2.40 can be described by

a differential or difference equation, then does equation (2.69) imply that the conditions given in Definition 2.1 hold?

To see that the answer to the question above is yes, we begin by simplifying notation as we did in Section 1.4; that is, with fixed zero input, initial time t_0 , and initial state $\mathbf{x}_0$, let $\mathbf{x}(t) = \phi(t, t_0, \mathbf{x}_0, \theta)$ ($\mathbf{x}_0 = \mathbf{x}(t_0)$) be the resulting state transition function under these fixed conditions. Our formulation here follows Padulo and Arbib ([PA], p. 572).

For the discrete case, where

$$\mathbf{x}(t+1) = \phi(t+1, t_0, \mathbf{x}(t_0), \theta) \text{ for all } t \geq t_0 \in \mathbb{Z} \text{ (integers),}$$

if $\mathbf{x}_e$ is an equilibrium state, then with $\mathbf{x}(t_0) = \mathbf{x}_e$ we have

$$\mathbf{x}(t+1) = \phi(t+1, t_0, \mathbf{x}_e, \theta) = \mathbf{x}_e \text{ for all } t \geq t_0$$

This last equation agrees with equation (2.2) for the special case of $T = \mathbb{N}$.

Turning to the continuous case, where

$$\mathbf{x}(t) = \phi(t, t_0, \mathbf{x}(t_0), \theta)$$

and

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \theta)$$

if $\mathbf{x}_e$ is an equilibrium state, then

$$\dot{\mathbf{x}}(t) = \frac{d}{dt} \phi(t, t_0, \mathbf{x}_e, \theta) = \frac{d}{dt} \mathbf{x}_e = \theta$$

whence $\mathbf{f}(t, \mathbf{x}_e, \theta) = \theta$ for all $t \geq t_0$, which agrees with equation (2.1) for the special case of zero input, i.e., $\mathbf{u}(t) \equiv \theta$.

Continuing with the use of the notation $\mathbf{x}(t) = \phi(t, t_0, \mathbf{x}_0, \theta)$, Padulo and Arbib give similar definitions to those given in Section 2.2 ([PA], pp. 574–575), so that the generalized types of stability are defined exactly as they were in Section 2.2.

Finally, an identical definition to Definition 2.9 in Section 2.2.2, of a Liapunov function for $\mathbf{x}_e$, can be given for deterministic systems if we continue with the notation $\mathbf{x}(t) = \phi(t, t_0, \mathbf{x}_0, \theta)$. However, the most practical way of checking that

$V(\mathbf{x}(t))$ is nonincreasing in t is to check similar conditions to those given in Definitions 2.10C or 2.14C, for the case of a continuous or discrete dynamical system given by a differential or difference equation, respectively. Hence, in the literature, it is usually assumed that the deterministic system is given by a differential or difference equation. However, in Sections 1.4.3 and 1.4.4 we considered differential and difference equations in more general vector spaces than $\mathfrak{R}^n$, so that a more general Liapunov theory is possible (see [PA], pp. 588–590 for more).

CHAPTER 3

EXAMPLES FROM THE LITERATURE

3.0 ABOUT THIS CHAPTER

In this third and final chapter, our purpose is twofold: (i) in Section 3.1, we perform an in-depth stability analysis of a linear model for eye movement that appeared in a paper written in 1980, and (ii) in Section 3.2, we present a very brief overview of some ideas in recent chaos theory which relate to dynamical systems.

The eye movement model is a classical linear time invariant system, and we will draw upon the theory developed in Section 2.3 when we consider the stability of its equilibrium state.

Much of recent research in systems theory over the last 25 years pertains to chaos theory, so it seems appropriate to discuss the essential ideas. As we will see, forming a discrete time sample from simple nonlinear dynamical systems

can lead to very irregular “chaotic” behavior and to very elegant mathematics created to describe this behavior. Indeed, we will see that the discrete one-dimensional dynamical system given by $x(t+1) = x(t)^2 - 2$ can give rise to very complicated dynamical behavior.

3.1 STABILITY ANALYSIS: A LINEAR MODEL FOR HUMAN MOVEMENT

In this section we present a time invariant continuous linear dynamical system that models horizontal eye movement. The basis of our work here is a paper by Bahill, Latimer, and Troost that appeared in the *IEEE Transactions on Biomedical Engineering* in 1980 [BLT]. The authors point out that while the focus of the paper was to model the jerky movements that occur when we scan our environment or read, which are called *saccades*, the model is applicable to other physiological movement systems whose control is governed by neurological input, by changing the parameters of the model. After developing the linear model, we will analyze its stability in terms of the parameters of the dynamical system.

3.1.1 Derivation of the Linear Model

The purpose of this subsection is to retrace the key steps made by the authors in deriving their linear model. They call their model *homeomorphic*, because there is a one-to-one correspondence between components of the model and elements or effects in the physiological system ([SB], p. 164). For example, muscle elasticities are modeled by springs, and muscle forces are modeled by force generators. The model is sufficiently complicated that we will first describe each major component of the model and then present the completed version.

Data collected by measurements on animal and human test subjects has revealed nonlinearities in length-tension and force-velocity relationships for the eye muscle. Nevertheless, most attempts to model such a system linearize it in some way. The authors of [BLT] discovered a new linear relationship between two quantities, over specified ranges. In particular, they linearize the relationship

between eye position and the muscle force required to hold the eye in a particular position and the force-velocity relationship for the eye muscle, for a certain specific range of muscle stimulations.

A *saccadic* movement of the eye occurs when the eye reacts quickly to brain input, for example, when we suddenly fix on different elements in a scene. It is well known that such motion is quick and jerky, as opposed to what is called a tracking motion of the eye, which is given, for example, by eye movement related to following a finger move slowly back and forth. Watching someone else's eye during these two types of eye movements will readily illustrate their differences.

For saccades within $\pm 20^\circ$ of a straight-ahead eye position, the model was found to simulate eye movement well, in the sense that the mean squared error between model saccades and human saccades was of the order 10^{-6} deg^2 .

Consider now the motor control system for the left eye. It is comprised of essentially two identical muscle systems, each with parameters inherent to its particular function in the larger system controlling saccadic eye movement. For an eye movement to the right (towards the nose), there is a muscle system which pulls the eye right, which will be called the agonist muscle (AG) and one which applies tension to the left, which will be called antagonist muscle (ANT). We will begin with an analysis of a simplified form of one such system and then consider the eye movement system as a whole. To account for passive elasticity of the eye muscle, an additional system will be added to the final version of the model.

Before describing the various parts of the model below, we state formally that all forces will be given in Newtons, distances in meters, time intervals in seconds, and eye position in degrees. The authors use two systems of units, but since we

are concerned here with the mathematics of the model, our assumption will simplify the presentation.

3.1.2 Model of the Agonist Muscle

The linearized length-tension relationship for the agonist muscle is illustrated in Figure 3.1. We will modify this figure later to include components related to the force-velocity relationship, but for now we neglect this relationship.

Figure 3.1 A simplified mechanical model of a human muscle

The various components in Figure 3.1 are given as follows:

- F = force generator, which assumes fixed values
- K_{SE} = series elasticity spring constant
- K_{LT} = length-tension spring constant
- L' = fixed reference length
- L = muscle length
- T = tension in the muscle = force exerted on the eye by the muscle

Using static equations for equilibrium, one can derive the following linear relationship between muscle tension T and muscle length L :

$$T = \frac{K_{LT}K_{SE}}{K_{LT} + K_{SE}}L + \frac{K_{SE}F}{K_{LT} + K_{SE}} \quad (3.1)$$

Let

$$K' = \frac{K_{LT}K_{SE}}{K_{LT} + K_{SE}} \quad (3.2)$$

denote the slope of this line. The value of K_{SE} was computed earlier by another author to be approximately 125 N/m. Human physiological data from two other studies was used to estimate a value for K' , obtaining $K' \approx 40 \text{ N/m}$. These two values can be used to compute K_{LT} using (3.2):

$$K_{LT} = \frac{K_{SE}K'}{K_{SE} - K'} \approx 60 \frac{\text{N}}{\text{m}} \quad (3.3)$$

The eyeball is assumed to be a spherical region having radius equal to 11mm. When the eye is in the straight-ahead position, $L = 0$ and the angular displacement from the straight-ahead position is $\phi = 0^\circ$. When a neural impulse from the brain causes the agonist muscle to contract a distance of $-L$ meters, this corresponds to a positive displacement for ϕ . The relationship between L and ϕ is given in Figure 3.2.

Figure 3.2 Eye displacement diagram for an agonist muscle contraction

The force T required to hold the eye in a position of ϕ degrees is defined to be the value T_{LOC} . Using the linearization, and again analyzing equilibrium, results in the following relationship between T_{LOC} and the desired final eye position ϕ :

$$T_{LOC} = 14 + 0.8\phi \quad (\phi \text{ in degrees}) \quad (3.4)$$

For a displacement of ϕ degrees, the muscle tension axis (i.e., T -axis) intercept, T_{INT} , is obtained from the linearization to be

$$\begin{aligned} T_{INT} &= T_{LOC} + K'\phi \\ &= 14 + 1.6\phi \end{aligned} \quad (3.5)$$

Using equation (3.1) T_{INT} is also calculated by letting $L = 0$:

$$T_{INT} = \frac{K_{SE}F}{K_{LT} + K_{SE}} \quad (3.6)$$

In any fixed position ϕ , the eye is assumed to be in a steady-state condition, and the force required for the agonist force generator F to hold the eye in this position is defined to be $N_{AG-STEP}$. Setting $F = N_{AG-STEP}$ in (3.6) and combining (3.5) and (3.6) gives

$$N_{AG-STEP} = (0.2 + 0.02\phi)N \quad (3.7)$$

3.1.3 Model of the Antagonist Muscle

The entire derivation above can be repeated for the antagonist muscle, and the following equations result:

$$\begin{aligned} T_{LOC} &= 14 + 0.3\phi \\ T_{INT} &= T_{LOC} - 0.8\phi = 14 - 0.5\phi \\ N_{ANT-STEP} &= (0.2 - 0.007\phi)N \end{aligned} \quad (3.8)$$

Notice the sign change that occurs in the equations (3.8) as compared to the equations (3.7) for the agonist muscle. This is because positive values of ϕ correspond to an agonist contraction; hence when an antagonist contraction occurs, the eye displacement is $-\phi$. Also, we have assumed that spring tensions in the agonist and antagonist muscles are identical, i.e., $K_{SE} = K_{SE-AG} = K_{SE-ANT}$ and $K_{LT} = K_{LT-AG} = K_{LT-ANT}$.

3.1.4 Extending the Model

Next we consider the force-velocity linearization and modify the model derived so far to accommodate this additional aspect of the system. The authors of [BLT] report that experimental data showed that the maximum velocity achievable by a muscle depends on the amount of stimulation of the muscle. Hence for each level of muscle stimulation, a corresponding force-velocity curve for the muscle was created. A single line was used to approximate the force-velocity relationship for the agonist muscle, corresponding to each level of muscle stimulation; moreover, the slopes of the linearizations were identical, which simplified the model. For a stimulation of the antagonist muscle, the resulting force-velocity relationship could not be modeled by a single line, and this complicated the model.

However, when the agonist muscle is active, the corresponding force applied by the antagonist is less than 2% of maximum muscle force, and the authors of [BLT] discovered that a single line could be used to approximate the force-velocity relationship for the antagonist muscle as well. This greatly simplified the model. They saw from experimental data that the slopes for the lines in the force-velocity linearization of the agonist and antagonist muscles are given by

$$B_{AG} = 2.36 \text{N} \frac{\text{s}}{\text{m}} \quad (3.9)$$

$$B_{ANT} = 1.12 \text{N} \frac{\text{s}}{\text{m}} \quad (3.10)$$

The force-velocity lines show that muscle force is greatest at lower velocities, and to account for this, a *dashpot*, i.e., a muscle dampening mechanism which is velocity dependent, is added to our model. Our improved muscle model that includes this dashpot is given in Figure 3.3.

Figure 3.3 Agonist muscle model with tension-dampening dashpot

This new illustration also includes features to account for neural impulses from the brain, in the form of τ_{AG} and N_{AG} . The symbols for the agonist muscle modeled in Figure 3.3 have the following meanings:

- B_{AG} = slope of agonist force-velocity line (will be used as a tension-dampening dashpot below)
- F_{AG} = agonist muscle force generator (assumes fixed values)
- K_{SE} = series elasticity spring constant

K_{LT} = length-tension spring constant

ϕ_1 = eye position

ϕ_2 = position of agonist node (i.e., muscle connection to the eye)

N_{AG} = force required for steady-state behavior of the agonist muscle

T_{AG} = force exerted on the eye by the agonist muscle

τ_{AG} = agonist muscle activation (AC)/deactivation (DE) time constant

A similar diagram can be given for the antagonist muscle, where the letters “AG” above become “ANT” in the antagonist muscle diagram.

Using the notation above, it follows from Figure 3.2 that $\phi_1 = -L$; substitution into equation (3.1) then gives

$$T_{AG} = \frac{K_{SE}F_{AG}}{K_{LT} + K_{SE}} - \frac{K_{LT}K_{SE}}{K_{LT} + K_{SE}}\phi_1 \quad (3.11)$$

To account for the fact that muscle tension is reduced by the velocity sensitive dashpot, the term $-B_{AG}\dot{\phi}_2$ is added to equation (3.11), yielding

$$T_{AG} = \frac{K_{SE}F_{AG}}{K_{LT} + K_{SE}} - \frac{K_{LT}K_{SE}}{K_{LT} + K_{SE}}\phi_1 - B_{AG}\dot{\phi}_2 \quad (3.12)$$

where T_{AG} is the muscle force that is applied to the eyeball. Similarly, T_{ANT} is given by the equation

$$T_{ANT} = \frac{K_{SE}F_{ANT}}{K_{LT} + K_{SE}} - \frac{K_{LT}K_{SE}}{K_{LT} + K_{SE}}\phi_1 + B_{ANT}\dot{\phi}_3 \quad (3.13)$$

The corresponding sign changes that occur between equations (3.12) and (3.13) are due to the fact that the antagonist dashpot B_{ANT} adds to the resistive force of the antagonist active state tension T_{ANT} . Observe that as muscle length increases the length-tension relationship dictates that increased muscle force will be

needed to continue stretching the muscle, which adds to the antagonist resistive force. An additional dashpot B_p and spring constant K_p are added to account for passive elasticities of the eye muscles. Figure 3.4 illustrates the completed model.

Figure 3.4 The linear homeomorphic model for eye movement

To get equations for the complete model, begin by assuming that

$$K_{SE} = K_{SE-AG} = K_{SE-ANT} \text{ and } K_{LT} = K_{LT-AG} = K_{LT-ANT}$$

The muscle forces pulling in opposite directions on the eyeball are given to be

$$T_{AG} = K_{SE}(\phi_2 - \phi_1) \quad (3.14)$$

$$T_{ANT} = K_{SE}(\phi_1 - \phi_3) \quad (3.15)$$

Setting equations (3.12) and (3.14) equal, we obtain

$$\frac{K_{SE}F_{AG}}{K_{LT} + K_{SE}} - \frac{K_{LT}K_{SE}}{K_{LT} + K_{SE}}\phi_1 - B_{AG}\dot{\phi}_2 = K_{SE}(\phi_2 - \phi_1) \quad (3.16)$$

Similarly, setting (3.13) and (3.15) equal gives

$$\frac{K_{SE}F_{ANT}}{K_{LT} + K_{SE}} - \frac{K_{LT}K_{SE}}{K_{LT} + K_{SE}}\phi_1 + B_{ANT}\dot{\phi}_3 = K_{SE}(\phi_1 - \phi_3) \quad (3.17)$$

An additional equation that accounts for the passive elasticity modeled by the values K_p and B_p is given to be

$$K_{SE}(\phi_2 - \phi_1) - K_{SE}(\phi_1 - \phi_3) = K_p\phi_1 + B_p\dot{\phi}_1 + J\ddot{\phi}_1 \quad (3.18)$$

The value J in this equation is inertial mass of the eyeball and is given to be $2.2 \times 10^{-3} \text{ N} \cdot \frac{\text{s}}{\text{m}^2}$.

We are now ready to write differential equations that will describe this model as a time invariant linear dynamical system

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \quad (3.19)$$

where $\mathbf{A}$ and $\mathbf{B}$ are constant matrices and input $\mathbf{u}(t)$ is constant. In terms of the variables given in the homeomorphic model (Figure 3.4), define the components of $\mathbf{x}(t)$ to be

$$\begin{aligned} x_1 &= \phi_1 && \text{position of the eye} \\ x_2 &= \phi_2 && \text{position of the agonist node} \\ x_3 &= \phi_3 && \text{position of the antagonist node} \\ x_4 &= \dot{\phi}_1 && \text{eye velocity} \\ x_5 &= F_{AG} && \text{agonist active state tension} \\ x_6 &= F_{ANT} && \text{antagonist active state tension} \end{aligned}$$

Observe first that according to the definitions above,

$$\dot{x}_1 = x_4 \quad (3.20)$$

This is the first differential equation of the system. Equation (3.16) is solved for $\dot{\phi}_2 = \dot{x}_2$, and in terms of the state variables above we obtain

$$\dot{x}_2 = \frac{K_{SE}^2}{(K_{LT} + K_{SE})B_{AG}}x_1 - \frac{K_{SE}}{B_{AG}}x_2 + \frac{K_{SE}}{(K_{LT} + K_{SE})B_{AG}}x_5 \quad (3.21)$$

Similarly, (3.17) is solved for $\dot{\phi}_3 = \dot{x}_3$:

$$\dot{x}_3 = \frac{K_{SE}^2}{(K_{LT} + K_{SE})B_{ANT}} x_1 - \frac{K_{SE}}{B_{ANT}} x_3 - \frac{K_{SE}}{(K_{LT} + K_{SE})B_{ANT}} x_6 \quad (3.22)$$

Solving (3.18) in terms of $\ddot{\phi}_1 = \dot{x}_4$ we obtain

$$\dot{x}_4 = \frac{-2K_{SE} - K_p}{J} x_1 + \frac{K_{SE}}{J} x_2 + \frac{K_{SE}}{J} x_3 - \frac{B_p}{J} x_4 \quad (3.23)$$

Two additional equations are assumed, which complete the description of this dynamical system:

$$\dot{x}_5 = \frac{N_{AG} - x_5}{\tau_{AG}} \quad (3.24)$$

and

$$\dot{x}_6 = \frac{N_{ANT} - x_6}{\tau_{ANT}} \quad (3.25)$$

Equations (3.20)–(3.25) can be written in the desired matrix form

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

where

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ \frac{K_{SE}}{B_{AG}} & -\frac{K_{LT}+K_{SE}}{B_{AG}} & 0 & 0 & \frac{1}{B_{AG}} & 0 \\ \frac{K_{SE}}{B_{ANT}} & 0 & -\frac{K_{LT}+K_{SE}}{B_{ANT}} & 0 & 0 & \frac{-1}{B_{ANT}} \\ \frac{-2K_{SE}-K_p}{J} & \frac{K_{SE}}{J} & \frac{K_{SE}}{J} & \frac{B_p}{J} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{-1}{\tau_{AG}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{-1}{\tau_{ANT}} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \frac{1}{\tau_{AG}} & 0 \\ 0 & \frac{1}{\tau_{ANT}} \end{bmatrix}$$

and

$$\mathbf{u}(t) = \begin{bmatrix} N_{AG} \\ N_{ANT} \end{bmatrix}$$

The initial state of the system, $\mathbf{x}_0 = [x_1(0) \ x_2(0) \ \dots \ x_6(0)]^T$, is given by

$$x_1(0) = x_4(0) = 0 \text{ (so } \phi = 0 \text{ initially)}$$

$$x_2(0) = -x_3(0) = 1.1\text{mm} = 5.6^\circ \text{ (agonist/antagonist node position)}$$

$$x_5(0) = x_6(0) = 0.2\text{N}$$

We will analyze the stability of this system in Section 3.1.6, but first we consider input to this system in detail in the next subsection.

3.1.5 A Discussion of Input to the Model

In this section we consider input to the model, which is comprised of two different phases: activation (AC) and deactivation (DE) of the agonist muscle. We also summarize the parameter values for the homeomorphic eye movement model in a table given at the end of this section.

The basic idea is that for a given saccade, the eye must move from its initial position at the beginning of the saccade to a final position according to a given saccade size of $\Delta\phi$. This process is accomplished by activating the agonist muscle, which pulls the eye right, and then deactivating the agonist muscle when the eye reaches its final position. At the same time, the antagonist first releases tension to allow the eye to move right and then applies just enough tension to counterbalance the tension exerted by the agonist, which holds the eye in its final position until new input is received from the brain.

As the agonist muscle begins to pull the eye right, the antagonist aids this process by decreasing the force applied to the eye by the antagonist active state tension generator. In the first phase, the agonist muscle is activated by brain input in the form of $N_{AG} = N_{AG-PULSE}$, which is converted into active state tension F_{AG} according to differential equation (3.24). Concomitant with this pulse is a deactivation (DE) of the antagonist muscle by brain input in the form $N_{ANT} = N_{ANT-PULSE}$, which is converted into active state tension F_{ANT} as described by differential equation (3.25). The relevant time constants for this process are $\tau_{AG} = \tau_{AG-AC}$ and $\tau_{ANT} = \tau_{ANT-DE}$.

For the second phase, the eye has reached its final position and the agonist muscle is deactivated through brain input in the form $N_{AG} = N_{AG-STEP}$ and the antagonist muscle is activated to counterbalance the active state tension in the

agonist muscle, by brain input in the form $N_{ANT} = N_{ANT-PULSE}$. The relevant time constants in this phase of input to the system are $\tau_{AG} = \tau_{AG-DE}$ and $\tau_{ANT} = \tau_{ANT-AC}$.

The use of the terms “step” and “pulse” can be readily understood by inspection of the graphs below in Figure 3.5, which describe the processes given above.

Figure 3.5 Input to the eye movement model

Finally, observe that we are actually describing two simpler, linked time invariant linear dynamical systems when we write $\dot{x}(t) = Ax(t) + Bu(t)$: one for the activation of the agonist muscle and one for the deactivation of the agonist muscle. The particular system is obtained by using the appropriate values of N_{AG} , N_{ANT} , τ_{AG} , and τ_{ANT} as described above.

To complete our work in this section, we provide the numerical values used by the authors of [BLT] for the model. We will use these values in the next section when we consider stability of the model.

Table 3.1 Values for parameters given in the model

Parameter	Value	Parameter	Value
K_{SE}	125 N/m	$N_{ANT-PULSE}$	0.005 N
K_{LT}	60 N/m	$N_{AG-STEP}$	$(0.2 + 0.02\phi)N$
K_p	25 N/m	$N_{ANT-STEP}$	$(0.2 - 0.007\phi)N$
B_p	$3.1 \text{ N} \cdot \frac{s}{m}$	τ_{AG-AC}	$(11.7 - 0.2\Delta\phi)ms$
B_{AG}	$2.36 \text{ N} \cdot \frac{s}{m}$	τ_{AG-DE}	0.2 ms
B_{ANT}	$1.12 \text{ N} \cdot \frac{s}{m}$	τ_{ANT-AC}	2.4 ms
J	$2.2 \times 10^{-3} \text{ N} \cdot \frac{s^2}{m}$	τ_{ANT-DE}	1.9 ms
$N_{AG-PULSE}$	$(0.54 + 0.11\Delta\phi)N$ *		

* for $\Delta\phi \leq 11^\circ$ (see [BLT] for values outside this range)

3.1.6 Stability Analysis of the Model

Recall that the model above takes the form of a continuous time invariant linear dynamical system

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

where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{u}(t) \equiv \bar{\mathbf{u}}$ are constant. The first step in the stability analysis of the model is to compute the equilibrium state, that is, to solve the equation

$$\mathbf{A}\mathbf{x}_e + \mathbf{B}\bar{\mathbf{u}} = \mathbf{0} \quad (3.26)$$

for $\mathbf{x}_e$. If $\mathbf{A}$ is invertible, then this equilibrium state is uniquely determined by the equation

$$\mathbf{x}_e = -\mathbf{A}^{-1}\mathbf{B}\bar{\mathbf{u}} \quad (3.27)$$

We could evaluate the constants in $\mathbf{A}$, $\mathbf{B}$, and $\bar{\mathbf{u}}$ using values from Table 3.1 and compute the determinant of $\mathbf{A}$, $\det(\mathbf{A})$, numerically; however, this would only tell us about this particular model and the determinant computed would only be approximately correct. A symbolic computation of $\det(\mathbf{A})$ is not too difficult and yields:

$$\det(\mathbf{A}) = \frac{-K_{SE}^2}{\tau_{AG}\tau_{ANT}B_{AG}B_{ANT}J(K_{LT} + K_{SE})} \left\{ 4K_{SE}^2 + 2K_{SE}K_{LT} + K_pK_{LT} + K_pK_{SE} \right\} \quad (3.28)$$

Since the physical context suggests that all of the constants in (3.28) will be positive for any neuromuscular system that is modeled using this model—and is certainly the case for the eye movement model—we see immediately that $\det(\mathbf{A}) < 0$ under these assumptions, so that $\mathbf{A}$ is invertible and the equilibrium state for this system will always exist and be uniquely determined by equation (3.27).

Now that we know there is a unique solution to (3.26), we can work with the entries of equation (3.26) (see equations (3.20)–(3.25)) and perform a little algebra to find $\mathbf{x}_e$. Let $\mathbf{x}_e = [x_1^e \ x_2^e \ \dots \ x_6^e]^T$ and observe that the resulting equation in the first component of (3.26) is

$$x_4^e = 0 \quad (3.29)$$

Recalling that x_4 is eye velocity, we see that at equilibrium the eye is not moving, which is what we would expect. Also the equations in the fifth and sixth entries are

$$\frac{N_{AG} - x_5^e}{\tau_{AG}} = 0 \Leftrightarrow x_5^e = N_{AG} \quad (3.30)$$

and

$$\frac{N_{ANT} - x_6^e}{\tau_{ANT}} = 0 \Leftrightarrow x_6^e = N_{ANT} \quad (3.31)$$

Equation (3.30) says that the agonist active state tension generator F_{AG} must apply the force required to hold the eye in a steady-state position, i.e., N_{AG} , at equilibrium, and (3.31) says this, similarly, for the antagonist tension generator.

The equations in the second, third, and fourth components of equation (3.26) can be all solved for x_1^e , respectively, and the resulting equations are

$$x_1^e = \frac{(K_{LT} + K_{SE})x_2^e - N_{AG}}{K_{SE}} \quad (3.32)$$

$$x_1^e = \frac{(K_{LT} + K_{SE})x_3^e + N_{ANT}}{K_{SE}} \quad (3.33)$$

$$x_1^e = \frac{x_2^e K_{SE}}{2K_{SE} - K_p} \quad (3.34)$$

Equate equations (3.32) and (3.34) and solve for x_2^e :

$$x_2^e = \frac{-N_{AG}(2K_{SE} - K_p)}{K_{SE}^2 - (2K_{SE} - K_p)(K_{LT} + K_{SE})} \quad (3.35)$$

Since the entries on the right side of equation (3.35) are all fixed constants, this equation uniquely determines x_2^e —the equilibrium position of the agonist node. By equating (3.33) and (3.34) we obtain the equilibrium position of the antagonist node, x_3^e :

$$x_3^e = \frac{x_2^e K_{SE}^2 - N_{ANT}(2K_{SE} - K_p)}{(2K_{SE} - K_p)(K_{LT} + K_{SE})} \quad (3.36)$$

Since x_2^e is uniquely determined by (3.35) this value is uniquely determined by equation (3.36), and similarly, equilibrium eye position, x_1^e , is given from (3.34) to be

$$x_1^e = \frac{x_2^e K_{SE}}{2K_{SE} - K_p}$$

During activation of the agonist muscle for a ten degree saccade (i.e., $\Delta\phi = 10$), the computed values based on the values given in Table 3.1 are

$$\begin{aligned} x_2^e &= \frac{-N_{AG-PULSE}(2K_{SE} - K_p)}{K_{SE}^2 - (2K_{SE} - K_p)(K_{LT} + K_{SE})} \\ &= \frac{-(0.54 + 0.11(10))(2 \cdot 125 - 14)}{125^2 - 236 \cdot 185} \\ &\approx 13.8mm \\ x_1^e &= \frac{x_2^e K_{SE}}{2K_{SE} - K_p} \approx 7.31mm \end{aligned}$$

and

$$x_3^e = \frac{x_2^e K_{SE}^2 - N_{ANT-PULSE}(2K_{SE} - K_p)}{(2K_{SE} - K_p)(K_{LT} + K_{SE})} \approx 4.91mm$$

Notice first that since the agonist's initial position was 1.1mm, we see that the agonist muscle has to move 12.7mm to move the eye 7.31mm to its equilibrium position. Of course, here we are assuming that the agonist remains in its activation state for all future time and hence equilibrium is occurring when the eye cannot move any further, which in this case occurs for about a 38 degree

horizontal movement from a straight ahead stare. It is interesting to note that the model is smart enough to know that the eye cannot keep moving and that there is a physical restriction on how far right (or left) that the eye can move. The third point of interest here is that initially $|x_2(0)| = |x_3(0)| = 1.1mm$, that is, the agonist and antagonist nodes are in the same relative positions, but by the time equilibrium occurs, the agonist muscle has moved more than twice the distance the antagonist muscle has moved.

The deactivation phase for a 10 degree saccade results in the following equilibrium values:

$$\begin{aligned}
 x_2^e &= \frac{-N_{AG-STEP}(2K_{SE} - K_p)}{K_{SE}^2 - (2K_{SE} - K_p)(K_{LT} + K_{SE})} \\
 &= \frac{-(0.2 + 0.02(10))(2 \cdot 125 - 14)}{125^2 - 236 \cdot 185} \\
 &\approx 3.37mm \\
 x_1^e &= \frac{x_2^e K_{SE}}{2K_{SE} - K_p} \approx 1.78mm
 \end{aligned}$$

and

$$\begin{aligned}
 x_3^e &= \frac{x_2^e K_{SE}^2 - N_{ANT-STEP}(2K_{SE} - K_p)}{(2K_{SE} - K_p)(K_{LT} + K_{SE})} \\
 &= \frac{x_2^e 125^2 - (0.2 - 0.007(10))236}{236 \cdot 185} \\
 &\approx 0.50mm
 \end{aligned}$$

We see that if the agonist muscle remains in the deactivation phase, the eye will end up in an equilibrium position of 1.78mm or about 9 degrees left (or right) of a straight ahead stare in the long run.

It should be pointed out that the equilibrium state considered above is temporary in that usually the agonist does not remain in one particular state, i.e., activation or deactivation, for any sustained period of time.

The obvious next question is what kind of equilibrium is this, that is, is the equilibrium state stable? Asymptotically stable? Unstable? Intuitively, based on our own experience, we expect the equilibrium state to be asymptotically stable, i.e., in the case of activation of the agonist muscle, no matter what position the eye is initially in, the agonist muscle will pull the eye to the equilibrium position and hold it there until new input from the brain is received. The question, then, is: Does the model agree with our intuition?

Based on our work in Section 2.3.2, we should check whether all eigenvalues of the matrix A have negative real parts; if so, the equilibrium is asymptotically stable. We will first show that all the coefficients of the characteristic polynomial of A are positive. Recall that by Theorem 2.29, this is a necessary condition for the eigenvalues of A to have negative real parts. Again, we will work with the symbolic constants rather than their particular numerical values for this particular model.

Observe that the characteristic polynomial of A , $p_A(\lambda)$, takes the particular form

$$p_A(\lambda) = \left(\lambda + \frac{1}{\tau_{AG}}\right)\left(\lambda + \frac{1}{\tau_{ANT}}\right)f(\lambda)$$

where $f(\lambda) = \lambda^4 + m_3\lambda^3 + m_2\lambda^2 + m_1\lambda + m_0$ and each m_i is given below. We claim that each m_i is positive. As an overriding assumption, it will be assumed that all symbolic constants below are positive—which is the case for the eye movement model.

To establish the claim we consider each coefficient individually. Evaluating $f(\lambda)$ symbolically yields

$$m_3 = \frac{K_{SE}}{B_{AG}} + \frac{K_{SE}}{B_{ANT}} + \frac{B_p}{J}, \quad (3.37)$$

which is clearly positive under the assumption above. Also,

$$m_2 = \frac{K_{SE}^2}{B_{AG}B_{ANT}} + \frac{2K_{SE} + K_p}{J} + \frac{K_{SE}B_p}{B_{AG}J} + \frac{K_{SE}B_p}{B_{ANT}J} > 0 \quad (3.38)$$

and

$$m_1 = \frac{-B_{ANT}K_{SE}^3 + K_{SE}(2K_{SE} + K_p)(K_{LT} + K_{SE})(B_{AG} + B_{ANT}) - K_{SE}^3B_{AG}}{(K_{LT} + K_{SE})B_{AG}B_{ANT}J} + \frac{K_{SE}^2B_p(K_{LT} + K_{SE})}{(K_{LT} + K_{SE})B_{AG}B_{ANT}J} \quad (3.39)$$

The only negative terms are ones involving K_{SE}^3 ; collecting these terms in the numerator and simplifying, we obtain $(B_{ANT} + B_{AG} + B_p)K_{SE}^3 > 0$. Since the remaining terms are all positive, it follows that $m_1 > 0$. Lastly,

$$m_0 = \frac{K_{SE}^2(2K_{SE}K_{LT} + K_pK_{LT} + K_pK_{SE})}{(K_{LT} + K_{SE})B_{AG}B_{ANT}J} > 0 \quad (3.40)$$

The complexity of the coefficients, as given by equations (3.37)–(3.40), suggest that finding the exact roots of $f(\lambda)$ in terms of the symbolic constants of the model will be difficult. However, we can verify for the eye movement model that all the eigenvalues of A have negative real part.

In the case of the activation phase of the model, for a 10 degree saccade,

$$A \approx \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 35.79 & -52.97 & 0 & 0 & 0.29 & 0 \\ 75.41 & 0 & -111.61 & 0 & 0 & -0.60 \\ -120,000 & 56,818 & 56,811 & -1409 & 0 & 0 \\ 0 & 0 & 0 & 0 & -103 & 0 \\ 0 & 0 & 0 & 0 & 0 & -526 \end{bmatrix}$$

The computed eigenvalues of A are $\lambda_1 = -103$, $\lambda_2 = -526$, $\lambda_3 \approx -1313.72$, $\lambda_4 \approx -170.05$, $\lambda_5 \approx -15.36$, $\lambda_6 \approx -74.45$. These calculated eigenvalues lie far enough in the left half-plane of the complex numbers to guarantee that the exact values of the real parts of the eigenvalues of the matrix are indeed negative and therefore the equilibrium point for the activation phase of the agonist muscle is asymptotically stable.

Similarly, for the deactivation phase, for a 10 degree saccade,

$$A \approx \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 35.79 & -52.97 & 0 & 0 & 0.29 & 0 \\ 75.41 & 0 & -111.61 & 0 & 0 & -0.60 \\ -120,000 & 56,818 & 56,811 & -1409 & 0 & 0 \\ 0 & 0 & 0 & 0 & -5000 & 0 \\ 0 & 0 & 0 & 0 & 0 & -417 \end{bmatrix}$$

where the only changes from above are in 5,5 and 6,6 entries of A. Because of the structure of A, the only changes in the eigenvalues are that $\lambda_1 = -5,000$, $\lambda_2 = -417$, and the remaining eigenvalues are identical to those given above. Thus, the equilibrium state for the deactivation phase of the agonist muscle is also asymptotic.

3.2 CHAOS IN DYNAMICAL SYSTEMS

Our discussion up to this point has been concerned with the classical theory and methodologies of dynamical systems analysis. This theory has been valid and useful for a long time, and the eye movement model above is just one illustration that this theory remains so today. However, in the last 25 years much of the modern research in systems theory has been concerned with the apparently chaotic behavior of nonlinear systems—even extremely simple ones. To quote Devaney ([Dev], p. 2), “One of the remarkable discoveries of twentieth-century mathematics is that very simple systems, even systems depending on only one variable, may behave just as unpredictably as the stock market, just as wildly as a turbulent waterfall, and just as violently as a hurricane. The culprit, the reason for this unpredictable behavior, has been called ‘chaos’ by mathematicians.”

While one of the first mathematicians to witness chaotic behavior was Henri Poincaré, whose work dates back over 100 years, much of current research has its origins in work done by mathematical biologists during the 1950s and 1960s. ([Dev] and [May]) In 1964, the Russian mathematician Sarkovskii published a profoundly important theorem (which now bears his name) that shows why iteration of a nonlinear function of a real variable so often creates discrete sequences with very complicated behavior, even when the functions are very simple ([Dev], pp. 137–138). Sarkovskii’s theorem was not yet known in the west when in 1975, Li and Yorke published a special case of it and then studied specific functions. Their paper is titled “Period Three Implies Chaos,” and this marked the first time the word “Chaos” appeared in scientific literature ([Dev], p. 134). Li and Yorke comment on the fact that very simple models from

mathematical biology and from the study of turbulent fluid flow (i.e., by Lorenz) [Lor] could exhibit highly complicated dynamical behavior [LY].

Models of phenomena such as population growth or epidemics are often based on simple quadratic or exponential functions of the form $F(x) = x^2 + c$, $F(x) = cx(1 - x)$, or $F(x) = cxe^{-ax}$, where a, b , and c are parameters and can take the particular form of a discrete dynamical system given by $x(t + 1) = F(x(t))$, where $F: X \rightarrow X$ is a continuous function and X is an interval. For a given initial value such a discrete trajectory can look quite chaotic, even though it is generated by a well-defined function. This type of behavior called “deterministic chaos” in [May].

Indeed, Li and Yorke showed that for such F there is an uncountably infinite subset of X that contains no periodic or eventually periodic points (see Definition 3.3 below) [LY]. It is in this context, where the orbit (see Definition 3.1 below) of a point $x \in X$ never settles into an identifiable pattern that the term “chaos” was first used by Li and Yorke.

That is, that a slight change in initial conditions can produce trajectories that bear no resemblance to each other, and, worse, exhibit dramatically different long term behavior. The rules defining the function are simple and deterministic, and in theory we can reconstruct a particular trajectory with starting point $x_0 \in \mathfrak{R}$ by iterating the function F . However, when done on a computer, floating point arithmetic involves some round off, and the result can be quite different from what it should be theoretically. This sensitivity to initial conditions is one of the defining properties of chaos.

One of the goals of this newly emerging theory is to study and classify the chaotic behavior exhibited by computed trajectories, and we will give some examples.

There is another aspect to chaos in dynamical systems. Knowing that many trajectories can look very chaotic, the natural alternate question is: Given a discrete time sample of data that jumps about and exhibits no discernible pattern, does this data exhibit truly random behavior, or is it deterministic chaos? That is, is it a trajectory of some deterministic model? One approach to this problem involves constructing a library of past patterns of discrete time samples from the dynamical system, and then making short term predictions based on this library. Attempts can then be made to explain dissimilarities between the actual trajectory and the predicted trajectory as a result of deterministic chaos occurring in the system, random noise, or measurement error. As an additional complication, some forms of higher-dimensional chaos can be viewed, equivalently, as low-dimensional chaos tainted with noise. [May]

In certain situations, for example, in modeling measles epidemics occurring in large cities, these nonlinear forecasting methods have been found to outperform linear models for making short term predictions, and, moreover, have been able to detect deterministic chaos that was invisible to the linear model. In more complicated situations, these methods are unable to distinguish deterministic chaos from apparently random behavior. [May]

We will give several examples to illustrate deterministic chaos, using simple real-valued functions, in the following subsections.

3.2.1 Definitions and Examples

Here we will illustrate how chaos occurs in simple discrete dynamical systems of the form $x(t+1) = F(x(t))$, where X is a subset of $\mathfrak{R}$, often a closed interval $[a, b]$, and $F: X \rightarrow X$. We will always take $t_0 = 0$. Our development follows [Dev], pp. 17–51. To simplify notation, we will write $x_{t+1} = F(x_t)$. Additionally, since we wish to relate later behavior to the initial value, x_0 , and $x_1 = F(x_0)$, $x_2 = F(x_1) = F(F(x_0)) = F^2(x_0)$, $x_3 = F^3(x_0)$, $\dots$, $x_n = F^n(x_0)$, we will refer to the function $F^n(x)$ as the *n*th iterate of F .

3.1 Definition (orbit, seed) Suppose $F: X \rightarrow X$ and $x_0 \in X$. The sequence of numbers $x_0, x_1 = F(x_0), \dots, x_n = F^n(x_0), \dots$ is called the **orbit of x_0 under F** , and the number x_0 is the **seed** of the orbit. ■

Next, we characterize different types of orbits that can occur. Examples of each type defined are given in Table 3.2 below. The simplest orbit is one where the seed is an equilibrium point of the dynamical system.

3.2 Definition (Fixed point) If the seed, x_0 , is an equilibrium point of the discrete dynamical system $x_{t+1} = F(x_t)$, then $F^n(x_0) = x_0$ for all $n \geq 1$, and we will refer to x_0 as a **fixed point** for the function F . ■

3.3 Definition (Periodic point; eventually periodic point) A point x_0 is a **periodic point** of $F: \mathfrak{R} \rightarrow \mathfrak{R}$ if $F^n(x_0) = x_0$ for some $n > 0$. If x_0 is not fixed or periodic, but some point x_k on the orbit of x_0 is a periodic point for F , then say that x_0 is an **eventually periodic point** of F . ■

Notice that every fixed point is a periodic point but not conversely.

3.4 Definition (periodic orbit or cycle; prime period) Suppose x_0 is a periodic point for F . Then the orbit of x_0 forms a repeating sequence of numbers, and is therefore called a **periodic orbit** or **cycle**. The least such n such that $F^n(x_0) = x_0$ is called the **prime period** of the orbit of x_0 . ■

Observe that if the prime period of the orbit of x_0 is n , then the orbit has the particular form

$$x_0, x_1, \dots, x_{n-1}, x_n = F^n(x_0) = x_0, x_1, \dots, x_{n-1}, \dots \quad (3.41)$$

and hence is a repeating sequence of n numbers. Also, any point x_k on the orbit in (3.41) gives rise to a periodic orbit of prime period n .

Table 3.2 Examples of Definitions

$F(x)$	fixed points	some periodic points	some eventually periodic points	periodic and eventually periodic orbits
$x^2 - 2$	$-1, 2$	$-1, 2,$ $q_{\pm} = \frac{1}{2}(-1 \pm \sqrt{5})$	$0, \sqrt{2}, 1$	$1, -1, -1, \dots$ $-1, -1, -1, \dots$ $\sqrt{2}, 0, -2, 2, 2, \dots$ $0, -2, 2, 2, \dots$ $q_+, q_-, q_+, q_-, \dots$
$x^2 - 1$	$p_{\pm} = \frac{1 \pm \sqrt{5}}{2}$	$p_{\pm}, 0, 1, -1$		$1, 0, -1, 0, -1, \dots$ $0, -1, 0, -1, \dots$
$\frac{-3}{2}x^2 + \frac{5}{2}x + 1$	$p_{\pm} = \frac{3 \pm \sqrt{33}}{6}$	$p_{\pm}, 0$	$-\frac{1}{3}$	$0, 1, 2, 0, 1, 2, \dots$ $-\frac{1}{3}, 0, 1, 2, 0, 1, 2, \dots$
$2x \bmod 1$	0	$0, \frac{1}{5}, \frac{1}{9}$		$\frac{1}{5}, \frac{2}{5}, \frac{4}{5}, \frac{3}{5}, \frac{1}{5}, \dots$
$\frac{1}{2}x$	0	0 (only)	0 (only)	only $0, 0, 0, \dots$; if $x_0 \neq 0$, $orbit(x_0) \rightarrow 0$ as $t \rightarrow \infty$

The examples above are deceptively simple. One might conjecture that the orbit of an arbitrary number $x_0 \in \mathfrak{R}$ under such functions F above would exhibit relatively simple behavior: the value x_0 could be a periodic point (resp. eventually periodic) for F , in which case the orbit will be an n -cycle (resp. eventually be an n -cycle), or x_0 is not a periodic point, but as $t \rightarrow \infty$ the orbit converges to a limiting value or diverges to infinity. We will see that this conjecture is not only incorrect, but that even the simple functions above may have very complicated orbits, whose behavior is non-cyclic and in the long run do not appear to approach a limiting value, nor diverge to infinity. This is exhibited in the following example.

3.5 Examples While one way to demonstrate the complexity of an orbit, as alluded to above, would be to tabulate values of $x_k = F^k(x_0)$, we can accomplish this task more vividly using a histogram. For example, suppose $F: X \rightarrow X$ and we divide X into several subintervals of equal length, which we will call cells. Let $x_0 \in X$. If x_0 is periodic of prime period n , the histogram will have will have n nonempty cells (if we choose the width of the cell small enough) of roughly identical height. An example of this is considered in part A below. If the orbit of x_0 approaches a limiting value or diverges to infinity, we would expect to see this in the histogram; for example, if the orbit approaches a limiting cycle, we would, eventually, expect to see a few cells being visited with high frequency—relative to the other cells in the histogram, because the orbit gets closer and closer to the limit cycle. This behavior is illustrated in part B below. However, if every cell has been visited with roughly equal frequencies, then the behavior of the orbit will appear random and chaotic. An example of this is given in part C.

In each histogram below, the functions map $[-2, 2]$ to itself, and the cells were created by dividing the interval $[-2, 2]$ into 400 equal subintervals of width 0.01. The plotted points are the frequencies of each nonempty cell in the histogram.

A. In this first example, we illustrate a histogram for 200 iterations of the periodic orbit of $x_0 = 0$, $F(x) = -\frac{3}{2}x^2 + \frac{5}{2}x + 1$: $0, 1, 2, 0, 1, 2, \dots$. The histogram is given in Figure 3.6 below.

Figure 3.6 Histogram for Example 3.5A

B. In this example, we consider the histogram for 200 iterations of the function $F(x) = \frac{1}{2}x$, using the seed $x_0 = 2$. This orbit converges to zero, and this behavior is illustrated in Figure 3.7. (After about 20 iterations, x_k looks like zero to the computer.)

Figure 3.7 Histogram for Example 3.5B

C. This example illustrates “deterministic chaos.” The function $F(x) = x^2 - 2$ was iterated 10,000 times, using the seed $x_0 = 0.001$. The histogram is given in Figure 3.8 below.

Figure 3.8 Histogram for Example 3.5C

One important question in this theory is: For a given continuous function $F : \mathbb{R} \rightarrow \mathbb{R}$, just how many periodic points exist for this function? Toward answering this, we cite Sarkovskii’s theorem, which was first proved in 1964 ([Dev], p. 137).

3.6 Theorem (Sarkovskii) Suppose $F: \mathfrak{R} \rightarrow \mathfrak{R}$ is continuous, F has a periodic point of prime period n , and k is any number which follows n in the *Sarkovskii ordering* of (all) the natural numbers:

$$\begin{array}{l} 3, 5, 7, 9, \dots \\ 2 \cdot 3, 2 \cdot 5, 2 \cdot 7, \dots \\ \vdots \\ 2^m \cdot 3, 2^m \cdot 5, 2^m \cdot 7, \dots \\ \vdots \\ \dots, 2^m, \dots, 2^3, 2^2, 2^1, 1 \end{array}$$

Then F also has a periodic point of prime period k . ([Dev], p. 137) ■

To understand how strong this conclusion is, we consider some implications of this theorem. First, the “period three implies chaos” result of the Li and Yorke paper [LY] is a special case of this theorem: since 3 precedes every other natural number in Sarkovskii’s ordering of the natural numbers, if F has a point of prime period 3, then F must have points of prime period k for all natural numbers $k \neq 3$, that is, F has points with orbits of every possible length. Second, since every row of natural numbers in Sarkovskii’s ordering is infinitely long, the only way F can have finitely many periodic points is if the prime periods for these points lie along a tail of finite length in the last row, that is, the period of every periodic point is a power of 2. Third, the theorem can be extended to continuous functions $F: [a, b] \subseteq \mathfrak{R} \rightarrow \mathfrak{R}$ by defining

$$\hat{F}(x) = \begin{cases} F(x) & \text{if } x \in [a, b] \\ F(a) & \text{if } x < a \\ F(b) & \text{if } x > b \end{cases}$$

([Dev], p. 138). For a proof of Sarkovskii’s Theorem, the reader is referred to [MS], pp. 82–85.

There is a converse for this theorem also, and this too is surprising:

3.7 Theorem For each n , there exists a continuous function $F: \mathbb{R} \rightarrow \mathbb{R}$ which has a cycle of period n , but no cycles of any period preceding n in the Sarkovskii ordering ([Dev], p. 138; for a proof, see [Ela]). ■

Recall that in Table 3.2 above, we showed that $F(x) = -\frac{3}{2}x^2 + \frac{5}{2}x + 1$ has a 3-cycle $0, 1, 2, 0, 1, 2, \dots$. It follows that this function has points lying along cycles of every possible length; in particular there are actually infinitely many periodic points for this function!

The obvious limitation of Sarkovskii's theorem is that it is not constructive, that is, it provides no clues as to where these periodic points are. For example, we know that $F(x) = -\frac{3}{2}x^2 + \frac{5}{2}x + 1$ has a 3-cycle, hence has cycles of every length n . To find a 4-cycle we must solve $F^4(x) = x$, a 16th degree polynomial equation, and then eliminate from the set of 16 roots, the non real points, the fixed points, and the prime period 2 points, which are periodic points of F but not of prime period 4. In the remaining set of roots there must be at least one real root, which will be the seed for a 4-cycle; moreover, since every (distinct) point on such a 4-cycle has prime period 4, there must be at least 4 real roots from the remaining set of roots of this polynomial equation. Of course, the first step, i.e., of finding exact roots of the 16th degree polynomial equation, cannot in general be performed—and finding points on 5-cycles, 6-cycles, etc., is even more difficult.

An ingenious method that bypasses the problem above is *symbolic dynamics*, developed by Stephen Smale, which, for certain continuous functions $F: \mathbb{R} \rightarrow \mathbb{R}$, sets up a one-to-one, onto, and continuous map between a symbolic dynamical system defined on a sequence space and the periodic points of the continuous function F . The periodic points of the symbolic model are easy to locate, and then

the inverse map can be used to assert their existence in the dynamical system defined by the real function F . The reader is referred to [Dev], pp. 97–113 for more.

To motivate the definition of chaos given below, consider that knowledge of infinitely many periodic points of every prime period does not, in general, give a reason for the behavior illustrated in Figure 3.8 of Example 3.5C. That is, even if we know a particular function has cycles of infinitely many different orders, this does not predict the bizarre, non-cyclic behavior of some orbits, as illustrated in 3.5C. As in [Dev], pp. 114–121, we will state three defining properties of a chaotic dynamical system. It should be pointed out that other definitions of chaos appear in the literature. There is no need to restrict the state spaces to be subsets of the reals in this definition. We only need to assume that our dynamical systems is given by $x_{t+1} = F(x_t)$, where $x_0 \in X$, $F: X \rightarrow X$, and X is any suitable state space (see Definition 1.46 in Section 1.4). After stating the definition, we will define the terms and discuss them in detail.

3.8 Definition A discrete dynamical system given by $x_{t+1} = F(x_t)$, where $F: X \rightarrow X$, and X is the state space of the dynamical system, is **chaotic** if:

- A. The periodic points form a dense subset of X .
- B. The dynamical system is transitive.
- C. The system depends sensitively on initial conditions.

([Dev], p. 119) ■

3.9 Definition A subset Y of a set X is **dense** in X , if for any $x \in X$ there exists a sequence of points $y_0, y_1, \dots$ in Y that converges to x . ■

For example the rationals form a countably infinite dense subset of $\mathfrak{R}$ since there exists a sequence of decimal approximations converging to every irrational number in $\mathfrak{R}$. Similarly, the irrational numbers form an uncountably infinite

dense subset of the real numbers. Notice that if the periodic points for some function $F: \mathbb{R} \rightarrow \mathbb{R}$ are dense in some interval $I \subseteq \mathbb{R}$, and the seed of an orbit of F is an element in I , then one might hope that in most cases the orbit will be an n -cycle. However, this is not true for a chaotic system. Even if x_0 is chosen to be periodic, because floating point arithmetic involves round off, numerical error can knock the orbit off of the cycle.

The second defining property of a chaotic dynamical system is *transitivity*.

3.10 Definition Suppose $\|\cdot\|$ is a metric for the state space X of a discrete dynamical system given by $x_{t+1} = F(x_t)$. The system is called **transitive** if for every pair $x, y \in X$ and every $\varepsilon > 0$, there exists a seed $x_0 \in X$ such that $\|x - x_0\| < \varepsilon$ and $\|F^k(x_0) - y\| < \varepsilon$ for some $k \geq 1$. ■

The idea of transitivity is that given any two points x and y , there exists an orbit of some seed that comes arbitrarily close to both points.

The third and final property of a chaotic dynamical system is *sensitivity*.

3.11 Definition A discrete dynamical system given by $x_{t+1} = F(x_t)$ is said to **depend sensitively on initial conditions** if there exists a $\beta > 0$ such that for every $x \in X$ and $\varepsilon > 0$ there exists a point $y \in X$ and $k \geq 1$ such that $\|x - y\| < \varepsilon$ but $\|F^k(x) - F^k(y)\| \geq \beta$. ■

A few remarks about Devaney's definition are in order. First, the sensitivity condition says that no matter how close to an arbitrary point $x \in X$ we choose the point $y \in X$, the corresponding orbits of x and y will always be at least β apart at some point in the progression of the orbit; moreover, the distance β does not depend on x . However the definition does not require that in the long term that the orbits must stay apart. Indeed, both orbits could converge to the same point as $t \rightarrow \infty$. Second, other definitions of sensitivity to initial conditions are possible.

May characterizes this sensitivity by the property that the orbits become farther apart as time increases [May]. In either case, sensitive dependence on initial conditions means that, for all practical purposes, numerical computation of the orbit of x_0 , where x_0 is in a region of sensitive dependence will probably bear no resemblance to the actual orbit of x_0 —even though we know the seed and the rule specifying how to compute the orbit. This is because no matter how many digits of accuracy are used in determining the value of the seed, it will never in general be exact; then because of sensitive dependence, the computed orbit may totally diverge from the actual orbit. Much of the behavior that is referred to as “chaotic” in the literature is ascribed to this property [BBCDS].

A surprising fact recently proved is that Devaney’s sensitivity condition follows from his first two conditions, that is, using a combination of the triangle inequality for the metric defined on the space of the dynamical system, the denseness of the periodic points, and transitivity, it has been shown that any transitive system with dense periodic points will exhibit sensitive dependence on initial conditions [BBCDS].

It is known that the restriction of the function $F(x) = x^2 - 2$ to the interval $[-2, 2]$ does map the interval to itself and satisfies Devaney’s definition of chaotic. Clearly the orbit of $x_0 = 0$ is eventually fixed: $0, -2, 2, 2, \dots$. As is apparent in the histogram of the orbit in 3.5C, the orbit of $x_0 = 0.001$ illustrates sensitive dependence on initial conditions, for clearly it visits every cell of the histogram repeatedly and with varying frequency—and bears no resemblance to the orbit of the nearby value $x_0 = 0$. The irregular behavior of such an orbit illustrates “deterministic chaos”: that is, although this behavior appears random, there is an

underlying simple function, so what we are seeing is not truly random behavior but rather behavior that is attributable to properties A–C of the Definition.

3.2.2 Classification of Fixed and Periodic Points

Even more curious than the chaotic behavior of orbits of $F(x) = x^2 - 2$ with seeds near zero, as described above, is the way the orbit behavior changes as the value of the parameter c changes for the one parameter family of quadratic functions $Q_c(x) = x^2 + c$, of which this function is a member. We will discuss that in Section 3.2.3 but we need a few more basic ideas, which we present in this section. Again, our development in this section is essentially that in [Dev], pp. 29–47.

A useful tool in the visualization of the orbit of a function is *graphical analysis*. The idea is to begin on the line $y = x$ at the point (x_0, x_0) , and then move vertically to the graph of F , which brings us to the point $(x_0, F(x_0))$. Now move horizontally to the line $y = x$ from here, which brings us to the point $(F(x_0), F(x_0))$. Repeat the entire process, moving vertically to the graph of F , etc. The resulting diagram forms a “cobweb” pattern. This technique reveals easily that for $c > \frac{1}{4}$ every orbit of $Q_c(x) = x^2 + c$ tends to infinity because the entire graph of Q_c lies above the line $y=x$. This is demonstrated for the particular value $c = \frac{1}{2}$ in Figure 3.9 below.

Figure 3.9 Graphical analysis of the orbit of $x_0 = 0$ for $F(x) = x^2 + 0.5$

Two important classifications for fixed points are those that are *attracting* and those that are *repelling*.

3.12 Definition Suppose $\bar{x}$ is a fixed point of $F: X \subseteq \mathfrak{R} \rightarrow \mathfrak{R}$ such that F is differentiable at $\bar{x}$. Then $\bar{x}$ is an **attracting fixed point** if $|F'(\bar{x})| < 1$, a **repelling fixed point** if $|F'(\bar{x})| > 1$, and a **neutral fixed point** if $|F'(\bar{x})| = 1$. ■

To understand why the terminology above is suggestive, consider that $|F'(\bar{x})|$ gives the magnitude of the slope of the tangent line T to the graph of F at the fixed point $(\bar{x}, F(\bar{x})) = (\bar{x}, \bar{x})$, and this is the point of intersection of T with the line $y = x$ at $x = \bar{x}$. It is not hard to verify using graphical analysis, after an appropriate translation so we may assume that $\bar{x} = 0$, that if m_T is the slope of T , then if $|m_T| < 1$ then all nonzero seeds near zero have orbits that converge to zero, so that zero is an “attracting” fixed point; and if $|m_T| > 1$, then nonzero seeds near zero have orbits that tend to infinity, whence zero is a “repelling” fixed point. This observation can then be used in studying the behavior of orbits with seeds nearby a fixed point of a nonlinear function F , as in the following example.

3.13 Example Consider the function $F(x) = 2x(1 - x) = 2x - 2x^2$, which has derivative $F'(x) = 2 - 4x$ and fixed points $\bar{x} = 0, \frac{1}{2}$. Since $F'(0) = 2 > 1$, $\bar{x} = 0$ is a repelling fixed point, and since $F'(0.5) = 0 < 1$, $\bar{x} = \frac{1}{2}$ is an attracting fixed point. Graphical analysis for seeds in the interval $(0, 1/2)$ confirms that orbits for such values are repelled away from the fixed point at zero and converge to the fixed point at $1/2$. ■

What about the case of a neutral fixed point? Unfortunately, in this case the situation is not as simple. For example, $F(x) = x$ fixes all points, and all of these fixed points are neutral, so none are attracting or repelling. Also, $F(x) = x - x^2$ has a neutral fixed point at $x = 0$ that is attracting from the right and repelling from the left.

It would be nice not to have to use graphical analysis every time we wish to study the behavior of orbits for seeds near fixed points. The following two theorems show how to do this when the fixed point is either attracting or repelling, and are easily proved using the Mean Value Theorem.

3.14 Theorem Suppose $\tilde{x}$ is an attracting fixed point for $F: X \subseteq \mathfrak{R} \rightarrow \mathfrak{R}$, and F' is continuous at $\tilde{x}$. Then there exists an interval I containing $\tilde{x}$ in its interior, such that the following is true: if $x \in I$, then $F^n(x) \in I$ for every $n \geq 1$, and $F^n(x) \rightarrow \tilde{x}$ as $n \rightarrow \infty$. ■

3.15 Theorem Suppose $\tilde{x}$ is a repelling fixed point for $F: X \subseteq \mathfrak{R} \rightarrow \mathfrak{R}$, and F' is continuous at $\tilde{x}$. Then there exists an interval I containing $\tilde{x}$ in its interior such that the following is true: if $x \in I$ and $x \neq \tilde{x}$, then there exists an integer $n \geq 1$ such that $F^n(x) \notin I$. ■

These theorems imply the following. If $\tilde{x}$ is an attracting fixed point, and, if the orbit of an arbitrary seed enters the interval of convergence, I , of Theorem 3.14, then the orbit will converge to $\tilde{x}$. In this context, an attracting fixed point is “visible” on a computer, in that if y_0 is sufficiently near an attracting fixed point $\tilde{x}$, we will see that the orbit of y_0 is approaching $\tilde{x}$. However, if $\tilde{x}$ is a repelling fixed point, nearby points are eventually forced away, so, it is not usually possible to “see” a repelling fixed point by studying calculated orbits, unless the orbit lands on the fixed point exactly, which rarely occurs because of round off errors. This is why repelling fixed point are not “visible” in the *orbit diagrams* in the next section.

Periodic points may also be classified as attracting, repelling, or neutral. The situation here is a little more complicated, since these points will be defined in reference to higher iterates, F^k , of the function F .

3.16 Definition Suppose x_0 is a periodic point of $F: X \subseteq \mathfrak{R} \rightarrow \mathfrak{R}$, having prime period n , that is, $F^n(x_0) = x_0$ with n least. Suppose further that F^n is differentiable at $x = x_0$. Then x_0 is

- A. an **attracting periodic point** if $\left| \frac{dF^n}{dx} \Big|_{x=x_0} \right| < 1$;
- B. a **repelling periodic point** if $\left| \frac{dF^n}{dx} \Big|_{x=x_0} \right| > 1$; and
- C. a **neutral periodic point** if $\left| \frac{dF^n}{dx} \Big|_{x=x_0} \right| = 1$.

■

We can also state theorems for classifying periodic points as attracting or repelling.

3.17 Theorem Suppose x_0 is an attracting periodic point for $F: X \subseteq \mathfrak{R} \rightarrow \mathfrak{R}$ having prime period n , and $\frac{dF^n}{dx}$ is continuous at x_0 . Then there exists an interval I containing x_0 in its interior, such that the following is true: if $x \in I$, then $F^{nj}(x) \in I$ for every $j \geq 1$, and $F^{nj}(x) \rightarrow x_0$ as $j \rightarrow \infty$.

■

Notice that the convergent sequence $F^n(x_0), F^{2n}(x_0), \dots$ in Theorem 3.17 is actually a subsequence of the orbit of x_0 . Below, we will consider an orbit that is bounded, but does not converge to a limiting value, but rather an attracting 2-cycle. By the Bolzano-Weierstrass Theorem from real analysis, we know this orbit has a convergent subsequence—but that theorem does not tell us how to find it. Theorem 3.17 not only gives us a particular convergent subsequence, but the limit point as well.

3.18 Theorem Suppose x_0 is a repelling periodic point for $F: X \subseteq \mathfrak{R} \rightarrow \mathfrak{R}$ having prime period n , and $\frac{dF^n}{dx}$ is continuous at x_0 . Then there exists an interval I containing x_0 in its interior, such that the following is true: if $x \in I$ and $x \neq x_0$, then there exists an integer $j \geq 1$ such that $F^j(x) \notin I$.

■

In the next example we consider the relationship between the orbits of F and the orbits of higher iterates of F corresponding to the same seed, as discussed following the statement of Theorem 3.17.

3.19 Example Let's consider the orbit behavior of the function $F(x) = x^2 - 1$. Recall that its fixed points are given in Table 3.2 to be $p_{\pm} = \frac{1}{2}(1 \pm \sqrt{5})$. These are the roots of the equation $F(x) - x = 0$. To find the periodic points of prime period 2, solve the equation

$$F^2(x) - x = 0 \Leftrightarrow x^4 - 2x^2 - x = 0 \quad (3.42)$$

Now p_+, p_- are fixed points of F and hence of F^2 , so they are roots of (3.42).

Hence, $(x - p_+)(x - p_-) = F(x) - x = x^2 - x - 1$ divides $x^4 - 2x^2 - x$. Observe that

$$\frac{x^4 - 2x^2 - x}{x^2 - x - 1} = x^2 + x = x(x + 1) \quad (3.43)$$

Thus, the roots of (3.42) are $p_+, p_-, 0, -1$, and since these are all distinct, it follows that $x = 0, -1$ are periodic points of F having prime period 2, and they form a unique 2-cycle for F : $0, -1, 0, -1, \dots$.

Next, we classify $x = 0, -1$ as attracting or repelling periodic points. In fact they are both attracting periodic points. To see this, calculate:

$$\begin{aligned} \frac{dF^2}{dx} &= \frac{d}{dx} F(F(x)) \\ &= F'(F(x))F'(x) \end{aligned} \quad (3.44)$$

by the chain rule. Our 2-cycle takes the general form $x_0, x_1, x_0, x_1, \dots$, and, since

$F(x_0) = x_1$, we have by (3.44) that

$$\left. \frac{dF^2}{dx} \right|_{x=x_0} = F'(F(x_0))F'(x_0) = F'(x_1)F'(x_0) \quad (3.45)$$

Similarly, since $F(x_1) = x_0$, we have

$$\left. \frac{dF^2}{dx} \right|_{x=x_1} = F'(F(x_1))F'(x_1) = F'(x_0)F'(x_1) \quad (3.46)$$

So the right sides of (3.45) and (3.46) are identical, and

$$\left. \frac{dF^2}{dx} \right|_{x=0} = \left. \frac{dF^2}{dx} \right|_{x=-1} = F'(-1)F'(0) = 0$$

so this 2-cycle is attracting.

To illustrate, consider the orbit with seed $x_0 = 0.5$. The graph of the second iterate, $F^2(x) = x^4 - 2x^2$, is given in Figure 3.10. Notice that by graphical analysis the orbit of $x_0 = 0.5$ for F^2 , which is a convergent subsequence of the orbit of $x_0 = 0.5$ for F , converges to the fixed point $x = 0$ for F^2 , which is one of the points of our attracting 2-cycle for F .

Figure 3.10 The graph of $F^2(x) = x^4 - 2x^2$ and an attracting fixed point

However, the orbit of $x_0 = 0.5$ for F is not a convergent sequence, but is rather an oscillating sequence which is attracted to the 2-cycle $0, -1, 0, -1, \dots$. Graphical analysis for this orbit is done in Figure 3.11.

Figure 3.11 The orbit of $x_0 = 0.5$ for $F(x) = x^2 - 1$

■

As a final point, notice the simple algebra and geometry that illustrate succinctly the more abstract concept of a sequence which does not converge but does contain a convergent subsequence.

In Example 3.19, we showed both points of the 2-cycle $\{0, 1\}$ are attracting.

The proof of that is easily generalized using the Chain Rule:

3.20 Theorem Suppose $x_0, x_1, \dots, x_{n-1}$ form a periodic orbit of prime period n , that is, an n -cycle for F , so $x_j = F^j(x_0)$, $j = 1, 2, \dots, n-1$. Then

$$\left. \frac{dF^n}{dx} \right|_{x=x_j} = F'(x_0)F'(x_1) \cdots F'(x_{n-1})$$

for every $j = 0, 1, 2, \dots, n-1$. Consequently, if any one point, x_k , on the cycle is attracting (resp. repelling, neutral), then all points on the cycle must be attracting (resp. repelling, neutral). ■

3.2.3 The Orbit Diagram for the Quadratic Family

In Example 3.5C we saw the chaotic behavior of orbits of $F(x) = x^2 - 2$ which have seeds near zero, even though the orbit of zero itself is eventually fixed: $0, -2, 2, 2, \dots$. This function is a special case of the quadratic family $Q_c(x) = x^2 + c$. Now instead of fixing the parameter c and varying the seeds, we will look at what happens when the seed is fixed but the parameter c is varied. The question is, how does the behavior of the orbit change as the parameter is varied? Our goal in this section is to consider this question using an intuitive approach, and the development is essentially that in [Dev], pp. 69–95.

We will use the graphical tool of the *orbit diagram*. The idea is to fix x_0 and track the eventual behavior of the orbit of Q_c . We will discard the first 100 or so iterations so as to avoid initial “transient behavior,” and then plot the values of $Q_c^n(x_0)$ above c , c being on the horizontal axis. Using the same seed, the parameter c is varied slightly and the eventual behavior of the orbit is plotted above the c -value. The resulting graph in the x, c -plane is the orbit diagram for $Q_c(x) = x^2 + c$. This is illustrated in Figure 3.12 below.

A first question is, why are only c -values in the interval $[-2, 0.25]$ used? First, recall that graphical analysis can be used to verify that for $c > 0.25$, all orbits of $Q_c(x) = x^2 + c$ tend to infinity, so the dynamics here are uninteresting.

Figure 3.12 Orbit diagram for $Q_c(x) = x^2 + c$

Also, in Section 3.2.1 we commented on the fact that $Q_{-2}(x) = x^2 - 2$ is chaotic on the entire interval $[-2, 2]$ and hence has dense periodic points in this interval. This is apparent in Figure 3.12: as the parameter decreases from 0.25 to -2, the simple dynamics associated with few or no periodic points gives way to chaos, that is, in addition to sensitive dependence on initial values of the seed, for a fixed seed, perturbations of the parameter c can also cause chaotic behavior. This is what is

illustrated in Figure 3.12. In addition, $Q_c(x) = x^2 + c$ is also chaotic for values $c < -2$, but it is known that the set in which chaotic points are dense for such c is not an interval as it is when $c = -2$. This set is discussed in detail in [Dev] and has many fascinating properties, but such a discussion requires symbolic dynamics and is beyond the scope of our treatment here. Since we are interested in how the “transition to chaos” occurs, the interval $[-2, 0.25]$ will suffice to illustrate this transition.

To discuss what happens as $c \rightarrow -2$, recall that fixed points of Q_c are solutions to

$$Q_c(x) = x \Leftrightarrow x^2 - x + c = 0 \quad (3.47)$$

which has solutions

$$p_{\pm}(c) = \frac{1}{2}(1 \pm \sqrt{1 - 4c}) \quad (3.48)$$

To simplify notation, we will suppress the c and write instead p_+ or p_- to represent the fixed points of Q_c for a given c .

To classify the fixed points above as attracting, repelling, or neutral, we must compute:

$$\left| Q'_c(p_{\pm}) \right| = \left| 1 \pm \sqrt{1 - 4c} \right| \quad (3.49)$$

as discussed in Section 3.2.1. Setting the quantity in (3.49) equal to one, less than one, and greater than one, and then solving the resulting equality or inequality in terms of c , we obtain the following results, which together with our observations so far, are referred to by Devaney as *The First Bifurcation* ([Dev], p. 54).

3.21 Proposition For the quadratic family $Q_c(x) = x^2 + c$ the following are true:

- A. if $c > 0.25$ then all orbits with $x_0 \in (-\infty, \infty)$ tend to infinity;
- B. when $c = 0.25$, Q_c has a single fixed point at $p_+ = p_- = 1/2$ that is neutral;
- C. if $c < 0.25$, then the fixed points are distinct, and the fixed point p_+ is always repelling. In addition,
 - a. if $-3/4 < c < 1/4$, then p_- is attracting; and
 - b. if $c < -3/4$, then p_- is repelling. ■

The term *bifurcation* means to divide or to fork into two branches. Proposition 3.21 says that this is what happens as c moves below $1/4$: the fixed point in 3.21B splits into the two distinct fixed points described in 3.21C.

It can be shown that (i) for $c \in (-\frac{3}{4}, \frac{1}{4})$, all orbits with seeds $x_0 \in (-p_+, p_+)$ tend to the attracting fixed point at p_- , and (ii) all orbits tend to infinity if $x_0 > p_+$ or $x_0 < -p_+$ ([Dev], p. 55). Since p_+ is strictly decreasing in c , we also have

$$[-p_+, p_+] \subsetneq [-2, 2] \text{ for } c \in (-2, \frac{1}{4}]$$

We see that for $c \in (-\frac{3}{4}, \frac{1}{4})$, the dynamics are quite tame: either an orbit tends to infinity, or an orbit tends to the fixed attracting fixed point p_- . In our example we chose seed $x_0 = 0$, so it follows that for $c \in (-\frac{3}{4}, \frac{1}{4})$ the orbit should approach p_- , and this is evident in Figure 3.12: p_- is the single plotted point in the orbit diagram for c -values in this range.

The information in the preceding paragraph, as well as some other useful information that will soon be relevant to us is summarized in Table 3.3 below.

Table 3.3 Seed/c-value summary of eventual orbit behavior

c-values	seed values, x_0	orbit(x_0)
$c \in \mathfrak{R}$	$x_0 < -p_+$ or $x_0 > p_+^*$	tends to infinity
$c \in [-2, \frac{1}{4}]$	$x_0 \in (-p_+, p_+)$	"transition to chaos"
$c \in (-\frac{3}{4}, \frac{1}{4})$	$x_0 \in (-p_+, p_+)$	tends to p_-

* If $c > \frac{1}{4}$, there are no fixed points and all orbits tend to infinity.

We now turn to more complicated behavior illustrated in Figure 3.12 for $c < -\frac{3}{4}$. For c-values just slightly below $-3/4$, we see something new in Figure 3.12: instead of the fixed point p_- that we have been plotting up to this point, now a 2-cycle appears; see the two plotted points above the c-axis for c-values slightly below $-3/4$. As remarked following Theorem 3.15 in Section 3.2.2, we do not see the repelling fixed point p_- . Instead, the orbit of zero is attracted to a 2-cycle. This accounts for the disappearance of p_- in the orbit diagram when $c < -\frac{3}{4}$.

To calculate the attracting 2-cycle, recall the method used to find the attracting 2-cycle in Example 3.19 for $Q_{-1}(x) = x^2 - 1$. The method can be generalized to Q_c to find the general 2-cycle $q_+, q_-, q_+, q_-, \dots$, where

$$q_{\pm}(c) = \frac{1}{2}(-1 \pm \sqrt{-4c - 3}), c \leq -\frac{3}{4} \quad (3.50)$$

Similar to the classification of the fixed points given in Proposition 3.21, the periodic points q_+ and q_- can be classified as attracting, fixed, or neutral. This information and more are summarized in what Devaney calls a *period-doubling bifurcation* ([Dev], p. 56).

3.22 Proposition For the quadratic family $Q_c(x) = x^2 + c$, the following is true:

- A. if $-\frac{3}{4} < c < \frac{1}{4}$, then Q_c has an attracting fixed point at p_- and no 2-cycles;
- B. if $c = -\frac{3}{4}$, then Q_c has a neutral fixed point at $p_- = q_{\pm} = -\frac{1}{2}$; and
- C. if $-\frac{5}{4} < c < -\frac{3}{4}$, then Q_c has repelling (i.e., invisible) fixed points at $p_{\pm}$ and an attracting 2-cycle at $q_{\pm}$. ■

This proposition explains what we see in Figure 3.12 for $c \leq -\frac{3}{4}$. The attracting fixed point p_- goes from attracting to neutral, and then to repelling—and hence becomes “invisible”—as c makes the transition from slightly above $-3/4$, to $-3/4$, and then to just below $-3/4$. Then, when p_- is neutral (i.e., $c = -3/4$) an attracting 2-cycle is born at p_- and we see a period-doubling bifurcation occur as this period-1 fixed point splits into an attracting 2-cycle.

A few comments about what we see for $-2 < c < -5/4$ are the following. Because of Sarkovskii’s Theorem, if there are finitely many periodic points, they must have prime period 2^n for some $n \geq 0$. This would explain why we see first a period-1 fixed point, a 2-cycle, then a 4-cycle, so that the periodic points appear to occur through a “succession of period-doubling bifurcations” ([Dev], p. 83). Also, amidst the dark regions there is a region of white separated by points at the top, middle, and bottom. This is the so-called “period-3 window” ([Dev], p. 86), labeled in Figure 3.12. Since the particular Q_c associated with this window have a 3-cycle, they must have infinitely many periodic points of all prime periods. Presumably, we see only the period-3 window because the other periodic points are repelling. In this sense, the orbit diagram is misleading.

Also, we know that Q_{-2} is chaotic on the interval $[-2, 2]$, in the sense of Definition 3.10, so by the time $c = -2$, the periodic points are dense in $[-2, 2]$. This would, in part, explain why there are regions of solid black, i.e., corresponding to

non-periodic or chaotic behavior of the orbit of $x_0 = 0$ in Figure 3.12, as the periodic points begin to multiply without bound—and Q_c exhibits apparent sensitive dependence on values of the parameter c . Of course the orbit of $x_0 = 0$ for Q_{-2} is eventually periodic: $0, -2, 2, 2, \dots$. What is curious here is that for values of the parameter in the neighborhood to the right of -2 on the c -axis, we also witness deterministic chaos, that is, not only is Q_c sensitive to slight perturbations of the seed, but also to slight perturbations of the parameter c .

For a discussion of why regions of the orbit diagram seem to be miniature copies of the orbit diagram itself and a more detailed discussion of the ideas introduced above, the reader is referred to [Dev].

LITERATURE CITED

The following reference list is ordered alphabetically by *author*. This ordering does not always coincide with an alphabetical ordering by *mnemonic code*.

- BLT A. Terry Bahill, J. R. Latimer, and B. T. Troost. "Linear Homeomorphic Model for Human Movement," *IEEE Transactions on Biomedical Engineering*, Vol. BME-27, No. 11, November (1980), 631–639.
- BBCDS J. Banks, J. Brooks, G. Cairns, G. Davis, and P. Stacey. "On Devaney's Definition of Chaos," *American Mathematical Monthly*, Vol. 99, April (1992), 332–334.
- BL P. Baxandall and H. Liebeck. *Vector Calculus*. Oxford University Press, New York, 1986.
- Dev R. L. Devaney. *A First Course in Chaotic Dynamical Systems*. Addison-Wesley, Reading, MA, 1992.
- Ela S. Elaydi. "On a Converse of Sharkovsky's Theorem," *American Mathematical Monthly*, Vol. 103, No. 5, May (1996), 386–392.
- Gle P. Glendinning. *Stability, Instability, and Chaos*. Cambridge University Press, Cambridge, 1994.
- HJ85 R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.
- HJ91 R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991.

- KB R. E. Kalman and J. E. Bertram. "Control System Analysis and Design Via the 'Second Method' of Lyapunov," *ASME Transactions*, Vol. 82, Ser. D, No. 2, June (1960), 371–400.
- LY Tien-Yien Li and J. A. Yorke. "Period Three Implies Chaos," *American Mathematical Monthly*, Vol. 82, December (1975), 985–992.
- Lor E. N. Lorenz. "Deterministic Nonperiodic Flow," *Journal of Atmospheric Science*, Vol. 20 (1963), 130–141.
- Lue David G. Luenberger. *Introduction to Dynamic Systems*. John Wiley and Sons, New York, 1979.
- Lya A.M. Lyapunov. *Collected Works of Academician A. M. Lyapunov*. Vols. I–II, Wright-Patterson Air Force Base, Ohio: Translation Division, Foreign Technology Division, 1967.
- May R. M. May. "Necessity and Change: Deterministic Chaos in Ecology and Evolution," *Bulletin of the American Mathematical Society*, Vol. 32, No. 3, July (1995), 291–308.
- MS W. de Melo and S. van Strien. *One – Dimensional Dynamics*. Springer-Verlag, New York, 1993.
- Mil K. S. Miller. *Linear Difference Equations*. Benjamin, Reading, MA, 1968.
- Mun J. R. Munkres. *Topology, a First Course*. Prentice Hall, Englewood Cliffs, NJ, 1975.
- PA L. Padulo and M. A. Arbib. *System Theory*. W. B. Saunders, Philadelphia, 1974.
- RS B. J. Rice and J. D. Strange. *Ordinary Differential Equations with Applications*. Brooks/Cole Publishing, Belmont, CA, 1989.

- Rod J. J. Rodden. "Applications of Lyapunov Stability Theory," unpublished Ph. D. dissertation, Stanford University, 1964.
- San D. A. Sanchez. *Ordinary Differential Equations and Stability Theory: An Introduction*. W. H. Freeman, San Francisco, 1968.
- SB F. Szidarovsky and A. Terry Bahill. *Linear Systems Theory*. CRC Press, Boca Raton, 1992.